IEMTRONICS
International Conference
Toronto, Canada

2021
CONFERENCE PROCEEDINGS

Date: 21st - 24th April, 2021

Editors:
Satyajit Chakrabarti, Rajashree Paul, Bob Gill,
Malay Gangopadhyay, Sanghamitra Poddar

IEM
INSTITUTE OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs

IEEE VANCOUVER SECTION

IEEE Advancing Technology for Humanity
TORONTO Section est. 1903

SMART

UEM
UNIVERSITY OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs

About the Conference

IEMTRONICS 2021

Continuing with the outstanding success of IEEE IEMCON, IEEE CCWC, IEEE UEMCON, IEMANTENNA we are proud to present IEMTRONICS 2021 (International IOT, Electronics and Mechatronics Conference) which will be held during **21st- 24th April, 2021 at Toronto, Canada, in online mode**. Keeping in mind the pandemic situation prevalent globally due to Covid 19 and following the legacy of organizing highly successful conferences, we have planned for the online conference. The conference aims to bring together scholars from different backgrounds to emphasize dissemination of ongoing research broadly in the fields of IOT, Electronics and Mechatronics. Research papers are invited describing original works in above mentioned fields and related technologies. The conference will include a peer-reviewed program of technical sessions, special sessions, tutorials and demonstration sessions.

All accepted papers which will be presented during the parallel sessions of the Conference will be submitted for publication in IEEE Xplore Digital Library (Scopus, DBLP, Ei Compendex, Web of Science and Google Scholar).

This conference will also promote an intense dialogue between academia and industry to bridge the gap between academic research, industry initiatives, and governmental policies. This is fostered through panel discussions, keynotes, invited talks and industry exhibits where academia is exposed to state-of-practice and results from trials and interoperability experiments. The industry in turn benefits by exposure to leading-edge research in networking as well as the opportunity to communicate with academic researchers regarding practical problems that require further research.

Our Reviewers

IEMTRONICS 2021 followed a rigorous triple-blind review process in order to identify suitable papers for both presentation and publication. This process helped the organizers to shortlist good quality papers from diverse regional areas and across various domains. A detailed review process was possible due to the excellent and enthusiastic support extended by the strong technical review team of IEMTRONICS 2021. For every stage of submission, IEMTRONICS had a specific template review procedure to analyze the submissions and provide suitable comments for the authors to incorporate. The review team which formed the technical backbone for the selection of submissions for the edited book and the conference presentation was supervised by:

NAME	AFFILIATION
Ranjay Hazra	NIT, Silchar
Arighna Deb	KIIT University
Sukomal Dey	Indian Istitute of Technology, Delhi
Rohit Singh	University of Colorado Denver
Mainak Adak	WBUT
Riya Agarwal	Manipal Institute of Technology
Navid Bin Ahmed	Independent University Bangladesh (IUB)
Ritesh Ajoodha	The University of the Witwatersrand, Johannesburg
Md L Ali	Rider University
Mohammad Anees	Xilinx
Mohamed Imran Mohamed Ariff	Universiti Teknologi MARA Perak Branch
Haissam Badih	Oakland University
Anindya Bal	BRAC University
Ke-Lin Du	Concordia University
Patricia Enrique	University of Waterloo
Vilas H Gaidhane	Birla Institute of Technology and Science Pilani, Dubai Campus
Sasirekha Gvk	Electronics City
William P. Haggerty	Georgia Southern University
Qaiser Ijaz	University of Burgundy
Gustavo Jamanca-Lino	Colorado School of Mines
Haruo Kobayashi	Gunma University
Anna Kuwana	Gunma University
Moises Levy	West Texas A&M University
Regina Lionnie	Universitas Indonesia

Soliman Mahmoud	University of Sharjah
John Joshua F. Montañez	Bicol State College of Applied Sciences and Technology
Kumar Rahul	XILINX
Sowmya Sanagavarapu	Anna University
Jiacheng Shang	Montclair State University
Mario E. B. G. N. Silva	University of Campinas
Shahab Tayeb	California State University, Fresno
Ashleigh Townsend	North-West University, Potchefstroom
Tri Minh Tran	Gunma University
Doina Bein	California State University, Fullerton
Aleksandr V. Belov	National Research University Higher School of Economics
Jason Brown	University of Southern Queensland
Bhabendu Kumar Mohanta	International Institute of Information Technology, Bhubaneswar
Gopal Venkata Tadepalli	College of Engineering, Guindy Campus
Soham Ghosh	University of Kansas
Kamran Hameed	Imam AbdulRahman Bin Faisal UNiversity
Maryam Heidari	George Mason University
Banage Kumara	Sabaragamuwa University of Sri Lanka
Sashank Sridhar	Anna University
Olusiji O Medaiyese	University of Louisville
Kanika Sood	California State University, Fullerton
Md Imtiaz Ahmed	Daffodil Institute of IT
Ahmed Ammari	INSAT - Carthage University Tunisia
Pratik Chattopadhyay	Indian Institute of Technology (BHU), Varanasi
Sangay Chedup	Jigme Namgyel Engineering College
Monica Isabel Costa	Polytechnic Institute of Castelo Branco
Ashiq Sakib	Florida Polytechnic University
Gameel Saleh	Imam Abdulrahman Bin Faisal University
Nikhilkumar Shardoor	MIT School of Engineering, MIT-ADT University
José Cornejo	Bioastronautics and Space Mechatronics Research Group
S Dhivya	VIT University
Rajib Kumar Halder	Jagannath Universuty
Rasha Kashef	Ryerson University
Banujan Kuhaneswaran	Sabaragamuwa University of Sri Lanka
Pravir Malik	Deep Order Technologies
Moses O. Onibonoje	Afe Babalola University, Ado Ekiti
Deep Roy	WBUT
Mrinal Sen	Indian Institute of Technology(ISM), Dhanbad
Qasem Abu Al-Haija	University of Petra (UoP)

Kehinde Adeniji	Afe Babalola University, Ado Ekiti
Vaibhav Anu	Montclair State University
Mouna Gassara	Sfax University
Dawei Li	Montclair State University
Nasar Aldian Shashoa	The Libyan Academy
Fadi Abusafat	University of Minho
Naheem Adesina	Louisiana State University
Sushree Bibhuprada B. Priyadarshini	SOA University Bhubaneswar
Harikrishna Bommala	KG Reddy College of Engineering and Technology Moinabad Hyderabad
Kevin Matthe Caramancion	University at Albany
John Hurley	Defense Intelligence Agency
Rania Majdoubi	Mohammed V University in Rabat
Arup Mohanty	SOA University Bhubaneswar
Asterios Mpatziakas	Centre for Research and Technology Hellas
Gedalia Nabe Razafindrobelina	ITS
Debabrata Singh	ITER, SOA University, Bhubaneswar
Mudrik Alaydrus	Universitas Mercu Buana
Enrico Angeles	Abu Dhabi Polytechnic
Narayana Darapaneni	Great Learning
Debashreet Das	GIET University Gunupur
Amin A. M. Fadlalla	King Fahd University of Petroleum & Minerals
Joaquin Gonzalez	Pontificia Universidad Catolica del Peru
Binayak Kar	National Taiwan University of Science and Technology
Asif Uddin Khan	Sillicon Institute of Technology, Bhubaneswar
Shahriar Khan	Independent University
Mohan K N	KLEF Deemed to be University
Sivaselvan N	Manipal Institute of Technology
Vikul Pawar	Government Engineering College Aurangabad (An Autonomous) M. S. INDIA
Mohammed Rajhi	University of California Santa Cruz
Santosh Kumar Sahoo	Trident Academy of Technology Bhubaneswar
Ahmed Shafkat	Fareast International University
Charles Shibu	Abu Dhabi Polytechnic
Sodessa Soma Shonkora	Arba Minch University
Bevek Subba	Jigme Namgyel Engineering College
Mohammad Nayeem Teli	University of Maryland
Sudarshan Kumar Babu Valluru	Delhi Technological University
Marvin S Verdadero	Bicol State College of Applied Sciences and Technology
Wael M.S Yafooz	Taibah University

Sponsors

- Society for Makers, Artists, Researchers and Technologists, Canada
- IEEE VANCOUVER SECTION
- IEEE TORONTO SECTION
- Institute of Engineering & Management, Kolkata
- University of Engineering & Management, Kolkata
- University of Engineering & Management, Jaipur

COPYRIGHT

2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).

CATALOG NUMBERS: CFP21Y72-ART

ISBN: 978-1-6654-4067-7

ORGANIZING COMMITTEE

General Chair:

Rajashree Paul

University of Engineering & Management, Kolkata, India

Technical Co-Chair:

Bob Gill

British Columbia Institute of Technology, Burnaby, Canada

Malay Gangopadhyay

Institute of Engineering & Management, Kolkata, India

Finance Chair:

Sanghamitra Poddar

Institute of Engineering & Management, Kolkata, India

Publicity Chair:

Fatima Hussain

Professor, Ryerson university, Canada, Editor IEEE Newsletter, IEEE Toronto section

ADVISORY COMMITTEE

Name	University
Dr. Chuck Easttom	University of Dallas, USA & Georgetown University, USA
Dr. Phillip Bradford	University of-Connecticut-Stamford, USA
Dr. Ronald F. DeMara	University of Central Florida, USA
Dr. Fatima Hussain	Professor, Ryerson university, Canada, Editor IEEE Newsletter, IEEE Toronto section
Dr. Ashutosh Datta	Johns Hopkins University, USA
Dr. Yang Hao	Queen Mary University, London
Dr. Vien Van	University of Alberta, Canada
Dr. Omar Ramahi	University of Waterloo, Canada
Dr. Yahia Antar	Royal Military College, Canada
Dr. Zhizhang (David) Chen	Dalhousie University, Canada
Dr. Detlef Streitferdt	Technische Universitat Ilmenau, Germany
Prof. Shahab Tayeb	California State University, Fresno.

TECHNICAL COMMITTEE

Name	University
Dr. Nabeeh Kandalaft	Grand Valley State University, USA
Dr. Alex "Sandy" Antunes	Capitol Technology University, USA
Dr. Izzat Alsmadi	Texas A&M, San Antonio, USA
Dr. Lo’ai Tawalbeh	Texas A&M University-San Antonio, USA
Dr. Pratik Chattopadhyay	Indian Institute of Technology (BHU), Varanasi
Dr. Doina Bein	California State University, Fullerton, USA
Dr. Hasan Yasar	Carnegie Mellon University, USA
Dr. Moises Levy	West Texas A&M University, USA
Dr. Christian Trefftz	Grand Valley State University, USA
Dr. Mrinal Sen	Indian Institute of Technology(ISM), Dhanbad
Dr. Petros Spachos	University of Guelph, Canada
Dr. Kanika Sood	California State University, Fullerton
Dr. Ke-Lin Du	Concordia University, Canada
Dr. Wenlin Han	California State University, Fullerton
Dr. Ashiq Adnan Sakib	Florida polytechnic University, USA
Dr. Morteza Modarresi Asem	Islamic Azad University, Iran
Dr. Md. Liakat Ali	Rider University, USA
Dr. Tarek El Salti	Sheridan College, Canada
Dr. Sukomal Dey	Indian Institute of Technology, Palakkad
Dr. Maysam Chamanzar	Carnegie Mellon University, USA
Dr. Kean Boon Lee	Sheffield University, UK

Track Topics:

IoT & Data Science:

- IoT and blockchain
- IoT and big data
- Next-generation infrastructure for IoT
- Cloud computing and IoT
- Edge computing and IoT
- IoT platforms, tools, and applications
- IoT systems development methodologies
- IoT applications

Electronics:

- Antenna and wireless communication
- Microwave Engineering
- Photonics
- Nano science & Quantum Technology
- VLSI and Microelectronic Circuit Embedded Systems
- System on Chip (SoC) Design
- FPGA (Field Programmable Gate Array) Design and Applications
- Electronic Instrumentations
- Sensors & Systems
- NEMS & MEMS
- Integrated circuits & power electronics
- Electronic Power Converters and Inverters
- Electric Vehicle Technologies
- Control Theory, Optimization and Applications
- Robotics and Autonomous Systems
- Intelligent,Optimal,Robust,Adaptive Control
- Linear and Nonlinear Control Systems
- Complex Adaptive Systems
- Industrial Automation and Control Systems Technology
- Modern Electronic Devices
- Biomedical devices & Imaging
- Energy Harvesting & Conversions
- Energy Efficient Hardware systems

Mechatronics:

- Sensing and Control Systems
- Mechatronics Systems
- Mechanical Systems
- Artificial Intelligence
- Applications of Robotics

Information Technology:

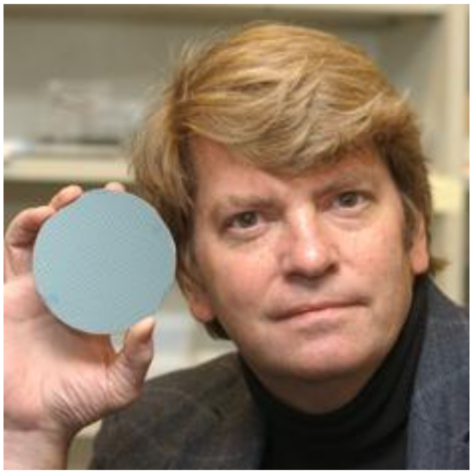
- Business Intelligence and Applications
- Computer Network
- Evolutionary Computation and Algorithms
- Intelligent Information Processing
- Information System Integration and Decision Support

- Image Processing and Multimedia Technology
- Signal Detection and Processing
- Technique and Application of Database
- Software Engineering
- Mobile Computing
- Distributed Systems
- Artificial Intelligence
- Visualization and Computer Graphic
- Natural Language Processing
- Deep Learning
- Machine Learning
- Internet of Things, Data Mining
- Data Science
- Cloud Computing in E-Commerce Scenarios
- E-Business Systems Integration and Standardization, E-government
- Electronic Business Model and Method
- E-Commerce Risk Management
- Recommender system
- Semantic Web Service Architecture for E-Commerce
- Service Oriented E-Commerce and Business Processes
- Data Analytics and Big Data
- Software defined networking
- Secured distributed systems

Mobile Communication:

- Ad hoc networks
- Body and personal area networks
- Cloud and virtual networks
- Cognitive radio networks
- Cyber security
- Cooperative communications
- Delay tolerant networks
- Future wireless Internet
- Local dependent networks
- Location management
- Mobile and wireless IP, Mobile computing
- Multi-hop networks
- Network architectures
- Network Security, Information Security, Encoding Technology
- Routing, QoS and scheduling
- Satellite communications
- Self-organising networks
- Telecommunication Systems
- Vehicular networks
- Wireless multicasting, Wireless sensor networks

Keynote speakers



Prof. Mark Arthur Reed

Professor, Department of Electrical Engineering, Yale University, USA

Bio: Mark Arthur Reed is an American physicist and professor at Yale University. He coined the term quantum dots, for demonstrating the first zero-dimensional electronic device that had fully quantized energy states. Reed does research in electronic transport in nanoscale and mesoscopic systems, artificially structured materials and devices, molecular electronics, biosensors and bioelectronic systems, and nanofluidics. He is the author of more than 200 publications, has given over 75 plenary and over 400 invited talks, and holds 33 U.S. and foreign patents on quantum effect, heterojunction, and molecular devices. He was the Editor in Chief of the journal Nanotechnology (2009-2019), is the present Editor in Chief of the journal Nano Futures, and holds numerous other editorial and advisory board positions.

Reed received his Ph.D. from Syracuse University in 1983. He worked at Texas Instruments from 1983 to 1990, where he demonstrated the first quantum dot device. He has been at Yale School of Engineering and Applied Science since 1990, where he holds the Harold Hodgkinson Chair of Engineering and Applied Science. Notable work there includes the first conductance measurement of a single molecule, the first single molecule transistor and the development of CMOS nanowire biosensors.

Reed has been elected to the Connecticut Academy of Science and Engineering and Who's Who in the World. His awards include; Fortune Magazine "Most Promising Young Scientist" (1990), the Kilby Young Innovator Award (1994), the Fujitsu ISCS Quantum Device Award (2001), the Yale Science and Engineering Association Award for Advancement of Basic and Applied Science (2002), Fellow of the American Physical Society (2003), the IEEE Pioneer Award in Nanotechnology (2007), Fellow of the Institute of Electrical and Electronics Engineers (2009), and a Finalist for the World Technology Award (2010).

Prof. Nicholas Bambos

Professor, Department of Electrical Engineering and Management Science, Stanford University, USA

Bio : Nick Bambos is a Professor at Stanford University, having a joint appointment in the Department of Electrical Engineering and the Department of Management Science & Engineering. He heads the Network Architecture and Performance Engineering research group at Stanford, conducting research in wireless network architectures, the Internet infrastructure, packet switching, network management and information service engineering, engaged in various projects of his Network Architecture Laboratory (NetLab). His current technology research interests include high-performance networking, autonomic computing, and service engineering. His methodological interests are in network control, online task scheduling, queueing systems and stochastic processing networks.

He has graduated over 20 Ph.D. students, who are now at leadership positions in academia (Stanford, CalTech, Michigan, GaTech, NYU, UBC, etc.) and the information technology industry (Cisco, Broadcom, IBM Labs, Qualcomm, Nokia, MITRE, Sun Labs, ST Micro, Intel, Samsung, TI, etc.) or have become successful entrepreneurs. From 1999 to 2005 he served as the director of the Stanford Networking Research Center, a major partnership/consortium between Stanford and information technology industries, involving tens of corporate members, faculty and doctoral students. He is now heading a new research initiative at Stanford on Networked Information Service Engineering.

He is on the Editorial Boards of several research journals and serves on various international technical committees and review panels for networking research and information technologies. He has been serving on the boards of various start-up companies in the Silicon Valley, consults on high technology development and management matters, and has served as lead expert witness in high-profile patent litigation cases in networking and computing.

Prof. Sanjay Lall

Professor, Department of Electrical Engineering, Stanford University, USA

Bio: Sanjay Lall is Professor of Electrical Engineering in the Information Systems Laboratory and Professor of Aeronautics and Astronautics at Stanford University. He received a B.A. degree in Mathematics with first-class honors in 1990 and a Ph.D. degree in Engineering in 1995, both from the University of Cambridge, England. His research group focuses on algorithms for control, optimization, and machine learning. From 2018 to 2019 he was Director in the Autonomous Systems Group at Apple. Before joining Stanford he was a Research Fellow at the California Institute of Technology in the Department of Control and Dynamical Systems, and prior to that he was a NATO Research Fellow at Massachusetts Institute of Technology, in the Laboratory for Information and Decision Systems. He was also a visiting scholar at Lund Institute of Technology in the Department of Automatic Control. He has significant industrial experience applying advanced algorithms to problems including satellite systems, advanced audio systems, Formula 1 racing, the America’s cup, cloud services monitoring, and integrated circuit diagnostic systems, in addition to several startup companies. Professor Lall has served as Associate Editor for the journal Automatica, on the steering and program committees of several international conferences, and as a reviewer for the National Science Foundation, DARPA, and the Air Force Office of Scientific Research. He is the author of over 130 peer-refereed publications.

Prof. Krishna Saraswat

Rickey/Nielsen Chair Professor, Stanford University, USA

Bio : Saraswat is working on a variety of problems related to new and innovative materials, structures, and process technology of silicon, germanium and III-V devices and interconnects for VLSI and nanoelectronics. Areas of his current interest are: new device structures to continue scaling MOS transistors, DRAMs and flash memories to nanometer regime, 3-dimentional ICs with multiple layers of heterogeneous devices, metal and optical interconnections and high efficiency and low cost solar cells.

Prof. Fakhri Karray

Professor, Department of Electrical and Computer Engineering, University of Waterloo, Canada

Bio: Prof. Fakhri Karray is the Loblaws Research Chair Professor in Artificial Intelligence in the department of electrical and computer engineering and is the founding co-director of the Waterloo AI Institute. He is the co-author of a textbook on applied artificial intelligence: Soft Computing and Intelligent Systems Design, Addison Wesley Publishing, 2004. He has authored extensively in his field of research (whether applied or theoretical), and has been issued 20 patents (US registered). He is the Associate Editor (AE) of the IEEE Transactions on Cybernetics, the IEEE Transactions on Neural Networks and Learning, and served as AE for the IEEE Transactions on Mechatronics, the IEEE Computational Intelligence Magazine. He also serves on the editorial board of the Elsevier Journal of Information Fusion, International Journal of Robotics and Automation, the Journal of Control and Intelligent Systems, and the Journal of Advances in Artificial Intelligence. Recent work of Fakhri and his research team's work on deep learning-based driver behaviour recognition and prediction has been featured on The Washington Post, Wired Magazine, Globe and Mail, CBC radio and Canada's Discovery Channel. He is a Fellow of the IEEE, a Fellow of the Canadian Academy of Engineering, a Fellow of the Engineering Institute of Canada and the President of the Association for Image and Machine Intelligence. He served as a Distinguished Lecturer for the IEEE and is a Fellow of the Kavli Frontiers of Science (a major research and symposium program of the US National Academy of Sciences)
Recent areas of research include:

Operational artificial intelligence and machine learning
Predictive analytics with application to virtual care
Multi-sensor data fusion
Cognitive robotics and autonomous machines
Smart mobility and big data analytics
Concept extraction and natural speech understanding

Prof. Raafat R. Mansour

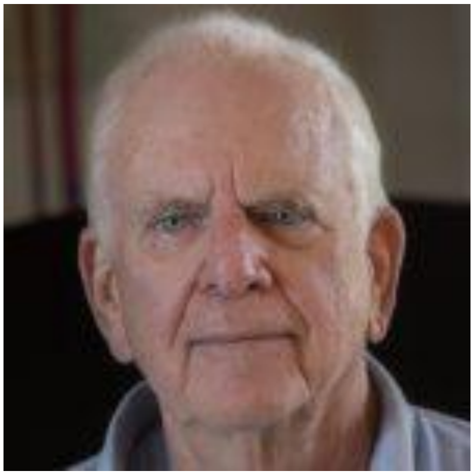
Professor, Canada Research Chair, Electrical and Computer Engineering Department, University of Waterloo, Canada

Bio: Dr. Mansour is a Professor of Electrical and Computer Engineering at the University of Waterloo and holds Tier 1 – Canada Research Chair (CRC) in Micro-Nano Integrated RF Systems. He held an NSERC Industrial Research Chair (IRC) for two terms (2001-2005) and (2006-2010). Prior to joining the University of Waterloo in January 2000, Dr. Mansour was with COM DEV Cambridge, Ontario, over the period 1986-1999, where he held various technical and management positions in COM DEV's Corporate R&D Department. Professor Mansour holds 37 US and Canadian patents and more than 380 refereed publications to his credit. He is a co-author of a 23-chapter Book published by Wiley and has contributed 6 chapters to four other books. Since joining the University of Waterloo in 2000, Professor Mansour has graduated 37 Ph.D, 32 M.Sc students and trained 14 Postdoctoral Fellows. His students hold key positions in academia and industry, including 5 holding faculty positions. Professor Mansour founded the Centre for Integrated RF Engineering (CIRFE) at the University of Waterloo https://uwaterloo.ca/centre-integrated-rf-engineering/. It houses a clean room and a state-of-the-art RF test and characterization laboratory. Professor Mansour has acted as a catalyst for ideas inspiring the next generation of Waterloo entrepreneurs to bring their work to market. Out of research carried out in his research Lab at the University of Waterloo, Professor Mansour and his graduate students co-founded two companies: AdHawk Microsystem http://www.adhawkmicrosystems.com/ and Integrated Circuit Scanning Probe Instruments (ICSPI-Corp) https://www.icspicorp.com/. Professor Mansour is a Fellow of IEEE, a Fellow of the Canadian Academy of Engineering (CAE), a Fellow of the Engineering Institute of Canada (EIC). He was the recipient of the 2014 Professional Engineers Ontario (PEO) Engineering Medal for Research and Development and the 2019 IEEE Canada A.G.L. McNaughton Gold Medal Award.

Prof. David Campbell

Professor, Department of Electrical and Computer Engineering, Boston University, USA

Bio: Professor David K. Campbell received his bachelor's degree in physics and chemistry from Harvard College in 1966, Part III Mathematics Tripos, with distinction, from Cambridge University in 1967, and his Ph.D. in theoretical physics and applied mathematics from Cambridge in 1970. He has pioneered the systematic study of inherently nonlinear phenomena throughout physics. The central theme of his work is the role of nonlinear excitations—solitons—in novel states of matter. His contributions span many distinct subfields of physics from high-energy field theory to condensed matter. Professor Campbell is a leader in the emerging field of nonlinear science. His influential overview articles and his direction of the flagship journal Chaos, of which he was the founding editor, have established key interdisciplinary organizing principles—the paradigms of solitons, chaos, and patterns—and have played a seminal role in defining the research agenda in nonlinear science.



Steven J. Davis

Founder, Golem Labs, USA

Bio : Steve Davis received his Bachelor of Electrical Engineering at the City College of New York, and a master's degree of Electrical Engineering from Drexel University in 1968, specializing in information theory and signal processing. His started his career doing research and development of highly compressed speech communication systems used in aircraft and later in deep space voice communications. He then was recruited by General DataComm Industries in 1978 to convert the original Bell modems and line interface devices, from mechanical logic devices to miniaturized solid state modems. This was achieved using custom integrated circuits with analog and digital elements on the same substrate.

Steve then in 1983, moved on to Warner Communications to be the chief technologist in implementing the first two-way interactive cable network for Warner Cable. After the completion of the system, Steve headed west to be director of the Atari Advanced Research Lab , located on the Burbank Studio lot. Here the first LAN network was developed for home computers using the Atari 800 device. An interactive video disk system was presented in Paris showing how the combination of video and a consumer computer could be used to create unique experiences. This work was part of the many projects Mr. Alan Kay directed.
Steven then started his own venture, Golem Labs . At Golem Labs a wireless games and data systems were developed using FM Sub Carrier technology. Steve has since moved on to bring several conceptual systems to reality. Navigation devices used in tunnels, hand held IED jamming, focused energy beams , wireless watering controllers, miniature ferrite antenna systems, planar antennas, and stealth communication under a DARPA contract. Presently Steve is consulting on secure cloud computing systems, and on secure handheld financial devices that are Quantum proof. He holds 14 patents.

Dr. Bonita Bhaskaran

Principal Engineer, NVIDIA, USA

Bio : With 13+ years of VLSI Design industry experience, **Dr. Bonita Bhaskaran** have worked cross-functionally on a few high-visibility projects @NVIDIA. A strong believer of "Where there is a will, there is way", she has not let impediments stop her in solving tough engineering problems. During her career span, a diverse background in Asynchronous Design, Design for Test, Signal Integrity, Power Integrity and Low Power VLSI has honed her skills as a Domain Expert in Low Power DFT.

Specialties – DFT, ATPG Tools, Verification, Silicon Power Measurements, Die-Pkg Co-Simulation, On-Chip Power Integrity

Prof. Sean Follmer

Department of Mechanical Engineering and Computer Science, Stanford University, USA

Bio : Sean Follmer is an Assistant Professor of Mechanical Engineering and Computer Science (by courtesy) at Stanford University. His Research in Human Computer Interaction, Haptics, and Human Robot Interaction explores the design of novel tactile physical interfaces and novel robotic devices. Dr. Follmer directs the Stanford Shape Lab and is a faculty member of the Stanford HCI Group. He is a core faculty member of the Design Impact masters program focusing on innovation and human centered design at Stanford.

Dr. Follmer received a PhD and a Masters from the MIT Media Lab in 2015 and 2011 (respectively) for his work in human-computer interaction, and a BS in Engineering with a focus on Product Design from Stanford University. His talk featured on TED.com was named one of the best science and tech TED talks of 2015 and has been viewed more than 1.4 million times. He has received numerous awards for his research and design work such as Best Paper Awards and nominations from premier conferences in human-computer interaction (ACM UIST and CHI conferences), Fast Company Innovation By Design Award, Red Dot Design Award, and a Laval Virtual Award. His work has been shown at the Smithsonian Cooper Hewitt Design Museum, Ars Electronica Center, and the Milan Design Week. Dr. Follmer also leads workshops and executive education around design thinking and innovation.

Content:

Algorithmic method of security of the Internet of Things based on steganographic coding

1st Anvar Kabulov
dept. of Information Security
National University of Uzbekistan
Tashkent, Uzbekistan
anvarkabulov@mail.ru

2nd Islambek Saymanov
dept. of Information Security
National University of Uzbekistan
Tashkent, Uzbekistan
orcid.org/0000-0003-3530-4488

3rd Inomjon Yarashov
dept. of Information Security
National University of Uzbekistan
Tashkent, Uzbekistan
orcid.org/ 0000-0002-0855-3318

4th Firdavs Muxammadiev
dept. of Information Security
National University of Uzbekistan
Tashkent, Uzbekistan
firdavsmukhammadiyev@gmail.com

***Abstract*—In the Internet of Things, it is more important than ever to effectively address the problem of secure transmission based on steganographic substitution by synthesizing digital sensor data. In this case, the degree to which the grayscale message is obscured is a necessary issue. To ensure information security in IoT systems, various methods are used and information security problems are solved to one degree or another. The article proposes a method and algorithm for a computer image in grayscale, in which the value of each pixel is one sample, representing the amount of light, carrying only information about the intensity. The proposed method in grayscale using steganographic coding provides a secure implementation of data transmission in the IoT system. Study results were analyzed using PSNR (Peak Signal to Noise Ratio).**

***Index Terms*—IoT, steganography, LSB, information security, grayscale metod, PSNR, stego image, histogram, RGB.**

I. Introduction

Internet of Things (IoT) incorporate Smart devices, sensor networks and wearable devices with the reason of trading data and administrations while sensor networks are the key for making smart environments. IoT systems are developing quickly due to the quick increase of remote networks and improved range of sensing devices. IoT technology deals with millions and billions of sensing objects, machines and virtual entities that interact with each other [1]. A new forecast from International Data Corporation (IDC) estimates that there will be 41.6 billion connected IoT devices, or ”things”, generating 79.4 zettabytes (ZB) of data in 2025 [2]. The Internet has become the most convenient and effective means of communication. Over the Internet, messages can be transmitted quickly and cheaply in various areas, such as government agencies, the private sector, the military, and medical fields. In many cases, the confidentiality of the transmitted message must be preserved [3]. To guarantee that the message is exchanged safely and securely over the network, an appropriate method is required. Steganography demonstrates as a trust able method for accomplishing this aim. In steganography, the information are covered up within the cover media. The cover medium can be within the form of image file, text record, video record, or sound file. Steganography is characterized as a science or craftsmanship of covering up the message interior a few covers medium (Figure 1). The word stenography is built up of two words of ancient Greek origin steganos meaning covered, concealed, or protected and ”graphie” meaning ”writing” [4]. The concept of steganography isn’t unused; its utilization can be seen within the past moreover. Chronicled records delineate that around 440 BC, Herodotus sent secret messages utilizing the concept of steganography.

IoT is the technology of the decade which moreover in turn has parts of potential focuses for a breach of data. This article explores the security issue and proposes an improved steganography scheme for IoT devices. The advancement of the IoT and inescapable computing makes a significant affect within the data and communication. The development rate of the gadgets appeared to be climbed in exponential way. In the general objective appears to be giving superior administrations for the conclusion client with the development of the new age gadgets. Steganography exists the total of conceal a file, message, image, or video interior another file. The benefit of steganography in overload of cryptography alone be to the planned secret communication doing not attract attention to itself as an object of security [5]. Doubtlessly able to be lively see encrypted messages, no matter how unbreakable they are, stimulate intrigued and may in themselves be there implicating in countries in which encryption is illicit. This investigates proposes a method, which can cover up the secret information in image layers, and steganography for IoT. The proposed technique is LSB substitution ciphers for the IoT. Experiments are conducted with different aspect ratio images, which show that proposed algorithms seem to work better. The proposed method performs well in specifically in RGB color model. The simple LSB method seems to have a lack which is rectified with the proposed method Vacillating LSB in terms of the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

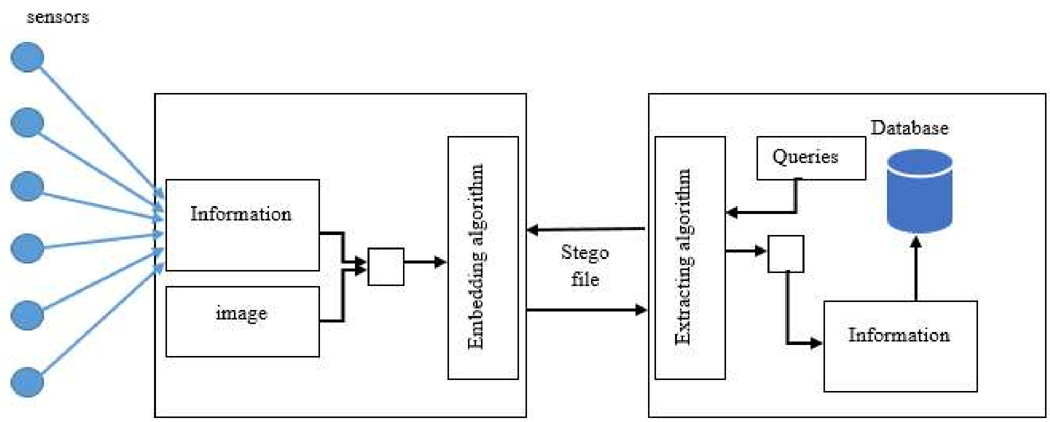


Fig. 1. Original pictures in BMP and JPEG format

MSE and PSNR. 3-3-2 LSB image steganography is analyzed using Grayscale methods.

II. Schematic of the RGB color model

Typically, a graphics palette consists of three color components such as red, green and blue. Each of the components has 256 shades. In total, all three components form a palette of 16,777,216 colors, which is a 24-bit color. Let's consider one of the color components. The shades of the color change linearly in accordance with the growth of the value itself, hence the close colors differ in the values of the least significant bits. The human eye as a whole does not distinguish between minimal differences in color shades. Based on these facts, if, to manipulate the values of the least significant bits of the three color palettes, then the whole picture will not change. It is suggested to use the three low-order bits of the red and green component and the two low-order bits of the blue component to write one message byte.

TABLE I
8 bits representing color

Bit	7	6	5	4	3	2	1	0
Data	R	R	R	G	G	G	B	B

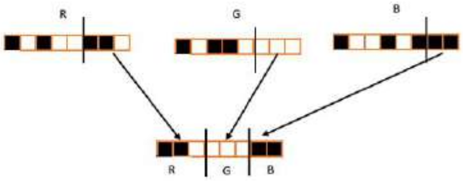


Fig. 2. 8-bit color, with 3 bits of red, 3 bits of green, and 2 bits of blue

A 24-bit color image is taken into account the most effective in accordance with the definition of RGB color model during which every color shows as in its primary spectral component of red, green and blue. This model is based on Cartesian coordinate system as shown in Figure 3. In this way, the RGB essential values are laid in three corners. The secondary colors acknowledge as cyan, magenta, and yellow, they're targeted at 3 alternative corners. Black color is at the origin and white is at the farthest corner from the origin. Equal values embrace red, green, blue are consisted the line that links 2 corners. Therefore, this produces various shades of grey.

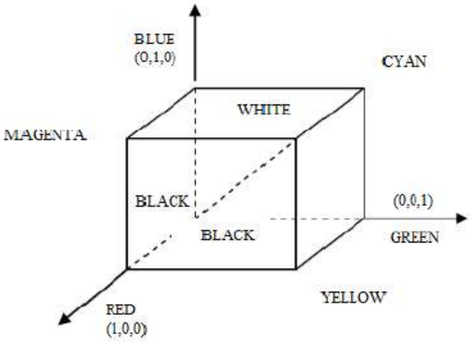


Fig. 3. Schematic of the RGB color model

The locus of these points is called the gray line. In fact, every pixel within the RGB model composes of RGB values. So, every of those colors needs 8-bit for its illustration. Consequently, every pixel signifies by twenty four bits in total. So the sum of possible colors with 24-bit RGB image reaches $(2^8)^3 = 16,777,216$ [6].

III. Gray scale Methods

The use of three-component RGB colors in BMP image format for graphics is not rational. More preferable to use JPEG format. The JPEG format has two key features. The first is the use of color subsampling, the second is the discrete cosine transform. At the same time, if we go to a single-component color, for example, a shade of gray, then in the background of the peculiarities of the JPEG format, further operations with the picture take place at the brightness level as shown in Figure 2 [7]. For example, the picture below with dimensions of 256x256 pixels in BMP format has a volume of 192 kB, and the same picture in grayscale JPEG format has a volume of 16.2 KB.

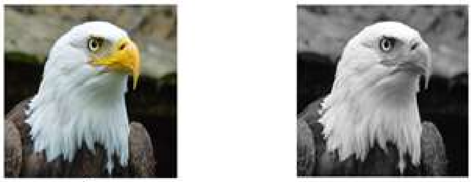


Fig. 4. Original pictures in BMP and JPEG format

To write a message to a grayscale picture, a similar method is used as for color. Three image pixels are used to write one byte of the message, in the ratio of 3 least significant bits + 3 least significant bits + 2 least significant bits. With this approach, a 10 kB message of English text can be placed in the specified picture.

IV. Proposed Algorithm

Let N be the image consist of $H * W$ elements, C be the M-bite safes note, y be the elements value of N and C be the bite of safes note, then the picture set of values can be show on (1) and C can be show on (2) [8]

$$N = \{y_{ij} | 1 \leq i \leq H, 1 \leq j \leq W, y_{ij} \in \{0, 1, ..., 255\}\} \quad (1)$$

$$C = \{C_M | 1 \leq M \leq m, C_M \in \{0, 1, ..., 255\}\} \quad (2)$$

$$L = \{255 * l_w + l_r | 0 \leq l_w \leq 255, 0 \leq l_r \leq 255\} \quad (3)$$

Let C is the note to be covert, Y is the picture, L is the length of notes (3), E and O are the encoding and decoding algorithms, accordingly and Y' is the steganographic document (4). Perhaps the encoding event is given by the resulting equation:

$$Y' = E(C, Y, L) \quad (4)$$

The encoding algorithm is show in Block scheme 1. At the other end, the inverse event is executed and the note is extracted applying the algorithm in Block scheme 2. The note is extract from the picture by (5). Perhaps the decoding event is given by the resulting equation:

$$X = O(Y', L) \quad (5)$$

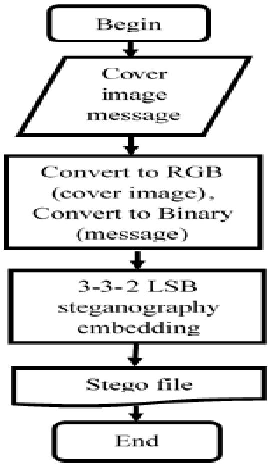

Begin
Cover image message
Convert to RGB (cover image), Convert to Binary (message)
3-3-2 LSB steganography embedding
Stego file
End

Fig. 5. Embedding algorithm

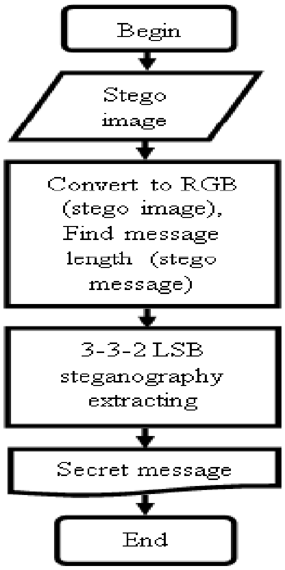

Begin
Stego image
Convert to RGB (stego image), Find message length (stego message)
3-3-2 LSB steganography extracting
Secret message
End

Fig. 6. Extracting algorithm

V. Experimental Results

The performance is tested on the base of two parameters, that is, PSNR and MSE [8]. The most effective grayscale method should be identified:

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (y_{ij} - {y'}_{ij})^2 \quad (6)$$

Where H and W show the measurements of the picture set of values y_{ij}, shows the original picture and ${y'}_{ij}$ shows the steganographic picture.

$$PSNR = 10 \log[\frac{255^2}{MSE}] \quad (7)$$

TABLE II
ANALYSIS OF THE EFFECTIVENESS OF METHODS USING PICTURES. SIZE OF THE INFORMATION: 2KB

Picture name	**PSNR**	**PSNR (GM1)**	**PSNR (GM2)**	**PSNR (GM3)**	**The most PSNR number**
	1	2	3	4	
Eagle	59.328	59.438	59.410	59.547	4
Cameraman	59.251	59.286	59.275	59.315	4
Climbing	59.500	59.528	59.356	59.388	2
Couple	59.129	59.422	59.395	59.458	4
Girl	59.383	59.588	59.266	59.471	2
Mountain	59.460	59.531	59.396	59.411	2
Average	59.342	59.466	59.350	59.432	

TABLE III
ANALYSIS OF THE EFFECTIVENESS OF METHODS USING PICTURES. SIZE OF THE INFORMATION: 4KB

Picture name	**PSNR**	**PSNR (GM1)**	**PSNR (GM2)**	**PSNR (GM3)**	**The most PSNR number**
	1	2	3	4	
Eagle	56.349	56.409	56.284	56.493	4
Cameraman	56.319	56.327	56.299	56.345	4
Climbing	56.300	56.432	56.431	56.302	2
Couple	56.264	56.406	56.360	56.422	4
Girl	56.273	56.430	56.212	56.385	2
Mountain	56.329	56.331	56.352	56.337	3
Average	56.306	56.389	56.323	56.381	

TABLE IV
ANALYSIS OF THE EFFECTIVENESS OF METHODS USING PICTURES. SIZE OF THE INFORMATION: 8KB

Picture name	**PSNR**	**PSNR (GM1)**	**PSNR (GM2)**	**PSNR (GM3)**	**The most PSNR number**
	1	2	3	4	
Eagle	53.328	53.388	53.313	53.383	1
Cameraman	53.225	53.306	53.374	53.304	3
Climbing	53.331	53.414	53.336	53.282	2
Couple	53.258	53.456	53.331	53.361	2
Girl	53.271	53.459	53.175	53.274	2
Mountain	53.332	53.318	53.395	53.354	3
Average	53.310	53.390	53.321	53.326	

TABLE V
ANALYSIS OF THE EFFECTIVENESS OF METHODS USING PICTURES. SIZE OF THE INFORMATION: 10KB

Picture name	**PSNR**	**PSNR (GM1)**	**PSNR (GM2)**	**PSNR (GM3)**	**The most PSNR number**
	1	2	3	4	
Eagle	52.480	52.423	52.367	52.420	1
Cameraman	52.365	52.340	52.408	52.358	3
Climbing	52.393	52.455	52.377	52.311	2
Couple	52.292	52.452	52.369	52.408	2
Girl	52.305	52.480	52.243	52.271	2
Mountain	52.326	52.361	52.400	52.367	3
Average	52.360	52.419	52.361	52.356	

VI. HISTOGRAM RESULTS FOR THE IMAGES

Multi-image histogram results are displayed using the suggested grayscale methods using an 8KB note. The original picture and steganographic picture together with their appropriate histograms are shown in Figures 7-14.

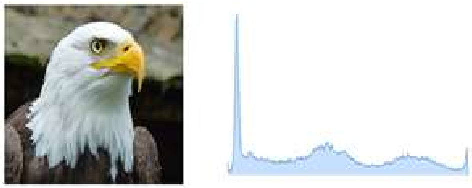


Fig. 7. Original picture and histogram of eagle

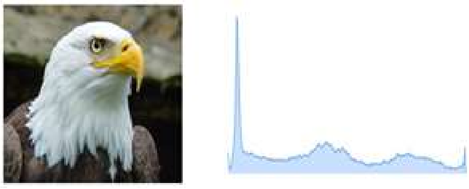


Fig. 8. Steganographic picture and histogram of eagle

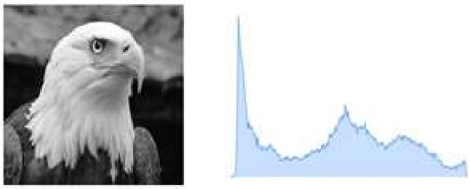


Fig. 9. Original (Grayscale Method1) picture and histogram of eagle

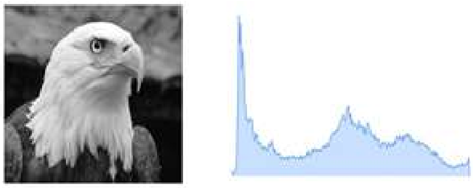


Fig. 10. Steganographic (Grayscale Method1) picture and histogram of eagle

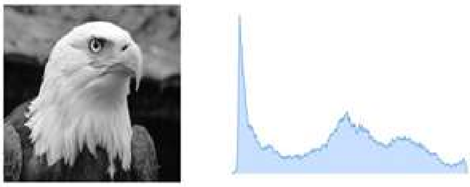


Fig. 11. Original (Grayscale Method2) picture and histogram of eagle

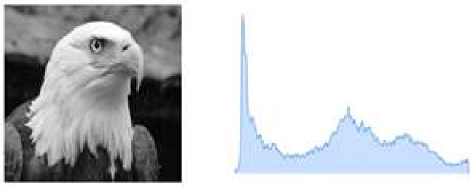


Fig. 12. Steganographic (Grayscale Method2) picture and histogram of eagle

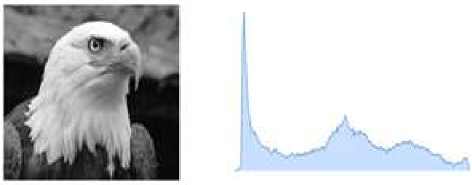


Fig. 13. Original (Grayscale Method3) picture and histogram of eagle

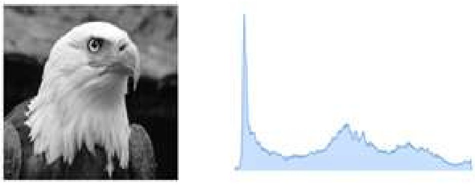


Fig. 14. Steganographic (Grayscale Method3) picture and histogram of eagle

VII. Conclusion

This article explores the security issue and proposes an improved steganography scheme for IoT devices. The paper proposes a steganography method using an improved LSB (least significant bit) replacement algorithm for a 24-bit color image capable of creating a hidden embedded image that is completely indistinguishable from the original image by the human eye. The steganographic grayscale technique allows data to be hidden within the grayscale medium. Moreover, each pixel stores eight message bits inside a pixel, while other methods, such as LSB, allow only two message bits inside a pixel. A fairly simple mathematical bit comparison function is used, which meets the basic requirement of steganography, which consists in sending a secret message inside the carrier of the image without creating much difference from the original image. Finally, the proposed steganography method provides a good PSNR.

References

[1] Z. Safdar, S. Farid, M. Pasha, and K. Safdar, “A security model for iot based systems,” *Technical Journal. University of Engineering and Technology (UET)*, vol. 22, no. 33, pp. 74–75, 2017.

[2] A. Kabulov, E. Urunboev, and I.Saymanov, “Object recognition method based on logical correcting functions,” in *Proc. IEEE International Conference on Science and Communications technologies Applications, Trends and Opportunities (ICISCT 2020)*, Tashkent, Uzbekistan, 2020, pp. 1–4.

[3] A.Kabulov, I.Normatov, A.Karimov, and E.Navruzov, “Algorithm of constructing control models of complex systems in the language of functioning tables,” *Advances in Mathematics: Scientific Journal*, vol. 9, no. 12, pp. 10 397–10 417, 2020.

[4] K. Joshi, S. Gill, and R. Yadav, “A new method of image steganography using 7th bit of a pixel as indicator by introducing the successive temporary pixel in the gray scale image,” *Journal of Computer Networks and Communications*, vol. 8, pp. 1–10, 2018.

[5] M. C and S.Prema, “Instigating improved steganography scheme for internet of things,” *International Advanced Research Journal in Science, Engineering and Technology*, vol. 5, no. 10, pp. 83–87, 2018.

[6] M. A. Majeed and R. Sulaiman, “An improved lsb image steganography technique using bit-inverse in 24 bit colour image,” *Journal of Theoretical and Applied Information Technology*, vol. 80, no. 2, pp. 342–348, 2015.

[7] A.Kabulov, I.Normatov, Sh.Boltaev, and I.Saymanov, “Logic method of classification of objects with non-joining classes,” *Advances in Mathematics: Scientific Journal*, vol. 9, no. 10, pp. 8635–8646, 2020.

[8] K. Joshi, R. Yadav, and S. Allwadhi, “Psnr and mse based investigation of lsb,” in *Proc. IEEE International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT 2016)*, New Delhi, India, 2016, pp. 1–4.

A Hybrid Blockchain Consensus Algorithm Using Locational Marginal Pricing for Energy Applications

Soham Ghosh
School of Engineering
University of Kansas
Overland Park, KS, USA
sghosh27@ieee.org, ORCID: 0000-0002-6151-8183

Raj Sekhar Dutta
College of Computing
Illinois Institute of Technology
Chicago, IL, USA
rdutta2@hawk.iit.edu, ORCID: 0000-0001-8647-1521

***Abstract*—Blockchain technology has recently witnessed a rapid proliferation across multiple industries and has the potential to evolve as a popular energy transaction platform for grid operators and general users. The current proof-of-work consensus algorithm used by many blockchain protocols suffer from vulnerabilities such as the majority attacks problem and strip mining. The algorithm often requires specialized application-specific hardware and involves energy-intensive computation, making it an unlikely candidate for power grid applications. On the other hand, the proof-of-stake consensus algorithm is susceptible to the well-known nothing-at-stake vulnerability issue, making it equally unlikely to be used in large-scale secured energy transactions. An energy-efficient locational marginal pricing-based hybrid proof-of-work, proof-of-stake consensus algorithm is proposed to mitigate such risks. The hybrid consensus algorithm uses an optimized market-based energy pricing model to select the mining and forging nodes. A detailed framework of the hybrid consensus algorithm is provided, along with an overview of its potential application to record large volumes of energy transactions and execute obligatory smart-contract agreements.**

Keywords—Energy blockchains, proof of work, proof of stake, hybrid consensus, locational marginal price, smart-contract, initial coin offering.

I. Nomenclature

Sets, indices, and parameters:

g	set of generators
n, m	set of nodes
k	set of branches (lines and/or transformers)
$k_{(\leftarrow n)}$	set of branches with n as the incoming node
$k_{(n\rightarrow)}$	set of branches with n as the outgoing node
P_k^{max}	maximum thermal rating of the branch k
P_g^{max}	maximum thermal rating of the generator g
C_g	the variable production cost of generator g
P_n	real power load at node n
Im(Y)	the imaginary part of admittance

Variables:

P_g	real power supply ∀ g
P_k	real power-flow through branch ∀ k
θ_m, θ_n	the voltage at node m, n
$\tau, \tilde{\alpha}, \overline{\alpha}, \overline{\overline{\beta}}, \overset{\frown}{\tilde{\beta}}$	dual variables corresponding to the primal constraints

II. Introduction

Blockchain technology has rapidly evolved ever since it was first introduced in 2008, with protocols as Bitcoin and Ethereum gaining significant popularity. In the original whitepaper [1], the author/working group demonstrated the concept of an immutable blockchain ledger that can be used to record timestamped network transactions using the hash-based proof-of-work (PoW) algorithm. The PoW algorithm searches for a unique nonce value in a block that will produce a definite number of leading zeros as defined by the protocol's difficulty level when hashed. The difficulty level of the cryptographic problem is periodically increased to compensate for the effects of new hardware being introduced with increasing hashing power. Increasing the level of difficulty over the years has caused the establishment of colossal blockchain mining pools, consuming enormous energy, significant enough to congest electrical flow-gates and affect energy prices at specific electrical nodes. As reported in [2], Bitcoin's energy consumption's upper bounds are estimated to be 125 TWh. Current 2020 projections indicate the Bitcoin network operating on the PoW algorithm at its core consumes energy at an amount comparable to the yearly consumption of the Netherlands [3]. In addition to the outrageous energy consumption issue, the PoW algorithm suffers from well-known vulnerabilities such as the 51% attack and strip mining [4, 5].

The proof-of-stake (PoS) consensus algorithm relies on validators staking their cryptocurrency token or voting tickets to improve their chances of being selected as the forger of the next block using a pseudo-random process. Chances of selection as the next forger in the pseudo-random selection process is linearly proportional to the amount each validator puts up as a stake. This linear relationship prevents large stakeholders from gaining competitive advantage through economies of scale, as often witnessed in the PoW consensus algorithm [6].

Though the PoS consensus algorithm has its merit, it still suffers from the nothing-at-stake problem, which exists whenever there is an accidental or malicious fork in the blockchain. Given that forging a block doesn't incur any significant cost, and a fork creates an equivalent amount of stake on the new blockchain, the block creators often have the incentive to create blocks on both the forks. The nothing-at-stake vulnerability could potentially result in a double-spend attack. The solution offered in most cases for the nothing-at-stake problem centers around an identity-based [7] or a slashing-based [8] staking mechanism.

There is a clear gap in the availability of a blockchain consensus algorithm for energy application that is efficient, relatively secured, and optimized based on the electric grid's

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

topology. In response to the limitations of both the PoW and PoS consensus algorithms and accounting for the network topology in terms of the generation, load, and the thermal limits of the transmission grid, a hybrid PoW-PoS consensus algorithm framework is proposed. This hybrid consensus algorithm selects mining and forging nodes based on the locational marginal pricing (LMP) of electricity and overcomes individual consensus algorithms' limitations. The proposed hybrid framework's energy efficiency and security features can be implemented for wide-scale energy applications, including mandatory smart-contract agreements.

The remainder of the paper is organized as follows: Section III provides a comprehensive discussion of the two most popular consensus mechanisms and their relative merits; section IV highlights the hybrid PoW-PoS consensus algorithm in conjunction with the LMP based nodes selection. This section also highlights necessary design considerations to be made in the blockchain protocol layer in terms of reward and difficulty adjustments, application-specific integrated circuit (ASIC) resistance, and selection of hashing algorithm, among other issues. The application of the proposed hybrid consensus in energy blockchain transactions and smart contract applications are discussed in section V. Finally, concluding remarks are provided in section VI.

III. Popular Consensus Algorithms in Use

There are several consensus algorithms documented in the blockchain technology literature [9, 10, 11], mostly serving two primary functions: guaranteeing that the proposed new block is the only valid version of the chain and preventing users with malicious intent from manipulating the records in the blockchain or introducing an unintentional chain fork.

Besides PoW and PoS, the two most popular consensus algorithm, some of the other less known algorithms include proof-of-burn (PoB), proof-of-capacity (PoC), delegated proof-of-stake (dPoS), and Cardano Ouroboros. Given the current trends and the relative merits, future consensus algorithms would likely be a hybrid or standalone derivative of either the PoW or PoS algorithm, and hence they require further discussion.

A. Proof-of-Work Algorithm

The PoW algorithm relies on proof that an adequate amount of resources in terms of computing power and energy has been spent to propose a value acceptable by the distributed network. This algorithm is popular with blockchain protocols as Bitcoin, Bitcoin Cash, Litecoin, and Monero. Most of these blockchain protocols implement a periodic difficulty adjustment algorithm (DAA) of the hashing target (h_{max}), effectively readjusting the feasible solution-space proportional to the total hashing power across the network. For a message (m) to be incorporated in a block, a nonce (n) is computed such that the hashed output (h_{output}) is below the target difficulty and (1) holds. Most of the cryptographically secured hashing functions commonly used are variants of the secured hash algorithm (SHA) [12]. The popular SHA-256 hashing algorithm yields a 64 character hexadecimal output, and the addition of every single leading zero as a target constraint effectively reduces the feasible solution space by sixteen times.

$$hash(m,n) = h_{output}\,; h_{output} < h_{max} \tag{1}$$

1) Advantages of the Proof-of-Work Algorithm

One of the great strengths of the PoW consensus algorithm is that the smallest change in the input results in a massive uncorrelated change in the hash output, giving all attempts equal probability. Also, the PoW consensus algorithm is a Poisson process [13], given the probability of the next block generated within a specific timeframe can be statistically predicted using the Poisson distribution function. A nonnegative random integer value (X), with an intensity parameter (λ), is said to follow a Poisson distribution if (2) holds.

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!};\ x = 0,1,\ldots \tag{2}$$

To demonstrate the nature of the discrete-event distribution, consider, on average, a single Bitcoin block is mined every ten minutes (λ = 0.1). Under a Poisson process, one can estimate using the generalization from (2) other outcomes, such as the probability of at least two blocks being mined in the next minute. As shown in (3), the event has a probability of 0.00468.

$$\begin{aligned} P(\text{at least two blocks in the next minute}) &= P(X \geq 2) \\ &= 1 - P(X < 2) \\ &= 1 - \frac{e^{-\frac{1}{10}}\left(\frac{1}{10}\right)^{0}}{0!} - \frac{e^{-\frac{1}{10}}\left(\frac{1}{10}\right)^{1}}{1!} \\ &= 0.00468 \end{aligned} \tag{3}$$

Another strength of the PoW consensus algorithm is its resistance to Sybil attacks [14], a type of attack where a single malicious user creates several fake identities on the system to gain influence. As long as the malicious user is not in control of more than 50% of the network's hashing power, any attempts to add a fake block will be rejected due to an insufficient amount of work.

2) Vulnerabilities of the Proof-of-Work Algorithm

The most well-documented vulnerability of the PoW consensus algorithm is the 51% attack, where a single malicious miner or mining group controls more than 50% of the network's hashing power. Having this dominant hashing power will allow the malicious party to mine and verify fraudulent transaction blocks at a rate faster than the regular miner, potentially leading to a gamed double-spend transaction scenario. This issue can be further aggravated, given that most mining pools for popular blockchain protocols are centralized in regions where there is access to cheap electric power. At present, no regulations prevent the biggest mining pools from merging to trigger a 51% attack.

As documented in [4, 15], another vulnerability of the PoW consensus algorithm is the issue of strip mining. Big mining pools can effectively game the blockchain protocol's current difficulty level to their advantage by timing their network participation close to the difficulty adjustment timeframe. For instance, a mining pool with significant hash power may quit from the network soon after a difficulty target readjustment, resulting in a significant increase in the mining time for the next blocks, and leading to a backlog of unconfirmed transactions, effectively chocking out the system.

B. Proof-of-Stake Algorithm

First introduced by Peercoin, the proof-of-stake consensus algorithm relies on the fact that well-invested users have a sufficient degree of vested interest in the network to prevent them from any malicious attempt. Under a PoS consensus, the probability of selecting a node as the next forger is linearly proportional to its staked cryptocurrency tokens (as in NXT) or voting tickets (as in Decred).

1) Advantages of the Proof-of-Stake Algorithm

The PoS algorithm is not computationally intensive and is thus resistant to centralization in regions where electric and hardware resources are cheaply available. Further, as block verification under a PoS algorithm doesn't follow a Poisson process, the algorithm can be scaled up to support much larger transaction volumes. Ethereum, which currently uses the PoW consensus algorithm, has proposed an iterative upgrade to Serenity by 2022. Serenity is a modified version of PoS [16] to support the platform's scalability.

2) Vulnerabilities of the Proof-of-Stake Algorithm

In PoW, the miners have an incentive to mine on the longest chain because of the higher likelihood of adding the next block to the longest chain. However, in the PoS consensus algorithm, under a theoretical gaming scenario, forgers will have the incentive to vote on all forks, to maximize their chances of receiving the transaction fee. This issue is referred to as nothing-at-stake vulnerability. Two of the well-known mitigation strategies implemented are:

- Peercoin [17]: Introduced the concept of coinage, which is consumed upon minting a block. Successful block validators need to wait for thirty days before their coins in the wallet requalify.
- Ethereum Slasher 2.0 [18]: Introduced a voter penalty concept for voting on the wrong fork.

IV. Hybrid Consensus Algorithm

The drawbacks of both the popular PoW and PoS consensus algorithm for energy application demands the development of an alternative consensus mechanism that is more energy-efficient and provides better security against majority attacks and systematic gaming. The PoW and PoS consensus algorithm's strengths can be combined to create a hybrid model where the PoW mining nodes are responsible for mining a new block for validation, and the PoS validators vote on each candidate block to be incorporated into the chain.

A crucial consideration in selecting the mining and validating nodes is the locational marginal price (LMP) of electricity at the grid's major electrical nodes. In a deregulated energy market, electric prices at different nodes often experience a price separation due to congestion in the grid. Economic benefits can be realized if nodes with lower real-time LMP are selected for mining the blocks using PoW consensus, while the high the higher-priced nodes may act as PoS validators. The following section presents a mathematical derivation of LMP from an optimal power-flow (OPF) formulation.

A. Locational Marginal Price Based Nodes Selection

The concept of LMP (also referred to as spot pricing) was originally introduced in [19], and is usually derived from either the AC or DC model of the OPF formulation. However, the existence of strong NP-hardness of the AC-OPF problem [20] and its divergence propensity [21] makes the DC-OPF model a better choice and is widely used in commercial power market software. Many variations of the DC-OPF are available in the literature [21, 22, 23, 24, 25], and the dual interpretation changes are subjected to the primal problem formulation. The below DC-OPF formulation draws its inspiration from [21, 22, 26], intending to develop the various components of LMP.

1) Locational Marginal Price Computation

A DC approximation of the AC-OPF problem typically involves removing the reactive power-flow component from problem formulation and assuming a small voltage angle difference sufficient for linear approximation [21]. These approximations yield the following DC-OPF formulation (4)-(9).

Objective function: $$min \sum_{\forall g}(C_g P_g) \qquad (4)$$

Subject to the following constraints:

$$\sum_{\forall k_{(\leftarrow n)}} P_k - \sum_{\forall k_{(n\rightarrow)}} P_k + \sum_{\forall g} P_g = P_n \,\forall n \qquad (\tau) \qquad (5)$$

$$-P_k \geq -P_k^{max} \,\forall k \qquad (\tilde{\alpha}) \qquad (6)$$

$$-P_k \geq P_k^{max} \,\forall k \qquad (\bar{\alpha}) \qquad (7)$$

$$Im(Y)\,(\theta_n - \theta_m) - P_k = 0 \qquad (\bar{\bar{\beta}}) \qquad (8)$$

$$-P_g \geq -P_g^{max} \qquad (\overbrace{\beta}) \qquad (9)$$

The objective function and optimization constraints of the DC-OPF formulation are explained as below:

Equation 4: *Objective function*: The objective function is to minimize generation cost while being subjected to the constraints (5) – (9).

Equation 5: *Node's real power balance constraint*: The constraint balances the real power consumption of node *n* with the incoming and outgoing real power-flow through the node's associated branches and the generation associated with the node. The dual variable for this constraint is τ.

Equations 6 and 7: *Branch power-flow constraint*: The power-flow through a branch is constrained to its maximum thermal limit. Given that power-flow is a direction vector, two sets of equations are derived. The dual variables for these two constraints are $\tilde{\alpha}$ *and* $\bar{\alpha}$.

Equation 8: *Power-flow constraint*: With sufficient small voltage angles between two nodes, real power-flow can be approximated, as shown in (8). The dual variable for this constraint is $\bar{\bar{\beta}}$.

Equation 9: *Generator constraint*: This equation denotes the generator's upper power limit, assuming that the generator can be ramped down to a minimum power level of zero. The dual variable associated with this constraint is $\overset{\sim}{\overset{\frown}{\beta}}$.

The objective function of the dual formulation based on the primal problem formulation (4)-(9) can be expressed as (10).

$$max\ (\textstyle\sum_n \tau\, P_n\ -\ \sum_k P_k^{max}\tilde{\alpha}\ +\ \sum_k P_k^{max}\bar{\alpha}\ - \sum_g P_g^{max}\overset{\sim}{\overset{\frown}{\beta}})\ (10)$$

Assuming no duality gap exists, optimality (11) will hold. As can be seen, the cumulative load payment of the system ($\sum_n \tau\, P_n$) is the cumulative sum of all the generation cost ($\sum_{\forall\, \mathrm{g}}(C_g P_g)$), congestion rent ($\sum_{\mathrm{k}} P_k^{max}\tilde{\alpha}\ -\ \sum_k P_k^{max}\bar{\alpha}$) and generation rent ($\sum_g P_g^{max}\overset{\sim}{\overset{\frown}{\beta}}$).

$$\textstyle\sum_n \tau\, P_n = \sum_{\forall\, g}(C_g P_g) + \sum_k P_k^{max}(\tilde{\alpha}\ -\ \bar{\alpha})\ + \sum_g P_g^{max}\overset{\sim}{\overset{\frown}{\beta}}\ (11)$$

The cumulative load payment of the system ($\sum_n \tau\, P_n$) is of particular interest, given this term is composed of the real power load at node (n), and the LMP (τ) at the node. As such, assuming strong duality holds, the LMP of each node accounts for generation cost, congestion cost, and generation rent and can be used as a suitable representation of the grid to select the mining and forging nodes.

2) *Characteristics of Locational Marginal Price*

LMP is the incremental cost to withdraw 1 MW of energy from an electrical node and reflects a price proportional to the power system's constraints [21]. An interesting property of LMP is its dynamic nature, in the sense that real-time LMP of electrical nodes varies over time based on the generation and load levels at individual nodes and ongoing outages of electrical flow gates. The dynamic nature of LMP proves to be valuable in preventing systematic gaming of the electric grid, as there are no probabilistic methods to predict node prices without solving the DC-OPF problem in real-time. The dynamic nature of LMP is demonstrated through an example case study.

Example Case: Using a three-node example, as shown in Fig. 1, while simulating a DC-OPF using AMPL, results in an LMP$_{(a)}$ at \$80/MWh, LMP$_{(b)}$ at \$45/MWh, and LMP$_{(c)}$ at \$115/MWh. Under the present network scenario, node B is ideal to be selected as a mining node, given the low nodal energy price, while node C may serve as a validator node.

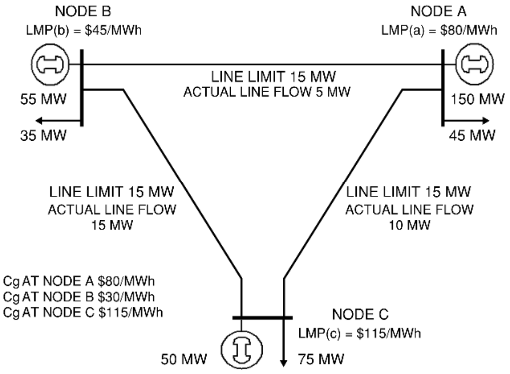


Fig. 1. LMP calculation through simulation for a three-node network

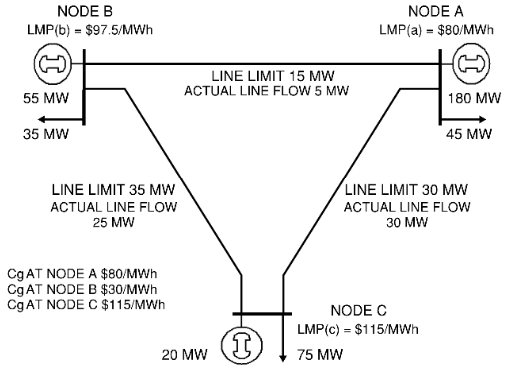


Fig. 2. LMP redistribution following a system transmission capacity increase

A mere transmission capacity upgrades on this three-node network for branch A-C and B-C will drastically change the nodal pricing, as shown in Fig. 2. With new thermal ratings of 30 MW for branch A-C and 35 MW for branch B-C, the LMP$_{(b)}$ changes to \$97.5/MWh. Under the modified network topography, node A turns out to be the cheapest node and best suited to serve as a mining node, with node C still qualifying as a validator node. For simulation purposes in both cases, the maximum generator real power output was constrained at 250 MW for nodes A and C, and 55 MW for node B.

3) *Selection of Miner and Validator Nodes from Locational Marginal Price*

A pricing distribution can be derived based on the LMPs of all the system nodes. Extreme value nodes beyond certain thresholds (both higher and lower values) may then be selected as forging and mining nodes, respectively. Fig. 3 shows the hourly day-ahead LMP distribution plot of all the system nodes of the US energy region ISO New England (ISO-NE) [27] for August 1st, 2020. The right-skewed frequency distribution represents a higher number of mining nodes than forging nodes, thus favorably providing sufficient hashing power. Distribution density plot analysis for all remaining months of the year reveals sufficient separation between the extreme values for distinct node segregation.

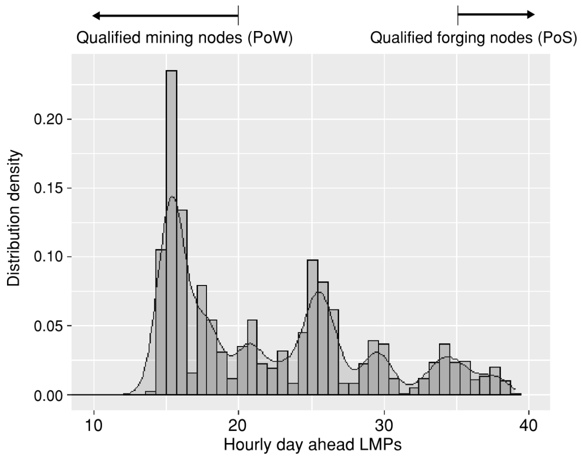


Fig. 3. Distribution density plot of the hourly day-ahead LMP for node segregation

4) Validity Assessment of Miners by Proof-of-Stake Voters

The nature of the LMP distribution plot will inherently enforce node segregation between miners and forgers, effectively preventing collusion. On the other hand, the forger (validator) needs to perform several checks on the newly minted block, such as validation of the last block hash, validation of the signature of the transactions, validation of the block count, and validation of the transactions against the publicly broadcasted set of transactions.

A potential form of network attack vector exists and has been reported in [28] where a malicious network node using the same public key and the broadcasting node may broadcast similar set transactions and the issue requires further discussion.

Given PoW mining follows a Poisson process, the fraudulent node may be successful in guessing the next mined block's timing and can time the broadcast of these fraudulent transactions. The PoS validators validating the PoW block transaction might notice these transactions as contradicting entries resulting in a majority vote to reject the mined block. A sustained network attack of this form could essentially create a backlog of transactions and freeze the network. To prevent this attack vector from gaining an advantage, the validators node needs a timing function. A timing function of this nature essentially segregates real transactions arriving from a node address (while some transactions from this node address are already present in the PoW block) versus fake transactions originating from this node aimed to prevent a block validation. The timing function essentially takes advantage of the fact that even with the network's broadcasting latency, real transactions from a node address are likely to be broadcasted with a smaller time difference than a set of malicious attack transactions, where the attack node guesses the time of the next mined block. A Gudermannian function (12) is thus proposed to act as a timing filter, with ΔT as the time difference between the present time and the discovery time of the transaction. If the timing filter's output score is more than a certain threshold, the validator blocks are made aware of an attack attempt and may choose to scrutinize the transactions' validation against the publicly broadcasted set of transactions before validation.

$$gd(\Delta T) = 2tan^{-1}\left[tanh\left(\frac{a}{2}\Delta T\right)\right] \quad where\ \Delta T \geq 0 \qquad (12)$$

Table I shows two scenarios and the Gudermannian output score being used to distinguish potential transactions from fraudulent ones originating from the same node under the scenario described above. Scenario 1 involves several transactions with the minimal time difference between the present network time and the transaction's discovery time and possibly indicates network latency. These transactions are most likely valid and should be added to the memory-pool to be picked up for mining by the next block. Gudermannian output score from scenario 2 involves a much larger time difference beyond possible network lags and is beyond the cumulative cutoff threshold. The transactions in this scenario indicate a possible network attack attempt. The blockchain validators much ignore these transactions, given these require further scrutiny. Network addresses with fraudulent, duplicate transactions detected by the Gudermannian function's output score may be blocked and flagged for further investigation.

TABLE I. GUDERMANNIAN OUTPUT SCORE FOR TRANSACTION CLASSIFICATION

Scenario 1[a]		*Scenario 2*[a]	
Time difference (ΔT)	***Gudermannian output score***	***Time difference (ΔT)***	***Gudermannian output score***
0.001	0.00399	0.039	0.11673
0.014	0.04198	0.055	0.16425
0.0162	0.04858	0.162	0.46792
0.052	0.15537	1.320	1.53267
$\sum \Delta T$	0.24893	$\sum \Delta T$	2.28159

[a] Assuming an 'a' value of 3 and an output cutoff score as 0.3

B. Design Considerations in the Blockchain Protocol Layer

A blockchain protocol layer running a hybrid consensus algorithm and aimed to serve energy applications need to account for a few design considerations, as discussed in this section.

1) Difficulty and Reward Adjustments

The proposed hybrid consensus is expected to serve as a transaction and smart-contract platform in major energy regions with commercial and private energy users participating in the network. Further, given the dynamic nature of the LMP spread, appropriate cutoff values can always be selected to only include a certain number of nodes as qualified mining nodes. Under these conditions, the blockchain's initial difficulty may be set such the average time for mining a block, under a Poisson process, is an integral multiple of the LMP refresh rate (δ). As such, most independent system operators (ISOs) in the US updates their LMP every five minutes [29].

If large electrical areas are integrated into the current operating region's electrical footprint, a difficulty adjustment might be necessary and can be accomplished using a process similar to (13a,b).

$$Difficulty = \frac{HEX(Max-target)}{HEX(Current-target)} \qquad (13a)$$

$$s.t.\ Pois(\lambda) = n\delta, \forall\, n \in I\ or\ \frac{1}{I} \qquad (13b)$$

Reward adjustment, on the other hand, should reflect a factor to include an average time required to set up new greenfield electrical transmission and distribution nodes (X). Assuming a new block under a hybrid consensus is mined every five minutes on average, two hundred and eighty-eight blocks are expected per calendar day. The current block reward (ω_n) can then be computed using the initial block reward at genesis (ω_i) and a decrement factor (ε), as shown in (14).

$$\omega_n = \omega_i(1-\varepsilon)^{\left(\frac{Block\ height}{288X}\right)}, \forall\, \varepsilon << 1 \qquad (14)$$

2) ASIC Resistance and Hashing Algorithm

Though it is implausible for any hashing algorithm to be utterly resistant against an ASIC implementation, a large memory utilization requirement capable of preventing parallelization or pipeline [30] is usually a strong deterrent. Two other approaches [31] in the construction of the ASIC-resistant

PoW algorithm consist of either using several hash functions to calculate the valid state of the block (multi-hash PoW) or selection of mathematic functions from a large pool for hashing, thereby rending specialized hardware useless (programmatic PoW).

A hashing algorithm for the given hybrid consensus needs to demonstrate some form of ASIC resistance to ensure no mining node enjoys an unfair advantage compared to CPU/GPU users. At present, Equihash, Ethash, SHA3, and its variants SHAKE 128 and SHAKE256 are among the popular algorithms that are reported to be reasonably ASIC-resistant [4, 32].

3) *Transaction Block Weight*

To prevent a denial-of-service attack (DoS attack), the legacy Bitcoin nodes required a block size to be less than 1 Mb. However, the block size specification severely restricted transaction volumes going into a single block. To accommodate larger transaction volumes and mitigate transaction malleability issues, Bitcoin introduces a soft fork in its transaction format, named Segregated Witness. A similar block architecture may be implemented for the hybrid consensus, where the block size is limited to 4 Mb, with 1 Mb of space given to store transactions and 3 Mb to hold the witness data. However, two separate size limitations would make the miner's transaction fee estimation extremely difficult. To solve this problem, block weight [33], as in (15), may be used as a selection factor.

$$Block\ weight\ =\ 3 \times \text{(Base TX size)} + \text{full TX size} \quad (15)$$

4) *Turing Completeness and Gas*

To implement a blockchain transaction environment, as shown in Fig. 4, with the capability of executing obligatory smart-contract agreements, a Turning complete programming language (able to process conditional loops) is required. A Turing complete machine such as the Ethereum Virtual Machine (EVM) also implements a taxation mechanism called 'Gas' to disincentivize programmers writing codes with a heavy looping algorithm.

Modern blockchain transaction platform featuring smart contract should implement some version of the EVM's taxation mechanism to provide incentives for developers to write efficient codes, resulting in faster execution time.

5) *Exponentially Weighted Moving Average of LMP to Account for Extreme Volatility*

Though relatively rare, real-time LMP may experience market volatility [34] triggered by factors such as fluctuations in fuel cost, short-term increase in peak load, and transmission line switching. Extreme volatility within a short time may interfere with the mining nodes' capacity to solve the hash puzzle. To prevent such interference, the node selection algorithm may choose to switch to an exponentially weighted moving average (EWMA) model of real-time LMP, as shown in (16), if high volatility is detected using typical statistical standard deviation tests.

$$LMP_t^{EWMA} = \begin{cases} LMP_1, \forall\ t = 1 \\ \alpha LMP_t +\ (1-\alpha)LMP_{t-1}^{EWMA}, \forall\ t > 1 \end{cases} \quad (16)$$

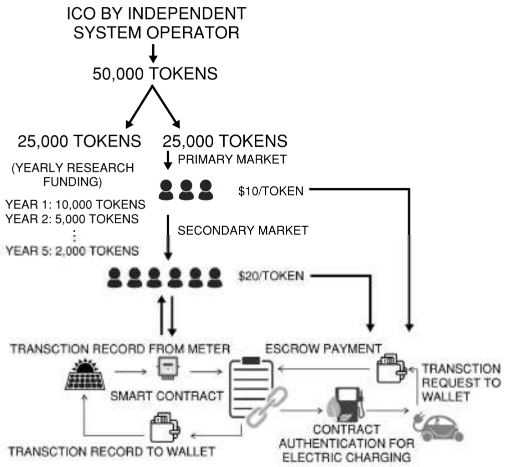

ICO BY INDEPENDENT SYSTEM OPERATOR
50,000 TOKENS
25,000 TOKENS
(YEARLY RESEARCH FUNDING)
YEAR 1: 10,000 TOKENS
YEAR 2: 5,000 TOKENS
⋮
YEAR 5: 2,000 TOKENS
25,000 TOKENS
PRIMARY MARKET
$10/TOKEN
SECONDARY MARKET
$20/TOKEN
TRANSCTION RECORD FROM METER
SMART CONTRACT
ESCROW PAYMENT
TRANSCTION REQUEST TO WALLET
CONTRACT AUTHENTICATION FOR ELECTRIC CHARGING
TRANSCTION RECORD TO WALLET

Fig. 4. Initial coin offering and the tokenization of energy

The first step of the hybrid consensus mechanism begins with the selection of mining and forging nodes. The proposed LMP based node selection method is outlined in Algorithm 1. LMP nodal pricing at a node (n) for a give time instance (t), LMP_n^t, are updated every five to ten minutes across all the major US ISOs. The upper and lower LMP cutoff values, C_{min} and C_{max}, may be established using the standard deviation of the distribution from its mean value.

```
Input: LMP nodal price at node n at time t ($LMP_n^t$), market-timer ($t_m$), upper LMP cutoff ($C_{max}$), lower LMP cutoff ($C_{min}$)
Output: Set of mining nodes ($l_n$), set of forging nodes ($u_n$)
LMP node selection:
Set $t_m$ to zero
Track market-timer for next set of nodal pricing output
while TRUE
        if $LMP_n^t$ != $LMP_n^{t-1}$
                c = 0
                $l_n$ = []
                $u_n$ = []
                if hist(sum of nodes($LMP_n^t$)) ≤ $C_{min}$
                        c = c+1
                        $l_n[c] = LMP_n^t$
                else if hist(sum of nodes($LMP_n^t$)) ≥ $C_{max}$
                        c = c+1
                        $u_n[c] = LMP_n^t$
        return $l_n[c]$, $u_n[c]$
end
```

Algorithm 1. LMP node selection from the real-time DC-OPF market solution

V. Application of Hybrid Consensus in Energy Blockchain and Smart Contract Applications

The tokenization of energy can accomplish energy exchange in a blockchain ecosystem. As demonstrated in Fig. 4, an independent system operator of the electric grid may generate a certain number of tokens, 50,000 tokens in this case, as a part of an initial coin offering (similar to an initial public offering for stock markets). Some token may be reserved for research and development, while the remaining might be exchanged in a primary cryptocurrency-exchange market. These tokens' primary holders may use them for smart contract-based obligatory transactions or exchange them in a secondary cryptocurrency-exchange market, very similar to how stocks are trader in the primary and secondary market.

Other energy-specific applications are highlighted in this section in addition to energy tokenization. The most impressive of these applications involve smart contract executions in day-ahead, real-time, and financial transmission rights (FTR) based energy markets.

A. *Microgrids for Local Energy Trading*

Microgrids' importance has recently received much attention in the US due to the growing number of transmission grid induces wildfires in states like California, Oregon, and Washington [25]. In a local community, the exchange of power, backed by roof-top solar energy with a slim backbone transmission grid, is gaining traction. Microgrid users may record their transactions and execute obligatory smart contracts using a blockchain platform with the secured and energy-efficient hybrid PoW-PoS consensus algorithm.

B. *Tracking of Carbon Credits*

Carbon credits have been used for a long time to limit pollution by companies producing energy from conventional energy resources. Companies not using enough of their carbon credits often sell them to companies in need of those credits. A blockchain platform can provide a secured immutable platform between the regulatory authority and the utilities, enabling the regulatory body to enforce emission limits using smart contracts and allowing utilities to exchange carbon credits on a secured platform.

C. *Energy Trade Executions Using Smart-Contracts*

Significant portions of the US energy grid are operated by independent system operators (ISOs) to check systematic gaming and monopoly and ensure competitive energy pricing. To reduce price volatility, market participants often commit to buy or sell energy in a day-ahead market, whereas the differences are fulfilled in the real-time energy market. All market transactions and contract agreements of this nature can be transitioned to a safe and secured blockchain platform, allowing the grid operators to transact with commercial or private users using a standard token system.

US-based ISOs also have a market around FTRs, where market participants transact financial instruments to hedge congestion cost or arbitrage the price difference. Like stock exchanges, these transactions records are stored in centralized databases, posing a security vulnerability. Blockchain technology can provide a secured decentralization platform for the FTR markets while reducing overhead fees and helping with regulatory reporting requirements.

VI. Conclusion

Currently, existing popular consensus algorithms have several limitations preventing the widescale implementation of blockchain technology for energy applications. A secured and energy-efficient consensus algorithm such as the one proposed in this paper has tremendous potential to help the proliferation of blockchain technology specific to energy grid applications. Locational marginal price is an integral part of many restructured power markets and can be used to select a set of qualifying nodes for blockchain mining and validation. From an implementation standpoint, design considerations such as reward and difficulty adjustments, the choice of a suitable hashing algorithm, and nodal pricing volatility must be made to provide a stable and scalable platform.

Among its many potential applications, the blockchain technology may be used in day-ahead and real-time market-based transactions, microgrid contract obligations, and escrow payments from electric vehicle users. An underlying hybrid consensus algorithm that is secured, energy-efficient, and capable of handling high volumes of transactions can play a crucial role in modern electric smart grids.

References

[1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 03 2009. [Online]. Available: https://bitcoin.org/bitcoin.pdf. [Accessed 12 2020].

[2] J. Sedlmeir, H. U. Buhl, G. Fridgen and R. Keller, "The Energy Consumption of Blockchain Technology: Beyond Myth," *Business & Information Systems Engineering,* vol. 62, pp. 599-608, 2020.

[3] University of Cambridge Judge Business School, "Cambridge Bitcoin Electricity Consumption Index," 01 2017. [Online]. Available: https://cbeci.org/cbeci/comparisons. [Accessed 12 2020].

[4] M. Harvilla and J. Du, *Prospective Hybrid Consensus for Project PAI,* arXiv preprint arXiv:1902.02469, 2019.

[5] C. Ye, G. Li, H. Cai, Y. Gu and A. Fukuda, "Analysis of Security in Blockchain: Case Study in 51%-Attack Detecting," in *2018 5th International Conference on Dependable Systems and Their Applications (DSA)*, Dalian, China, 2018.

[6] D. Mingxiao, M. Xiaofeng, Z. Zhe, W. Xiangwei and C. Qijun, "A Review on Consensus Algorithm of Blockchain," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, AB, Canada, 2017.

[7] W. Li, S. Andreina, J. M. Bohli and G. Karame, "Securing Proof-of-Stake Blockchain Protocols," in *Conference: European Symposium on Research in Computer Security International Workshop on Data Privacy Management Cryptocurrencies and Blockchain Technology*, 2017.

[8] V. Buterin, "Slasher: A Punitive Proof-of-Stake Algorithm," Ethereum Blog, 01 2014. [Online]. Available: https://blog.ethereum.org/2014/01/15/slasher-a-punitive-proof-of-stake-algorithm/. [Accessed 12 2020].

[9] L. M. Bach, B. Mihaljevic and M. Zagar, "Comparative Analysis of Blockchain Consensus Algorithms," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2018.

[10] Z. Zheng, S. Xie, H. Dai, X. Chen and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," in *2017 IEEE International Congress on Big Data (BigData Congress)*, Honolulu, HI, USA, 2017.

[11] N. Chaudhry and M. M. Yousaf, "Consensus Algorithms in Blockchain: Comparative Analysis, Challenges and Opportunities," in *2018 12th International Conference on Open Source Systems and Technologies (ICOSST)*, Lahore, Pakistan, 2018.

[12] Federal Information Processing Standards Publication: FIPS PUB 180-4 , "Secure Hash Standard (SHS)," 08 2015. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf. [Accessed 12 2020].

[13] G. Casella and R. L. Berger, Statistical Inference, New Delhi, India: Cengage, 2018.

[14] I. Bashir, Mastering Blockchain, Birmingham, UK: Packt Publishing Ltd., 2020.

[15] G. Bissias, D. Thibodeau and B. N. Levine, "Bonded Mining: Difficulty Adjustment by Miner Commitment," *Data Privacy Management, Cryptocurrencies, and Blockchain Technology,* pp. 372-390, 2019.

[16] ConsenSys, "The Roadmap to Serenity," ConsenSys, 05 2019. [Online]. Available: https://media.consensys.net/the-roadmap-to-serenity-bc25d5807268. [Accessed 12 2020].

[17] S. King and S. Nadal, "PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake," 08 2012. [Online]. Available: https://whitepaper.io/document/139/peercoin-whitepaper. [Accessed 12 2020].

[18] V. Buterin, "Proof of Stake: How I Learned to Love Weak Subjectivity," 11 2014. [Online]. Available: https://blog.ethereum.org/2014/11/25/proof-stake-learned-love-weak-subjectivity/. [Accessed 12 2020].

[19] F. C. Schweppe, M. C. Caramantis, R. D. Tabors and R. E. Bohn, Spot pricing of electricity, Springer Science & Business Media, 2013.

[20] D. Bienstock and A. Verma, "Strong NP-hardness of AC power flows feasibility," *arXiv preprint arXiv:1512.07315,* 2015.

[21] H. Liu, L. Tesfatsion and A. A. Chowdhury, "Locational Marginal Pricing Basics for Restructured Wholesale Power Markets," in *IEEE Power & Energy Society General Meeting*, Calgary, AB, Canada, 2009.

[22] T. Gregory, C. Li, M. Zhang and K. W. Hedman, "The Effects of Extended Locational Marginal Pricing in Wholesale Electricity Markets," in *2013 North American Power Symposium (NAPS)*, Manhattan, KS, USA, 2013.

[23] J. A. Momoh, R. Adapa and M. E. El-Hawary, "A Review of Selected Optimal Power Flow Literature to 1993. I. Nonlinear and Quadratic Programming Approaches," *IEEE Transactions on Power Systems,* vol. 14, no. 1, pp. 96-104, 1999.

[24] J. A. Momoh, M. E. El-Haary and R. Adapa, "A Review of Selected Optimal Power Flow Literature to 1993. II. Newton, Linear Programming and Interior Point Methods," *IEEE Transactions on Power Systems,* vol. 1, no. 105-111, p. 14, 1999.

[25] S. Ghosh and S. Dutta, "A Comprehensive Forecasting, Risk Modelling and Optimization Framework for Electric Grid Hardening and Wildfire Prevention in the US," *International Journal of Energy Engineering,* vol. 10, no. 3, pp. 80-89, 2020.

[26] M. Saleh, K. Saikumar and K. W. Hedman, "Pricing Implications of Transmission Security Modeling In Electric Energy Markets," in *2019 North American Power Symposium (NAPS)*, Wichita, KS, USA, 2019.

[27] ISO New England, "Pricing Reports for Markets and Operations," ISO New England, 12 2020. [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/lmps-da-hourly. [Accessed 12 2020].

[28] A. Mackenzie, "Memcoin: A Hybrid Proof-of-Work, Proof-of-Stake Crypto-Currency," 2013. [Online]. Available: https://decred.org/research/mackenzie2013.pdf. [Accessed 11 2020].

[29] ISO New England, "FAQs: Locational Marginal Pricing," ISO New England, 12 2020. [Online]. Available: https://www.iso-ne.com/participate/support/faq/lmp#e. [Accessed 12 2020].

[30] Stack Exchange, "What does it mean for a cryptocurrency to be ASIC-resistant?," 12 2020. [Online]. Available: https://bitcoin.stackexchange.com/questions/29975/what-does-it-mean-for-a-cryptocurrency-to-be-asic-resistant. [Accessed 12 2020].

[31] H. Cho, "ASIC-Resistance of Multi-Hash Proof-of-Work Mechanisms for Blockchain Consensus Protocols," *IEEE Access,* vol. 6, pp. 66210-66222, 2018.

[32] A. R. Zamanov, V. . A. Erokhin and P. S. Fedotov, "ASIC-Resistant Hash Functions," in *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, Moscow, Russia, 2018.

[33] E. Lombrozo, J. Lau and P. Wuille, "BIP: 141 Segregated Witness (Consensus layer)," 12 2015. [Online]. Available: https://github.com/bitcoin/bips/blob/master/bip-0141.mediawiki#cite_note-3. [Accessed 12 2020].

[34] "LMP Electricity Markets: Market Operations, Market Power, and Value for Consumers," 02 2006. [Online]. Available: http://www.synapse-energy.com/sites/default/files/SynapseReport.2007-02.APPA_.LMP-Electricity-Markets.06-060-Report.pdf. [Accessed 12 2020].

Design of an Accurate, Cost-effective Radix-4 Booth Multiplier

Riya Agarwal, Sanjana Jayakrishna and Sivaselvan N
Dept. of Computer Science & Engineering, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal 576104, India
{riya.agarwal | sanjana.jayakrishna}@learner.manipal.edu, siva.selvan@manipal.edu

***Abstract*—Multiplication is a key process in various applications. Consequently, the multiplier is a principal component in several hardware platforms. For multiplication of signed integers, radix-4 booth multipliers are widely used as they reduce the number of partial products to half. Several approximate multipliers for radix-4 booth multiplication have been presented in recent times to achieve energy efficiency. However, the errors displayed by these multipliers make them unsuitable for lossless applications. For these applications, accurate radix-4 booth multiplication is much needed. In this paper, we introduce an optimized algorithm and multiplier for accurate radix-4 booth multiplication. The algorithm reduces the number of bits that participate in the addition process during multiplication. At the same time, the algorithm has sufficient storage area for reduced design complexity in its multiplier. The multiplier gives promising results of 100% when tested for accuracy for a wide spectrum and all possible cases of multiplicand-multiplier pairs. Besides, the multiplier is shown to be cost-effective. Lastly, we compare our multiplier with the current radix-4 booth multipliers in terms of key error parameters to highlight its accuracy.**

***Index Terms*—Accurate, Cost-effective, Radix-4, Booth multiplier, Lossless applications.**

I. Introduction

Several machine learning and digital signal processing applications are highly influenced by the multiplication process. For instance, multiply-accumulate operations account for more than 90% of the computations in convolutional neural networks. Hence, the multiplier is a key component in various hardware platforms. Radix-4 booth multiplication is widely used for the multiplication of signed binary integers since it reduces the number of partial products by half compared to the conventional multiplication process.

Research in radix-4 booth multiplication made a shift from accurate to approximate computing design paradigm which constructs energy-efficient circuits at the cost of a little accuracy. Researchers have devised several radix-4 booth multipliers based on this emerging paradigm [1]–[10]. However, the errors featured by these multipliers are not tolerable by lossless applications. Besides, cost is another important parameter in the context of multiplier design. Therefore, there is a need for a cost-effective multiplier for accurate radix-4 booth multiplication.

In an attempt to address the above requirement, we introduce an accurate, cost-effective radix-4 booth multiplier in this paper. Initially, we devise an optimized algorithm for accurate radix-4 booth multiplication by reducing the number of bits that participate in the addition process during radix-4 booth multiplication while having sufficient storage area for reduced hardware design complexity. The accuracy of the multiplier is tested for a wide range and all possible cases of multiplicand-multiplier pairs. The cost of the multiplier circuit and its operations are critically analyzed. Besides, the proposed multiplier's accuracy is compared with the current radix-4 multipliers in key error parameters viz., Mean Relative Error Distance (MRE/MRED), Normalized Mean Error Distance (NMED), and Probability Relative Error Distance (PRED).

The remainder of the paper is organized in the following manner: The current radix-4 booth multipliers are discussed in Section II. The proposed algorithm and multiplier for accurate radix-4 booth multiplication are presented in Section III. Section IV critically analyzes the accuracy and cost-effectiveness of the multiplier. Besides, the proposed multiplier's accuracy is compared with the current radix-4 booth multipliers. Section V summarizes this work.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. Literature Review

This section presents the summary of the current research work regarding radix-4 booth multiplication.

Liu *et al.* [1] presented ”two approximate radix-4 booth encoding algorithms” viz., ”R4ABE1 and R4ABE2”. They also designed ”two approximate booth multipliers: R4ABM1 and R4ABM2”, based on the algorithms with an approximate wallace tree structure in place. In this work, the energy efficiency of the multipliers is decided by a parameter called the approximation factor. Among the two multipliers, R4ABM2 produces relatively more error. For the approximation factor 2, it produces MRED, NMED, and PRED of 0.344×10^{-2}, 0.2543×10^{-5}, and 99.79% respectively. Whereas, for the approximation factor 32, it produces the highest error with MRED, NMED, and PRED of 2.52×10^{7}, 17879×10^{-5}, and 0.123% respectively.

In [2], the authors devised ”an approximate hybrid high radix encoding” to generate the partial products for signed multiplication. In this technique, ”accurate radix-4 encoding” is used to encode the Most Significant Bits (MSBs), and ”approximate higher radix encoding” is used to encode the Least Significant Bits (LSBs). ”Three approximate multipliers: RAD64, RAD256 and RAD1024”, are presented. Of the three multipliers, RAD1024 produces the highest error with an MRE of 0.93% and PRED of 93.26%. The authors did not consider evaluating NMED in this work.

Ansari *et al.* [3] presented "an improved, approximate 4:2 compressor", in which generate and propagate signals are used for encoding the inputs. This has resulted in the reduced number of flawed rows in the truth table of the compressor. Based on this compressor, the authors designed "an approximate multiplier called CABM". CABM produces MRED and NMED of 0.014 and 0.18 respectively.

"Three approximate radix-4 booth multipliers: ABM-M1, ABM-M2, and ABM-M3", are proposed in [4]. In each of the multipliers, a unique approximation technique is employed which encompasses reduction in the logic complexity of "booth partial product generator" and "partial product accumulation". Among the three multipliers, ABM-M2 generates the highest error with MRED and NMED of 2.689×10^{-2} and 1087.270×10^{-6} respectively.

The authors of [5] designed "two approximate multipliers: LOBO10 and LOBO12", using "radix-4 booth encoding" and "logarithmic product approximation". The multipliers encode the MSBs using accurate radix-4 booth encoding and approximate the higher radix partial products arising from the LSBs using logarithmic approximation. Of the two multipliers, LOBO12 generates the highest error with MRE, NMED, and PRED of 0.44-96.09%, 18.45×10^{-4}, and 90.99% respectively.

[6] introduces "two approximate multipliers: HLR-BM1 and HLR-BM2", based on hybrid low radix encoding. In this work, the odd multiples of radix-8 are approximated to their nearest powers of two to fix the low-performance issue of radix-8 booth encoding. Among the two multipliers, HLR-BM1 produces the highest error with MRE, NMED, and PRED of 0.03%, 0.79×10^{-5}, and 98.35% respectively.

The aforementioned approaches are inclined to approximate computing design paradigm which aims to construct energy-efficient circuits at the cost of a little accuracy. Not all applications are inherently resilient to accuracy loss. For instance, lossless applications such as medical image compression have zero tolerance to accuracy loss. Besides, the approximate encoding techniques discussed in the aforementioned approaches are based on accurate radix-4 encoding. This signifies the importance of accurate radix-4 encoding in the context of signed multiplication. We present a novel algorithm and cost-effective hardware design for accurate radix-4 booth multiplication in the next section.

III. Proposed Algorithm and Multiplier

In this section, we introduce our algorithm and multiplier for accurate radix-4 booth multiplication.

A. Algorithm

The proposed algorithm is illustrated in Algorithm 1. We have the input variables $'m'$ and $'q'$, two decimal integers converted into two's complement binary integers of length $'n'$ bits each, where $'n'$ is calculated as the maximum of the number of bits used to represent $'m'$ and $'q'$ as binary numbers.

Algorithm 1 Optimized radix-4 booth multiplication.

```
Input: Two integers $'m'$ and $'q'$ in decimal form.
Output: Product of $'m'$ and $'q'$.
1. Convert $'m'$ and $'q'$ into binary numbers of $'n'$ bits
2. if $(n\%2 = 0)$, then
      $qlen \leftarrow n + 1$
   else
      $qlen \leftarrow n + 2$
   end if
3. Declare arrays $M[n]$, $C[1]$, $A[n]$, $Q[qlen]$
4. $A \leftarrow 0$
5. $M \leftarrow$ binary equivalent of $'m'$
6. $L \leftarrow ceil(n/2)$
7. $C \leftarrow 0$
8. $Q[0 : qlen - 2] \leftarrow$ binary equivalent of $'q'$
9. $Q[qlen - 1] \leftarrow 0$
10. Check whether, in $Q$, groups of 3 bits are being
      formed or not. If not, do sign extension
11. Compute the values of $+2M$, $-2M$, $-M$ and store them
      in arrays $Plus2M[n + 1]$, $Minus2M[n + 1]$, and
      $MinusM[n]$
12. while $L > 0$, do
      if $Q[qlen - 3 : qlen - 1] = 000$ or $111$, then
         $c \leftarrow 0$
         $a \leftarrow A[0]$
      else if $Q[qlen - 3 : qlen - 1] = 001$ or $010$, then
         $c \leftarrow M[0]$
         $a \leftarrow A[0]$
         $A \leftarrow A + M$
      else if $Q[qlen - 3 : qlen - 1] = 011$, then
         $c \leftarrow Plus2M[0]$
         $a \leftarrow A[0]$
         $A \leftarrow A + Plus2M$
      else if $Q[qlen - 3 : qlen - 1] = 100$, then
         $c \leftarrow Minus2M[0]$
         $a \leftarrow A[0]$
         $A \leftarrow A + Minus2M$
      else if $Q[qlen - 3 : qlen - 1] = 101$ or $110$, then
         $c \leftarrow MinusM[0]$
         $a \leftarrow A[0]$
         $A \leftarrow A + MinusM$
      end if
      $Cout \leftarrow$ carry out after binary addition
      $C \leftarrow Cout \oplus a \oplus c$
      if $C = 0$, then
         $LRS(A\$Q)$ by 2 times
      else
         $ARS(C\$A\$Q)$ by 2 times
      end if
      $L \leftarrow L - 1$
    end while
13. $res \leftarrow (A\$Q[0 : qlen - 2])$
14. $prod \leftarrow$ decimal form of $res$
15. return $prod$
```

The following arrays are used: M ($'n'$ bits) to hold the binary representation of $'m'$, C (1 bit) to hold the carry, and A ($'n'$ bits) to hold the intermediate summation results of the partial products in binary representation. The array Q to hold the binary representation of multiplier $'q'$ has some additional components. After converting $'q'$ to a binary number, an implicit zero is added to the right of the LSB.

The crux of radix-4 booth multiplication is to convert the multiplier in its binary form into a recoded multiplier by checking whether successive triplets (groups of 3 bits) are being formed or not, starting from LSB. Therefore, after adding the implicit zero, the aforesaid condition is checked, and if it's not satisfied, sign extension is done to complete the triplet. The process of forming the triplets is shown in Fig. 1.

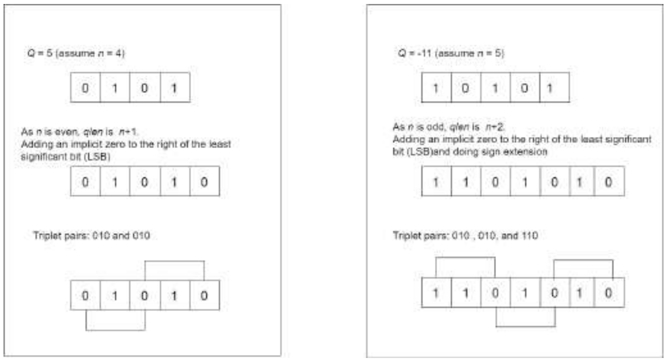

Q = 5 (assume n = 4)
0 1 0 1
As n is even, qlen is n+1.
Adding an implicit zero to the right of the least significant bit (LSB)
0 1 0 1 0
Triplet pairs: 010 and 010
0 1 0 1 0
Q = -11 (assume n = 5)
1 0 1 0 1
As n is odd, qlen is n+2.
Adding an implicit zero to the right of the least significant bit (LSB)and doing sign extension
1 1 0 1 0 1 0
Triplet pairs: 010 , 010, and 110
1 1 0 1 0 1 0

Fig. 1: Triplets formation.

Sign extension will be needed only when $'n'$ is odd. Thus, Q is $n+1$ bits long if $'n'$ is even, and $n+2$ bits long if $'n'$ is odd.

To save repeated computations, the values of $+2M$, $-2M$, $-M$ are computed beforehand as follows:

- $-M$ is the 2's complement of multiplicand $'m'$ and is stored as a binary number in array $MinusM$ ($'n'$ bits)
- $+2M$ is the binary representation of multiplicand $'m'$, with an added zero at the LSB, and is stored in array $Plus2M$ ($n+1$ bits)
- $-2M$ is the binary representation of the 2's complement of multiplicand $'m'$, with an added zero at the LSB, and is stored in the array $Minus2M$ ($n+1$ bits)

In each iteration, the last 3 digits of Q determine which operation will take place. The variable $Cout$ will have the carry-out value after the addition of A and $M/MinusM/Plus2M/Minus2M$. The variable C takes three intermediate variables and performs eXclusive-OR (XOR) operation on them, denoted by $\oplus$ operator in the algorithm. C determines whether to carry out Logical Right Shift (LRS) or Arithmetic Right Shift (ARS). The final product is the concatenation (denoted by \$ operator) of the values in arrays A and Q, where the higher-order half is stored in A and the lower-order half in Q, except for the last bit of Q.

B. Multiplier

Fig. 2 introduces the multiplier for the proposed radix-4 booth multiplication algorithm, for multiplication of two signed integers, both having $'n'$ bits in their binary representation.

The hardware mainly consists of an 8 × 1 Multiplexer (Mux), an n-bit adder, a carry logic, a control sequencer and registers C (1 bit), A ($'n'$ bits), M ($'n'$ bits), $Plus2M$ ($n+1$ bits), $Minus2M$ ($n+1$ bits), $MinusM$ ($'n'$ bits), q_{n+1} (1 bit) and Q ($n+1$ bits). A and Q are shift registers, concatenated as shown in Fig. 2. These registers are analogous to the arrays declared in the algorithm.

The register Q holds the binary representation of the multiplier, including the implicit zero to the right of LSB. If $'n'$ is odd, then the input and output lines to and from the register q_{n+1} are enabled, and the shift line from a_0 to q_n is disabled. In this case, q_{n+1} holds the sign extension bit for Q. If $'n'$ is even, then the input and output lines to and from q_{n+1} are disabled, and the shift line from a_0 to q_n is enabled.

The register M holds the binary representation of the multiplicand, register $MinusM$ the 2's complement representation of M, register $Plus2M$ the binary representation of M after logical left shift by one-bit position, and register $Minus2M$ the binary representation of $MinusM$ after logical left shift by one-bit position.

At the start, the multiplier and the implicit zero are loaded into register Q, the multiplicand into register M, and registers C and A are cleared to 0. The carry logic is essentially an XOR gate whose output is directed to C and the control sequencer.

The product is computed in L cycles (whose value is $ceil(n/2)$), for which the underlying logic is present in the control sequencer. At the beginning of each cycle, the last 3 bits (starting from q_0) of Q are sent to the control sequencer to determine which operation will take place.

A signal from the control sequencer to determine whether the addition of M, $Plus2M$, $MinusM$, $Minus2M$ or no addition takes place, is sent to the 8 × 1 Mux and accordingly M, $Plus2M$, $MinusM$, $Minus2M$ or 0 is sent as the output of the 8 × 1 Mux to the n-bit adder, and its MSB is sent to the carry logic. The MSB of the current value in register A is sent to the carry logic, and A is sent as input to the n-bit adder. The sum from the n-bit adder is sent to A, and the carry-out is sent to the carry logic. The output of the carry logic is sent to C.

Depending on the output of the carry logic, a signal is sent from the control sequencer to C, A, and Q. If the output of carry logic is "0", and if $'n'$ is even, then the logical right shift of A and Q takes place. If $'n'$ is odd, then the logical right shift of A, q_{n+1}, and Q by two-bit positions takes place. If the output is "1", and if $'n'$ is even, then the arithmetic right shift of C, A, and Q takes place. If $'n'$ is odd in this case, then the arithmetic right shift of C, A, q_{n+1}, and Q by two-bit positions takes place.

Together, registers A and Q hold the intermediate partial product while the last 3 bits of Q (starting from LSB) generate the add/no-add signal, which determines the next operation that will take place.

After L cycles, the higher-order half of the product is held in register A, and the lower-order half in register Q (q_n to q_1 bits) when $'n'$ is even, and in registers q_{n+1} and Q (q_n to q_1 bits) when $'n'$ is odd.

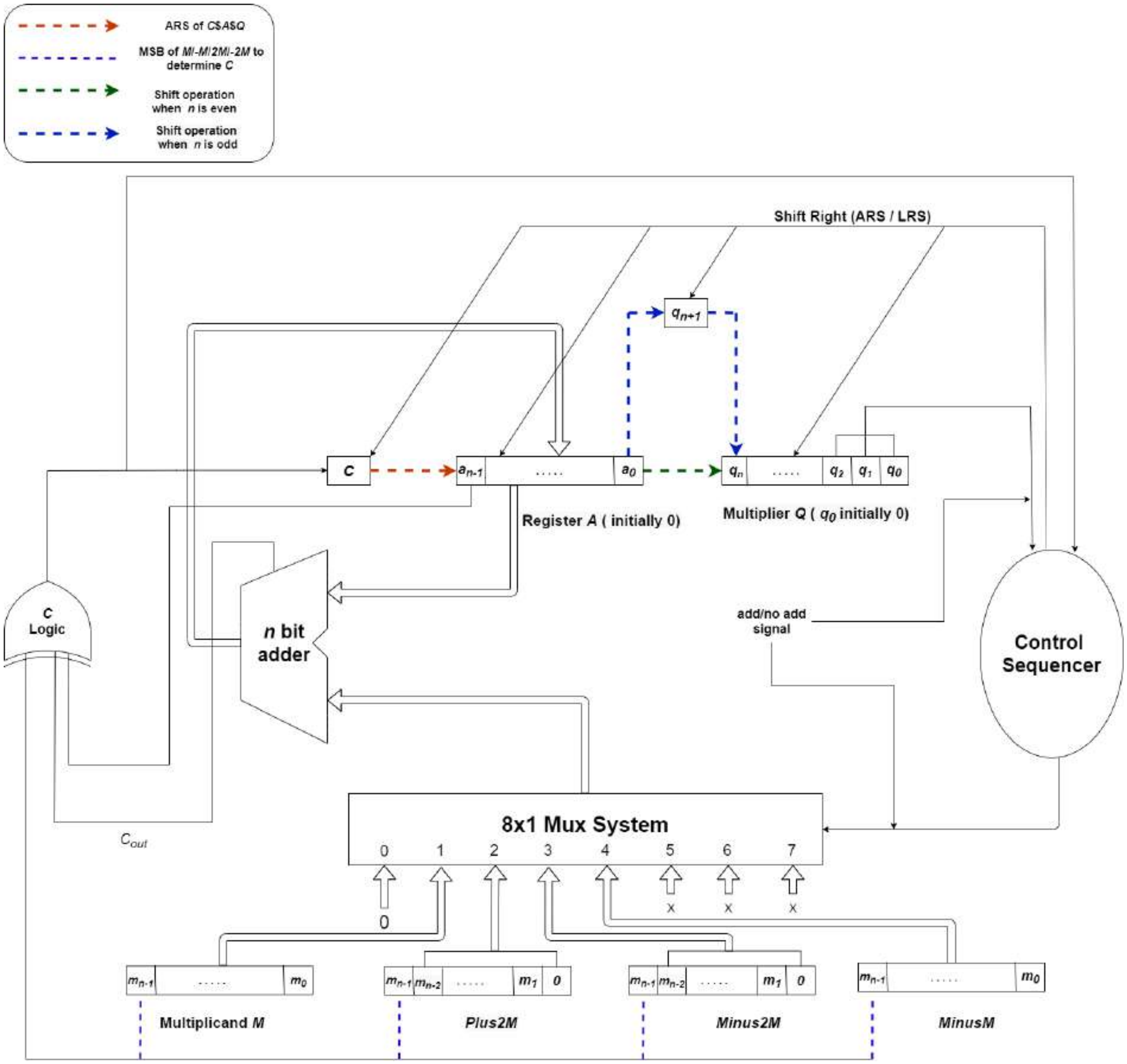


Fig. 2: Proposed radix-4 booth multiplier.

IV. Results and Discussions

The proposed algorithm and multiplier are evaluated for accuracy and cost-effectiveness in this section. Besides, the proposed multiplier's accuracy is compared with the current radix-4 booth multipliers.

A. Accuracy

Fig. 3 presents the code snippet where inputs over a range of integers are sent to the algorithm to check the algorithm's accuracy.

As shown in Fig. 4, the algorithm's accuracy is tested for 50 random integer pairs, each pair consisting of a multiplicand and a multiplier, in the range of -5000 to 5000.

At first, the product for each integer pair is calculated by the computer. Then, the integer pairs are sent to the algorithm to calculate their products. Finally, the products from the algorithm are compared with the actual products to test the accuracy of the algorithm.

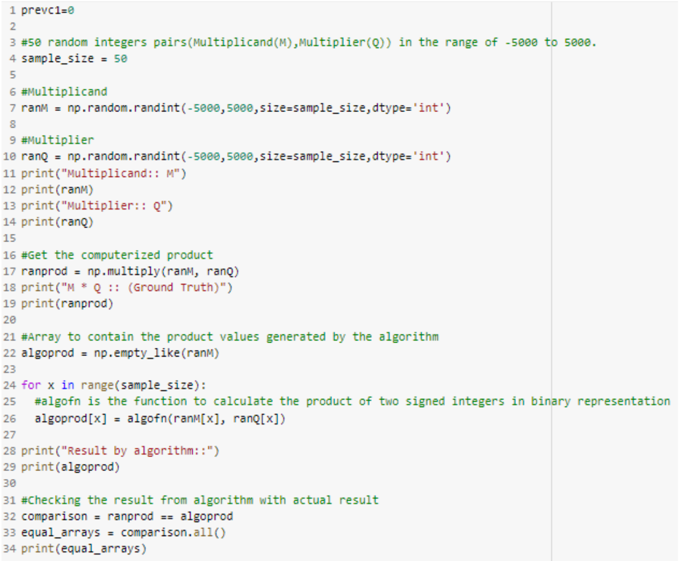
The accuracy of the algorithm is also tested for 4 different cases of multiplicand and multiplier values as presented in Figures 5, 6, 7, and 8 respectively.

To test the accuracy of the algorithm for a wide range of inputs, the sample size is increased to 10,000, for which the algorithm gave the correct results.

B. Cost of the circuit and operations

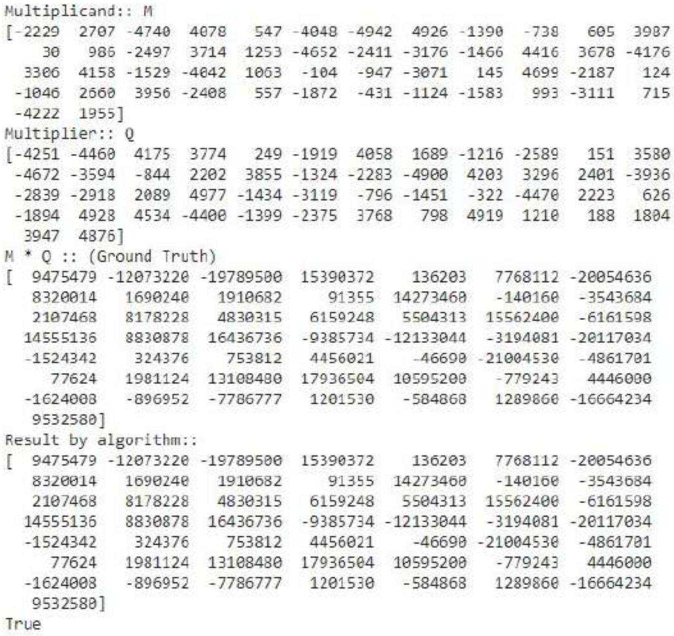
Our algorithm presents a cost-effective strategy for the implementation of radix-4 booth multiplication.

The original algorithm makes use of $n+1$ bits for the calculation of the subsequent partial products. In our algorithm, the C logic is a 3-input XOR gate, which reduces the size of registers A and M from $n+1$ to $'n'$ bits, thereby, reducing the number of flip flops by one in each of these registers.



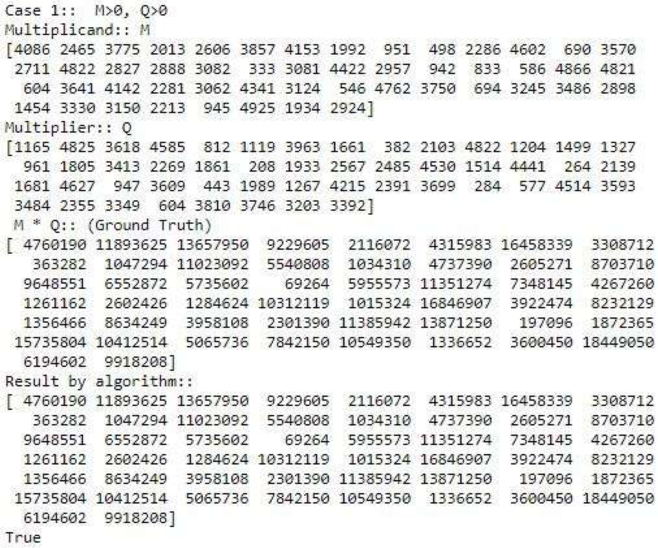
```python
prevc1=0

#50 random integers pairs(Multiplicand(M),Multiplier(Q)) in the range of -5000 to 5000.
sample_size = 50

#Multiplicand
ranM = np.random.randint(-5000,5000,size=sample_size,dtype='int')

#Multiplier
ranQ = np.random.randint(-5000,5000,size=sample_size,dtype='int')
print("Multiplicand:: M")
print(ranM)
print("Multiplier:: Q")
print(ranQ)

#Get the computerized product
ranprod = np.multiply(ranM, ranQ)
print("M * Q :: (Ground Truth)")
print(ranprod)

#Array to contain the product values generated by the algorithm
algoprod = np.empty_like(ranM)

for x in range(sample_size):
  #algofn is the function to calculate the product of two signed integers in binary representation
  algoprod[x] = algofn(ranM[x], ranQ[x])

print("Result by algorithm::")
print(algoprod)

#Checking the result from algorithm with actual result
comparison = ranprod == algoprod
equal_arrays = comparison.all()
print(equal_arrays)
```

Fig. 3: Code snippet to check accuracy of the algorithm.



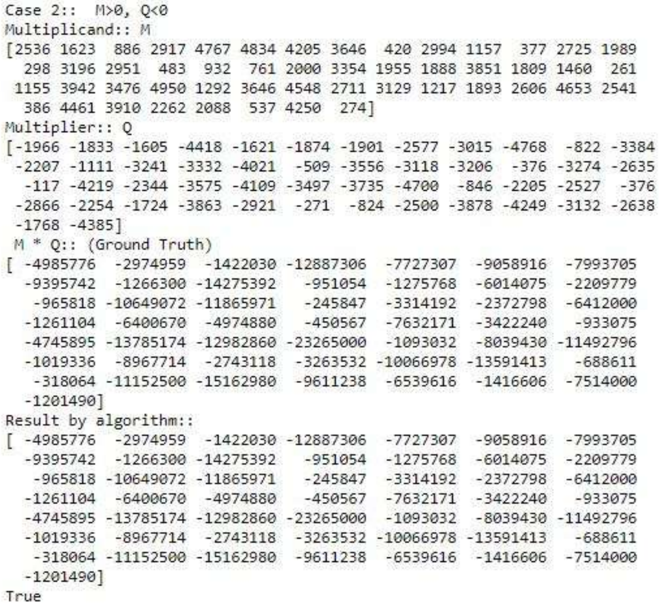
```
Multiplicand:: M
[-2229  2707 -4740  4078   547 -4048 -4942  4926 -1390  -738   605  3987
    30   986 -2497  3714  1253 -4652 -2411 -3176 -1466  4416  3678 -4176
  3306  4158 -1529 -4042  1063  -104  -947 -3071   145  4699 -2187   124
 -1046  2660  3956 -2408   557 -1872  -431 -1124 -1583   993 -3111   715
 -4222  1955]
Multiplier:: Q
[-4251 -4460  4175  3774   249 -1919  4058  1689 -1216 -2589   151  3580
 -4672 -3594  -844  2202  3855 -1324 -2283 -4900  4203  3296  2401 -3936
 -2839 -2918  2089  4977 -1434 -3119  -796 -1451  -322 -4470  2223   626
 -1894  4928  4534 -4400 -1399 -2375  3768   798  4919  1210   188  1804
  3947  4876]
M * Q :: (Ground Truth)
[  9475479 -12073220 -19789500  15390372    136203   7768112 -20054636
   8320014   1690240   1910682     91355  14273460   -140160  -3543684
   2107468   8178228   4830315   6159248   5504313  15562400  -6161598
  14555136   8830878  16436736  -9385734 -12133044  -3194081 -20117034
  -1524342    324376    753812   4456021    -46690 -21004530  -4861701
     77624   1981124  13108480  17936504  10595200   -779243   4446000
  -1624008   -896952  -7786777   1201530   -584868   1289860 -16664234
   9532580]
Result by algorithm::
[  9475479 -12073220 -19789500  15390372    136203   7768112 -20054636
   8320014   1690240   1910682     91355  14273460   -140160  -3543684
   2107468   8178228   4830315   6159248   5504313  15562400  -6161598
  14555136   8830878  16436736  -9385734 -12133044  -3194081 -20117034
  -1524342    324376    753812   4456021    -46690 -21004530  -4861701
     77624   1981124  13108480  17936504  10595200   -779243   4446000
  -1624008   -896952  -7786777   1201530   -584868   1289860 -16664234
   9532580]
True
```

Fig. 4: Testing accuracy for 50 random integer pairs in the range -5000 to +5000.



```
Case 1::  M>0, Q>0
Multiplicand:: M
[4086 2465 3775 2013 2606 3857 4153 1992  951  498 2286 4602  690 3570
 2711 4822 2827 2888 3082  333 3081 4422 2957  942  833  586 4866 4821
  604 3641 4142 2281 3062 4341 3124  546 4762 3750  694 3245 3486 2898
 1454 3330 3150 2213  945 4925 1934 2924]
Multiplier:: Q
[1165 4825 3618 4585  812 1119 3963 1661  382 2103 4822 1204 1499 1327
  961 1805 3413 2269 1861  208 1933 2567 2485 4530 1514 4441  264 2139
 1681 4627  947 3609  443 1989 1267 4215 2391 3699  284  577 4514 3593
 3484 2355 3349  604 3810 3746 3203 3392]
 M * Q:: (Ground Truth)
[ 4760190 11893625 13657950  9229605  2116072  4315983 16458339  3308712
   363282  1047294 11023092  5540808  1034310  4737390  2605271  8703710
  9648551  6552872  5735602    69264  5955573 11351274  7348145  4267260
  1261162  2602426  1284624 10312119  1015324 16846907  3922474  8232129
  1356466  8634249  3958108  2301390 11385942 13871250   197096  1872365
 15735804 10412514  5065736  7842150 10549350  1336652  3600450 18449050
  6194602  9918208]
Result by algorithm::
[ 4760190 11893625 13657950  9229605  2116072  4315983 16458339  3308712
   363282  1047294 11023092  5540808  1034310  4737390  2605271  8703710
  9648551  6552872  5735602    69264  5955573 11351274  7348145  4267260
  1261162  2602426  1284624 10312119  1015324 16846907  3922474  8232129
  1356466  8634249  3958108  2301390 11385942 13871250   197096  1872365
 15735804 10412514  5065736  7842150 10549350  1336652  3600450 18449050
  6194602  9918208]
True
```

Fig. 5: Testing accuracy for both M and Q being positive in the range of [0,5000].



```
Case 2::  M>0, Q<0
Multiplicand:: M
[2536 1623  886 2917 4767 4834 4205 3646  420 2994 1157  377 2725 1989
  298 3196 2951  483  932  761 2000 3354 1955 1888 3851 1809 1460  261
 1155 3942 3476 4950 1292 3646 4548 2711 3129 1217 1893 2606 4653 2541
  386 4461 3910 2262 2088  537 4250  274]
Multiplier:: Q
[-1966 -1833 -1605 -4418 -1621 -1874 -1901 -2577 -3015 -4768  -822 -3384
 -2207 -1111 -3241 -3332 -4021  -509 -3556 -3118 -3206  -376 -3274 -2635
  -117 -4219 -2344 -3575 -4109 -3497 -3735 -4700  -846 -2205 -2527  -376
 -2866 -2254 -1724 -3863 -2921  -271  -824 -2500 -3878 -4249 -3132 -2638
 -1768 -4385]
 M * Q:: (Ground Truth)
[ -4985776  -2974959  -1422030 -12887306  -7727307  -9058916  -7993705
  -9395742  -1266300 -14275392   -951054  -1275768  -6014075  -2209779
   -965818 -10649072 -11865971   -245847  -3314192  -2372798  -6412000
  -1261104  -6400670  -4974880   -450567  -7632171  -3422240   -933075
  -4745895 -13785174 -12982860 -23265000  -1093032  -8039430 -11492796
  -1019336  -8967714  -2743118  -3263532 -10066978 -13591413   -688611
   -318064 -11152500 -15162980  -9611238  -6539616  -1416606  -7514000
  -1201490]
Result by algorithm::
[ -4985776  -2974959  -1422030 -12887306  -7727307  -9058916  -7993705
  -9395742  -1266300 -14275392   -951054  -1275768  -6014075  -2209779
   -965818 -10649072 -11865971   -245847  -3314192  -2372798  -6412000
  -1261104  -6400670  -4974880   -450567  -7632171  -3422240   -933075
  -4745895 -13785174 -12982860 -23265000  -1093032  -8039430 -11492796
  -1019336  -8967714  -2743118  -3263532 -10066978 -13591413   -688611
   -318064 -11152500 -15162980  -9611238  -6539616  -1416606  -7514000
  -1201490]
True
```

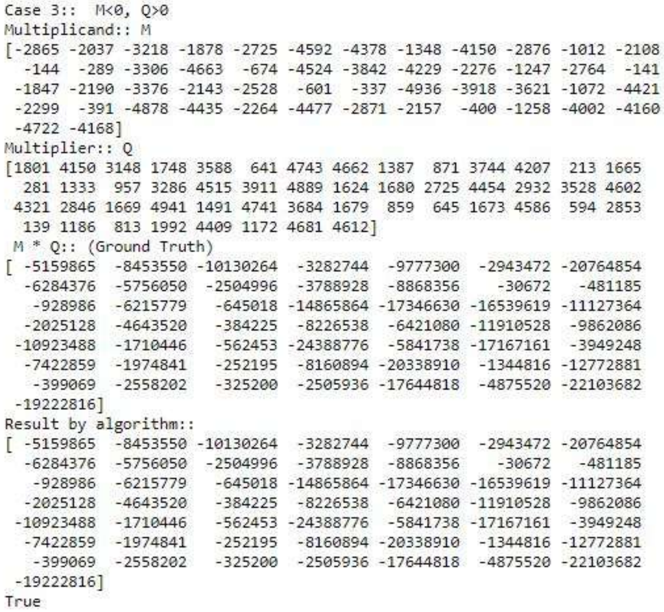
Fig. 6: Testing accuracy for M being positive in the range [0,5000] and Q being negative in the range [-5000,0].

Consequently, we avoided the binary addition on that extra bit taking place in each iteration. Hence, the total number of binary additions avoided is equal to the total number of iterations in the multiplication process. This C logic is also the reason for the size of the adder to be $'n'$ bits. Thus, the proposed C logic reduces the cost of the circuit and operations.

Secondly, in contrast to the original algorithm, where the values $-M$, $+2M$, and $-2M$ are computed every time they are needed, in the proposed architecture, they are computed only once irrespective of the number of times they are required for the multiplication process because they are stored in the dedicated registers viz., $MinusM$, $Plus2M$, and $Minus2M$. The presence of these dedicated registers also reduces the design complexity of the architecture.

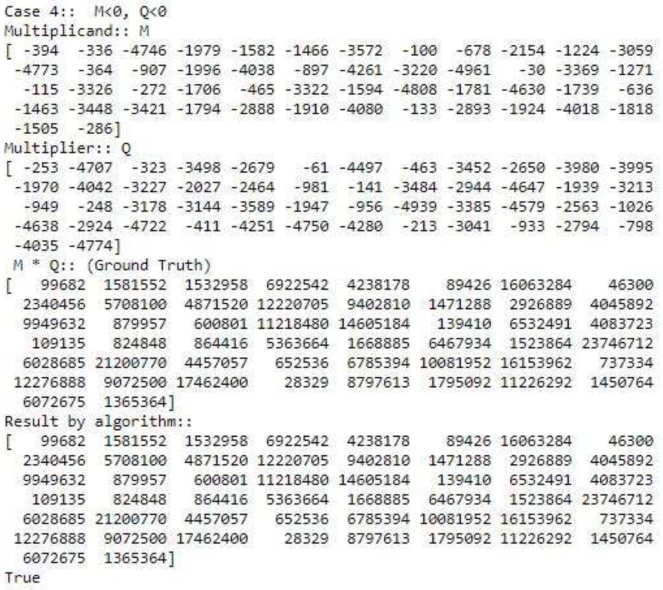
C. Comparison of different radix-4 booth multipliers

Table I compares the different radix-4 booth multipliers in terms of key error parameters namely MRE/MRED, NMED, and PRED. It is evident from the table that, the proposed multiplier is very accurate compared to the current multipliers.



```
Case 3::  M<0, Q>0
Multiplicand:: M
[-2865 -2037 -3218 -1878 -2725 -4592 -4378 -1348 -4150 -2876 -1012 -2108
  -144  -289 -3306 -4663  -674 -4524 -3842 -4229 -2276 -1247 -2764  -141
 -1847 -2190 -3376 -2143 -2528  -601  -337 -4936 -3918 -3621 -1072 -4421
 -2299  -391 -4878 -4435 -2264 -4477 -2871 -2157  -400 -1258 -4002 -4160
 -4722 -4168]
Multiplier:: Q
[1801 4150 3148 1748 3588  641 4743 4662 1387  871 3744 4207  213 1665
  281 1333  957 3286 4515 3911 4889 1624 1680 2725 4454 2932 3528 4602
 4321 2846 1669 4941 1491 4741 3684 1679  859  645 1673 4586  594 2853
  139 1186  813 1992 4409 1172 4681 4612]
 M * Q:: (Ground Truth)
[ -5159865  -8453550 -10130264  -3282744  -9777300  -2943472 -20764854
  -6284376  -5756050  -2504996  -3788928  -8868356    -30672   -481185
   -928986  -6215779   -645018 -14865864 -17346630 -16539619 -11127364
  -2025128  -4643520   -384225  -8226538  -6421080 -11910528  -9862086
 -10923488  -1710446   -562453 -24388776  -5841738 -17167161  -3949248
  -7422859  -1974841   -252195  -8160894 -20338910  -1344816 -12772881
   -399069  -2558202   -325200  -2505936 -17644818  -4875520 -22103682
 -19222816]
Result by algorithm::
[ -5159865  -8453550 -10130264  -3282744  -9777300  -2943472 -20764854
  -6284376  -5756050  -2504996  -3788928  -8868356    -30672   -481185
   -928986  -6215779   -645018 -14865864 -17346630 -16539619 -11127364
  -2025128  -4643520   -384225  -8226538  -6421080 -11910528  -9862086
 -10923488  -1710446   -562453 -24388776  -5841738 -17167161  -3949248
  -7422859  -1974841   -252195  -8160894 -20338910  -1344816 -12772881
   -399069  -2558202   -325200  -2505936 -17644818  -4875520 -22103682
 -19222816]
True
```

Fig. 7: Testing accuracy for M being negative in the range [-5000,0] and Q being positive in the range [0,5000].



```
Case 4::  M<0, Q<0
Multiplicand:: M
[ -394  -336 -4746 -1979 -1582 -1466 -3572  -100  -678 -2154 -1224 -3059
 -4773  -364  -907 -1996 -4038  -897 -4261 -3220 -4961   -30 -3369 -1271
  -115 -3326  -272 -1706  -465 -3322 -1594 -4808 -1781 -4630 -1739  -636
 -1463 -3448 -3421 -1794 -2888 -1910 -4080  -133 -2893 -1924 -4018 -1818
 -1505  -286]
Multiplier:: Q
[ -253 -4707  -323 -3498 -2679   -61 -4497  -463 -3452 -2650 -3980 -3995
 -1970 -4042 -3227 -2027 -2464  -981  -141 -3484 -2944 -4647 -1939 -3213
  -949  -248 -3178 -3144 -3589 -1947  -956 -4939 -3385 -4579 -2563 -1026
 -4638 -2924 -4722  -411 -4251 -4750 -4280  -213 -3041  -933 -2794  -798
 -4035 -4774]
 M * Q:: (Ground Truth)
[   99682  1581552  1532958  6922542  4238178    89426 16063284    46300
  2340456  5708100  4871520 12220705  9402810  1471288  2926889  4045892
  9949632   879957   600801 11218480 14605184   139410  6532491  4083723
   109135   824848   864416  5363664  1668885  6467934  1523864 23746712
  6028685 21200770  4457057   652536  6785394 10081952 16153962   737334
 12276888  9072500 17462400    28329  8797613  1795092 11226292  1450764
  6072675  1365364]
Result by algorithm::
[   99682  1581552  1532958  6922542  4238178    89426 16063284    46300
  2340456  5708100  4871520 12220705  9402810  1471288  2926889  4045892
  9949632   879957   600801 11218480 14605184   139410  6532491  4083723
   109135   824848   864416  5363664  1668885  6467934  1523864 23746712
  6028685 21200770  4457057   652536  6785394 10081952 16153962   737334
 12276888  9072500 17462400    28329  8797613  1795092 11226292  1450764
  6072675  1365364]
True
```

Fig. 8: Testing accuracy for M being negative in the range [-5000,0] and Q being negative in the range [-5000,0].

V. Conclusion

In this work, we introduced an accurate, cost-effective radix-4 booth multiplier. We optimized the method to multiply two signed binary integers using radix-4 booth multiplication, by reducing the number of bits that participate in the addition process while having sufficient storage area for reduced hardware design complexity. We also designed a high-level architecture for the same. Using the proposed algorithm as the blueprint, we compiled a program to compute the product of two signed integers (in Python) and hence evaluate the algorithm's viability. We achieved promising results i.e 100% accuracy when we tested it against a sample size of 10,000 random integer pairs in the domain of (-5000,5000). We also compared it with the current radix-4 booth multipliers in specific key error parameters and showed that our algorithm performed better in terms of accuracy. We also showed how our algorithm is successful in being more cost-effective as compared to the original methodology by critically analyzing the cost of the multiplier circuit and its operations. The proposed algorithm can be used in several machine learning and digital signal processing applications like convolutional neural network computations. It can also be used in lossless applications which are highly intolerable of errors. Future work is to simulate the hardware architecture and evaluate the same in terms of energy efficiency.

TABLE I: Comparison of error parameters of different radix-4 booth multipliers.

Multiplier	**MRE/MRED**	**NMED**	**PRED (%)**
R4ABM2 [1]	0.344×10^{-2}	0.2543×10^{-5}	99.79
	2.52×10^{7}	17879×10^{-5}	0.123
RAD1024 [2]	0.93%	NE	93.26
CABM [3]	0.014	0.18	NE
ABM-M2 [4]	2.689×10^{-2}	1087.270×10^{-6}	NE
LOBO12 [5]	0.44-96.09%	18.45×10^{-4}	90.99
HLR-BM1 [6]	0.03%	0.79×10^{-5}	98.35
Proposed	-	-	-

* NE: Not Evaluated

References

[1] W. Liu, L. Qian, C. Wang, H. Jiang, J. Han, and F. Lombardi, "Design of Approximate Radix-4 Booth Multipliers for Error-Tolerant Computing," *IEEE Transactions on Computers*, vol. 66, pp. 1435–1441, 2017.

[2] V. Leon, G. Zervakis, D. Soudris, and K. Pekmestzi, "Approximate Hybrid High Radix Encoding for Energy-Efficient Inexact Multipliers," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, pp. 421–430, 2018.

[3] M. S. Ansari, H. Jiang, B. F. Cockburn, and J. Han, "Low-Power Approximate Multipliers Using Encoded Partial Products and Approximate Compressors," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, pp. 404–416, 2018.

[4] S. Venkatachalam, E. Adams, H. J. Lee, and S.-B. Ko, "Design and Analysis of Area and Power Efficient Approximate Booth Multipliers," *IEEE Transactions on Computers*, vol. 68, pp. 1697–1703, 2019.

[5] R. Pilipovic and P. Bulic, "On the Design of Logarithmic Multiplier Using Radix-4 Booth Encoding," *IEEE Access*, vol. 8, pp. 64 578–64 590, 2020.

[6] H. Waris, C. Wang, and W. Liu, "Hybrid Low Radix Encoding based Approximate Booth Multipliers," *IEEE Transactions on Circuits and Systems II: Express Briefs*, to be published.

[7] G. Jain, M. Jain, and G. Gupta, "Design of Radix-4,16,32 Approx Booth Multiplier Using Error Tolerant Application," in *6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2017.

[8] S. Venkatachalam, H. J. Lee, and S.-B. Ko, "Power Efficient Approximate Booth Multiplier," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018.

[9] T. Zhang, W. Liu, J. Han, and F. Lombardi, "Design and Analysis of Majority Logic Based Approximate Radix-4 Booth Encoders," in *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2019.

[10] N. R. Varghese and S. Rajula, "High Speed Low Power Radix 4 Approximate Booth Multiplier," in *3rd International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech)*, 2019.

Security Procedures for User-Centric Ultra-Dense 5G Networks

Mohammed Rajhi
Department of Computer Science and Networking
Jazan University
mrajei@jazanu.edu.sa

***Abstract*— The super-compressed network is a forthcoming 5G that can without difficulty attend to network system sufficiency concerns. To be able to attain the uncontrolled and diminutive base station placement, the need for UDN (Ultra-Dense Network) architecture can't be overemphasized. This paper takes the route of the UDN architecture plan, including the remonstrance's it comprises of with adding the security demands and procedure to deal with them. The new security procedure is fashioned after the implicit certificate or IC scheme, which decides the APS's security issues. The security process involvesthe APG (APs Group) security, network security domain, AP (Access Point) access security, UE (User Equipment) to UUDN (User-centric Ultra-Dense Network) access network security, dual authentication in addition to expanded keys.**

***Index Terms: Ultra-Dense network, User Centric, Security Challenges*, 5G.**

I. Introduction

Data traffic is about to witness great innovations in years not so distant away. End-users' device and wireless systems are growing rapidly, which will allow the introduction of new and better mobile applications. Due to this, a rising number of flexible organizations limit will be substantial for previous 2020 social orders, not particularly to achieve the excellent climb of client traffic acquired from these new compact broadband administrations most likely to also put intoconsideration the surplus supply of devices needed for their use [1]. To demonstrate the 5G network, the most suitable rundownis the ultra-dense networks (UDNs). UDNs works with the utmost traffic necessities through infrastructure densification. (9-1). Bestirred by the inability of new introductions to effectuate such requirements, the researched network has during the recent decade been working effectively towards establishing the understructure of the current Fifth Generation (5G) for the wireless networks. UDNs, in comparison to the latest traditional deployments, make use of denser arrangement for the access point [1]. The organizational disposition of more BSs or reception apparatuses per unit territory constitute the assets' spatial reuse. In the aftermath, more assets can be distributed toall client, bettering the previous information rates communicated to them. The last constituent remodels UDN into a doubtlessly suitable condition for the application of Millimeter-Wave, where a large percentage of data transfer capacity enables exceedingly high throughput under LOS situations [2][1]. To increase network adequacy, densification has been proven to be the most direct mechanism. Nevertheless, there are some factors that limit the dense network, as they tendto cause a high volume of inference. So, there is the need for further studies to figure out the limits by making use of new models. The application of DAS (Distributed Antenna Systems) is another technology that can maximize network capability. It is like a natural facilitator for the Millimeter-Wave for indoor placement.

Details	Traditional Cellular Network	Ultra-Dense Network
Access Point Density	Lower than user density	Comparable to the user density
Coverage	Hundred meters and more	Almost 10 m
Characteristics of Coverage	Regular cell, single layer	Heterogeneous, irregular
Backhaul	Wired, ideal	Ideal, non-ideal, wired, wireless
User Density	Low	High
Mobility	High	Low
Spectrum Bands	Low, limited	Higher, wider
Data Rate Requirement	Lower, Medium	High
Deployment	Wider coverage	Indoor, hotspot
Access point type	Macro, Macro Base Station	Small cell, relay, UErelay

Table 1.1 Comparison between Traditional Cellular Network and UDN.

A. *Ultra-Dense Network*

UDN is intent on giving a comparatively high data usage to every consumer indoor and in hotspot areas, including offices, packed residential places, arenas, and open-air-converging points, and many more [7]. Area spectrum efficiency becomes one of the key metrics to determine future mobile systems for extremely dense traffic demands [8], and then influences system capacity. A type of UDN consist of the static virtual cell that has quite the number of access points by which a small static cell can be created and can likewise accord the same coverage. User-centric is a different type of UDNs that can integrate with the local control center same as with the user, and relying on the unit of the single user, the virtual combined cell is elucidated. There are some standards and procedures that are illustrated simply reliant on architecture design. Localization and flatten is a practical way to handle 100 times excess traffic. Localization relies on backing from the flattened architecture and therefore reduces the expenses used in the transmission. METIS and NGMN put forward the localization of the data path [13] [14]. UDNs varies from the traditional cellular network which is why the user-centric principle is vital to apply. On the other hand, RAN (Radio access network) is the centralized access network, and the architecture of UDN is crucial here.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

It is able to give a preferable joint processing and thus deliver greater spectrum efficiency, but there is certainly still high potentiality. Apportioned architecture is far more adaptable for the placement process of the network. So the architecture that is farmore adaptable is usually preferred to connect with APS.

B. User-Centric Ultra-Dense Network

The 5G network faces more extensive and intricate security peril compared with the now 3G and 4G networks. It customary security dangers are remembered for it many UEs' diversity and the clarity of the remote channel. Additionally, it also makes use of new security instability from the upgraded usability in different designs that uses the merging between improved heterogeneous remote organizations, the open organization foundation is reliant on the IP system, and the enhanced business carrier with different trust-appraisals. In the UUDN, the system organizes a dynamic APs group (APG) that relies on each UE's condition. It delivers unnoticed and seamless service to the user all through the APG dynamic, reloading as an unobserved network coverage going together with user mobility [2]. Importantly, in user-centric UDNs (UC-UDNs), the user-centric collections are created by categorizing the more preferable number of APs collectively, therefore leading to a user-centric clustering architecture [9].

C. 5G Overview

The precise statement of meaning for 5G has always been argued. In a GSMA Intelligence report that was published recently, illustrated two aspects of the 5G technology that exist today: the hyper-connected vision and the next-generation radio access technology [10]. The latter which is the hyper-connected invention, is depicted as a recent technology in which all the advantageous features from the current generations like the 2G, 3G, and 4G are scratched out for the enablement of a better system that can provide users with wider implications than the 4G and LTE that exist now [10]. The UE dispatches a fundamental access demand. At that juncture, the AP gets the requisition and conveys it to the nearest organization framework. As per the requisition's initial situation, the nearest organization framework requires that the center organization framework compare network layer authentication vector and output. At that moment, based on the resulting network layer security boundary, the nearest organization framework begins the organization layer's standard authentication measure with the UE (like the 4G EPS-AKA measure).

II. SECURITY CHALLENGES OF UUDN

Soon as the organization layer's average verification of the UE is done, a static virtual cell or an APG is charged to the UE by the UDN. The framework of the network creates network access validation vectors to the UDN control layer which is reliant on the requisition boundary acquired from the UE. It coordinates the entrance layer (a particular virtual cell or APG) standard verification measure on the UE. When the UE does theshared validation measures toward both the organization layer and the entrance layer, the security access is refined [4]. APs are interlinked and arranged depending on many structural datalike it also relies on the static virtual cell. So, the UDN should actualize the access services irrespective of what the APS organization is. Like channel interference or any other type of interference, some factors can be erased within the UDN. Each network element requires to be generally verified, and the two-sided substance should make use of their different private keys. The security system needs to be used to ascertain that the two sides can receive essential information. Each comparable element need to have an alternative when acquiring the shared keys during information exchanges. The different correspondence comes together to make use of the uniquely shared keys. The security measures for encoded information is reliant on the shared keys and should uphold the dynamic collaboration or disengagement of correspondence elements. All entities should be bounded together by the command from the organization administrator. The longevity of divided keys among corresponding elements must be equal to the significant directions. The security implement must uphold the different legitimate channels amongst uniform sources or objections and keep a strategic distance from replicating the keystream. The security system must be dynamic to give an assurance of speedy reactions and adapt to the element's demonstration and data transfer capabilities in different corresponding capacity. When we make use of the traditional cellular network architecture, it can result to issues like too many functions being centralized in the core network. UDN wouldn't be efficient with high traffic throughput because it emanates from the overhead signaling and long data transmission between the network and APS [3]. Moreso, the highest layer functions of every allocated AP need to decline for a smoother support in interference management because the higher-layer procedures cannot in a long shot be appropriate for AP. When the coverage of AP is minimal, it can cause more handover. This will lead to ineffectiveness and less flexibility within both macro and UDN.

III. INTELLIGENT NETWORKING

Researches have been carried out on ultra-thin intelligent systems administration, which brought about change and allocates framework boundaries and techniques through the organization's clever perspective of remote climate andbusiness needs. Naturally developing the information data for each client and then using that data for the radio asset of the board, UUDN should maintain new perception in joining the executive's plane, client plane, and control plane.

	3GPP SCE	3GPP HeNB	UUDN
Flexible Backhauling	No	No	Yes
NFV & SDN base	No	No	Yes
U/C separation	Yes	No	Yes
User-centric	No	No	Yes
Localization and flatter	LIP & SIPTO	LIP & SIPTO	Localized (user plane)

Table 3.1 The differences between the networks

The main objective is to enhance the range productivity and consumer experience, achieve low energy or low force usage of the framework, lessen the network actions, and minimize standard upkeep costs. From the net-work and the board's perspective, to decrease the capital consumption and operating cost, UUDN should maintain self-systems organization and self-streamlining. To sustain the organization's adaptability, self-backhauling is also critical for UDN. The common cell network configuration places around isolating access, access configuration and the executives. In the UUDN scheme, the executives' plane of the network management is conducted to build up a universal point of view on the organization, to achieve the natural merging among the board control and user plane, and to recognize intelligent systems administration [6].

IV. Literature Review

More recently, user-centric UDNs have attracted substantial consideration. This is because of their capacity to satisfy each UE's TQoSin dense environs regardless of the location. In particular, some research efforts have focused on the user-centric assembling design issues [12]. In UUDN, the system can intelligently recognize the user's wireless interaction environments and then with flexibility organize the needed APG and abilities to help the user; subsequently, it becomes like a network is always following the user [11].

A. *Physical access security of entity APS*

The framework of UUDN is much flatter than the TCN, and the ultra - dense APs will be intensely launched to the clients [11]. The APS will be initiated by the users and then out of the joint launching by the operators. Consequently, it isn't easy to assure of a secure physical transmission environment [11]. The attacker can illegitimately acquire or make a copy of the digital certificate and the sensitive information stored in the APS. Because of this, the cloned APs may get into the UUDNnetwork and release security concerns into the network [11]. Malicious apps may attack CN entities and compromise the UE, such as spying on packages' operation.

B. *Dynamic network resources organization security*

Access point group is the key characteristic of the UUDN. Many APs in UUDN gets aptly organized into an APG to observe the user's movement and supply data transmission as required. The dynamic adjustment of APG will lead to some security concerns for UUDN [11]. The multiple numbers of APs connected to the network will have interacted with each other. It will make its security more difficult to realize self-configuration, self-optimization, and self-healing [11]. Together with the user movement, the APG will be entrenched and refreshed [11]. The security of the APG structure and refreshing is the key upset to UUDN [11].

C. *User information and privacy protection security*

The protection of user information and privacy is another vital concern for UUDN based on all IP networks linking the Internet [11]. Today and in the future of 5G, it is a common thing for individuals to keep their mobile devices physically around them. Assuming an UE's traffic is hijacked in a place of residence by an assailant [11]. If this is the case, it may be understood if a particular individual is present (or not) at one particular location, thus making an individual's privacy vulnerable to attacks [11]. The privacy data security contains the information prolateness and data wholeness of the user in question [11].

D. *Handover validation security*

Many handovers and verifications will often bring delays and a high probability of authentication failure. Thus, significantly reducing the user's experience [11]. Unmistakably, this traditional authentication mode cannot be supported entirely by UUDN [11]. To extenuate the lousy experience and adapt the APG mode, we find that a new authentication and key agreement (AKA) method, such as a group authentication procedure approach, should be adopted strongly in UUDN [11]. The group validation result should be easily inherited, recognized, and confidential among the group participants [11].

A different type of UDN is the user-centric UDN (UUDN), which has a local access center orchestrated with the user, and the virtually joined cell is created based on the unit of a single user [5] [2]. Moreover, more investigation of the spectral efficiency and energy efficiency and proposed energy-efficiency and user-centric cross-tier assembling solution can be subject to a minimum spectral efficiency constraint [11].

V. Proposed Methods

The structure of UDN and UUDN are flatter than common cellular networks. Furthermore, data communication between networking bodies is random and in a way not permanent. We will demonstrate some security domain procedures to achieve security. Those security domains are network domain security, AP access security, APG security, Dual authentication, UE to UUDN access network security and expanded key.

1) **Network Security domain**: APSs are interlinked and arranged according to various structural data like it also relies on the static virtual cell. So, the UDN has to actualize the access services irrespective of what the APS organization is. Some factors, like cochannel interference or an interference of any kind, can be erased within the UDN. Each network entityshould be generally verified, and the two-sided elements shouldmake use of their different private keys. The security system should be used to ensure that the two sides can get considerable information. Each correspondence entity should have the option to obtain the shared keys in information exchanges. Themany various correspondence joining's make use of distinctive shared keys. The security instruments for information encodedreliant on shared keys should maintain the dynamic joining or exit of correspondence elements. All entities should be bound together by the commands from the organization administrator.

2) **AP Access Security**: The APs are very close to the user and may even be launched by the user. The security concerns of AP deployment, which UUDN encounters, are same as that of Home eNB in the LTE network. To restrict the attacker from using HNB as an advantage to penetrate into the LTE network, the HeNB supports a device verification approach with certification-based or USIM-based consensual authentication. Same as the APS also requires mutual validations among the APs and UUDN network when the APs access the UUDN [3]. When the validation is successful, the APs can then access the operational status.

3) **APG Security**: At the point when the organization layer usual authentication of the UE is done, a static virtual cell or an APG is attached to the UE by the UDN. Currently, the network framework creates network access confirmation vectors to the UDN control layer reliant on the requisition boundary obtained from the UE. It directs the entrance layer (a particular virtual cell or APG) common validation measure on the UE.

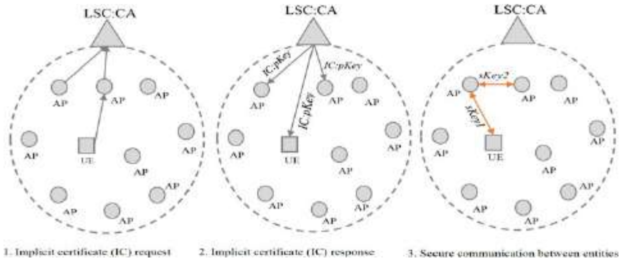

LSC:CA
LSC:CA
LSC:CA
AP
UE
IC:pKey
IC:pKey
IC:pKey
sKey2
sKey1
1. Implicit certificate (IC) request
2. Implicit certificate (IC) response
3. Secure communication between entities

fig 1 APG refresh and UE’s movement

During the period when the mutual validation leans toward both the organization layer and the entrance layer are concluded by the UE, the security access is refined [3]. In respect to the security policy, when the APG is unable to ascertain the member's AP or the AP shuts down and reports to APG, the key of the AP acquired in LSC will be eliminated. LSC will erase the AP from the APG index list. Simultaneously, the LSC will transfer some configuration adjustment commands to the APG and its members. If the AP rejoins APG, the validation procedure is required.

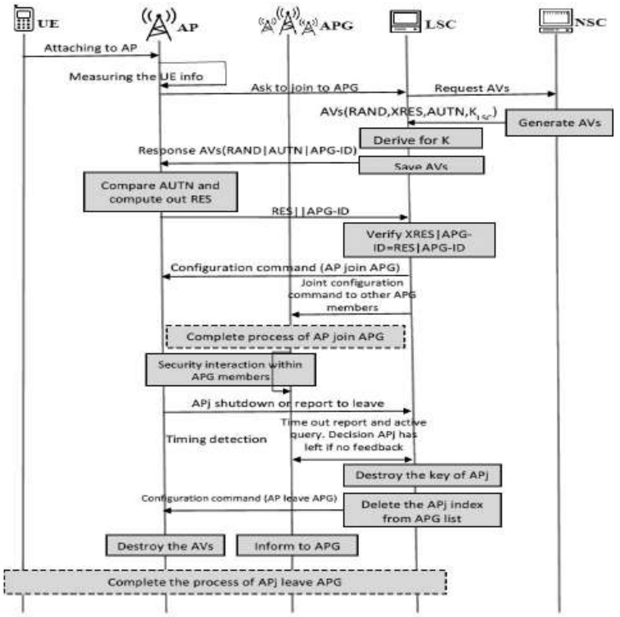

UE
AP
APG
LSC
NSC
Attaching to AP
Measuring the UE info
Ask to join to APG
Request AVs
AVs(RAND,XRES,AUTN,K$_{LSC}$)
Generate AVs
Derive for K
Response AVs(RAND|AUTN|APG-ID)
Save AVs
Compare AUTN and compute out RES
RES||APG-ID
Verify XRES|APG-ID=RES|APG-ID
Configuration command (AP join APG)
Joint configuration command to other APG members
Complete process of AP join APG
Security interaction within APG members
APj shutdown or report to leave
Timing detection
Time out report and active query. Decision APj has left if no feedback
Destroy the key of APj
Configuration command (AP leave APG)
Delete the APj index from APG list
Destroy the AVs
Inform to APG
Complete the process of APj leave APG

Fig 2 The APG-AKA in UUD

4) **Dual authentication**: The duration of divided keys amongst correspondence entities should be as per the cogent directions. The security implements should maintain different legitimate channels within similar sources or objections and sustain a strategic distance from replicating the keystream. The security system should be dynamic to assure of speedy reactions and adapt to the entity's demonstration and data transference capability in different correspondence undertakings. The network architecture design also becomes affected because of the new ways of the SDN and NFV for the 5G network architecture can predetermine the primary functionswith the aid of decoupling and transmission.

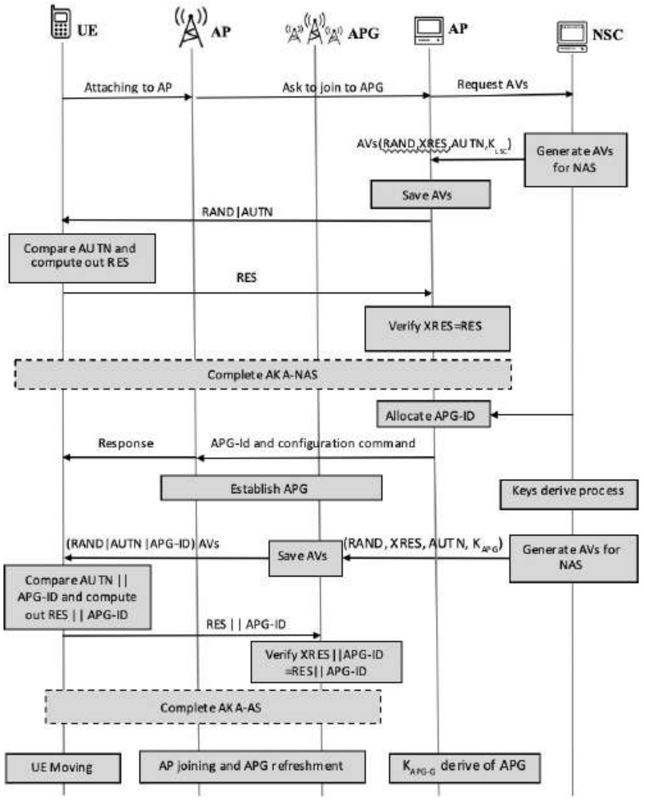


fig 3 Process of APG-AKA

It makes the architecture far more flexible toward the 5G dual connections and user plan is an efficient way to give higher user data rate by dense mini-cells with no mobility or any connectivity experience [4].

5) **UE to UUDN Access Network Security:** Due to the restoring of APG, many security concerns have arisen. APG members will be replaced, and attackers may latch onto the AP wireless connections. For example, when a data transference between the UE and APs (APG), the signaling and user data can be accessed illegitimately or can be exploited by the attackers. The user's movement between the APs may also be tracked or acquired at the actual user's locations. These can pose serious concerns for user's privacy. It efficiently shields the user's private data with the keys based on specific encryption algorithms.

6) **Expanded key:** To ensure network safety, the user data are linked to an encryption algorithm. Each process requires a different key. Without a doubt, the key for integrity procedures or ciphering shouldn't include transmitting data over the network directly. On the other hand, there is a risk that the key can be accessed illegitimately or even upstaged. A a more reliable approach is to ensure that the temporary key derives the keys for ciphering or integrity, and then the keys never exit the security elements. The key that is derived by hierarchy should be designed to adapt to the network architecture. It illustrates the different keys and the other keys in the key hierarchy system in LTE from which they are obtained. Characteristically, the design of an expanded key obtained from the hierarchy of UUEN and it has its own similarities. There may be multiple APGs in UUDN, even in similar network area, that operates in accordance to different UEs. An AP may serve various UEs at the same time, and an AP may also originate from different APGs. Even then, there are still concerns that the APG may be abused or misused by malicious APs.

VI. Conclusion

Meeting the high traffic maximum necessary of the 5G, UDN is a very impressive innovative bearing. In this work, a UUDN is the proposed idea. For this reason, we gave a network design for the UUDN idea. An architecture is structured with the probability of limitation, compliment, U/C partition, keenness, client-driven and adjustable systems administration. Due to the new structure, challenges and observations, many key innovations are familiar with giving high QoE, high region capability, productivity, minimal cost, and green correspondence. Four different practical cutting-edge subtitles are discussed in headings which includes; intelligent networking, interference management, and security.

VII. Future Work

For further studies, there is a massive range of areas to explore in relation to DAPGing, high level interference management, and many security clarifications, along with their disadvantages. In order to make UUDN a reality in the foreseeable future for deployment, a lot of areas needs to be covered for further studies. Another issue that needs to be appropriately addressed is the cooperative networking and heterogeneous networking. To support UUDN which is full of various complicated scenarios and inconsistent coverage will surely prove to be a big challenge.

VIII. Abbreviations

AKA	Authentication and key agreement
3GPP	3rd Generation Partnership Project
AP	Access Point
APG	Aps Group
BS	Base Station
DAPGing	Dynamic Aps Grouping
NFV	Network Function Virtualization
U/C	User/control plane
UUDN	User-centric Ultra-Dense Network
UDN	Ultra-dense network
APG-ID	APG-identity
LTE	Long term evolution
EPS	Evolved Packet System
NSC	Network service center
UE	User Equipment
QoE	Quality of experience
MeNB	Master evolved Node B
LSC	Local service center

IX. Acknowledgment

I am grateful to my supervisor and all CS department**,** for assisting me in this project. His skills and knowledge were indispensable, we also appreciate his efforts in giving technical piece of admonition and consultation. Secondly, it is a great privilege for me to be a recipient of the scholarship that got me here and I will like to acknowledge the efforts of my scholarship provider **Jazan University** and all of our friends and colleagues in the computer communication lab as well.

X. References

[1] S. G. ´. C.´as, "Ultra Dense Networks Deployment for beyond 2020 Technologies," Departamento de Comunicaciones Universitat Polit`ecnica de Val`encia, pp. 110-118, 2017.

[2] S. Chen, "A Security Scheme of 5G Ultradense Network Based on the Implicit Certificate," Recent Advances in Cloud-Aware Mobile Fog Computing, pp. 345-355, 2018.

[3] Rost, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine,* pp. 68-76, 2014.

[4] F. Qin, "User-Centric Ultra-Dense Networks (UUDN) for 5G: Challenges, Methodologies and Directions," *IEEE Wireless Communication Magazine,* pp. 78-85, 2016.

[5] S. Chen, "User-Centric Ultra-Dense Networks (UUDN) for 5G: Challenges, Methodologies and Directions,," IEEE Wireless Communication Magazine, pp. 85-98, 2016.

[6] S. Chen, F. Qin, B. Hu, X. Li, Z. Chen and J. Liu, User Centric Ultra dense networks for 5G, Springer, 2018.

[7] ITU-R Report M.2320, "Future Technology Trends of Ter-restrial IMT Systems," Nov. 2014.

[8] M. S. Alouini and A. J. Goldsmith, "Area Spectral Effi-ciency of Cellular Mobile Radio Systems," IEEE Trans. Vehic. Tech., vol. 48, no. 4, July 1999, pp. 1047–66.

[9] Lin, Yan, R. Zhang, L. Yang, Chunguo Li and L. Hanzo. "User-centric clustering aided design of ultra dense networks." (2019).

[10] Paudel, Prabesh, and Abhi Bhattarai. " 5G Telecommunication Technology: History, Overview, Requirements and Use Case Scenario in Context of Nepal," May 19, 2018.

[11] Chen, Zhonglin, S. Chen, H. Xu and B. Hu. "Security architecture and scheme of user-centric ultra-dense network (UUDN)." Trans. Emerg. Telecommun. Technol. 28 (2017): n. pag.

[12] Lin, Y., Rong Zhang, L. Yang and L. Hanzo. "Secure User-Centric Clustering for Energy Efficient Ultra-Dense Networks: Design and Optimization." IEEE Journal on Selected Areas in Communications 36 (2018): 1609-1621

[13] Uwe Doetsch, Nico Bayer, et al., "Final Report on Architecture," METIS Deliverable D6.4, January 2015.

[14] 7. Rachid EI Hattachi, Javan Erfanian, et al., "NGMN 5G Initiative White Paper," February 2015, http://www.ngmn.org/5g-white-paper.html.

An Ultra-Low Power MOS_2 Tunnel Field Effect Transistor PLL Design for IoT Applications

Naheem Olakunle Adesina
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
nadesi1@lsu.edu

Ashok Srivastava
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
eesriv@lsu.edu

Md Azmot Ullah Khan
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
mkhan42@lsu.edu

Jian Xu
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
jianxu1@lsu.edu

***Abstract*—This work presents the implementation of analytical transport model of MoS_2 tunnel field effect transistor (TFET) using Verilog-A in Cadence/Spectre. The parameters of the model are extracted, and most notable is its high I_{ON}/I_{OFF} ratio, which makes it suitable for low power design, and IoT applications. Subsequently, we employ the TFET model in the design of ultra-low power phase locked loop (PLL). Various components of PLL are examined in terms of their structures and functions. The results show that the voltage controlled oscillator (VCO) operates from 0.5 GHz to 2.9 GHz with a tuning range of 82.8%. It also consumes 1.91 µW power at 2 GHz carrier and has a phase noise of -117.3 dBc/Hz at 1 MHz offset frequency. The average power consumed by phase locked loop is 0.021 mW. In addition, the operation of the PLL is not sensitive to temperature variations.**

Keyword—phase locked loop (PLL), phase noise, tuning range, tunnel field effect transistor (TFET), voltage controlled oscillator (VCO)

I. INTRODUCTION

It is already established that tunneling field effect transistor (TFET) is a good choice for designing low power circuits. This is because of its ability to operate with low supply voltage (sub-0.5 V), which drastically reduces the static power consumption. The miniaturization of MOS transistor has reached the limit where we experience short channel effect (SCE) such as drain induced barrier lowering. In this case, the OFF-state current (I_{OFF}) increases as a result of lowered potential barrier for electrons from source to drain. This makes short channel MOSFET unsuitable for digital logic applications because of its low I_{ON}/I_{OFF} ratio. The maximum sub-threshold slope (SS) achievable by a typical MOS transistor is 60 mV/decade at room temperature. Tunneling field effect transistor has shown much better performance in terms of SS, leakage current, power consumption etc. Applying an appropriate bias voltage to the gate-source region of TFET causes the valence band to align with the conduction band and injecting the charge carriers. Similarly, under reverse bias, the two bands are misaligned and there is no injection of carriers. This effect is referred to as band-to-band tunneling [1 - 3]. Unlike thermionic-based MOSFET, the principle of operation of TFET is less dependent on temperature. Thus, process, voltage, and temperature (PVT) variations are not serious concerns in TFET-based circuits. Although the dynamic power decreases linearly in CMOS technology, the static power continues to increase because the threshold voltage cannot be scaled proportionally with the supply voltage. This causes the leakage current to increase, which consequently increases the leakage power. In order to achieve ultra-low power or energy-efficient design in analog, digital or mixed signal circuits, it is desirable to employ a transistor with different mechanisms of injecting charge carriers other than thermionic or temperature dependent. Some other benefits of TFET is good compatibility CMOS transistor and can also be integrated with the current FinFET device. Nevertheless, it does not show a better performance than CMOS when operated beyond a certain voltage. So, it cannot completely replace CMOS transistor in high performance applications [4]. In addition, it does not have a symmetrical structure like MOSFET because the source and drain are made of and doped with different materials.

There are various choices of material that are suitable for TFET design. Some of the factors considered are low band gap and low effective mass that aids the tunneling probabilities of electrons and holes carriers. It ensures that TFET produces high I_{DS} even at low voltage supply. Graphene is a promising candidate and it satisfies most of the requirements in material choice selection. It is regarded as zero band-gap material because there's no band gap between conduction and valence bands. At Dirac's point, graphene is considered massless because its effective mass is zero. However, studies have shown that transistor made of graphene, in its pristine form, has high I_{OFF} current. One of the ways to overcome this challenge is by creating nanoribbon with the graphene, which eventually opens up the band gap and lowers I_{OFF} current. The graphene nanoribbon (GNR) TFET provides the required I_{ON}/I_{OFF} ratio

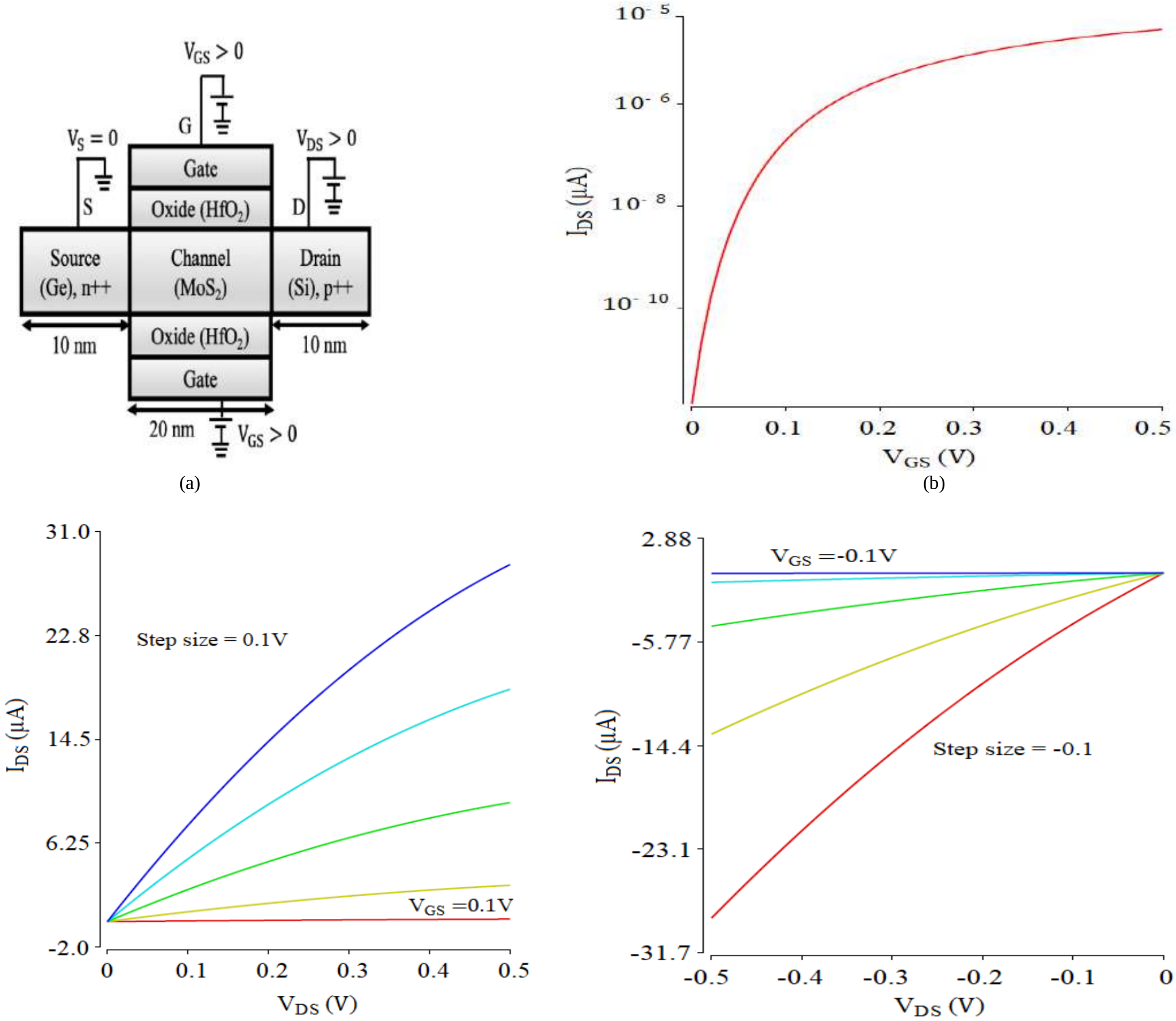
978-1-6654-4067-7/21/$31.00 ©2021 IEEE

for digital applications [5]. Transition metal dichalcogenide (TMDC) are 2-D materials that have also been employed in the fabrication of TFET. Unlike graphene, TMDCs have infinite band gaps, which usually ranges from 1 to 2 eV. Although transistor made of TMDCs has a low I_{ON} current when compared with graphene-based transistor, it has better I_{ON}/I_{OFF} ratio because of its low OFF-state leakage current. This makes it more suitable for logic design. Phase locked loop operates based on feedback control mechanism whereby the external reference is expected to track the feedback signal from frequency divider. It has various applications such as data clock recovery, frequency synthesizer etc. PLL also has been widely used in the field of power electronics and communication. In this work, we have designed a novel low power phase locked loop with tunneling field effect transistor. Its performances in terms of tuning range, center frequency, power consumption, and phase noise are also presented and compared with related work in other transistor technologies.

This paper is organized as follows; we present the brief description of transistor model in Section II, Section III examines the design of each component of PLL and discusses the results, and Section IV is the conclusion.

II. MODEL DESCRIPTION

The model shown in Fig. 1(b) is an heterostructure, which consists of source, drain, and double gates. The source is made of heavily doped Ge and the drain is silicon. The transistor channel is 0.65 nm thick MOS_2 with a band gap of 1.79 eV. For the gate oxide, HfO_2 is chosen because of its high dielectric. The source and drain are doped to ensure proper band tunneling and silicon is chosen as drain material because of its higher band gap, which lowers the flow of ambipolar current.



(a)

(b)

(c)

(d)

Fig. 1. (a) nTFET structure (b) $I_D - V_{GS}$ logarithmic plot for nTFET (c) $I_D - V_{DS}$ plot for n-type TFET (d) $I_D - V_{DS}$ plot for p-type TFET.

Similarly, the p-type TFET also has the same structure, but with reverse doping for the source and drain. The polarity of V_{GS} is also reversed for appropriate bias of gate-source junction. We will not go into the details of our TFET model because it is already presented in [6].

The behavior of both n and p-type transistors are summarized in Figs. 1(b) – 1(d). As shown in Fig. 1(b) , $I_D - V_{GS}$ logarithmic plot shows that the n-type TFET has high I_{ON}/I_{OFF} ratio of 5.27×10^7 for $V_{GS} = V_{DS} = 0.5$V. The output characteristics of nTFET is presented in Fig. 1(c) with $V_{DS} =$ 0.5V and V_{GS} ranges from 0.1 to 0.5V with a step size of 0.1 V. Similarly, $I_D - V_{DS}$ plot for p-type is shown in Fig. 1(d). In this case, V_{DS} = -0.5V and the values of V_{GS} ranges from -0.1V to -0.5V with a step size of -0.1V. In addition, the transistor model achieves SS as low as 10 mV/decade, which is six order of magnitude better than MOSFET. We can conclude that the TFET is suitable for low power design, such as internet of things (IoT) applications.

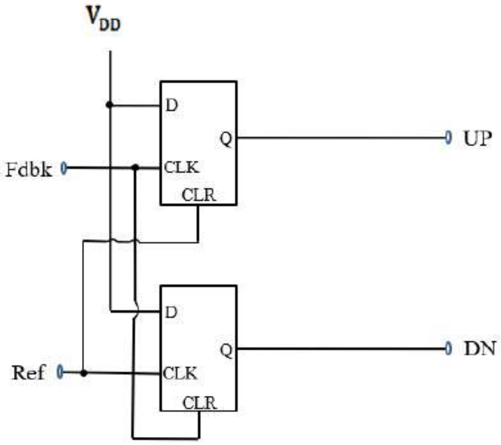
III. TFET-BASED PHASE LOCKED LOOP DESIGN

Phase locked loop consists of different building blocks that are used to generate a stable output frequency, which is a multiple of the input reference. The usual components of PLL are voltage controlled oscillator (VCO), phase frequency detector (PFD), charge pump, loop filter (LF), and frequency divider. PFD produces an error signal from the differential phase and frequency of both input reference and feedback signals. Depending on which of the two signals is leading, the frequency of VCO is increased or decreased by the error signal. The loop capacitors are charged and discharged through charge pump output current. The loop passive filter is responsible for eliminating high frequency components and reference spurs. LF also determines the bandwidth and stability of phase locked loop. It generates the control voltage, which alters the output of voltage-controlled oscillator. The frequency divider takes in the output of VCO, samples it, and convert it to a fractional output, which is fed back to PFD. The phase detector we used in this work has quite similar structure with the conventional tri-state PFD. It employs two D flip flops, but without NAND gate. The reset path is modified to reduce dead zone and power consumption. Figures 2(a) and 2(b), respectively, show the structure and output of PFD when the reference and feedback signals are in phase. It is shown in Fig. 2(b) that the UP and DN outputs have low glitches, which makes it more energy efficient and useful for IoT devices (e.g. IoT sensor nodes).

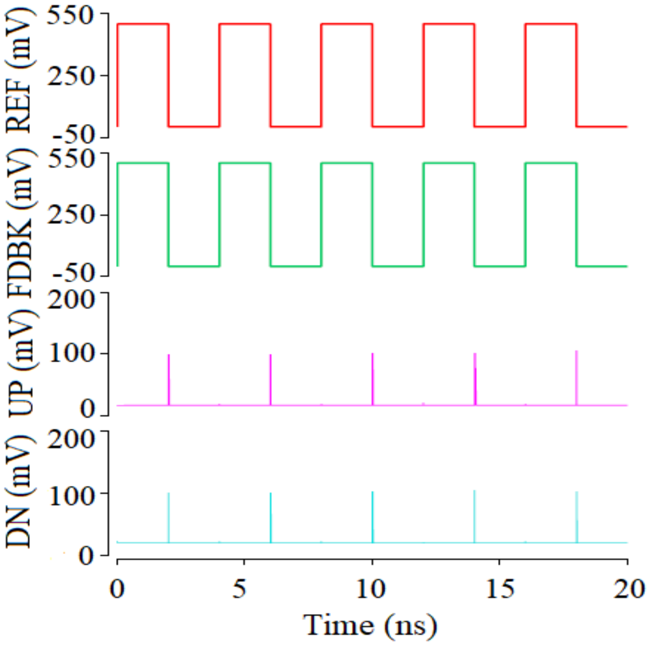
The charge pump (CP) takes the two outputs (UP and DN signals) from PFD and convert to charge pump current that charges and discharges the capacitors of the loop filter. LF is a second order low pass filter and it controls the dynamics and transient response of PLL. The CP/LF configuration in [7] provides the needed control signal, V_{ctrl}, for voltage controlled oscillator. More so, the loop filter prevents instability of the loop in lock condition from slight variations in the input. This is achieved by blocking higher order harmonics from PFD. The transfer function of LPF is given as:

$$F(s) = \frac{Vctrl(s)}{I_{cp}(s)} = \frac{1+sRC_1}{s(C_1+C_2)+s^2RC_1C_2} \qquad (1)$$

where R, C_1, C_2 are the loop filter resistance and capacitors, respectively.



(a)



(b)

Fig. 2. (a) PFD (b) UP and DOWN outputs.

The current starved voltage-controlled oscillator (CSVCO) is preferred to LC VCO because of its wide tuning range, low cost, small silicon area, and low power consumption. However, LC voltage controlled oscillator can achieve a better phase noise than CSVCO through high Q inductor. Furthermore, it is widely used in high frequency applications such as RF. The control voltage from CP/LF provides bias current for each inverter stage, which charges their load capacitances. The biasing transistors control the amount of current flowing through the inverter. The charging time of the capacitance varies depending on the amount of current., which then results in changes in frequency of the oscillator. The oscillation frequency is expressed as:

$$f_{osc} = \frac{I_D}{NC_L V_{DD}} \qquad (2)$$

where I_D is drain current, V_{DD} is the supply voltage, N is the number of delay stage, and C_L is the cumulative capacitance of the delay stages.

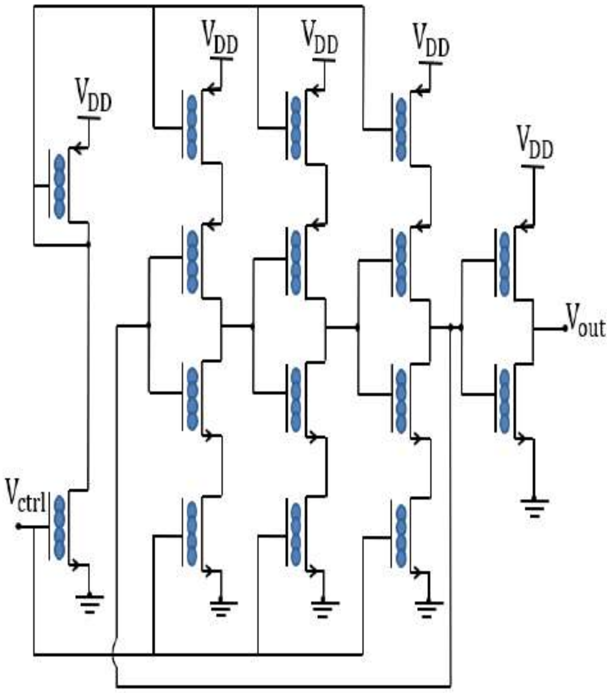
It is evident that ultra-low power design is mainly achieved by keeping I_D low and maintaining the desired oscillation. This is one of the major reasons we have employed TFET in our design because it can easily be operated with low V_{DD} . So, the downward scaling of bias current does not affect the oscillation frequency. Power analysis is a very important aspect of analog or digital circuits, and it can be categorized into two; static and dynamic. Although there is also short circuit power, which occurs whenever the pull-up and pull-down transistors are both conducting. It is considered infinitesimally small, so it is ignored in this study. According to (3), the dynamic power reduces greatly by scaling down V_{DD} , minimizing the parasitic capacitance of the transistor, or decreasing frequency. Since we want the PLL to be suitable for high frequency applications, we have employed the model because of its inherent low capacitance. This reduces the delay and invariably increases its speed of operation. TFET has also solved the problem of high leakage current in CMOS device, thus, the static power is drastically reduced. The power of current starved voltage controlled oscillator is estimated as 1.91 μW at oscillation frequency of 2 GHz, which is quite low when compared to the oscillator in [8]. This implies that the TFET-based VCO has a wide tuning range of 82.8% and the estimated phase noise is -117.3 dBc/Hz measured at 1 MHz offset frequency.

$$P_{dyn} = \alpha \times f \times V_{DD}^{2} \times C_L \qquad (3)$$

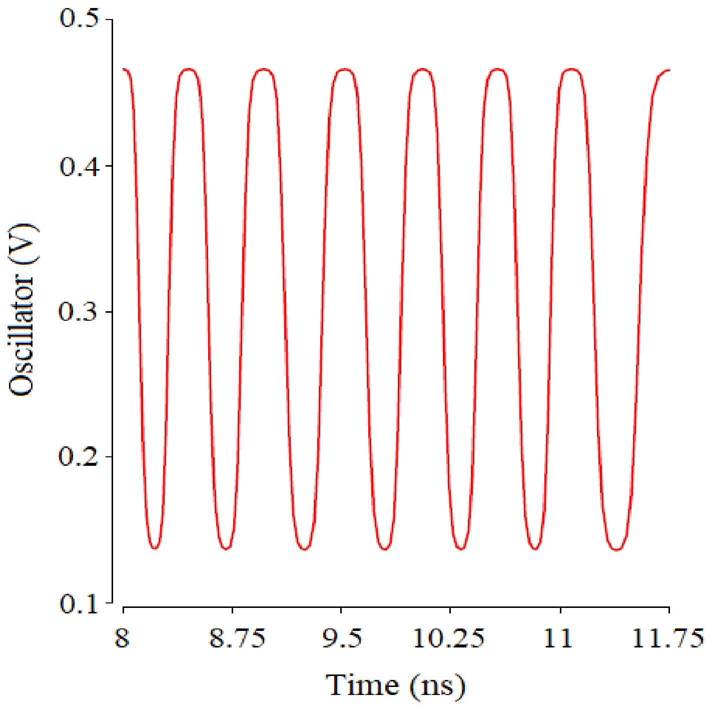
where f is the frequency, α is the activity constant, and C_L is the load capacitance.

We equally performed temperature analysis between -55 ˚C to 125 ˚C. The results show that the output of voltage controlled oscillator does not vary with temperature and there is no degradation in its performance. The final block of PLL is frequency divider, and it is implemented with D flip flop divide by two circuit.

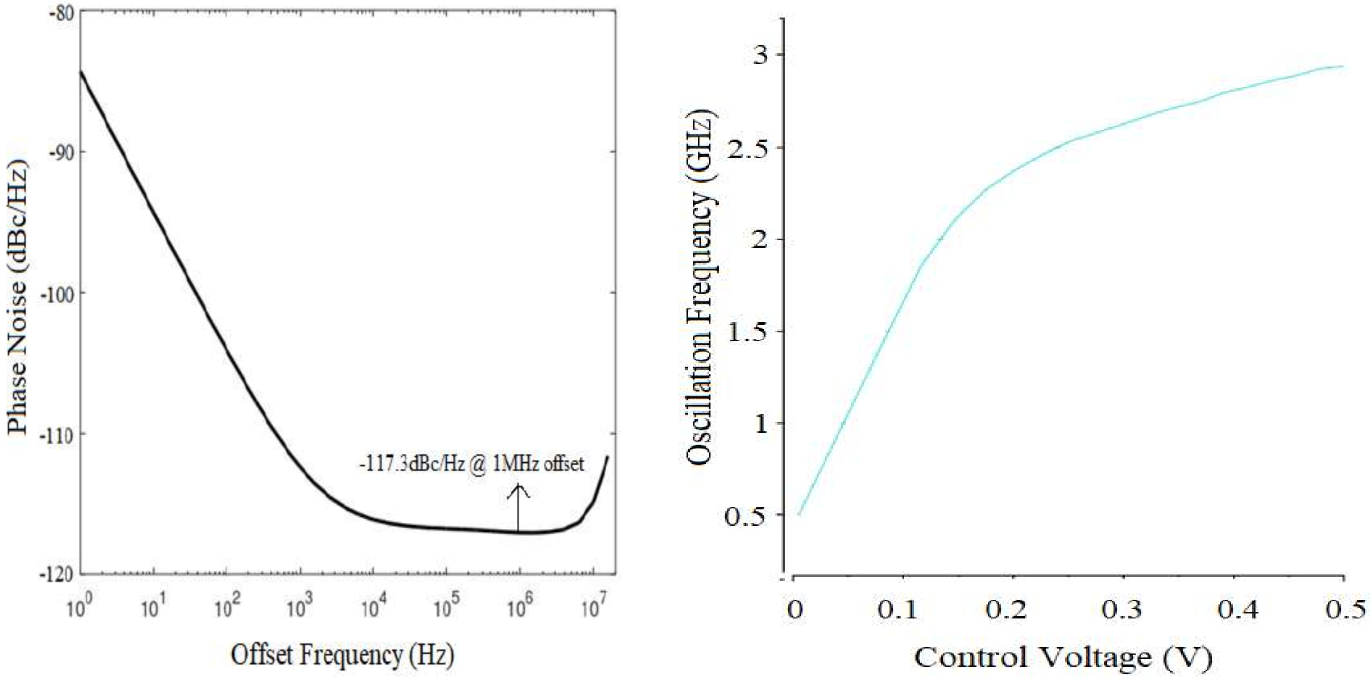
The VCO performance comparison is presented in Table 1. The designed oscillator operates between from 500 MHz to 2.9 GHz with a tuning range higher than the ring oscillators presented in [9, 10]. Furthermore, our designed oscillator consumes ultra-low power of 0.0019 mW, which is much smaller than the power consumed by CMOS, FINFET, and graphene nanoribbon (GNR) TFET-based oscillators [11 - 13].



(a)



(b)



(c)

(d)

Fig. 3. (a) Three-stage current starved VCO (b) Transient response of oscillator (c) Open loop phase noise (d) Tuning characteristics.

IV. CONCLUSION

Since the Internet-of-Things (IoT) is gaining more interests and phase locked loop is one of its most widely employed circuits, it is desirable to design ultra-low power PLL. This work presents ultra-low power phase locked loop design for IoT and biomedical applications. The design of each functional block of PLL is examined, and we have employed MoS_2 TFET because of its low sub-threshold and propensity for low power analog and digital electronics. We characterized the current starved VCO and compared it with the related work. The results show that the PLL operates in the range of 0.5 GHz to 2.90 GHz with a wide tuning range of 82.8%. The VCO has a phase noise of -117.3 dBc/Hz at 1 MHz offset frequency and consumes a very low power of 1.91 µW. The total power consumed by phase locked loop is 0.021 mW. In addition, the output of PLL shows resistance to temperature variations, which make it suitable for Internet of Things (IoT), medical, industrial, and military applications.

ACKNOWLEDGEMENTS

The work is supported by the Louisiana State University (LSU) Economic Development Assistantships (002128), and the Office of Research & Economic Development (ORED) Grant (004475).

TABLE I. PERFORMANCE COMPARISON TABLE

Parameter	**This work**	CMOS [9]	CMOS [10]	CMOS [11]	FINFET [12]
Technology (nm)	**20**	90	180	180	45
Power supply (V)	**0.5**	1.2	1.8	1.8	1
Power consumption (mW)	**0.0019**	7.3	0.27	7.49	2.05
Frequency (GHz)	**0.5 - 2.9**	3.5 – 7.1	0.35 - 1.1	1 – 3.9	0.25 – 1.60
Phase noise at 1 MHz (dBc/Hz)	**-117.3**	-105	-94	-80.17	-135.2
Tuning range (%)	**82.8**	72	76.5	74.4	81.2

REFERENCES

1. M. Alioto and D. Esseni, “Tunnel FETs for Ultra-Low Voltage Digital VLSI Circuits: Part II– Evaluation at Circuit Level and Design Perspectives,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 22, no. 12, pp. 2499-2512, Dec. 2014.
2. B. Sedighi, X. S. Hu, L. Huichu, J. J. Nahas, and M. Niemier, “Analog Circuit Design Using Tunnel-FETs,” Circuits and Systems I: Regular Papers, IEEE Trans. on, vol.62, no. 1, pp.39- 48, Jan. 2015.
3. Z. Lin *et al*., “Challenges and Solutions of the TFET Circuit Design,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 12, pp. 4918-4931, Dec. 2020.
4. B. Gopireddy, D. Skarlatos, W. Zhu and J. Torrellas, “HetCore: TFET-CMOS Hetero-Device Architecture for CPUs and GPUs,” 2018

ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA), Los Angeles, CA, 2018, pp. 802-815.

5. M. S. Fahad, A. Srivastava, A. K. Sharma, and C. Mayberry, “Analytical Current Transport Modeling of Graphene Nanoribbon Tunnel Field-Effect Transistors for Digital Circuit Design,” IEEE Transactions on Nanotechnology, vol. 15, no. 1, pp. 39–50, Jan. 2016.
6. M. A. U. Khan, A. Srivastava, C. Mayberry, and A. K. Sharma, “Analytical Current Transport Modeling of Monolayer Molybdenum Disulfide-Based Dual Gate Tunnel Field Effect Transistor,” IEEE Transactions on Nanotechnology, vol. 19, pp. 620–627, 2020.
7. W. Abbas, Z. Mehmood, and M. Seo, “A V-Band Phase-Locked Loop with a Novel Phase-Frequency Detector in 65 nm CMOS,” Electronics, vol. 9, pp. 9, 2020.
8. N. O. Adesina, and A. Srivastava, “Memristor-Based Loop Filter Design for Phase Locked Loop,” J. Low Power Electron. Appl., vol. 9, no. 3, pp. 24, 2019.
9. S. Min, T. Copani, S. Kiaei, and B. Bakkaloglu, “A 90-nm CMOS 5-GHz ring-oscillator PLL with delay discriminator-based active phase-noise cancellation,” IEEE Journal of Solid-State Circuits, vol. 48, no. 5, pp. 1151- 1160, 2013.
10. N. O. Adesina, and A. Srivastava, “Threshold Inverter Quantizer Based CMOS Phase Locked Loop with Improved VCO Performance,” IEEE VLSI Circuits and Systems Lett., vol. 6, no. 3, pp. 1-13, 2020.
11. A. Mishra and G. K. Sharma, “Design of power optimal, low phase noise three stage Current Starved VCO,” 2015 Annual IEEE India Conference (INDICON), New Delhi, pp. 1-4, 2015.
12. N. O. Adesina and A. Srivastava, “A 250 MHz-to-1.6 GHz Phase Locked Loop Design in Hybrid FinFET-Memristor Technology,” 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, pp. 0901-0906, 2020.
13. N. O. Adesina, A. Srivastava and M. A. U. Khan, “Evaluating the Performances of Memristor, FinFET, and Graphene TFET in VLSI Circuit Design,” 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, pp. 0591-0595, 2021.

The development of a routing protocol based on Reverse-AODV by considering an energy threshold in VANET

1stNabe Gedalia Razafindrobelina
dept. of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
nabe.19051@mhs.its.ac.id

2ndRadityo Anggoro
dept. of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
onggo@if.its.ac.id

3rdAry Mazharuddin Shiddiqi
dept. of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ary.shiddiqi@if.its.ac.id

***Abstract*—Many approaches based on the improvement of ad hoc routing protocols are proposed due to the issue of the mobile node in VANET (Vehicular Ad Hoc Networks). Indeed, before a source node transfers data, it sends a request message RREQ to all other neighbor nodes in the network where all nodes participate in transmitting the packet. After the message RREQ (route request) comes to the destination, it is replied to by a message RREP (route reply). If an intermediate nodes' energy went out due to the battery, it drops a data packet RREP message after a new recovery must be started, and it can affect the performance of the protocol. This paper proposed two news protocols E-AODV and RE_AODV, with an energy threshold and a new "minimum residual energy" for VANET. The E-AODV approach improves network performance because of node battery depletion and RE_AODV because of the RREP messages' broadcasting and filtering it by the energy threshold.**

Keywords—AODV, energy threshold, RAODV, VANET, E-AODV.

I. Introduction

With intelligent traffic management systems, the streets are much safer, and the energy consumption for vehicle mobility can be significantly reduced. Further pedestrian traffic with a battery-powered device like smartphones or smartwatches can be incorporated in the future. Much information data is necessary to realize these traffic control systems. VANET routing protocols are Ad hoc self-organizing routing protocols without using access points or pre-existing network infrastructure. Proactive routing protocols establish routing paths between all nodes periodically and continuously, affecting high data transfer of control messages, also called overhead. So, much energy and network resources are wasted with an often-changing network topology. Reactive protocols are on-demand routing protocols, in which the routing path from the source to the destination is established when the source needs to send data.

Nevertheless, the established route can have a link breakage caused by the battery run out of a node. It can be avoided by enhancing the RREQ message with an energy threshold and a variable to record the nodes' minimum residual energy, which forms the route for sending the data packet. Additionally, the RE_AODV protocol proposes to replace the RREP message with an enhanced R-RREQ message to face the problem when a node moves out of transmission range while sending the RREP message. The enhanced R-RREQ message is also broadcasted like the RREQ message to find the source node by multi-hop. This work is performed to know the impact of modifying the RREP packet delivery mechanism to a broadcast mechanism and adding an energy threshold to the RREQ packet delivery mechanism in the route discovery stage on the overall data packet delivery performance of the VANET network. The paper is continuous as follows: section 2 in Previous literature study, section 3 Methodology, section 4 Implementation, section 5 Result, and the last section conclusion and future work.

II. Previous Literature Study

The ad hoc networks' dynamic change conducts many problems like loss of packet data during transmission. Many types of research are proposed to overcome the issue so that the paper introduces one solution against the issue mentioned. Chonggun Kim, Elmurod Talipov, and Byoungchul Ahn proposed a new approach, "a Reverse AODV Routing Protocol in Ad Hoc Mobile Networks," [2] where it improves the AODV protocol named reverse AODV or R-AODV, and it is a reactive protocol, too. This new protocol increases the output of data transmissions. It is providing a new path for better performance, throughput, end-to-end delay, etc. To skip the loss, the route reply message. Indeed, the Ad hoc AODV protocol requires the successful delivery of the RREP message. The routing performance retrogrades due to the RREP packets' loss due to the nodes' high mobility. They found out that the performance of AODV is improving by R-AODV for packet delivery, End to end delay, and energy consumption.

The modified AODV, introduced by Landge and Nigavekar, is also motivated by increasing data throughput and energy consumption. To achieve these goals [1], they integrate a pre-calculated energy parameter into the RREQ packet. The pre-calculated energy is for a specific event for sending to the destination. With this new energy field, a connection break can

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

be avoided by ensuring that the intermediate nodes have sufficient residual energy. After the reception of the RREQ message by the intermediate node, it compares its recent energy with the value inside the energy field. If the remnant energy is lower than the pre-calculated energy, then it discards the RREQ message. Otherwise, to the next neighbor, the message is forwarded. It guarantees that the intermediate nodes will be alive as long as needed for the data transfer. Data package loss due to this reason is avoided. After the reception of the RREQ messages by the destination node, it will choose the route with the lowest hop number and thus with the lowest energy consumption. The procedure for route reply is the same as in the AODV protocol specified.

III. Methodology

A. *Vehicular Ad Hoc Networks*

The biggest obstacle in VANETs than MANETs is the nodes' movement, resulting in highly dynamic network topology changes. Routing protocols in VANETs face this problem using Ad hoc networks, which means a route will be found after a source demands a data transmission and the maintenance of temporarily stored routes in the routing table. The routing maintains the routing and the protection of the communication of 2 nodes with them routing messages. That means the reliability of the transmission of the messages depends on the technique of the protocol. Routing protocols in VANETs also take into account the lack of infrastructure.

B. *Ad hoc On-demand Distance Vector*

The AODV-protocol (Ad hoc On-demand Distance Vector) is a reactive routing protocol that further develops DSDV. This protocol does not create a routing table in advance, but it reacts if a node demands a data transfer and start a route discovery. An established routing path will be stored for a specific time in a routing table. The RREQ (route request message) and the RREP (route reply message) are used to establish a routing path for the data transfer. To maintain and repair a route inside the routing table, the RERR (route error message) and periodically "HELLO"-messages are transmitted.

Causing by a data transfer demand of a source node, the RREQ message is transmitted by flooding into the network, and it is forwarded until it arrives at the destination or an intermediate node that has a valid path to the destination in the routing table, see Fig. 1.

Following an RREP message is forwarded back through the chosen data transfer route by unicast transmission, according to Fig. 2. A data transfer is started when the RREP message arrives back at the source node. The discovered route will be stored for a limited time in the routing table of the nodes.

C. *Reverse- AODV*

In R-AODV(Reverse- AODV), the route request is the same as in AODV, but the reverse route request is not unicast to find the source node. After receiving the RREQ by the destination node, an R-RREQ (reverse route request) is prepared. It reduces a failing data transfer because of link breakage by being also broadcasted by flooding into the network. The possibility of finding a routing path is increased.

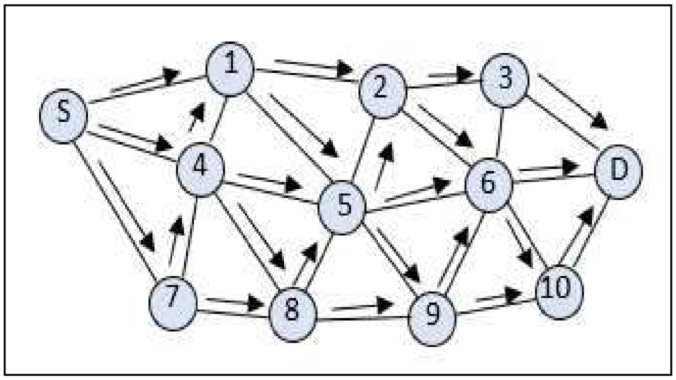


Fig. 1: Forwarding of RREQ message

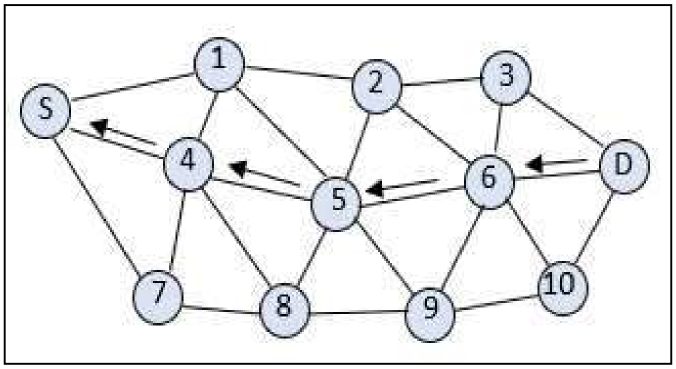


Fig. 2 Unicasting of RREP message

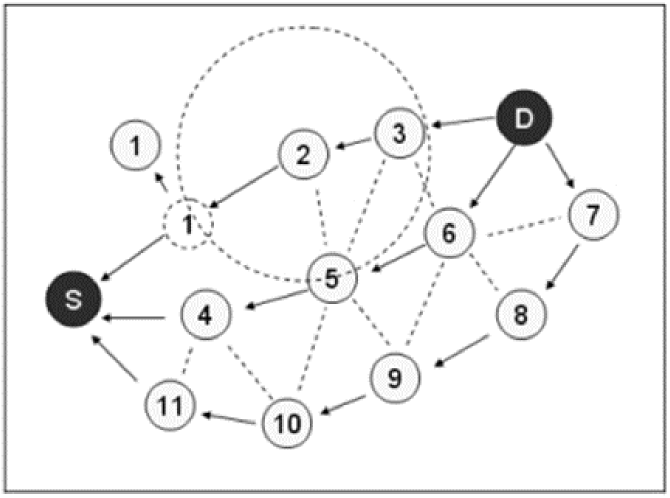
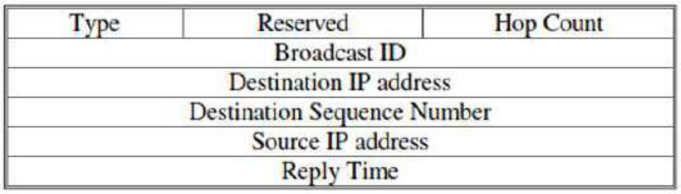


Fig. 3: R-RREQ from destination to source node [2]

Table 1 R-RREQ Message Format[2]



Type	Reserved	Hop Count
Broadcast ID		
Destination IP address		
Destination Sequence Number		
Source IP address		
Reply Time		

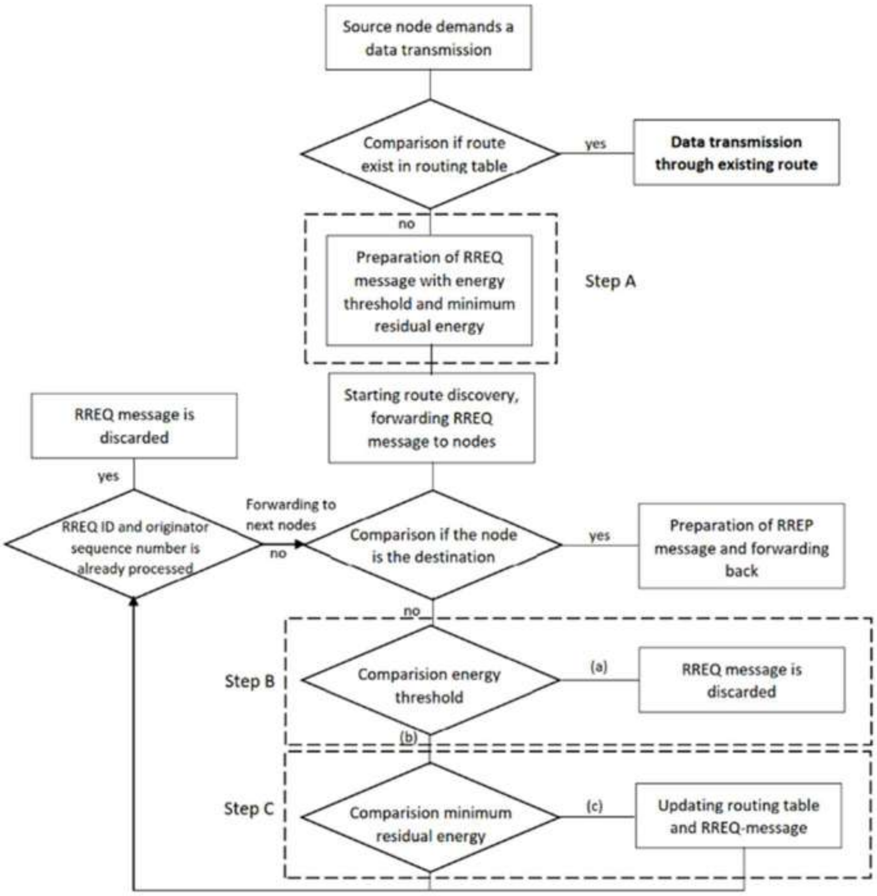

Source node demands a data transmission
Comparison if route exist in routing table
yes
Data transmission through existing route
no
Preparation of RREQ message with energy threshold and minimum residual energy
Step A
Starting route discovery, forwarding RREQ message to nodes
RREQ message is discarded
yes
RREQ ID and originator sequence number is already processed
Forwarding to next nodes
no
Comparison if the node is the destination
yes
Preparation of RREP message and forwarding back
no
Step B
Comparision energy threshold
(a)
RREQ message is discarded
(b)
Step C
Comparision minimum residual energy
(c)
Updating routing table and RREQ-message

Fig. 4 Flowchart of the proposed algorithms E-AODV

This last discovered route is used to transfer the data package. If a discovered route via node 1-2-3 (see Fig.3) fails because node one is moving out from the route, then the R-RREQ can discover another route, too. The structure of the R-RREQ message can be seen in Table 1, containing the following information.

D. *The proposed approach*

The RREP and R-RREQ message are used like in the AODV and R-AODV protocol for the E-AODV protocol. The RREQ message is enhanced by an energy threshold and a value for the "minimum residual energy" of an intermediate node taking part inside the data packet delivering route. A percentage of the initial energy provides the energy threshold. The minimum residual energy is the initial energy at starting the route discovery process and will be updated during route discovery. When the energy of the RREQ reception node is greater than the threshold, then the RREQ message will be forwarded, and when the residual energy is lower than the "minimum residual energy" value, then this value is updated in the RREQ message and the routing table of the node.

TABLE 1 MODIFIED RREQ-MESSAGE STRUCTURE

Packet type	Reserved	Hop Count
Broadcast ID		
Destination IP Address		
Destination Sequence Number		
Source IP Address		
Source Sequence Number		
Threshold Energy		
Minimum Residual Energy		
Timestamp		

Fig. 4 shows the process for the enhanced RREQ message. In general, the process follows the AODV approach. The enhancements are marked inside the dashed boxes with the next explained Steps.

Step A) The RREQ-message is prepared by adding the energy threshold and a "minimum residual energy." The energy can be chosen by a specific percentage of the initial energy. Further, at the preparation of the RREQ-message by the source node, the value of the minimum residual energy is filled with the initial energy of the source node. The changes of the RREQ message structure can be depicted in Table 1.

Step B) The receiving intermediate node compares the own residual energy with the energy threshold specified by the RREQ-message. If the own energy of the receiving intermediate node is lower than the threshold, then the RREQ message will be (a) dropped. In this case this intermediate will not be part of the data transmission route. However, when the own energy is higher than the energy threshold, it continues (b), see Fig. 4 with forwarding the modified RREQ message to the next nodes.

Step C) After the intermediate nodes' energy is compared to the value of "minimum residual energy." If the intermediate nodes' energy is lower than the value inside the RREQ-message, then the routing table information of the node and the RREQ message will be updated (c) with the new residual energy and hop count. Otherwise, the routing table and the RREQ-message will not be updated with residual energy of the node, and the process continues.

When the destination node receives more than one route. The change is made by adding the minimum residual energy, affects the selection of the main route. If there is another route with a minimum energy value on that route greater than the minimum residual energy value in the routing table, that route will be selected as the main data transmission route.

The RE_AODV protocol combines the E-AODV approach, with the enhanced RREQ message, by the energy threshold and the "minimum residual energy" and the R-RREQ message of the R-AODV protocol, developed by Chonggun Kim, Elmurod Talipov, and Byoungchul Ahn, X. Zhou et al.[2] . The R-RREQ message is forwarded back to the source by broadcast transmission.

IV. Implementation

This section describes the simulation environment and simulation parameters to compare the performance between the original AODV, the modified AODV -> E-AODV, R-AODV, and the hybrid E-AODV -> RE_AODV.

A. Simulation Environment

Three types of environments will be used in the simulation, namely 50 nodes for sparse environments, 100 and 150 nodes for medium environments, and 200 nodes for dense environments.

B. Simulation parameters

The NS-2.35 network simulator is used with the different protocols protocol to perform the simulation. The simulation area is determined by 1000m × 1000m in size with two types of geometrical topologies, see Fig. 5.

Each plot has an edge length of 125m x 125m. First is a grid map with a 9 x 9 intersection and 64 plots, the cutout from the real map around Jl. Soetomo in Surabaya is the second map. This map is converted from OpenStreetMap.

Ten random mobility route-sets for the nodes are created and converted into mobility files for both geometric topologies. These mobility files are used for the simulation in NS2. Complete information can be seen in Table 2

C. Simulation and evaluation

At first, a set of simulations is carried out to determine the suitable value for the energy threshold. The values 50%, 25%, 20%, 15%, 10%, and 5% were tested on the manual created grid map with 50 nodes. The packet delivery ratio results with ten random mobility scenarios can be depicted in Table 3, which are visualized in Fig. 5. The packet delivery ratio between 5%, 10%, and 15% are stable, and the decrease is minimal.

TABLE 2 SIMULATION PARAMETERS

No.	Parameter	Specification
1	Network simulator	NS-2.35
2	Routing protocol	AODV E-AODV R-AODV RE_AODV
3	Simulation time	200 seconds
4	Simulation area	1000 m × 1000 m
5	Number of nodes	50, 100, 150, 200
6	Transmission radius	400m
7	Default speed of nodes	50 m/s

TABLE 3 RESULTS OF THRESHOLD DETERMINATION

Test results						
Thres-hold	50%	25%	20%	15%	10%	5%
PDR	0.361	0.613	0.629	0.710	0.731	0.741

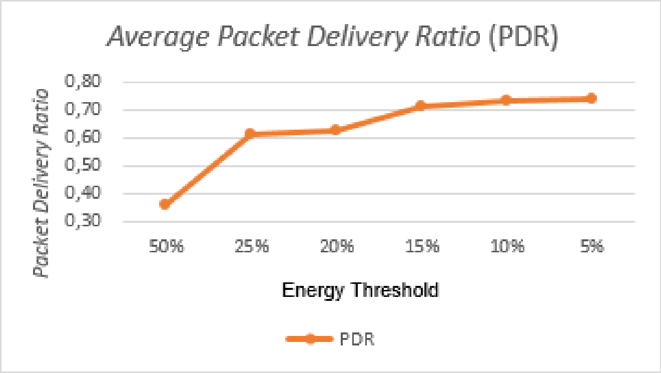


Fig. 5 PDR Graph in Threshold Determination

The drop of the PDR increases significantly for larger values. For 50%, 25%, and 20%, the number of control messages is vastly limited, so the threshold of 15% is considered suitable.

V. Results

At first, the simulations are carried out by comparing the protocols RE_AODV, R-AODV, and E-AODV to determine the AODV routing protocols' competitors.

The results are based on the grid map with 200 nodes and ten sets of the nodes' mobility. Fig. 6, 7, 8, 9, 10 are visualizations of Packet Delivery Ratio PDR, End-to-End delay E2E, Routing Overhead RO, Energy Consumption CE, Throughput, and Table 4 contains a summary of the average results.

The PDR values in Fig. 6 showing very high stability of R-AODV compared to E-AODV and RE_AODV. Otherwise, E-AODV has very fluctuating results for different mobility scenarios. However, on average, RE_AODV has the highest PDR. The stability of E-AODV is very low compared to RE_AODV and R-AODV, but the changes in the curves are similar between RE_AODV and R-AODV. The lowest End to End delays are achieved by the RE_AODV, followed by R-AODV and the last E-AODV protocol, as shown in Fig. 7.

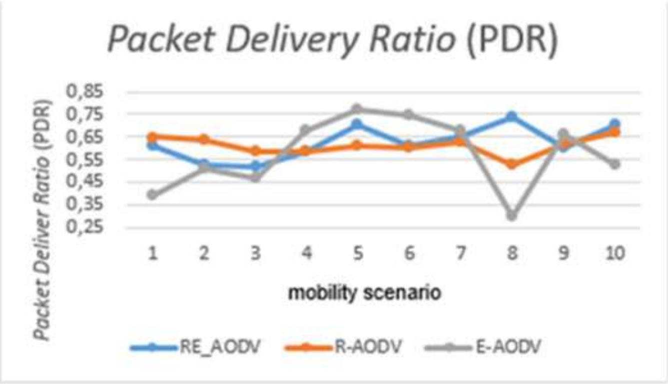


Fig. 6 PDR for RE_AODV, R-AODV, E-AODV

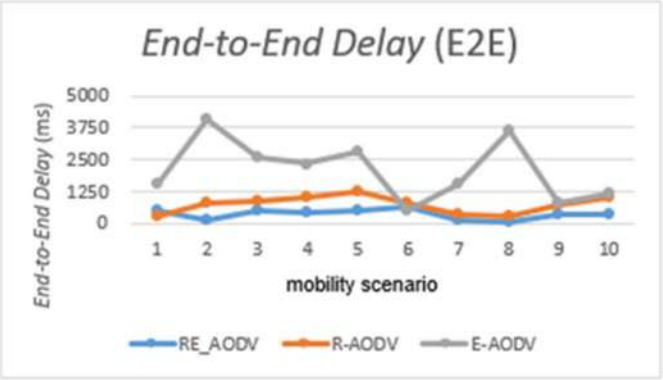


Fig. 7 E2E for RE_AODV, R-AODV, E-AODV

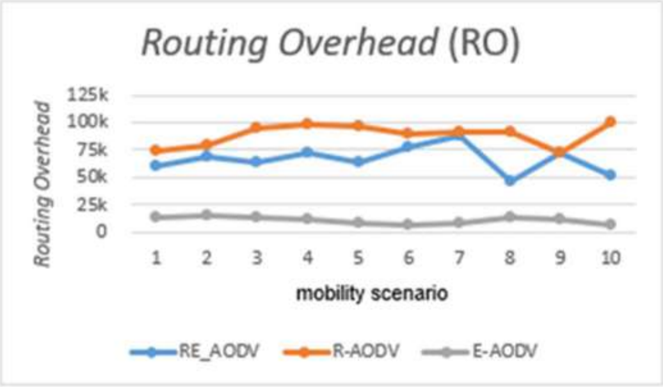


Fig. 8 RO for RE_AODV, R-AODV, E-AODV

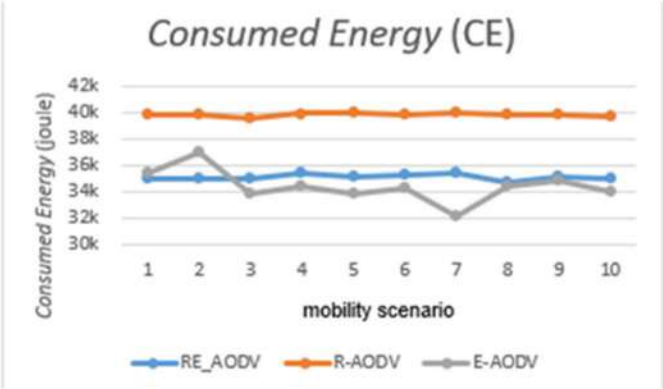


Fig. 9 CE for RE_AODV, R-AODV, E-AODV

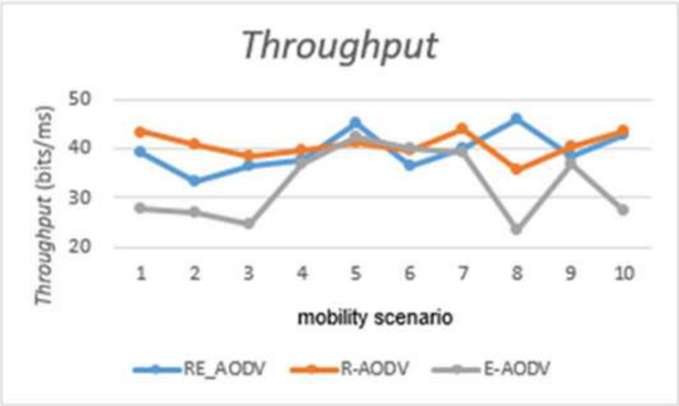


Fig. 10 Throughput for RE_AODV, R-AODV, E-AODV

TABLE 4 RESULTS OF RE_AODV, R-AODV, E AODV

	RE_AODV	R-AODV	E-AODV
PDR	0.63	0.61	0.57
E2E in ms	384	759	2,113
RO in %	66,080	88,620	10,640
CE in J	35,058	39,805	34,405
Throughput	39.54	40.63	32.54

TABLE 5 QUALITATIVE RESULTS OF R-AODV, RE_AODV, AND E-AODV

Metric	RE_AODV	R-AODV	E-AODV
PDR	Stable	Stable	Unstable
E2E	low, Stable	low, Stable	high, unstable
RO	Intermediate	high	low
CE	low, Stable	high, Stable	low, unstable
Throughput	high, Stable	high, Stable	low, unstable

The routing overhead for RE_AODV is decreased compared to R-AODV but vastly higher than E-AODV, see Fig. 8. E-AODV also shows the highest stability for the overhead routing values. The lowest energy consumption is achieved by the E-AODV protocol, in which RE_AODV is only slightly higher but much more stable, as showed Fig. 9. R-AODV consumed the most energy in the simulation. Overall the energy consumption is at the same level.

R-AODV accomplishes the highest throughput and is negligibly followed by RE_AODV, see Fig. 10. The Throughput for E-AODV is significantly lower, between the scenarios is much more fluctuating.

Qualitatively the results can be summarized according to Table 5. Based on the high throughput, packet delivery, and lowest End to end delay, along with the stable results, the RE_AODV is determined as the competitor of AODV.

The simulation of AODV and RE_AODV is executed in the before mentioned real scenario and grid scenario. The comparison between the standard AODV protocol and modified-AODV (RE_AODV) are shown as short summarization in Table 5 and Table 6. Both scenario tests are executed with sparse (50 nodes), medium (100, 150 nodes), and high (200 nodes) density environments. For the movement of the nodes, also ten sets of random mobility scenarios are applied. By use of an AWK script, the results of the Packet Delivery Ratio (PDR), End-to-end Delay (E2E), Routing Overhead (RO), Consumed Energy (CE), and Throughput are acquired. The energy threshold is determined as 15%, as discussed before.

In Table 6, a summary of the real scenario simulation results is presented, followed by the grid scenario simulations' summarized results; see Table 7. On average, the RE_AODV protocol obtains better results than the AODV protocol, except for the vastly higher Routing Overhead. In the real scenario, the Packet Delivery Ratio of RE_AODV is 2.42% higher, the end-to-end delay is 56.1% lower, the Routing overhead is 314.59% higher, the energy consumption is 2.68% lower, and the throughput is 8.66% higher. For the grid scenario, RE_AODV acquired a 7.59 higher PDR, a 68.64% lower E2E, a 336.5% higher RO, a 7.27% lower CE, and an increased throughput by 14.96%.

When analyzing the results, it can not be determined in which environment the RE_AODV protocol performs better. The results for energy consumption, throughput and end to end delay differs between the real and grid scenario, as well as for the node density. In the real scenario with a node density of 50, the RE_AODV protocol performs best with less energy consumption, but the throughput and end to end delay is best at 150 nodes. In the grid scenario the energy consumption, end to end delay and throughput is best at 200 nodes, but with 50 nodes the throughput is even lower, than AODV. This shows that it can not be generalized in which environment the RE_AODV protocol perferms best, but in average it obtain better results than AODV. It can be assumed, that the random mobility and thus the distribution of the nodes have a significant effect on the performance results, which have to considered for further studies or applications.

VI. CONCLUSION AND FUTURE WORK

The impact that occurs when implementing modifications to the RREP packet delivery mechanism on the network is a more stable Packet Delivery Ratio (PDR), low End-to-End Delay (E2E) with a high level of stability, very high Routing Overhead (RO). high, Consumed Energy (CE), which tends to be high with a reasonably good level of stability, as well as high and stable network throughput so that these modifications function to increase PDR stability, reduce E2E values, increase CE stability, and increase network throughput values in combined modifications (RE_AODV)

Meanwhile for the addition of energy filters in the RREQ delivery process is an unstable PDR, a high enough E2E with a low level of stability, a very low RO, a very low CE, and a network throughput that tends to be low so that these modifications play a role in reducing the RO and CE values in the network on combined modification (RE_AODV).

The number of Routing Overhead (RO) or the number of control packet deliveries that significantly increased on the modified AODV with the hybrid method (RE_AODV) increased RO by up to 336.51% in the simulation grid and 314.59% in real simulation. It is caused by the modification in the RREP sending mechanism from unicast to broadcast, causing massive RREP packet delivery.

TABLE 6 AVERAGE REAL SCENARIO METRICS

		Metric Attributes				
Environment type	Protocol	PDR	E2E ms	RO packages	CE J	Throughput bits/ms
50 node	AODV	0,79	586,95	3246,80	8233,02	47,10
	RE_AODV	0,80	297,55	5590,00	7809,22	47,62
Difference		-0,01	289,40	-2343,20	423,80	-0,52
Difference (percentage)		-2,18%	49,31%	-72,17%	5,15%	-1,12%
100 node	AODV	0,74	1063,02	8224,60	18885,22	44,86
	RE_AODV	0,74	494,73	32298,30	18304,60	45,66
Difference		-0,01	568,29	-24073,70	580,62	-0,79
Difference (percentage)		-0,77%	53,46%	-292,70%	3,07%	-1,78%
150 node	AODV	0,67	1607,17	12505,10	29171,96	36,70
	RE_AODV	0,70	503,39	62219,00	28633,77	43,83
Difference		-0,03	1103,77	-49713,90	538,19	-7,12
Difference (percentage)		-4,92%	68,68%	-397,55%	1,84%	-19,40%
200 node	AODV	0,64	1515,38	13638,20	37998,23	37,83
	RE_AODV	0,65	712,81	81276,80	37744,69	42,50
Difference		-0,01	802,56	-67638,60	253,54	-4,67
Difference (percentage)		-1,79%	52,96%	-495,95%	0,67%	-12,36%
Average - Difference (percentage)		-2,42%	56,10%	-314,59%	2,68%	-8,66%

TABLE 7 AVERAGE GRID SCENARIO METRICS

		Metric Attributes				
Environment type	Protocol	PDR	E2E ms	RO packages	CE J	Throughput bits/ms
50 node	AODV	0,87	412,73	2175,40	8681,03	46,80
	RE_AODV	0,85	187,16	8464,50	8106,35	45,95
Difference		0,01	225,56	-6289,10	574,68	0,85
Difference (percentage)		2,00%	54,65%	-289,10%	6,62%	1,82%
100 node	AODV	0,70	1225,58	6640,40	18459,53	38,57
	RE_AODV	0,74	310,93	31067,90	17425,58	44,21
Difference		-0,03	914,65	-24427,50	1033,95	-5,64
Difference (percentage)		-5,29%	74,63%	-367,86%	5,60%	-14,62%
150 node	AODV	0,62	1513,85	10832,30	28084,32	37,24
	RE_AODV	0,63	629,18	51203,60	26368,05	40,24
Difference		-0,02	884,66	-40371,30	1716,27	-3,00
Difference (percentage)		-3,06%	58,44%	-372,69%	6,11%	-8,08%
200 node	AODV	0,50	2920,60	15870,00	39272,28	28,45
	RE_AODV	0,62	383,96	66079,60	35058,00	39,53
Difference		-0,12	2536,63	-50209,60	4214,28	-11,08
Difference (percentage)		-24,03%	86,85%	-316,38%	10,73%	-38,96%
Average - Difference (percentage)		-7,59%	68,64%	-336,50%	7,27%	-14,96%

A. *Analysis*

In PDR, E2E and throughput can be seen, that the results of RE_AODV and R-AODV are significantly better than AODV but quite similar to each other. The effect of the energy threshold is low, and the slightly better performance of RE_AODV compared to R-AODV is probably based on the lower number of control messages.

The forwarding of the control messages by broadcasting ensures that the control messages arrive at the source and destination. Which effect that data transmission can be started, and the robustness of the network is increased. Otherwise, it can be concluded that a too high number of control messages can decrease the network performance because of traffic congestion. The intermediate nodes' queue will overflow, and packets will be dropped, resulting in reduced PDR, E2E, RO, and Throughput performance. The broadcast delivery mechanism has an increased performance impact on the network if enough intermediate nodes to forward the control messages are available. If the number of control messages is too high in relation to the number of nodes, then the performance decreases, which can be seen, when the number of nodes reducing, especially for the 50 nodes scenario.

In conclusion, the number of control messages must be adjusted according to the number of nodes inside the network. The goal for future work has to be to find an optimal number of control messages by finding suitable filters for broadcasted request and reply messages. The optimal solution would be an adjusted filter according to the density of nodes, in which the energy consumption can also be reduced.

B. Suggestions and future work

As a result of the investigation and analysis of this project, the following ideas can be suggested the use of dynamic filter of intermediate nodes by using "HELLO"-messages relating to the density of nodes in the network. The target is to ensure that the source and destination receive the control messages but simultaneously adjust the number of control messages to prevent an overflow of the nodes' queue. Moreover, Adding a node forwarding limitation not only when sending RREQ packets but also when sending R-RREQ packets to reduce the routing overhead and to select the route for the data packet, also for the hybrid RE_AODV protocol.

References

[1] Poonam Landge*, Prof. Atul Nigavekar, “Modified AODV protocol for energy-efficient routing in manet,” ISSN: 2277-9655 (I2OR), March 2016

[2] Chonggun Kim, Elmurod Talipov, and Byoungchul Ahn, X. Zhou et al., “A reverse AODV routing protocol in ad hoc mobile networks” (Eds.): EUC Workshops, LNCS 4097, pp. 522 – 531, 2006.

[3] Michael Behrisch, Laura Bieker, Jakob Erdmann, Daniel Krajzewicz "SUMO – simulation of urban mobility, an overview," the Third International Conference on Advances in System Simulation, 2011.

[4] P. Visalakshi, Mahesh Mishra, Yusuf H, and Snehasish Maity, "Ad hoc network- an overview," International Journal of Modern Engineering Research (IJMER) ISSN: 2249-6645, pp-27-29, 2013.

[5] Mohammed Abdulhameed Al-Shabi, "Evaluation the performance of MAODV and AODV protocols in VANETS models," International Journal of Computer Science and Security (IJCSS) Volume (14): Issue (1), 2020.

[6] Awerbuch, Baruch, and Dr. Amitabh Mishra, "Advanced topics in wireless network: ad-hoc on-demand vector (AODV) routing protocol," "CS: 647".

[7] F. Nutrihadi. R. Anggoro and R. M. Ijtihadie.. “Studi Kinerja VANET Scenario Generators: SUMO dan Vanet Mobisim untuk Implementasi Routing Protocol AODV menggunakan Network Simulator 2 (NS-2)”, 2016.

[8] S. N. Ferdous dan M. S. Hossain, "Randomized energy-based AODV protocol for wireless Ad-hoc network," IEEE, 2016.

[9] C. Perkins, E. Belding-Royer, S. Das "Ad-hoc On-Demand Distance Vector (AODV) Routing" Network Working Group, 2003

Design and Implementation of ZigBee Based Low-Power Wireless Sensor and Actuator Network (WSAN) for Automation of Urban Garden Irrigation Systems

Asrorbek Eraliev
Politecnico di Torino
Torino, Italy
asrorbek.eraliev@studenti.polito.it

Giovanni Bracco
Politecnico di Torino
Torino, Italy
giovanni.bracco@polito.it

***Abstract*— The implementation of automated micro-irrigation in urban gardens using wireless sensor and actuator networks (WSAN) is being more popular. However, the use of WSAN in automated irrigation brings various technical challenges, for example, achieving low cost, long life battery, adequate communication range and compactness. In this paper, a WSAN is designed and implemented for controlling drip irrigation of urban garden that can achieve long battery life, low cost, compactness with the sufficient range of communication. The designed WSAN consists of end-devices and a coordinator, each of which is fitted with single-chip microcontroller along with wireless transceiver. Each end-device is equipped with a microcontroller, a water valve, a soil moisture sensor, a temperature sensor, a rain sensor, battery charger through solar panel and a ZigBee transceiver. The end-device reads the garden temperature, soil moisture level, battery charge status and rain presence and controls the water valve based on these data values. At the same time, the read data is transmitted to the coordinator station through ZigBee ad-hoc network. The coordinator receives the data from each end-device and presents to the interfaced user application. Due to being powered up by an energy limited resource, special techniques and algorithms are developed and only low-power electronic components are selected carefully. This paper includes implemented hardware and software results for the wireless nodes.**

Keywords*— *ZigBee; WSAN; low power; drip irrigation

I. Introduction

Control of Drip Irrigation

As a type of micro-irrigation, drip irrigation is being used widely owing to its excellence to save water and increase yields. In drip irrigation, water is delivered by the network of pipes, tubings and emitters and dripped directly to the root zone of the plant. Generally, control of water dripping to plants works in on-off mode and controlled by from the starting point of the pipes network. Gardener can switch on the water flow and after certain period of time switches of the irrigation. Furthermore, this process can be controlled by time-based automation. This type of control is acceptable if all the irrigated plants are in the same type and having the same root zone area and shape. However, different plants require different amount of time to be irrigated. In that case, manual control of drip irrigation or time-based automation cannot satisfy the precision requirements. The most appropriate solution is to build fully automated irrigation system in which dripping emitters are controlled separately, based on the moisture level of each irrigated plant. Establishment of this control requires installation of particular sensors at plant root zones and special actuators as dripping emitters. Each actuator drips water according to the collected data of the respective sensors.

The most suitable approach to build the intended irrigation control is applying WSAN (Wireless Sensor and Actuator Network) to the system. However, there are some issues related to power consumption and control algorithms of actuators should be figured out.

A. Related Works

Many systems have been proposed to control drip irrigation using wireless technologies so far. For instance, in this paper [1], low power valve controller for drip irrigation is implemented based on JN5139 low power wireless module. The JN5139 module integrates microcontroller and radio transceiver inside itself and therefore the designed valve controller achieved low cost and compact size. It was designed to run for three months on two alkaline batteries, under different periods of sleep. The frequency of activation was assumed to be 4-5 times per week. The valve controller was designed to precisely control the solenoid valve, collect information on the status of the solenoid valve, and provide real-time feedback. The implemented controller was tested in field and the results showed that it can operate continuously at least 3 months with two alkaline batteries.

In this paper [2], authors proposed smart irrigation system based on wireless sensor nodes and coordinator hub. The whole system is solar powered and each sensor node was a part of a ZigBee network in a Star topology around a hub concentrator.

The designed wireless sensor nodes powered up by 9V rechargeable batteries and connected to 1 Watt solar panels. The communication between wireless nodes and coordinator

978-1-6654-4067-7/21/$31.00 ©2021 European Union

hub is realized by XBee modules in the range of 28 square meters. Arduino boards are used in sensor nodes and hub coordinator because of the easy interface to standard Wi-Fi shields, USB, I2C and power regulation. Sensor nodes read moisture, temperature and humidity by SHT1x sensors and send the collected data to coordinator hub wirelessly. The battery life-time of wireless nodes was calculated 6 to 9 months by putting them into sleep modes frequently.

Similarly, this irrigation system [3] was proposed utilizing ZigBee wireless technology in master-slave architecture. Each slave node includes temperature sensor, soil moisture sensor, water valve, microcontroller and ZigBee module. The microcontroller of the slave node reads and frames the air temperature and soil moisture of the garden. After that, the frame is transmitted to the master station through the ZigBee module. The master station employs an embedded fuzzy logic irrigation algorithm in order to provide the grass and trees with water based on a set of rules. Furthermore, the master station is interfaced with a home web server to access remote monitoring and operation. The detailed characteristics of power consumption were not highlighted. Because, the main objective of the proposed system was allowing the operation by the use of internet and smartphone.

In this experimental study [4], authors developed energy efficient wireless sensor mote for low-data-rate applications such as, control of drip irrigation systems. The designed platform is called DZ50. It consists of ATmega328P microcontroller and RFM12b wireless transceiver. Due to being in sleep mode for long period of time and operating with low power, this platform achieved very high energy efficiency. The experimental results demonstrated that the mote life-time can be extended more than 7 years.

Furthermore, several techniques have been implemented through scheduling [5],[6], and adaptive radio-frequency in wireless nodes and selecting a network configuration [7] in order to achieve higher energy efficiency.

One of the major power consuming component in a wireless node is a radio module. So, various wireless standards have been built with medium access control protocols that provide multitask support, data transmission and also energy performance [8], in particular the wireless local area network standards, IEEE 802.11b - Wi-Fi [9] and wireless personal area network, IEEE 802.15.1 – Bluetooth [10] and IEEE 802.15.4 – ZigBee [11].

B. *Objectives*

In this paper, the design and implementation of a ZigBee based low power wireless sensor and actuator network (WSAN) for the control of drip irrigation is presented. The main objective is to achieve low-cost, compactness and high energy efficiency.

Main tasks have been:

- To design and implement hardware for wireless sensor control nodes that capable to control actuator and monitor air temperature, soil moisture level, rain presence and battery status;
- To develop firmware for the wireless nodes;

The designed WSAN should satisfy following main specifications:

- A minimum of two years battery life with single AA type battery, supposing that each end-device reads every sensors and actuates solenoid valve maximum ten times per day;
- A minimum 100 meters of communication range between end-devices and central coordinator;
- A low cost and compact volume
- No actuation in rainy time

II. System Design and Implementation

A simple and low cost network architecture can be applied to build WSAN for the control of drip irrigation system. That’s why, typical master-slave approach is chosen, in which central control unit is master node and end-devices are slave nodes, respectively (Fig. 1.). Master node exchanges data with slave nodes only one-by-one through ZigBee modules and only slave nodes can initiate the communication. Master node is in receive mode continuously, since it is powered up by robust and unlimited power supply being in receive mode constantly does not make a problem from energy point of view. The slave nodes are not required to enter receive mode continuously or periodically, and this dramatically increases their battery-life.

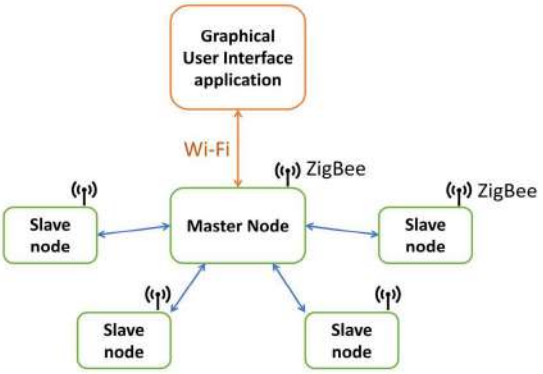


Fig. 1. Proposed architecture of the designing WSAN

A. *Main Components*

The specifications and requirements mentioned above demands to choose components very carefully.

As the wireless transceiver, XBee S2C ZigBee Standard module (from Digi) has been chosen due to its low power consumption and sufficient-range of communication. XBee S2C operates based on ZigBee protocol and direct-sequence spread spectrum modulation method. This module works on one of the 16 channels (channels from 11 to 26) in the 2.4GHz frequency band. It can reach up to 1200 meters of communication range in the line-of-sight and 250 kbps data rate. XBee S2C module works at 3.3V with the current consumption of 33mA (45 mA in boost mode) in transmission mode and 28mA in receive mode and less than 1 µA in sleep mode [12].

As a main controller unit for all the electronic components in a wireless node, Atmega328P microcontroller has been chosen. It is proposed to put microcontroller and XBee module into sleep mode in order to extend battery life. In Atmega328P microcontroller, built-in watchdog timer is available and we use it to wake-up from the deep sleep mode. But, the maximum predefined watchdog timer value is 8 seconds. Therefore, the duration of sleep period can be multiplication of 8 seconds, e.g. 450*8seconds=3600 seconds. What it means that microcontroller wakes up after every 8 seconds and counts one more, than enters into sleep mode immediately. When the count reaches 450, microcontroller begins to work in wake-up mode. In sleep mode, the microcontroller consumes 44µA at 3.3V at average.

The most power consuming element of the designing WSAN is the actuator which is used to open and close water flow. Generally, solenoid valves are employed to do this function. Our project requirements and specifications demands very low-power solenoid valve in order to save energy and reach longer battery-life. So, it is decided to choose latching solenoid valve as actuator due to its controllability with very low amount of energy spending. Latching solenoid valve, sometimes called bistable solenoid valve, does not require constant application of power to keep its close or open state. The state can be changed by applying very short electrical pulse to the coil of the bistable solenoid. In this project, it is decided to choose the latching solenoid valve with the following parameters (Table 1).

TABLE I. ELECTRICAL CHARACTERISTICS OF THE DEPLOYED VALVE

Nominal voltage of the pulse	Pulse duration	Coil resistance
3.3 V	30 ms	9 Ω

B. Coordinator (Master Node)

Master node integrates a Wi-Fi module in order to link the designing WSAN with gardener's user application. Therefore ESP-01 Wi-Fi module is integrated into master node owing to its very low cost, small size and ease of use. ESP-01 is configured to work in Access Mode for enabling gardener's devices to connect to the Wi-Fi access point of the module.

Master node performs following main functions:

- Receiving information of slave nodes wirelessly via XBee modules
- Sending the received information to user application using Wi-Fi module
- Controlling a water pump according to the signals of water level sensors in order to provide water tank with water in time

The program flow of the coordinator is illustrated in Fig. 2.

The master node integrates microcontroller, Wi-Fi module, XBee module, water tank level sensors and water pump. The initial input voltage of master node PCB is 12V. Among the main components of master node, microcontroller, water level sensors and water pump operate at 5V. These parts are powered up by one voltage regulator L7805 which has a wide input range between 7 – 35V [13]. Besides, the output current of this voltage regulator is sufficient to provide microcontroller and water pump.

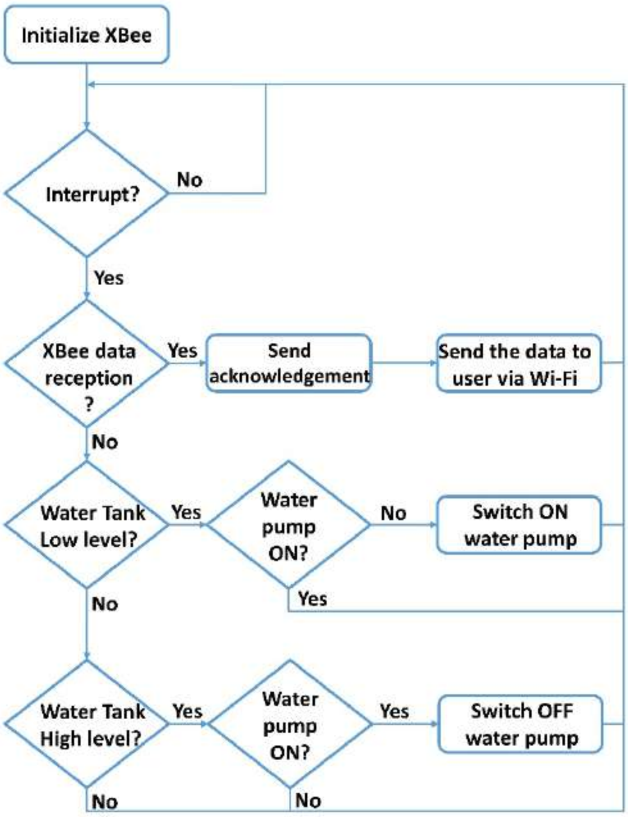


Fig. 2. Program flow of master node

MCP1700 low dropout voltage regulator is chosen to power up the components which work at 3.3V. This voltage regulator is very low cost (few cents) and consumes very low quiescent current (1.6 µA) [14]. Wi-Fi module and XBee module operate at 3.3V, but they cannot be supplied from one MCP1700. Because, ESP-01 module consumes 215mA and XBee modules consumes up to 33mA, and if both modules operate simultaneously the total current consumption is 248mA which almost equal to the maximum output current of the voltage regulator. Therefore, it is decided to dedicate separate voltage regulators for both modules in order to avoid overload on MCP1700. Figure 3 presents the implemented PCB of the master node.

C. End Device (Slave Node)

The end-device is designed to be capable to perform following main functions:

- Carrying out measurements of temperature, soil moisture level, rain presence and battery charge status
- Sending the measured values to coordinator wirelessly
- Resending the measured values to coordinator in case of not receiving acknowledgement from the coordinator device
- Automatically controlling actuator according to the measured values

- Automatically charging back-up battery from solar panel

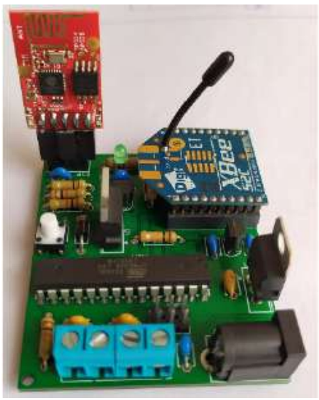


Fig. 3. Implemented PCB for the master node

The most significant parameters which must be under control are battery charge status and soil moisture level. Because, the back-up battery may be damaged due to overcharging or irrigating plant may die from draught owing to low level of soil moisture. Therefore, these parameters need to be measured frequently through the corresponding sensors and control actions should be performed. In this project, it is decided to measure them in every 8 minutes, but this period of time can be changed based on the type of irrigating plants. During 8 minutes, the microcontroller and radio module are put into sleep mode and all the sensors are disabled. In this time, only watchdog timer of microcontroller operates in active by consuming 10µA current at average. When the battery state is full of charge microcontroller terminates charging process, if not full, charging operation continues. If the soil moisture level increases up to the value that equals to the upper-threshold plus some hysteresis microcontroller closes solenoid valve and if the soil moisture level decreases down to the threshold microcontroller opens the solenoid valve. Besides that, in every hour end-device measures and transmits temperature, soil moisture level, battery charge status and rain presence values to the coordinator device. Fig. 4 presents the operation flow of the slave node.

The power section of the end-device is designed differently from the coordinator device. It is decided to involve all the electronic components that operate at the same voltage level in order to minimize number of voltage level conversions from one level into another. Considering the specification of being low-power of this work, it is proposed that all the electronic components of the end-device operate at the fixed 3.3Volts. Whole the circuit is powered up by the rechargeable battery with 3.7Volts and MCP1700 voltage regulator is employed to convert the input voltage into 3.3V. This voltage regulator is chosen due to its very low quiescent current of 1.6µA and maximum output current of 250mA [14]. However, the actuator of the end-device consumes 367mA, so it is decided to use separate voltage regulator for it. AP7366-33W5-7 fast transient low dropout regulator is chosen to power up actuator. Its maximum output current is 600mA at the fixed output voltage of 3.3V. But this regulator's quiescent current is 60µA that means it consumes 60µA current always whether it is used or not. This problem is solved that microcontroller disables AP7366-33W5-7 when the solenoid valve is not operating. In disabled mode the voltage regulator consumes only 0.05µA which is acceptable. When the solenoid valve needs to operate, microcontroller enables it by sending active high signal to its EN port.

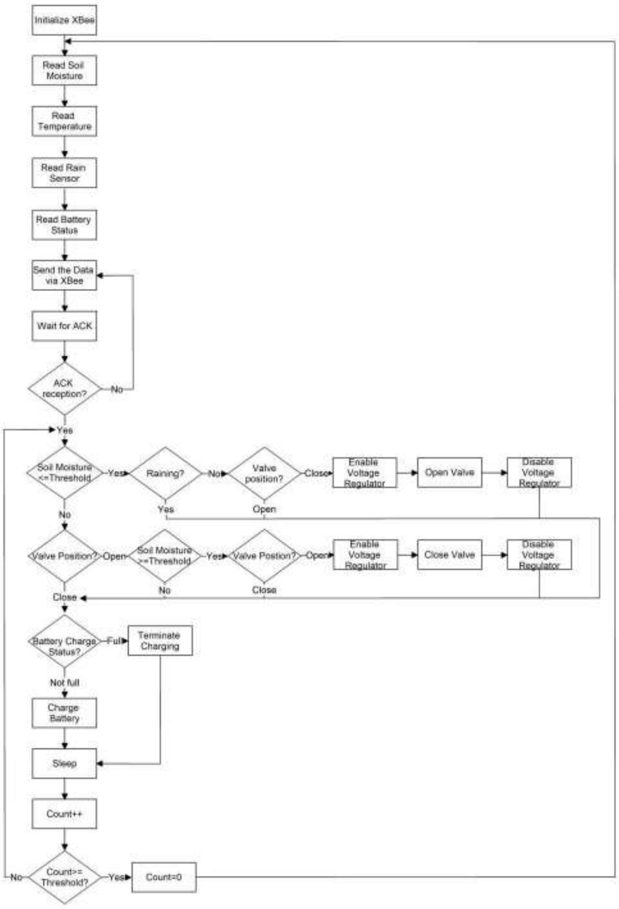


Fig. 4. The program flow of the end devices

In order to open and close the latching solenoid valve, the sign of control signal should be reversible. It can be executed by H-bridge drivers. TC78H651FNG bridge driver integrated circuit from Toshiba is chosen for the valve driver in this work. This IC operates at 3.3V and can output up to 1.5A of current at 3.3V [15]. Furthermore, this driver integrates two H-bridge circuits that means with TC78H651FNG the end-device is capable to control two actuators independently. Another advantage of this driver is its 0µA current consumption in standby mode. Because all the circuits in this integrated circuit configure with complementary metal oxide semiconductor elements that decrease intensely the standby current [16]. Besides, its small outline package (TSSOP) with the dimensions of 5mm X 4.4mm contributes overall PCB size to be compact. Fig. 5 shows the picture of the implemented PCB.

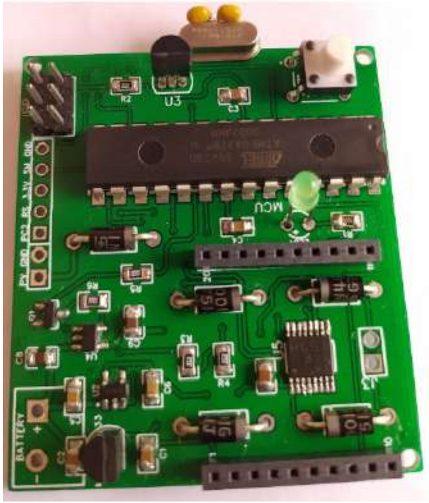


Fig. 5. End device PCB

The open-close states of the latching solenoid valve is executed according to the input and output functions of the H-bridge driver presented in the Table 2.

TABLE II. INPUT/OUTPUT FUNCTIONS OF H-BRIDGE DRIVER

Channel control inputs		**Channel outputs**		**Mode**
IN1	***IN2***	***OUT1***	***OUT2***	
HIGH	LOW	HIGH	LOW	Channel opening
LOW	HIGH	LOW	HIGH	Channel closing

The designed driver circuit for controlling the actuator is presented in Fig. 6.

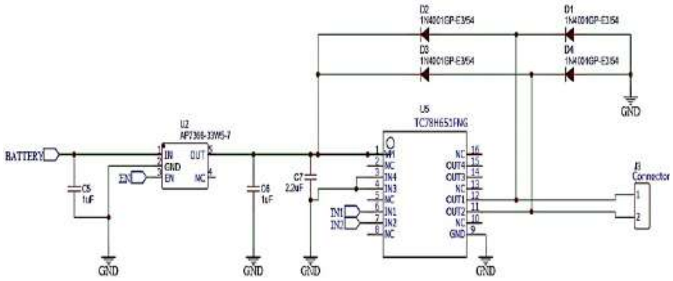


Fig. 6. Designed driver to control solenoid actuator

According to the requirements, the implemented end-device of WSAN should operate 2 years with the only one battery. Therefore, it is decided to design battery charging section by solar panel. Battery charging section should have following requirements:

- Over charge protection
- Reverse discharge protection
- Controllability by microcontroller
- Faster charge current regulation
- Compact dimension

Those requirements can be met by the use of MCP73831 charge management controller IC for Lithium-Ion and Lithium-Polymer batteries. This controller is very suitable to use in cost sensitive and space limited applications. It has fast charge constant-current mode and in this mode battery is charged at the programmed value of the current. The charge current value is formed inserting one external program resistor between PROG pin and V_{SS} pin. The programmed charge current is calculated using (1) [17]:

$$I_{reg} = \frac{1000}{R_{prog}} \tag{1}$$

where, R_{prog} is program resistor (in kΩ) and I_{reg} is programmed charge current (in mA). In this work, the charge controller is programmed to charge battery at 100mA constant current at 4.2V. In order to program it for 100mA charge current, 10kΩ program resistor is connected to the PROG pin. Besides, MCP73831 integrates the input overvoltage protection and reverse discharge protection.

TABLE III. STATE OF THE OUTPUT STATUS DURING CHARGING CYCLE [19]

Charge cycle state	**STAT1**	
	MCP73831	***MCP73832***
Shutdown	High Z	High Z
No Battery Present	High Z	High Z
Preconditioning	L	L
Constant-Current Fast Charge	L	L
Constant Voltage	L	L
Charge Complete-Standby	H	High Z

The charge status of battery can be read from STAT pin of MCP73831 by microcontroller. Table 3 shows the output signal on STAT pin of charge controller corresponding the state of charge cycle. The microcontroller reads STAT pin and makes decision to terminate or start the charging process by sending signal to PROG pin of MCP73831 through a transistor as depicted in Fig. 7, below.

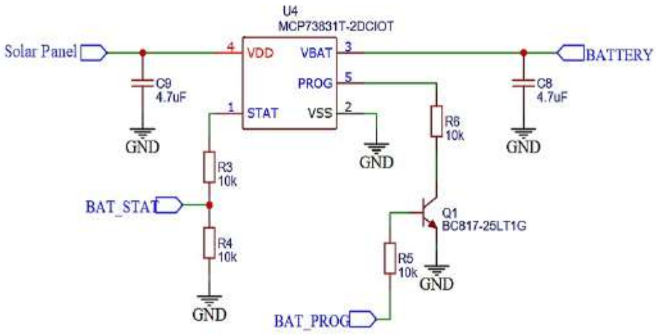


Fig. 7. Battery charging section of the end device

III. RESULTS AND DISCUSSION

The WSAN was tested in a rooftop garden and performed successful operation. Each end-device is powered up by single Li-Ion LIR14500 rechargeable battery. This battery has 3.7V nominal voltage and 800mAh nominal capacity. The end of discharge voltage of this battery is 2.8V [20]. In order to estimate the initial available energy amount of single battery following simple calculation can be used (2):

$$E_{init} = V \bullet I \bullet t \qquad (2)$$

$$E_{init} = 3.7V \bullet 800mA \bullet 3600\sec = 10656 Joules$$

Now, we can calculate total energy consumption by all the parts of an end-device.

One state change of latching solenoid valve spends 36.3mJ energy that is calculated from (3).

$$E_{state\_change} = \frac{V^2}{R_{coil}} \bullet T_{pulse} = \frac{3.3^2}{9} \bullet 0.03 = 36.3mJ \; (3)$$

Assuming the number of daily actuations is 10, and 60% of efficiency in power conversion, the total energy consumed to control one actuator per day is calculated as:

$$E_{daily\_actuator} = 10 \bullet E_{state\_change} \bullet \frac{100}{60} = 0.605 Joules$$

The maximum energy consumption values of the radio module in one cycle of wireless communication are presented in Table IV. One communication cycle is executed in every hour and spends 18.31*10^{-3} Joules from the back-up battery. In one day, the wireless communication cycle occurs 24 times and total energy consumption by XBee module is estimated as:

$$E_{daily\_XBee} = 24 \bullet 18.31 \bullet 10^3 = 0.44 Joules$$

TABLE IV. XBEE MODULE ENERGY CONSUMPTION OF SINGLE CYCLE ACTIVE COMMUNICATION

Action	**Duration (max)**	**Average current**	**Energy (Joules)**
Device is in sleep	≈ 1 hour	1µA	$11.88*10^{-3}$
Transition to active mode	10 ms	50 µA	$1.7*10^{-6}$
Data transmission and re-transmissions	50 ms (max)	33 mA	$5.5*10^{-3}$
Waiting and receiving acknowledgement	10 ms (max)	28 mA	$0.93*10^{-3}$
Device goes back to sleep			
Total energy for one hour activities:			$18.31*10^{-3}$

In one hour, microcontroller operates maximum $T_{hourly\_operation} \approx 78.5ms$ (1ms for reading sensors, 70ms for XBee active communication time, 1ms for 8-minute cyclic checking of soil moisture level and battery charge status) and sleeps almost $T_{hourly\_sleep} \approx 3600\sec$. Besides that, microcontroller is in active mode while the solenoid valve is actuating. In one day, solenoid valve actuates maximum 10 times and it lasts $T_{daily\_solenoid} \approx 10 \bullet 30ms = 300ms.$ That means, microcontroller also spends 300ms time per day in active mode. So, daily operation time of microcontroller is calculated as shown in (5).

$$T_{daily\_operation} = 24 \bullet T_{hourly\_operation} + T_{daily\_solenoid} \quad (5)$$

In active mode, the daily maximum energy consumption of microcontroller is

$$E_{daily\_active} = 3.3V \bullet 10mA \bullet 2.184\sec \approx 72.1 \bullet 10^{-3} Joules$$

In sleep mode, the daily maximum energy consumption of microcontroller is

$$E_{daily\_sleep} = 3.3V \bullet 44\mu A \bullet 24 \bullet 3600\sec = 12.55 Joules$$

The total daily energy consumption of microcontroller is sum of them:

$$E_{daily\_MCU} = E_{daily\_active} + E_{daily\_sleep} \approx 12.62 Joules$$

Moreover, voltage regulators also consume some energy consumption. The MCP1700 voltage regulator spends energy with the quiescent current of 1.6 µA during 24 hours in a day. Another voltage regulator AP7366 consumes energy with the quiescent current of 60µA only when the solenoid valve actuates (30ms). The total daily energy consumption of voltage regulators is

$$E_{daily\_VR} \approx 512.5 \bullet 10^{-3} Joules$$

In addition, all the sensors and H-bridge driver integrated circuit consume less than 1 Joule of energy in a day. Because, they are powered up only when microcontroller should measure the parameters or change the state of the actuator.

All in all, the total daily energy consumption of an end-device is calculated as:

$$E_{daily\_total} \approx E_{daily\_actuator} + E_{daily\_XBee} + \\ + E_{daily\_MCU} + E_{daily\_VR} + 1 = 14.5 Joules$$

As a result, the initial available energy of 10656 Joules a battery lasts approximately 734 days, which means more than 2 years of battery-life. However, it is crucial to take into consideration that there are some factors, e.g. temperature fluctuations and self-discharging phenomenon decrease the overall battery life time. That’s why, battery charging section is designed in the end-device PCB, and the charging can at least compensate those effects of discharging or increases the battery-life. Consequently, the requirement of 2 years battery life time is satisfied. Furthermore, the size of end-device’s PCB is 50mm to 45mm, so another requirement of being compactness is also satisfied.

IV. Conclusion

The solution to design and implement of low-power, low-cost and compact wireless sensor and actuator network to build automated irrigation systems in urban gardens has been presented in this paper. It utilizes ZigBee wireless technology to establish radio link between wireless nodes. In general, this work can have great impact for the automation of urban garden irrigation systems.

References

[1] Haijiang Tai, Nannan Wen and Daoliang Li. "A Wireless Low Power Valve Controller for Drip Irrigation Control Systems." Sensors & Transducers, Vol. 26, March 2014, pp. 7-16.

[2] Esther Ososanya, Sasan Haghani, Wagdy H Mahmoud and Samuel Lakeou. "Design and Implementation of a Solar-Powered Smart Irrigation System." 122nd ASEE Annual Conference & Exposition, (June, 2015) Seattle, WA.

[3] AbdelRahman Al-Ali, Murad Qasaimeh, Mamoun AI-Mardinia, Suresh Radder and I. A. Zualkernan. "ZigBee-Based Irrigation System for Home Gardens." 2015 International Conference on Communications, Signal Processing, and their Applications (ICCSPA'15), Sharjah, 2015, pp. 1-5

[4] Abdelraouf Ouadjaout, Noureddine Lasla, Miloud Bagaa, Messaoud Doudou, Cherif Zizoua, Mohamed Amine Kafi, Abdleouahid Derhab, Djamel Djenouri and Nadjib Badache. "DZ50: Energy-Efficient Wireless Sensor Mote Platform for Low Data Rate Applications", 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks - 2014, pp. 189-195

[5] J. Lin, W. Xiao, F. L. Lewis, and L. Xie, "Energy-efficient distributed adaptive multisensor scheduling for target tracking in wireless sensor networks", IEEE Transactions on Instrumentation and Measurement., vol. 58, no. 6, pp. 1886–1896, Jun. 2009.

[6] P. Györke and B. Pataki, "Energy-aware measurement scheduling in WSNs used in AAL applications," IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 5, pp. 1318–1325, May 2013.

[7] R. Yan, H. Sun, and Y. Qian, "Energy-aware sensor node design with its application in wireless sensor networks," IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 5, pp. 1183–1191, May 2013.

[8] P. Suriyachai, U. Roedig, and A. Scott, "A survey of MAC protocols for mission-critical applications in wireless sensor networks," IEEE Communications Surveys and Tutorials, vol. 14, no. 2, pp. 240–264, Apr./Jun. 2012.

[9] Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band, IEEE Standard 802.11b, 1999.

[10] Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs), IEEE Standard 802.15.1, 2002.

[11] Wireless Medium Access Control (MAC) and Physical Layer(PHY) Specifications for Low-Rate Wireless Personal Area Networks(LR-WPANs), IEEE Standard 802.15.4, 2003.

[12] DIGI. "XBee/XBee-PRO S2C ZigBee RF module" user guide. Digi International 2020

[13] STMicroelectronics. "L78xx – L78xxC/L7805xxAB – L7805AC datasheet" September 2010

[14] Microchip. "MCP1700 datasheet" – Microchip technology, 2018

[15] Toshiba "TC78H651FNG datasheet". Toshiba Electronic Devices & Storage Corporation, 2018

[16] Reference manual. "TC78H651FNG Application Note". Toshiba Electronic Devices & Storage Corporation, 2018

[17] Microchip. "MCP73831/2 datasheet" Microchip Technology 2014 EEMB Co., LTD. "LIR14500 Datasheet Li-ion battery" 2006

LoRa Cloud-based platform for Internet of Things Applications

Mouna GASSARA
Computer & Embedded Systems Laboratory,
University of Sfax,
Tunisia
gas.mouna@gmail.com

Manel ELLEUCHI
Computer & Embedded Systems Laboratory,
University of Sfax,
Tunisia
Digital Research Center of Sfax,
University of Sfax,
Tunisia
manelelleuchi@gmail.com

Mohamed ABID
Computer & Embedded Systems Laboratory,
University of Sfax,
Tunisia
Digital Research Center of Sfax,
University of Sfax,
Tunisia
mohamed.abid_ces@yahoo.fr

***Abstract*— Long-range radio technologies can ensure Low Power Wide Area Network deployment at a very low-cost intended for a huge range of Internet of Things applications. Recently, there are some attempts to implement Long-range cloud connection in this field. Nevertheless, several issues must be considered when setting up Long-range Internet of Things solutions. The paper proposes a novel solution based on Long-range Cloud implementation on The Things Network platform for the Water Pipeline monitoring application. This solution helps the researchers in terms of preparing their test-bed environment. For validation purposes, an experimental Long-range Cloud system has been established by the means of low-power and low-cost electronics using accessible open sources to deploy Long-range nodes, Long-range gateway, and The Things Network. The system proves its efficiency through its harmonic and synchronized operation mode.**

Keywords—IoT; LoRa; Low power; Cloud; The Things Network; Water Pipelines Monitoring application

I. Introduction

Network connectivity is provided on small edge devices by the Internet of Things (IoT) and new services are offered at reasonable prices. The major defy in the IoT applications and systems conception is the reduction of the implementation and maintenance costs of edge nodes having an increasing number concerning hand-held devices number.

By the dint of the wireless communication protocols specified for IoT applications, hardware complexity and power consumption of edge nodes can be decreased. In a similar context, Thanks to the cloud technology providing the common service frameworks, costs of IoT systems maintenance can be minimized. There are two important types of wireless communication protocols conceived for IoT: short- and long-range communication protocols. The first type includes Bluetooth, Zigbee, and WiFi as short-range communication protocols. These IoT protocols are appropriate for the indoor environment. These IoT systems are founded on application servers and wired networks excluding the short wireless link connecting edge nodes and wireless gateways. The second type integrates long-range wireless communication protocols for IoT like Sigfox [1] and LoRa [2]. The edge nodes related to this kind of systems do not necessitate battery change for many years. Additionally, these nodes have a communication range of tens of kilometers. These protocols are suitable for the outdoors permitting agricultural data sensing or civil infrastructure monitoring. However, adding network servers is essential to supervise the radio resource connecting wireless gateways and application servers taking into consideration the channel states of multiple wireless gateways. This fact creates the difference between long and short-range IoT systems. There are some low-power, long-range, operator-based technologies such as Sigfox^{TM} which cannot be utilized in an ad-hoc manner. Nevertheless, other technologies such as LoRa^{TM} proposed by Semtech radio manufacturers can be used in private. Deploying such technology according to the newly proposed LoRaWAN^{TM} specifications [2] is possible for large-scale interoperability or based completely on ad hoc solutions. These solutions enable us to perform customization toward specific application profiles.

In accordance with a previous survey on IoT cloud platforms [3], there are no less than 49 IoT cloud platforms on the market to satisfy the requirements of various individual users and groups such as enterprises, governments, farmers, healthcare providers, transport operators, manufacturing services and communications, etc. [4]. The best used IoT cloud platform [5] in LoRa Cloud projects is The Things Network (TTN) [6].

This paper focuses on the water pipeline monitoring as an IoT application, and more particularly, it presents a novel test-

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

bed LoRa Cloud platform based on the TTN which serves to detect the leaks in these pipelines since water is a vital raw material for the human being. It guarantees low-power and long-range communication with the cloud. Also, our proposed system is conceived with LoRa nodes and gateway, connected to the cloud via TTN.

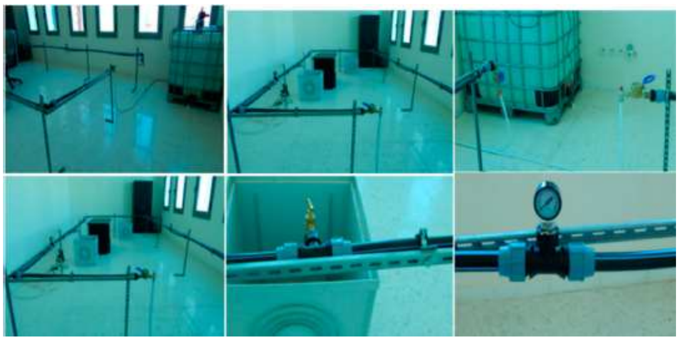


Fig 1. Water Pipeline Monitoring application [7]

The remainder of this paper is arranged as follows: Section II introduces the LoRaWAN, cloud computing, and TTN concepts. It treats different key issues with three notions. Section III presents the literature works related to the IoT cloud platforms suitable for the LoRa protocol and highlights some works in the water monitoring field. The realized test-bed structure is described deeply in section IV. Finally, section V concludes the paper.

II. PREREQUISITES

We define, in this section, three important concepts used throughout the paper, that is LoRaWAN, cloud computing, and The Things Network. First, we will identify the basics of the notion LoRaWAN by focusing on the principle characteristics that justify the LoRa protocol choice. Secondly, we present the cloud computing notion by emphasizing on its proven efficiency.

A. *LoRAWAN*

LoRa defines a low-power, long-range wireless communication platform designed as an efficient platform for small, low-powered devices. It is not only appropriate for battery-operated devices but also for IoT deployment [8].

The LoRaWAN system parts are illustrated in Figure 2. These parts are terminal, gateway, network server, and application server.

- **The terminal** is considered as an edge device deployed on remote sites.
- **The gateway** permits the transfer of packets coming from the LoRa terminals to the LoRa network servers and vice versa. Additionally, the gateway permits the transmission from the LoRa physical layer protocol to the protocols of the backhaul network.
- **The network server** operates with protocols of the upper-layer for IoT systems. Furthermore, control signaling messages of the MAC/L2/L3 layers are exchanged between this server and the terminals.

As a final point, **the application server** is responsible to handle the data collected from IoT terminals. Generally, different organizations work with the application server. Hence, it is independent of the IoT network protocols.

B. *Cloud computing*

Cloud computing presents the on-demand computer system resourcesavailability, especially computing power and data storage, with no active and direct managing of the user [9]. Broadly speaking, this term defines the data centers availability to numerous users through the Internet. Actually, huge clouds, having functions distributed throughout several emplacements from central servers, preponderate. If the connection to the user is completely closed, it can be considered an edge server.

There are three types of cloud[10]. Enterprise clouds are limited to a particular organization. Public clouds are available to various organizations. The hybrid cloud combines the first two types, i.e. enterprise and public clouds.

The cloud computing concept is based on resource sharing to reach scale economies and a high degree of coherence. Avoiding (or only minimizing) up-front IT infrastructure costs are permitted to companies by advocates of public or hybrid clouds. Furthermore, cloud computing allows enterprises to get their applications and run them faster, with ameliorated manageability and diminished maintenance costs. That demands from IT teams to rapidly regulate resources to accommodate unpredictable and fluctuating requirements. The “pay-as-you-go” is a model, typically used by cloud providers. Once administrators are not familiarized with cloud-pricing models [11], these models can bring unexpected operating expenses.

The pervasive availability of high-capacity networks induces the remarkable growth in cloud computing. The autonomic and efficient computing, as well as the low-cost computers and the ubiquity of high-storage devices,are considered another contributing factor in this growth.

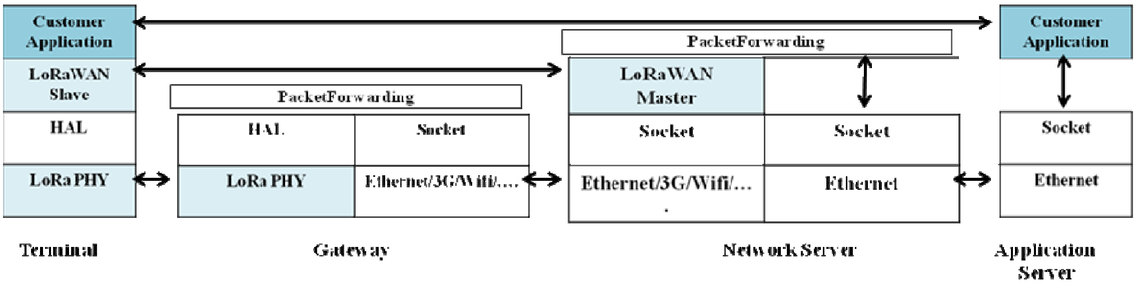


Fig 2. LoRaWAN system structure

Besides, hardware virtualization and the widespread implementation of service-oriented architectures are counted as a third factor that helped to improve cloud computing growth.

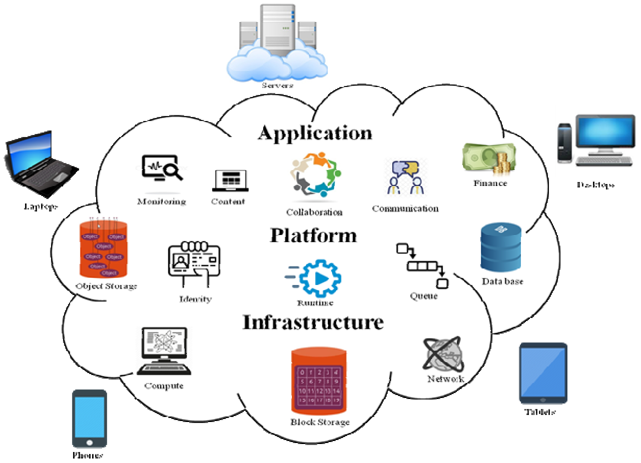


Fig 3. A general cloud computing model

C. *The Things Network*

The Things Network (TTN)[6] defines an IoT cloud platform suitable for the LoRa protocol. It allows low power devices to utilize long-range gateways for connecting to a decentralized network to exchange data with applications.

III. Related works

In the literature, several related works are developed using the LoRa Cloud Platform. The authors in [12] proposed a solution based on the LoRa platform. They deploy ad hoc IoT test-beds using Thingspeak [13] platform account for implementing the different sensors and the gateway. The data, also, is visualized in real-time.

ThingSpeak is an open-source API and application for the "Internet of Things", allowing storing and collecting data from connected objects using the HTTP protocol over the Internet or a local network. With ThingSpeak, we can create sensor data logging applications, location tracking applications and a social network for connected objects, with status updates. Moreover, it operates with several IoT platforms like Arduino, ioBridge, Twitter, Twilio, MATLAB, and ThingHTTP.

Pham et al. [14] explained how these applications are being utilized in the African rural and the sub-Saharan context. They detailed the deployment of sensors and gateway which can be made by small farms based on the free account created on the ThingSpeak platform. They explained a low-cost, long-range IoT framework that takes into account the cost of hardware and services, rapid ownership, flexibility, and customization by third parties.

WAZIUP project was proposed by Dupont et al. [15] to remote, largely isolated and rural, areas in the developing countries, particularly those located south of the Sahara Desert. The main objective is to deploy an open innovation IoT platform capable of bridging the North/South technological divide with Big Data capabilities.Three different cloud layers are used by WAZIUP IoT, these are the Infrastructure as a Service (IaaS), the Container as a Service (CaaS), and the Platform as a Service (PaaS). The first layer is provided by OpenStack [16] to provide virtual machines (VMs) running the full platform. The second layer is used to serve containers (e.g. Docker containers) to WAZIUP services and applications and provided by Kubernetes [17]. The compilation and deployment of an application are ensured by the third Cloud layer which is provided by Deis [18]. Indeed, Deis allows us to compile all the applications pushed by the users and to host them in containers on Kubernetes.

In addition to that, OpenStack is among the best known open source cloud platforms [16]. It provides essential functions for creating application services in any cloud environment. OpenStack provides a collection of interfaces for a service application program. It also ensures the automation of management operations.

The system services offered by OpenStack was used by So and his colleagues [19] in the real restructuring of the operations of the LoRa network server to make it scalable and flexible.Indeed, Four Virtual Machines are used on the function of the LoRa network server on the OpenStack platform. First, the gateway agent VM defines the protocol of communication between the LoRa network server and the LoRa gateway. The second is the LoRa data VM which ensures the operations of the data plane. The third VM is the application server which makes the LoRa network server connects with the application server using the efficient transport protocol. And the last one is the LoRa control that ensures the control plane.

The authors in [20] detail the implementation of the Thethings.io which is an industrial IoT cloud platform. By using a simple API, developers can have a whole back-end solution using the Thethings.io. It is considered hardware agnostic and permits to connect any device able to use WebSockets, HTTP, CoAP, or MQTT protocols. It supports supervision, monitoring, and analysis functions.

The project in [21] was based on TTN and develop a single channel LoRa gateway using Raspberry Pi 2 and other shields [22] open sources [23]. In this same type, the authors in [24] present a project that used the ESP8266 module [25] and the Radio RFM95/96. To test the developed single-channel LoRa gateway three nodes have been used. The first one is tested with an ESP8266-based LoRa node with RFM95 transceivers. The second one is with an ArduinoPro Mini. Then, the third one is based on a TeensyLC connected to HopeRF RFM 95 radio.

Saravanan et al. [26] proposed a solution formonitoring water level related to storage reservoirs in Water Distribution Networks (WDN). This solution is based on LoRa IoT enabled devices to remote actuation of valves. A gateway based on Raspberry Pi gathers the level data from all LoRa nodes. Then, it submits the data to the local server and control signals to the actuator. Grafana platform is utilized for data visualization and offers a human-machine interface serving to manipulate the valve remotely.

Another solution conceived for Smart Village Projects is proposed by Manoharan et al. [27]. Thanks to LoRa devices with sensors placed in water tanks at villages,the realized system supervises the level and quality of water in all tanks. The whole water distribution system is displayed and controlled from one place.

IV. LoRa IoT Cloud Platfrorm for water monitoring application

Our contribution consists in deploying a cloud-based platform for LoRa Internet of Things in Water Pipelines Monitoring Application (Fig. 4).

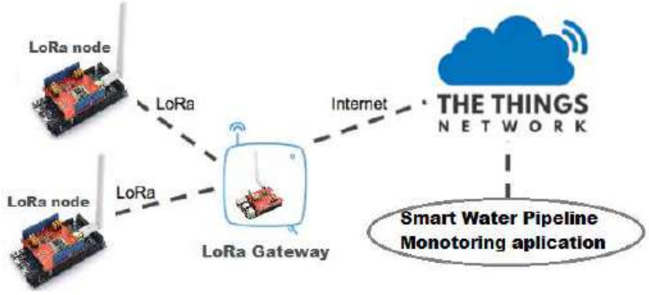


Fig 4. LoRa Cloud-based platform

On the one hand, we use two LoRa end-devices to collect water flow monitoring data based on YF-S201 Water Flow Sensor. We use the LoRa/ GPS HAT for RPI v1.4 [28] as the LoRa Radio module to deploy nodes. The first LoRa node is realized by the means of Arduino UNO R3 [29] and the second one is based on Arduino Mega 2560 [30] as shown in Fig.5.

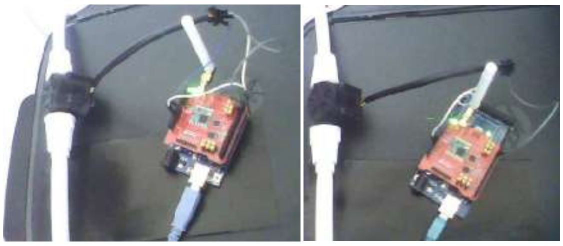


Fig 5. LoRa end-devices

On the other hand, we use a Raspberry Pi 2 model 3 [31] plugged into the LoRa/ GPS HAT for RPI v1.4 (Fig. 6). We program this component to be a LoRa gateway which receives data from LoRa nodes and sends it to the cloud. The realized component, having an identifier “eui-b827ebffff4f8ed9”, is characterized by its low-cost and low-power comparing with other electronic components. The whole platform is illustrated in Fig. 7.

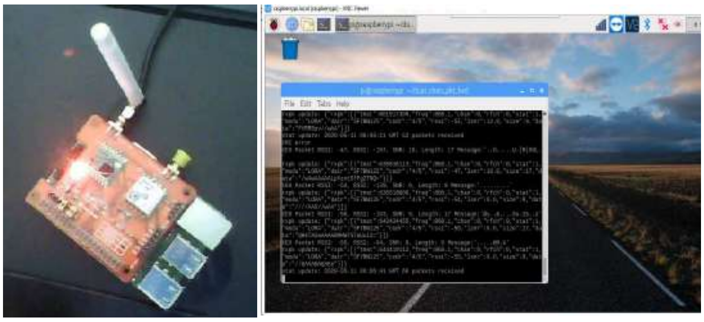


Fig 6. LoRa gateway

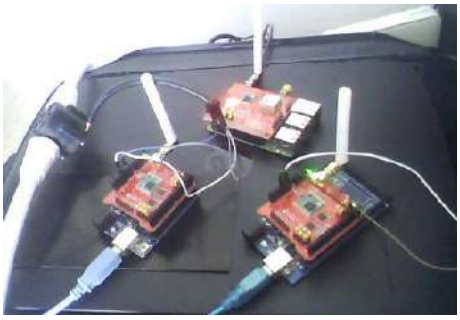
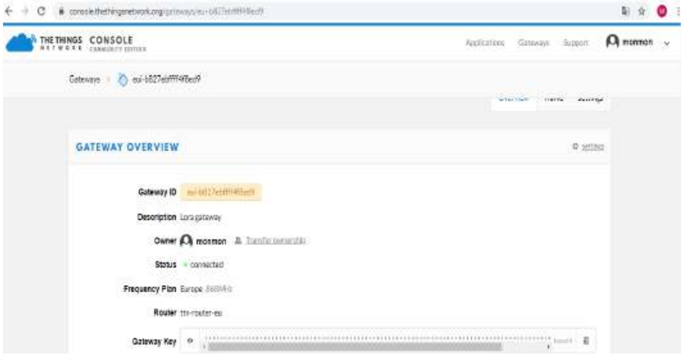


Fig 7. LoRa deployment

After ensuring the right function of our LoRa network (LoRa end-device sand gateway), we will register the gateway device in The Things Network (TTN) platform to connect to the Cloud. We add the LoRa component to TTN devices, that gives it active status (green color) as portrayed in Figure 8.This figure illustrates an overview of the added gateway. A specified ID is associated with the new component giving a detailed description. The connected status proves the successful connection of the created gateway.



GATEWAY OVERVIEW

Gateway ID eui-b827ebffff4f8ed9

Description Lora gateway

Owner monmon Transfer ownership

Status connected

Frequency Plan Europe 868MHz

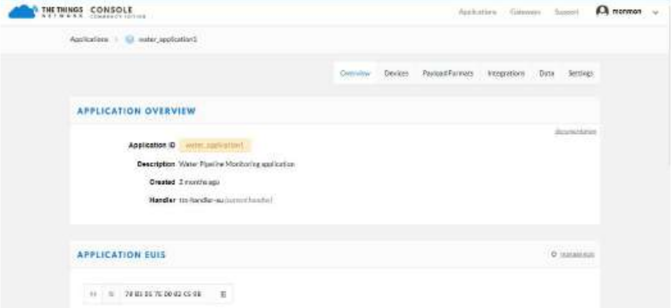
Router ttn-router-eu

Gateway Key

Fig 8. LoRa gateway registration in the TTN

While the gateway is ready to receive information, it is essential now we create an application in TTN for sending and receiving data packets from LoRa end-devices through the connected gateway. Figure 9 shows our Water Pipeline Monitoring application which is distinguished by its

“Application EUI” equal to “70 B3 D5 7E D0 02 C5 9B”.



THE THINGS NETWORK CONSOLE COMMUNITY EDITION

Applications > water_application1

Overview | Devices | Payload Formats | Integrations | Data | Settings

APPLICATION OVERVIEW

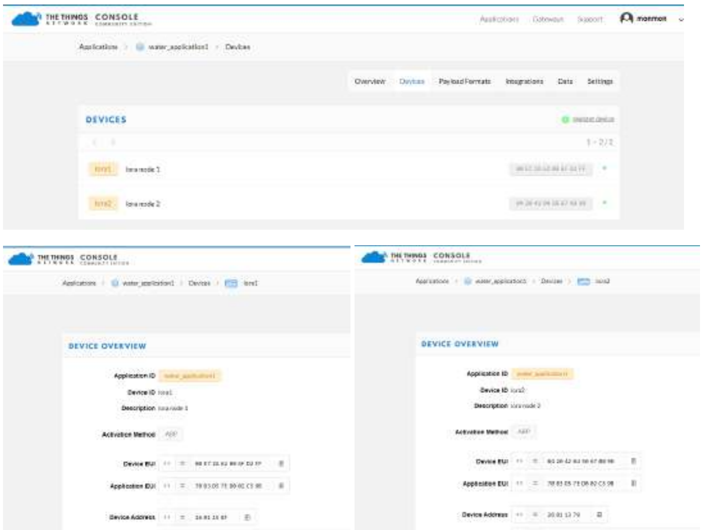
Application ID water_application1
Description Water Pipeline Monitoring application
Created 2 months ago
Handler ttn-handler-eu (current handler)

APPLICATION EUIS

70 B3 D5 7E D0 02 C5 9B

Fig 9. Creation of Water Pipeline Monitoring application in the TTN

It is essential now to add application devices to define its nodes correctly. For this, we add the two realized LoRa end-devices to the created application as shown in Figure 10. Therefore, every end-device is characterized by its device address. In our case, LoRa node based Arduino UNO has an address equal to “26 01 15 8F” and the node based on Arduino MEGA 2560 has an address equal to “26 01 13 79”.



THE THINGS NETWORK CONSOLE COMMUNITY EDITION

Applications > water_application1 > Devices

Overview | Devices | Payload Formats | Integrations | Data | Settings

DEVICES

1 - 2/2

lora1 lora node 1
lora2 lora node 2

Applications > water_application1 > Devices > lora1

DEVICE OVERVIEW

Application ID water_application1
Device ID lora1
Description lora node 1
Activation Method ABP
Device EUI
Application EUI 70 B3 D5 7E D0 02 C5 9B
Device Address 26 01 15 8F

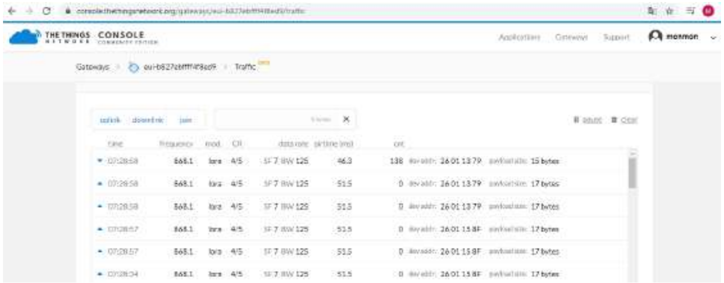
Applications > water_application1 > Devices > lora2

DEVICE OVERVIEW

Application ID water_application1
Device ID lora2
Description lora node 2
Activation Method ABP
Device EUI
Application EUI 70 B3 D5 7E D0 02 C5 9B
Device Address 26 01 13 79

Fig 10. Adding LoRa end-devices to the Water Pipeline Monitoring application in the TTN

After finalizing the conception of our proposed LoRa cloud platform on the TTN, it is important to validate its operating mode. We circulate a water flow in the pipeline and the sensors start sends LoRa data to the gateway so we can observe uplink packets in the TTN Gateway Data. These packets are coming from nodes having addresses “26 01 15 8F” and “26 01 13 79” as illustrated in Fig. 11.



THE THINGS NETWORK CONSOLE COMMUNITY EDITION

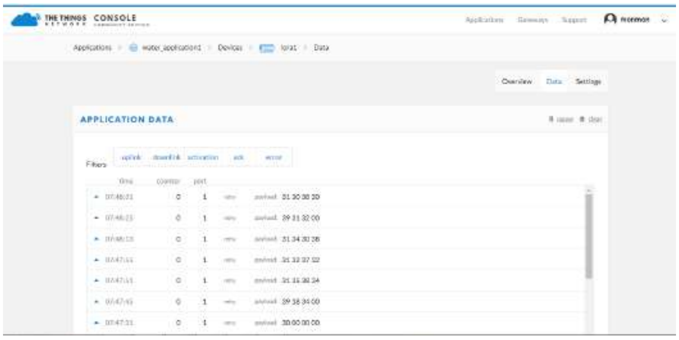
Gateways > eui-b827ebffff4f8ed9 > Traffic

uplink | downlink | join

time	frequency	mod.	CR	data rate	airtime (ms)	cnt		
07:28:58	868.1	lora	4/5	SF 7 BW 125	46.3	138	dev addr: 26 01 13 79	payload size: 15 bytes
07:28:58	868.1	lora	4/5	SF 7 BW 125	51.5	0	dev addr: 26 01 13 79	payload size: 17 bytes
07:28:58	868.1	lora	4/5	SF 7 BW 125	51.5	0	dev addr: 26 01 13 79	payload size: 17 bytes
07:28:57	868.1	lora	4/5	SF 7 BW 125	51.5	0	dev addr: 26 01 15 8F	payload size: 17 bytes
07:28:57	868.1	lora	4/5	SF 7 BW 125	51.5	0	dev addr: 26 01 15 8F	payload size: 17 bytes
07:28:34	868.1	lora	4/5	SF 7 BW 125	51.5	0	dev addr: 26 01 15 8F	payload size: 17 bytes

Fig 11. LoRa gateway data in the TTN

After receiving packets, LoRa gateway sends data to the Water Pipeline Monitoring application in TTN. We concentrate in this paper on monitoring the parameter water flow value since any degradation of its variation under a fixed threshold (a constant fixed experimentally) means leakage detection. In Fig. 12 and Fig. 14, we can perceive water flow values (L/hour) in hexadecimal form in the TTN application data of the first and the second end-devices.



THE THINGS NETWORK CONSOLE COMMUNITY EDITION

Applications > water_application1 > Devices > lora1 > Data

Overview | Data | Settings

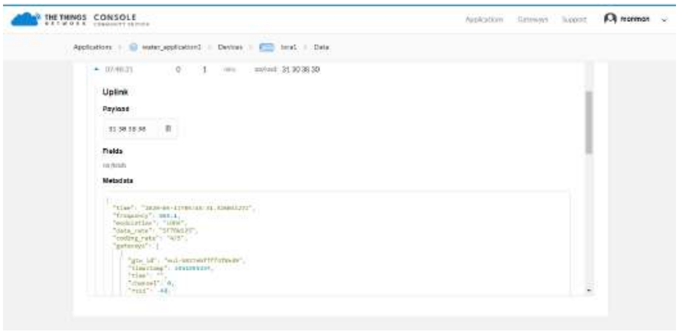
APPLICATION DATA

Filters: uplink | downlink | activation | ack | error

time	counter	port		
07:48:21	0	1	retry	payload: 31 30 38 30
07:48:21	0	1	retry	payload: 39 31 32 00
07:48:13	0	1	retry	payload: 31 34 30 38
07:47:41	0	1	retry	payload: 31 32 37 32
07:47:41	0	1	retry	payload: 31 31 38 34
07:47:41	0	1	retry	payload: 39 38 34 00
07:47:31	0	1	retry	payload: 30 00 00 00

Fig 12. TTN application data of the first LoRa end-device

Fig. 13 and Fig. 15 detailed data related to the water flow value “31 30 38 30” (1080 L/hour) and “37 30 34 00” (704 L/hour) coming from the first and the second LoRa end-device respectively. The more relevant information is that the used gateway is the same realized in our proposed platform having “eui-b827ebffff4f8ed9” as an identifier. So, we can affirm the good operation mode of our conceived system.



THE THINGS NETWORK CONSOLE COMMUNITY EDITION

Applications > water_application1 > Devices > lora1 > Data

07:48:21 0 1 retry payload: 31 30 38 30

Uplink

Payload

31 30 38 30

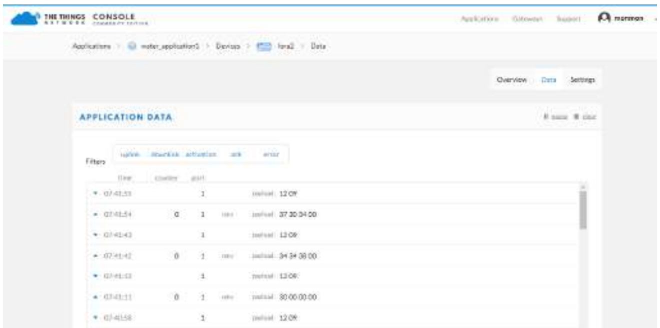
Fields

no fields

Metadata

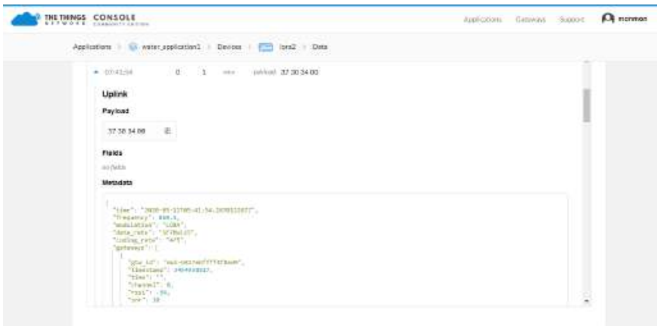
```
{
  "time": "2020-04-11T05:48:21.[illegible]Z",
  "frequency": 868.1,
  "modulation": "LORA",
  "data_rate": "SF7BW125",
  "coding_rate": "4/5",
  "gateways": [
    {
      "gtw_id": "eui-b827ebffff4f8ed9",
      "timestamp": [illegible],
      "time": "",
      "channel": 0,
      "rssi": -48,
```

Fig 13. Detailed information on water flow value related to the first LoRa end-device



time	counter	port		
07:41:53		1		payload: 12 09
07:41:54	0	1	retry	payload: 37 30 34 00
07:41:43		1		payload: 12 09
07:41:42	0	1	retry	payload: 34 34 38 00
07:41:13		1		payload: 12 09
07:41:11	0	1	retry	payload: 30 00 00 00
07:41:58		1		payload: 12 09

Fig 14. TTN application data of the second LoRa end-device


THE THINGS NETWORK CONSOLE
COMMUNITY EDITION
Applications
Gateways
Support
mohmon
Applications > water_application1 > Devices > lora2 > Data

07:41:54 0 1 retry payload: 37 30 34 00

Uplink

Payload

37 30 34 00

Fields

no fields

Metadata

```
{
  "time": "2020-05-11T05:41:54.[illegible]",
  "frequency": 868.1,
  "modulation": "LORA",
  "data_rate": "SF7BW125",
  "coding_rate": "4/5",
  "gateways": [
    {
      "gtw_id": "eui-[illegible]",
      "timestamp": [illegible],
      "time": "",
      "channel": 0,
      "rssi": -59,
      "snr": 10
```

Fig 15. Detailed information on water flow value related to the second LoRa end-device

We confirm from these results the well operating mode of our realized platform. To conclude, we achieve to conceive a successful LoRa cloud test-bed composed of two end-devices and a gateway connected to The Things Network for Water Pipeline Monitoring application.

V. Conclusion

The number of Internet of Things (IoT) applications has increased significantly in recent years. Thus, these require low-power operation and long-range communication which are provided by the application of LoRa/LoRaWAN. The connected devices generate big data that are processed by system architectures based on the cloud server. In this paper, we deploy a real LoRa Cloud platform on The Things Network for the Water Pipeline Monitoring application. Our platform is realized with materials guarantying low-cost, low-power, and low-consumption characteristics. Our contribution is very essential for researchers in terms of preparing their test-bed environment for any contribution related to the LoRa cloud communication in different IoT applications. As future work, we will develop a new version of the proposed platform based on several connected devices and that can be deployed in a large scale application.

References

[1] Sigfox. [Online]. Available: http://www.sigfox.com/en/

[2] LoRaAlliance, “*LoRaWAN specification, v1.0*,” 2015.

[3] Partha Pratim Ray,”*A survey of IoT cloud platforms*”. Future Computing and Informatics Journal 1, 2016, pp. 35.

[4] https://profitbricks.com.

[5] Mouna Gassara, Manel Elleuchi, and Mohamed Abid, “*LoRa IoT cloud platforms: Survey*,” unpublished (submitted to ClusterComputing journal).

[6] https://www.thethingsnetwork.org/docs/

[7] Manel Elleuchi, Manel Boujeleben & Mohamed Abid, “*Energy-efficient routing model for water pipeline monitoring based on wireless sensor networks*”, International Journal of Computers and Applications, 2019, pp. 7. DOI: 10.1080/1206212X.2019.1682239

[8] Shuhaizar Daud, Teoh Shi Yang, Muhamad Asmi Romli, Zahari Awang Ahmad, Norfadila Mahrom and Rafikha Aliana A. Raof, “*Performance Evaluation of Low Cost LoRa Modules in IoT Applications*”. IOP Conference Series: Materials Science and Engineering, 2018.

[9] https://en.wikipedia.org/wiki/Cloud_computing

[10] Heyong Wang, Wu He, and Feng-Kwei Wang, “*Enterprise Cloud Service Architectures*”. Information Technology and Management, 2012, pp. 445-54.

[11] https://www.forbes.com/sites/centurylink/2014/02/27/wheres-the-rub-cloud-computings-hidden-costs/#4cd0bd995f00

[12] Congduc Pham, “*Building low-cost gateways and devices for open LoRa IoT test-bed*”. Journal of Test beds and Research Infrastructures for the Development of Networks and Communities, 2017.

[13] https://thingspeak.com.

[14] Congduc Pham, Abdur Rahim, and Philippe Cousin, “*Low-cost, Long-range open IoT for smarter rural African villages*”. 2016 IEEE International Smart Cities Conference (ISC2), Trento, 2016, pp. 1-6.

[15] Corentin Dupont, Tomas Bures, Mehdi Sheikhalishahi, Congduc Pham, Abdur Rahim, “*Low-Cost IoT, Big Data, and Cloud Platform for Developing Countries*”. Economics of Grids, Clouds, Systems, and Services - 14th International Conference, GECON 2017, Biarritz, France, September 19-21, 2017, Proceedings, Sep 2017, Biarritz, France. pp.285-299.

[16] https://www.openstack.org.

[17] http://kubernetes.io.

[18] http://deis.io.

[19] Jaeyoung So, Daehwan Kim, Hongseok Kim, Hyunseok Lee, and Suwon Park, “*LoRaCloud: LoRa platform on OpenStack*”. 2016 IEEE NetSoft Conference and Workshops (NetSoft), Seoul, 2016, pp. 431-434.

[20] https://thethings.io/

[21] https://www.instructables.com/id/Use-Lora-Shield-and-RPi-to-Build-a-LoRaWAN-Gateway/

[22] https://www.dragino.com/products/module/item/102-lora-shield.html

[23] https://github.com/tftelkamp/single_chan_pkt_fwd

[24] https://github.com/things4u/ESP-1ch-Gateway-v5.0

[25] https://fr.aliexpress.com/item/32368848967.html.

[26] Saravanan Chinnusamy, Prasanna Mohandoss, Partha Paul, Rohit R, N Murali, S Murty Bhallamudi, Shankar Narasimhan, Sridharakumar Narasimhan,“*IoT Enabled Monitoring and Control of Water Distribution Network.*” (2018).

[27] A. M. Manoharan and V. Rathinasabapathy, "Smart Water Quality Monitoring and Metering Using Lora for Smart Villages," *2018 2nd International Conference on Smart Grid and Smart Cities (ICSGSC)*, Kuala Lumpur, 2018, pp. 57-61, doi: 10.1109/ICSGSC.2018.8541336.

[28] https://www.dragino.com/products/lora/item/106-lora-gps-hat.html

[29] https://store.arduino.cc/arduino-uno-rev3

[30] https://www.raspberrypi.org/products/raspberry-pi-3-model-b/

[31] https://codebender.cc/sketch:273109#thethingsnetwork-send-v1.ino.

[32] https://github.com/bokse001/dual_chan_pkt_fwd.

COVID-19 Identification in CLAHE Enhanced CT Scans with Class Imbalance using Ensembled ResNets

Sowmya Sanagavarapu
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
sowmya.ssanagavarapu@gmail.com

Sashank Sridhar
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
sashank.ssridhar@gmail.com

Prof. T.V. Gopal
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
gopal@annauniv.edu

***Abstract*—The occurrence of imbalanced datasets in medical imaging has proven to be a challenge for the development of models to analyze and evaluate the underlying condition. In this paper, the bias of the chest CT scan dataset is handled by taking discrete splits and employing ResNets to detect COVID-19 in each split. The scraped images were pre-processed using CLAHE histogram for comparison with low contrast images. Multiple ResNets were extended to form an ensemble neural network model using ANNs which handles the class imbalance. The system has an overall accuracy of 87.23% and the performance is assessed for each class. The image features identified are visualized using the GradCAM algorithm and some of the commonly found clinical features in the CT scan images of the patients suffering from this disease are summarized for better understanding the working of the model.**

Keywords—Imbalanced Data, COVID-19 detection, Ensemble ResNets, Chest CT-scans

I. Introduction

Deep neural network architectures have been widely used in the classification of medical images for the detection of the underlying condition [1]. These systems have the ability to identify and learn extensive features from the image datasets for discerning the clinical features detected in them.

COVID-19 [2] is a disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 or (SARS-CoV-2). It is a highly communicable disease with people experiencing respiratory illness and recovery expected with effective and appropriate treatment methods. The most commonly used test for the disease is the RT-PCR (Reverse Transcription Polymerase Chain Reaction) [3]. The identification of this disease can also be done by analyzing the chest CT scans of the patients for some of the widely observed clinical features given below [4]:

1) Ground Glass Opacities (or GGOs)

One of the most common findings in chest CT scans of people diagnosed with COVID-19 infections is GGOs [5]. Usually multifocal, bilateral and peripheral but in the early phase of this disease. They initially start developing from the inferior lobe of the right lung as a unifocal lesion.

2) Air space consolidation or opacification

This refers to the condition in which there is replacement of the usually present gas by fluids or solids in the lung parenchyma (small airways present in the lungs) [6].

3) Crazy paving appearance

This occurs in the more later stages of the infection. Here, thickened interlobular and intralobular lines in combination with a ground glass pattern are visible in the CT scans of the patients [7].

4) Bronchovascular thickening in the lesion

This refers to the widened vessels observed through the chest CT scans. The bronchial walls become thicker due to inflammation owing to the accumulation of liquid or mucus [8].

5) Traction bronchiectasis

This is a common condition observed in patients that arises when the airways to lungs may get damaged around GGOs [9]. Due to this, there will be accumulation of the mucus in the lungs, leading to a variety of complications including buildup of bacteria causing other infections.

6) Formation of subpleural bands

This is a comparatively rare condition found in patients where there is a formation of thin curvilinear opacities between the surface of the lungs and the chest walls [10].

The need for a model to handle highly imbalanced dataset would help in the identification of the disease with its estimated severity for timely treatment. With the spread of the disease to over 214 countries [11] of the world with widely varying severity and the low detection rates of the disease due to its asymptomatic nature, the identification of the disease is proving to be a challenge.

Residual deep neural networks or ResNets [12] follow the architecture of Convolutional Neural Networks (or CNNs) [13]. They exhibit high control over the information that flows from one cell to the next in the cell using a parameterized forget gate. These networks have a more refined residual block, a pre-activation variant of residual block for simplifying the flow of learned gradients in the network.

In the field of auxiliary medical diagnosis technology, ensembled deep neural networks [14] have performed extremely

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

capably for the detection of clinical features in the medial images. Ensemble network architectures [15] use the activations and the feature weights learned from the individual deep neural models and their learnings are enhanced for increasing the efficiency of their performance.

Images available for training may not be of uniform contrast or brightness for the neural network to identify features within them [16]. To improve the contrast improvement index, entropy and measure of enhancement [17], Contrast Limited Adaptive Histogram Equalization (CLAHE) is used for their normalization. The images are split into regions and histograms are calculated for each region which are then interpolated in between.

Gradient-weighted Class Activation Mapping (or GradCAM) [18] algorithms are a generalization of the Class Activation Mapping [19] algorithms used for producing visual explanations for trained deep neural models. The better localization and clear class discriminative saliency maps allow the understanding of the gradient weights built up at the final layer in the deep neural architecture [20]. They are combined with existing pixel-shaped visualizations to create high-resolution class-discriminative visualization.

In this paper, class imbalance is handled by creating splits of equal ratios of images across both COVID-19 and non-COVID-19 images. Each split is trained using a ResNet model and the outputs of the individual models are ensembled using a feed forward neural network. The effect of increasing contrast within the image by using CLAHE enhancement is analyzed. The features identified by the trained network are visualized using the GradCAM algorithm.

The rest of the paper is organized as follows. Section II gives the summary of some of the best deep neural networks used for medical image classification. The system design of the developed deep neural network is given in Section III, followed by the implementation details in Section IV. Results obtained from the model along with their analysis is presented in Section V. Section VI presents the conclusion of the paper with Section VII summarizing the proposed future works with improvements for the paper.

II. Related Works

The summary of some of the state-of-the-art work done in the field of classification of medical images, ensemble networks and imbalanced datasets is given below with a detailed discussion on its advantages and comparison with our developed model.

Ensemble deep neural network architectures are generally applied for the optimization of the performance of a single constructed network. Mporas and Naronglerdrit [21] evaluated the existing well-known pretrained models identification of COVID-19 in X-Ray images with publicly available datasets. Rafi [22] implemented an ensemble system for the detection of COVID-19 in X-Ray images. This network aimed to combine two state-of-the-art models, ResNet and Deep CNN, that used transfer learning techniques to achieve more accurate results. A SoftMax optimization function was used to further optimize over the trained models with pre-trained weights for faster concurrence.

In our model, the dataset contained highly imbalanced chest CT scans which were normalized using CLAHE algorithms for enhancing their features. These images were resampled and fed to a cluster of ResNet models for creating an ensemble architecture for improving the efficiency of the system and to balance the dataset by eliminating the high bias between the classes.

Liu et al [23] developed an ensemble architecture for highly imbalanced data using under sampling methods giving high performance with highly overlapped and skewed distributions. The use of classification hardness distribution concept helped to overcome the difficulties even with the presence of noise and missing values in the chosen dataset. Cahyana et al [24] expanded their minority class by oversampling to fix the size of the majority class for increasing the number of its instances for training the model for classification.

In our model, the dataset contained imbalance distribution of images in the binary classes and this was studied to improve the performance of our model by combining the attribute weights of each model trained with proportional images distributed in the classes and combined using an ensemble method. This has helped to avoid the oversampling of the minor class and ensure the high performance of the model for classification.

Deep neural architectures that implement supervised learning algorithms suffer from high prediction bias due to an imbalanced dataset resulting in poor performance and low computation efficiency. Aggarwal et al [25] used Active Learning (AL) algorithm to increase the effectiveness of labelling using an acquisition function. This function trains on the labelled dataset before picking more from the unlabeled dataset until the entire budget is spent.

Our model handles the imbalanced dataset by splitting the class with the higher images into multiple sets and training each set of the images with the images from the minor class. The performance of each deep neural model with a set of the images paired with the minor class is then saved with its learnt feature weights after training. The activations from each of the models is then used to build an ensemble network with high performance over the individual models.

Wang et al [26] performed the detection of COVID-19 in X-ray images dataset with data enhancement to 17 times the size for classification into COVID-19 positive, normal and viral pneumonia labels. They used transfer learning to fusion multiple trained models for dynamically improving their weight ratio during the training process.

In our paper, the enhancement of the dataset images were performed using CLAHE histogram for the ensemble ResNet-ANN model to extract the features for the classification of COVID-19 chest X-Ray images into COVID-19 positive and negative. This method would be applicable to study the imbalanced dataset with no oversampling to improve the quality of the training classification model. The learning of the model was visualized by using the Grad-CAM algorithm for observation.

In summary, an imbalanced COVID-19 dataset with chest CT-scans was chosen for binary classification. To improve the performance of the model, an ensemble architecture with multiple ResNets was developed with the major label split into sub-datasets for training with the minor dataset individually and the trained weights were fed into various meta classifiers for categorization. The visualizations of the trained attributes was done using the Grad-CAM algorithm and the analysis presented.

III. System Design

This section gives the details of the ensemble ResNet system for the binary classification of the COVID19 dataset with imbalanced classes.

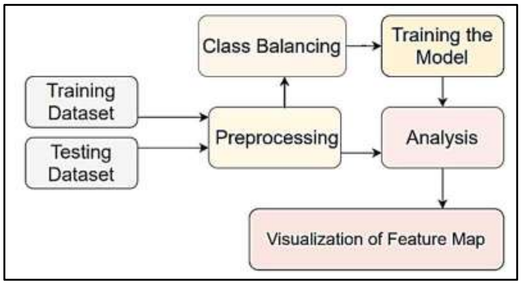


Fig. 1. Overall Architecture Diagram

Figure 1 describes the overall design which is adopted to handle class imbalance in the number of scans and to detect COVID-19 without bias towards any one class. The training and testing sets are preprocessed by using Contrast Limited Adaptive Histogram Equalization (CLAHE) to equalize the images. The training set is then balanced by splitting the set into equal parts and an ensemble of classifiers is trained. The testing set is analyzed against the trained model and the features identified by the model are visualized using GradCAM algorithm.

A. *Dataset Description*

The dataset contains chest CT-scans scraped from the Web [27] with the distribution given in Table I. Class 0 refers to the images that are obtained from those who have not tested positive for COVID-19 and Class-1 images are collected from those tested COVID-19 positive. The dataset has a high imbalance of positive images.

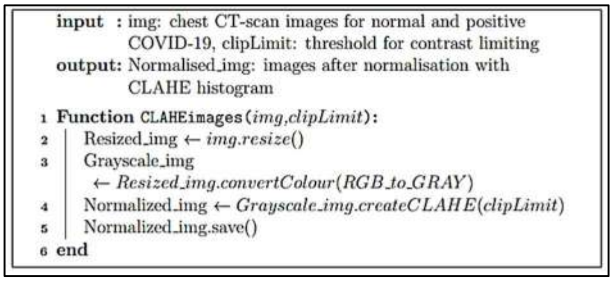
TABLE I. Images Per Class Distribution

Severity	Class 0	Class 1
Training set	566	5810
Testing set	50	100

B. *Preprocessing*

The CT Scan images are preprocessed by applying the CLAHE filter [28] which is a variant of Adaptive Histogram Equalization (AHE) [29]. AHE works on the principle of contrast amplification [30] and has a disadvantage of over amplifying noise when there are near-constant regions. This is overcome by using CLAHE which clips the histogram at a clip limit thereby limiting amplification to a value between 3 and 4. CLAHE performs histogram equalization locally, pixel by pixel. Algorithm 1 is used for applying CLAHE normalization on the collected CT-scans and to improve the contrast of the images.

Algorithm 1: CLAHE Normalization for pre-processing the images

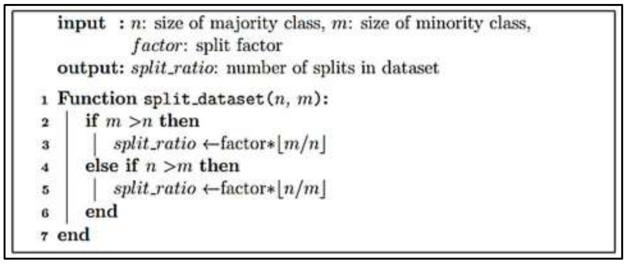


```
input  : img: chest CT-scan images for normal and positive
         COVID-19, clipLimit: threshold for contrast limiting
output: Normalised_img: images after normalisation with
         CLAHE histogram
Function CLAHEimages(img,clipLimit):
    Resized_img ← img.resize()
    Grayscale_img
      ← Resized_img.convertColour(RGB_to_GRAY)
    Normalized_img ← Grayscale_img.createCLAHE(clipLimit)
    Normalized_img.save()
end
```

C. *Class Balancing*

Imbalance in dataset [31] classes can result in the model overfitting the data of the major class and not being able to predict the minor class effectively. To resolve this, resampling [32] of the dataset using replication of minority classes is performed by calculating the ratio with which the dataset should be split as seen in Algorithm 2. This ensures the majority class is balanced with the minority class as in each split, all the samples of the minority class are present and the same number of differing samples of majority class are present. The split factor determines the balance of major and minor classes.

Algorithm 2: Determining Dataset Split Ratio



```
input  : n: size of majority class, m: size of minority class,
         factor: split factor
output: split_ratio: number of splits in dataset
Function split_dataset(n, m):
    if m > n then
        split_ratio ←factor∗$\lfloor m/n \rfloor$
    else if n > m then
        split_ratio ←factor∗$\lfloor n/m \rfloor$
    end
end
```

D. *Modelling Each Data Split*

Once the dataset is split into divisions with equal ratios of major and minor classes, each division is modelled using the ResNet model. ResNets [33] are a deep neural model that solve the vanishing gradient problem by using “Residual blocks” which are seen in Figure 2.

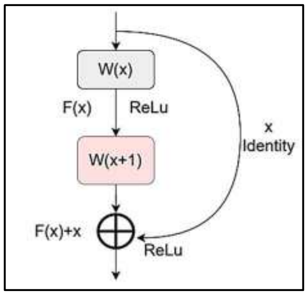


Fig. 2. Structure of ResNet Block

The core aspect of residual blocks are “skip connections” [34]. Using skip connections the output $H(x)$ can be defined as,

$$H(x) = F(x) + x \tag{1}$$

ResNets resolve the vanishing gradient problem by creating many small networks that are ensembled together to create a deeper network. The output of a previous layer is added to a deeper layer and this process is repeated throughout the network ensuring that there is a continuous propagation of values.

E. Ensemble ResNet Architecture

Once ResNets are built for each dataset split, an ensemble network is created. Ensembling [35] involves modelling the predictions of individually trained models in order to improve the performance of detection. Each individual ResNet is trained on a particular dataset split and then the entire training set is run against the trained ResNets as seen in Figure 3. The output predictions of the ResNet models are then trained using a Level-1 classifier achieving a two-step optimization in learning. The Level-1 classifier is chosen as an ANN based on the overall performance of the classification model.

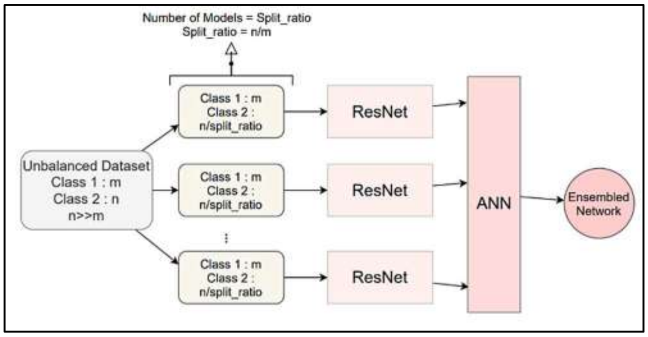


Fig. 3. Ensemble Network to Resolve Class Imbalance

IV. System Implementation

The neural models are implemented using Google Colab with Intel(R) Xeon(R) CPU @ 2.20GHz Processor, 13 GB RAM and 12GB NVIDIA Tesla K80 graphics processor.

A. Dataset Preprocessing

The images from the dataset contain chest CT-scans from both patients tested positive for COVID-19 and those who tested negative. These images undergo normalization using Contrast Limited Adaptive Histogram Equalization (CLAHE) [36] normalization for over-amplification of the contrast. This procedure is applied on the luminance channel of the images and the results are after equalizing only the luminance channel [37] of an Hue Saturation Value (HSV) image. The clip limit [38] which acts as the threshold for contrast limiting is taken as 5. Figure 4 a) and b) show how CLAHE histogram equalization increases the contrast within the image and shows how the features within the image get highlighted with better equalization. Once CLAHE histogram equalization is applied, the CT scan images are augmented using Keras's Image Data Generator [39]. Augmentation parameters used are specified in Table II.

B. Balancing the Dataset

The processed data set is then resampled to avoid class imbalance. In this dataset collected, positive COVID-19 images are the major class and negative COVID-19 images are the minor class. The split factor of the dataset is taken as 2 i.e. the number of positive COVID-19 images in each dataset split is twice the number of negative COVID-19 images in that split.

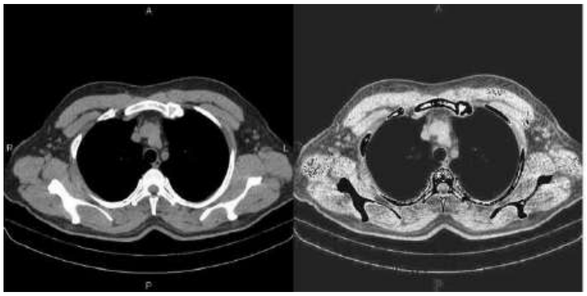


Fig. 4. CT Scan Image of COVID-19 Positive Patient a) Original Image b) with CLAHE Histogram Equalization

TABLE II. Image Data Generator Parameters

Parameter	Value
Rotation Range	40 degrees
Width_shift_range	0.2
Height_shift_range	0.2
Rescale	1/255
Shear_range	0.2
Zoom_range	0.2
Horizontal_flip	True

TABLE III. Training Parameters of Individual ResNet Models

Parameter	Value
ResNet Version	1
Number of Layers	20
Loss	Categorical Cross Entropy
Optimizer	Adam
Learning Rate	0.01
Batch Size	64
Epochs	10

The number of dataset splits is calculated to be 5 according to Algorithm 2. Each of the 5 dataset splits has 566 non COVID-19 images and 1162 positive COVID-19 images.

C. ResNet Model

The ResNet model was built using Keras and TensorFlow [40]. The model has an initial convolutional layer with 16 filters of 32 x 32 dimensions applied. The outputs of the initial convolutional layer are fed to a batch normalization layer followed by an activation layer. Next, 64 filters of 32 x 32 dimensions are applied followed by layers of batch normalization, activation and convolution. This process is repeated by reducing the dimensions of the filters by half and doubling the number of the filters until the number of filters reaches 256 and the dimensions are 8 x 8.

At each stage of the ResNet there are residual blocks that concatenate the output of the previous block with the output of the current block using a concatenation layer. The activation function used in these hidden layers is ReLu [41]. The final stage of the ResNet involves average pooling and flattening the vectors to form a 1D tensor. The output layer has 2 nodes which are activated using SoftMax activation [41]. The ResNet models are trained using the parameters described in Table III.

D. Ensemble Neural Network

Once the ResNet models are trained, each image of the training set is run through each of the 5 models and the output activations are noted and then fed into a Level-1 meta classifier. The Level-1 classifier learns the features from the already optimized activations and tries to determine if the CT scan images have COVID-19 or not.

Different machine learning algorithms are implemented as a part of determining the best meta classifier. It is observed from Figure 5 that the Artificial Neural Network (ANN) performed as the best Level-1 classifier with the accuracy of the predicted images at 87.13%. This can be attributed to the size of the dataset being smaller along with the presence of imbalanced data classes enabling the deep neural model to execute its own feature engineering to search for features and converge faster during the training of the model.

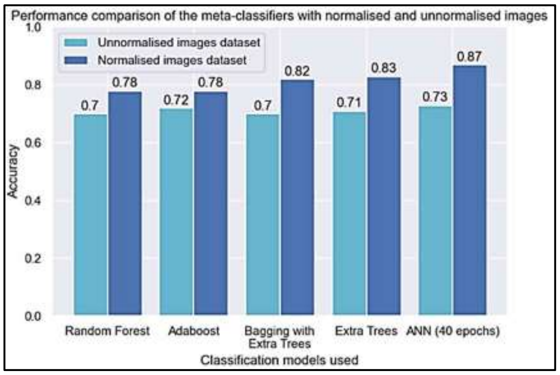


Fig. 5. Performance Comparison of Meta-Classifiers with Normalized and Unnormalized Images

The Level-1 ANN model has 10 input nodes corresponding to the output probabilities of 5 level-0 models. It is followed by 7 hidden layers with 8,16,32,64,64,32,16 and 8 nodes each activated by ReLu activation function. The output layer has 1 node activated by Sigmoid [41] activation function. The Level-1 ANN classifier is trained using the parameters given in Table IV.

TABLE IV. TRAINING PARAMETERS OF LEVEL-1 ANN CLASSIFIER

Parameter	Value
Loss	Binary Cross Entropy
Optimizer	Adam
Learning Rate	0.01
Batch Size	32
Epochs	50

V. RESULTS AND ANALYSIS

This section presented the results obtained from the classification of COVID-19 CT-scan images using the ensembled ResNet model developed.

A. Training plot of the ensemble ResNet for the classification of COVID-19 CT scans with and without CLAHE histogram

The training plot of the ensemble ResNet-ANN model using the ANN as meta-classifier is given in Figure 6 along with the variation of accuracy and loss. This helps to identify the learnings of the model to stop the training before overfitting the data but with enough time to avoid underfitting the data.

The accuracy is observed to have reached a peak at about 92.34% with the loss at 17.49%.

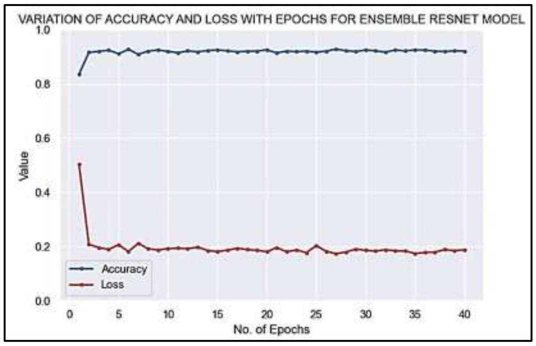


Fig. 6. Variation of Accuracy and Loss with Epochs for Ensemble ResNet Model

B. Confusion matrix of the ensemble ResNet model trained with images normalised with CLAHE histogram

The confusion matrix of the ensemble ResNet-ANN network in Figure 7 is observed to understand the performance of the network for the binary classification of images. The identification of the classes by the developed architecture proves the ability of the model to handle imbalanced datasets.

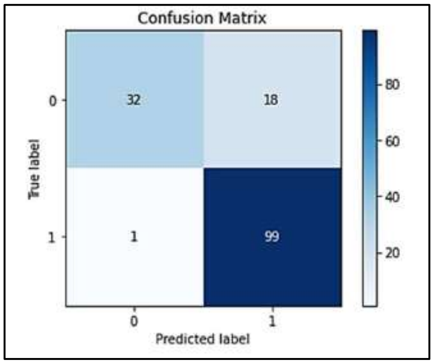


Fig. 7. Confusion Matrix for the Ensembled ResNet Performance Model

C. Performance comparison of the ensemble deep neural model for normalised and unnormalised images

The performance of the trained ensemble model for binary classification is analyzed with accuracy, precision recall and F1-score values.

Accuracy [42] gives the ratio of the correctly classified instances to the total number of instances.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \qquad (2)$$

Recall [42] value gives the proportion of actual positive instances to the total number of predicted positive instances.

$$Recall = TP/TP + FN \tag{3}$$

Precision [42] gives the proportion of predicted positive values that are actually positive and its variance with no. of dense layers.

$$Precision = TP/TP + FP \tag{4}$$

F1 score [42] combines the precision and recall value of the trained deep neural model using a harmonic mean.

$$F1\,Score = 2 * Precion * Recall/(Precision + Recall) \tag{5}$$

F1 Score is a better measure to seek a balance between Precision and Recall. This is highly useful in cases such as here where there is an uneven class distribution (large number of Actual Negatives).

TABLE V. PERFORMANCE OF THE ENSEMBLED RESNET MODEL

Trained over CLAHE normalized images				
Class	**Precision**	**Recall**	**F1-Score**	**Accuracy**
0	0.97	0.64	0.77	0.87
1	0.85	0.99	0.91	
Avg.	0.91	0.81	0.84	
Trained over Unnormalized images				
Class	**Precision**	**Recall**	**F1-Score**	**Accuracy**
0	0.99	0.23	0.33	0.73
1	0.71	0.99	0.83	
Avg.	0.86	0.73	0.67	

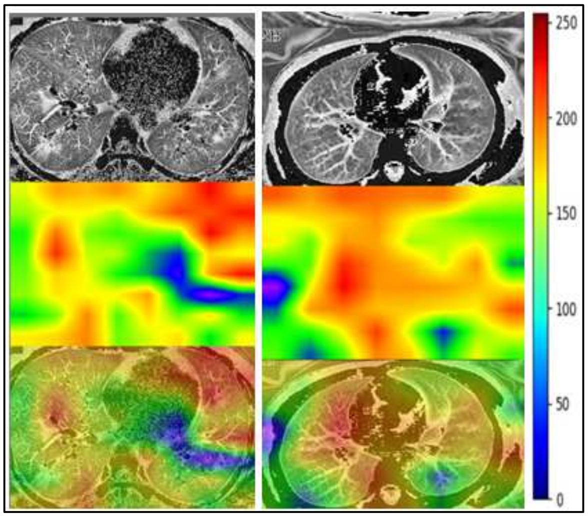


Fig. 8. GradCAM Visualization of CT Scan Images of Patient Tested Positive for COVID-19 with the Prediction Score at a) 0.51 and b) 0.79

From Table V, it is observed that the images that the model performed better when the input images were normalized with the CLAHE than the unnormalized images. The identification of the Class 0 images was performed inadequately when the model was trained with unnormalized images. An accuracy of 0.87 was obtained when the ensemble model was trained with images pre-processed with CLAHE histogram.

D. *Visualisation if the activations of the trained Ensemble ResNet model over normalised images using GradCAM*

GradCAM visualisation of the trained weights of the ensembled ResNet model on the CT scans dataset for Class-1 images is done here. This helps to provide a visual explanation for the trained deep neural model on the input dataset. From Figure 8 a) and b), the GradCAM visualization of the image is observed to possibly identify the areas of clinical features such as Ground Glass Opacities, air space opacification in the CT scan images for their classification as belonging to Class-1 or COVID-19 positive.

VI. CONCLUSION

An ensemble method is introduced to deal with class imbalance while dealing with identification of positive COVID-19 CT scans. The normalization of the images is achieved by applying CLAHE histogram over the images for enhancing their features before splitting the dataset into equally balanced parts and feeding the images to the Level-0 ResNet models for training. The output containing the activations with feature weights learned by these multiple ResNet models are used to create an ensemble network architecture for enhancing their performance and improving the existing bias in the binary classes of the images. The model performed with an accuracy of 87.23% with CLAHE enhanced images and the learning of the trained models are visually represented using GradCAM algorithm.

VII. FUTURE WORKS

The application of augmentation in the deep neural networks for the localization of the symptoms found in the visualization on CT scan images can be explored further for the diagnosis any disease. This would enhance the results presented for proper diagnosis and treatment of the patients for their recovery at the earliest.

ACKNOWLEDGEMENT

We thank Dr. Manjula Jayakumar, M.B.B.S, D.O., D.N.B, F.I.C.O. (UK), Chief Pediatric Ophthalmologist, Dr. Agarwal’s Eye Hospital and Research Centre, Chennai for her invaluable insights for the completion of the project.

REFERENCES

[1] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, “Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts,” IEEE Access, vol. 8, pp. 130820–130839, 2020.

[2] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, “COVID-19 infection: origin, transmission, and characteristics of human coronaviruses,” Journal of Advanced Research, vol. 24, pp. 91–98, 2020.

[3] P. Simoni, A. Bazzocchi, G. Boitsios, A. De Leucio, M. Preziosi, and M. P. Aparisi Gómez, “Chest computed tomography (CT) features in children with reverse transcription‐polymerase chain reaction (RT‐PCR)‐confirmed COVID‐19: A systematic review,” Journal of Medical Imaging and Radiation Oncology, vol. 64, no. 5, pp. 649–659, 2020.

[4] “The Radiology Assistant : COVID-19 Imaging findings,” radiologyassistant.nl, 2020. https://radiologyassistant.nl/chest/covid-19/covid19-imaging-findings (accessed Jan. 25, 2021).

[5] R. M. M. Ali and M. B. I. Ghonimy, “Semi-quantitative CT imaging in improving visualization of faint ground glass opacities seen in early/mild

coronavirus (covid-19) cases,” Egyptian Journal of Radiology and Nuclear Medicine, vol. 51, no. 1, 2020.

[6] R. Yasin and W. Gouda, “Chest X-ray findings monitoring COVID-19 disease course and severity,” Egyptian Journal of Radiology and Nuclear Medicine, vol. 51, no. 1, 2020.

[7] O. M. Sultan et al., “Pulmonary ct manifestations of COVID-19: changes within 2 weeks duration from presentation,” Egyptian Journal of Radiology and Nuclear Medicine, vol. 51, no. 1, 2020.

[8] G. Nino, J. Zember, R. Sanchez-Jacob, M. J. Gutierrez, K. Sharma, and M. G. Linguraru, “Pediatric Lung Imaging Features of Covid-19: A Systematic Review and Meta-Analysis,” Pediatric Pulmonology, vol. 56, pp. 252–263, 2020.

[9] Y. Zheng, L. Wang, and S. Ben, “Meta‐analysis of chest CT features of patients with COVID‐19 pneumonia,” Journal of Medical Virology, vol. 93, pp. 241‐249, 2020.

[10] Z. Lim et al., “Variable computed tomography appearances of COVID-19,” Singapore Medical Journal, vol. 61, no. 7, pp. 387–391, 2020.

[11] Worldometer, “Coronavirus Toll Update: Cases & Deaths by Country of Wuhan, China Virus - Worldometer,” Worldometers.info. https://www.worldometers.info/coronavirus/ (accessed Jan. 25, 2021).

[12] X. Cai, Y. Wang, X. Sun, W. Liu, Y. Tang, and W. Li, “Comparing the performance of ResNets on COVID-19 diagnosis using CT scans,” in 2020 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–4, 2020.

[13] D. Haritha, N. Swaroop, and M. Mounika, “Prediction of COVID-19 Cases Using CNN with X-rays,” in 2020 5th International Conference on Computing, Communication and Security (ICCCS), pp. 1–6, 2020.

[14] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, “An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification,” IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 31–40, 2017.

[15] S. Sridhar and S. Sanagavarapu, “Detection and Prognosis Evaluation of Diabetic Retinopathy using Ensemble Deep Convolutional Neural Networks,” in 2020 International Electronics Symposium (IES), pp. 78–85, 2020.

[16] Y. S. Moon, B. G. Han, H. S. Yang, and H. G. Lee, “Low Contrast Image Enhancement Using Convolutional Neural Network with Simple Reflection Model,” Advances in Science, Technology and Engineering Systems Journal, vol. 4, no. 1, pp. 159–164, 2019.

[17] G. Siracusano, A. La Corte, M. Gaeta, G. Cicero, M. Chiappini, and G. Finocchio, “Pipeline for Advanced Contrast Enhancement (PACE) of Chest X-ray in Evaluating COVID-19 Patients by Combining Bidimensional Empirical Mode Decomposition and Contrast Limited Adaptive Histogram Equalization (CLAHE),” Sustainability, vol. 12, no. 20, 2020.

[18] M. Lucas, M. Lerma, J. Furst, and D. Raicu, “Heatmap Template Generation for COVID-19 Biomarker Detection in Chest X-rays,” in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 438–445, 2020.

[19] T. Tagaris, M. Sdraka, and A. Stafylopatis, “High-Resolution Class Activation Mapping,” in 2019 IEEE International Conference on Image Processing (ICIP), 4514–4518, 2019.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

[21] I. Mporas and P. Naronglerdrit, “COVID-19 Identification from Chest X-Rays,” in 2020 International Conference on Biomedical Innovations and Applications (BIA), pp. 69–72, 2020.

[22] T. H. Rafi, “An ensemble deep transfer-learning approach to identify COVID-19 cases from chest X-ray images,” in 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–5, 2020.

[23] Z. Liu et al., “Self-paced Ensemble for Highly Imbalanced Massive Data Classification,” in 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 841–852, 2020.

[24] N. Cahyana, S. Khomsah, and A. S. Aribowo, “Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting,” in 2019 5th International Conference on Science in Information Technology (ICSITech), pp. 217–222, 2019.

[25] U. Aggarwal, A. Popescu, and C. Hudelot, “Active Learning for Imbalanced Datasets,” in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1417–1426, 2020.

[26] N. Wang, H. Liu, and C. Xu, “Deep Learning for The Detection of COVID-19 Using Transfer Learning and Model Integration,” in 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 281–284, 2020.

[27] “Radiopaedia.org, the wiki-based collaborative Radiology resource,” Radiopaedia.org, 2019. https://radiopaedia.org/ (accessed Jan. 29, 2021).

[28] E. D. Carvalho, E. D. Carvalho, A. O. de Carvalho Filho, A. D. de Sousa, and R. de Andrade Lira Rabulo, “COVID-19 diagnosis in CT images using CNN to extract features and multiple classifiers,” in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 425–431, 2020.

[29] A. Kaur and C. Singh, “Contrast enhancement for cephalometric images using wavelet-based modified adaptive histogram equalization,” Applied Soft Computing, vol. 51, pp. 180–191, 2017.

[30] D. Sonker, “Comparison of Histogram Equalization Techniques for Image Enhancement of Grayscale images in Natural and Unnatural light,” International Journal of Engineering Research and Development, vol. 8, no. 9, pp. 57–61, 2013.

[31] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, “A survey on addressing high-class imbalance in big data,” Journal of Big Data, vol. 5, no. 1, 2018.

[32] A. Puri and M. K. Gupta, “Comparative Analysis of Resampling Techniques under Noisy Imbalanced Datasets,” in 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 1–5, 2019.

[33] D. Hernandez, R. Pereira, and P. Georgevia, “COVID-19 detection through X-Ray chest images,” in 2020 International Conference Automatics and Informatics (ICAI), pp. 1–5, 2020.

[34] Y. Furusho and K. Ikeda, “Theoretical analysis of skip connections and batch normalization from generalization and optimization perspectives,” APSIPA Transactions on Signal and Information Processing, vol. 9, no. e9, pp. 1–7, 2020.

[35] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, “Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-rays,” IEEE Access, vol. 8, pp. 115041–115050, 2020.

[36] R. M. James and A. Sunyoto, “Detection Of CT - Scan Lungs COVID-19 Image Using Convolutional Neural Network And CLAHE,” in 2020 3rd International Conference on Information and Communications Technology (ICOIACT), pp. 302–307, 2020.

[37] A. Ajmal, C. Hollitt, M. Frean, and H. Al-Sahaf, “A Comparison of RGB and HSV Colour Spaces for Visual Attention Models,” in 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6, 2018.

[38] P. Patil and H. Patil, “X-ray Imaging Based Pneumonia Classification using Deep Learning and Adaptive Clip Limit based,” in 2020 IEEE 4th Conference on Information & Communication Technology (CICT), pp. 1–4, 2020.

[39] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.

[40] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition,. O’Reilly Media, Inc., 2019.

[41] M. A. Mercioni and S. Holban, “The Most Used Activation Functions: Classic Versus Current,” in 2020 International Conference on Development and Application Systems (DAS), pp. 141–145, 2020.

[42] T. B. Alakus and I. Turkoglu, “Comparison of deep learning approaches to predict COVID-19 infection,” Chaos, Solitons & Fractals, vol. 140, 2020.

Dynamic analysis of Quadcopter with a novel controller

Manas Kumar Sahoo
PhD, Dept. of Mechanical Engg.
(IIT Delhi)
New Delhi, India

Dr. J. K. Dutt
Professor, Dept. of Mechanical Engg.
(IIT Delhi)
New Delhi, India

Dr. S. K. Saha
Professor, Dept. of Mechanical Engg.
(IIT Delhi)
New Delhi, India

***Abstract*— Quadcopter, also known as quadrotor, is an unmanned Ariel vehicle (UAV). This paper presents the dynamic analysis of a quadcopter for propulsion and flight control with system noise. The quadcopter is maneuvered by adjusting the angular velocities of the rotors which are controlled by the DC motors. In this paper, a novel controller is proposed that uses the properties of four-element viscoelastic material. To make the system robust some random noise is generated which is absorbed by the controller. The performance of the novel controller is compared with the conventional PID controller by making both the controller optimum. The robustness of the novel controller is compared with the PID controller by using the force-displacement hysteresis curve.**

***Keywords*— unmanned aerial vehicles; UAVs; Quadcopters; PID controller; vertical takeoff and landing; VTOL**

I. INTRODUCTION

There are many types of multicopters which use different platforms such as Tricopter, Quadcopter, Hexacopter etc. Quadcopter has received significant attention of researchers as it gives rise to several areas of interest, which are very complex in nature. The Quadcopter has evolved into more complex system [1] [2]. The design requirements for various usage of UAVs have given various problems in the field of stability. Hence researchers have developed various control techniques to achieve stable system. There are different control systems applied for robust control of a Quadcopter, including PID controllers [3] [4] [5], LQR controllers [6], the backstepping control [7] and other nonlinear controllers [8]. The Quadcopters are used for various applications such as civilian applications, including remote sensing, aerial imaging, firefighting, environmental measurement, disaster relief and emergency management, situational awareness, infrastructure surveys, and several other military and commercial applications. While in operation the Quadcopters experience various forces, which can destabilize it. One of the main problems with the linear control system is adaptability with random forcing functions that come into action when a Quadcopter is flying. Most of the linear controllers fail when these forcing functions dominates. In this paper we tried to develop a novel control technique that will take care of the above problem. The controller is based on the response of a four-element (FE) viscoelastic material.

This is achieved by developing a mathematical model for a Quadcopter and then the controller is optimized and applied to the system. The results are compared with a conventional PID system.

Mathematical Model of Quadcopter:

THE FIGURE BELOW SHOWS THE BODY FRAME OF A QUADCOPTER MODEL.

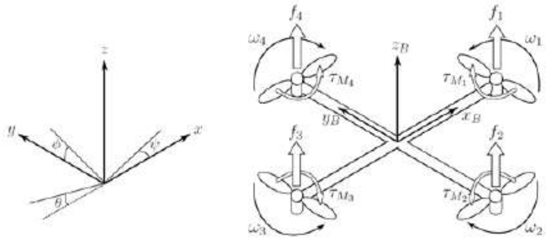


Figure 1Inertia and Body frame of a Quadcopter

In inertial frame the absolute position of the Quadcopter is defined as ζ. Similarly, the angular positions in inertial frame is defined as η. Pitch angle θ determines the rotation of the quadcopter around the y-axis. Roll angle Φ determines the rotation around the x-axis and yaw angle Ψ around the z-axis. The variable q contains both the linear and angular positions.

$$\zeta = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\eta = \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix}$$

$$q = \begin{bmatrix} \zeta \\ \eta \end{bmatrix}$$

The Centre of mass of the Quadcopter is considered as the origin of the body frame. Vb and are considered as the linear velocities and the angular velocities respectively in body frames.

$$V_b = \begin{bmatrix} v_{xb} \\ v_{yb} \\ v_{zb} \end{bmatrix}$$

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

$$w_b{}^b = \begin{bmatrix} p \\ q \\ r \end{bmatrix}$$

The rotation matrix is given by Q as follows:

$$Q = \begin{bmatrix} C\psi C\theta & C\psi S\theta S\phi - S\psi C\phi & C\psi S\theta C\phi + S\psi S\phi \\ S\psi C\theta & S\psi S\theta S\phi + C\psi C\phi & S\psi S\theta C\phi - C\psi S\phi \\ -S\theta & C\theta S\phi & C\theta C\phi \end{bmatrix}$$

Where Sx=Sin(x) and Cx=Cos(x). Here Q is an Orthogonal matrix, Hence $Q^{-1} = Q^T$.

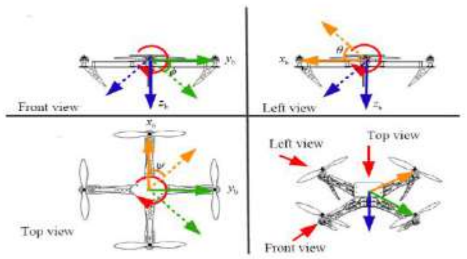


Figure 2Representation of Euler angles

Suppose W is the transformation matrix for angular velocities from Inertial frame to the body frame. Then W is given by:

$$W = \begin{bmatrix} 1 & 0 & -S\theta \\ 0 & C\phi & C\theta S\phi \\ 0 & -S\phi & C\theta S\phi \end{bmatrix}$$

$$w_b{}^b = \begin{bmatrix} p \\ q \\ r \end{bmatrix} = W \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix}$$

Here the Quadcopter is assumed to be symmetric about X and Y axes. Hence the inertia matrix is a diagonal matrix, which is given by:

$$I = \begin{bmatrix} I_{xx} & 0 & 0 \\ 0 & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{bmatrix}$$

The angular velocity of the rotor n, denoted as ω_n gives force F_n which is given by:

$$F_n = k\omega_n{}^2$$

Similarly, the torque due to the above rotor n is given by:

$$\tau_n = b\omega_n{}^2$$

Where k and b are lift and drag coefficient respectively. The combined force generated is in Z direction and is given by:

$$\text{F=} \sum_{n=1}^{4} F_n$$

Now the torque is given by:

$$\tau = \begin{bmatrix} \tau_\phi \\ \tau_\theta \\ \tau_\psi \end{bmatrix} = \begin{bmatrix} lk(-\omega_2{}^2 + \omega_4{}^2) \\ lk(-\omega_1{}^2 + \omega_3{}^2) \\ b\sum_{n=1}^{4} \omega_n{}^2 \end{bmatrix}$$

Where l is the distance between center of mass and the rotor.

Newton-Euler equations:

In the body frame the body is moving with an acceleration $\dot{V}_b{}^b$ for which the force required is m$\dot{V}_b{}^b$. The centrifugal force is given by ($w_b{}^b \times (mV_b{}^b)$). Now these two forces will be equal to the gravity force and thrust force created due to the rotors. Hence the equation is given as follows:

$$m\dot{V}_b{}^b + w_b{}^b \times (mV_b{}^b) = Q^T G + F^b \text{ ....(I)}$$

The centrifugal force is nullified in the inertial frame hence the above equation is reduced to:

$$m\ddot{\zeta} = mG + QF^b$$

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = -g \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \frac{F}{m} \begin{bmatrix} C\psi S\theta C\phi + S\psi S\phi \\ S\psi S\theta C\phi - C\psi S\phi \\ C\theta C\phi \end{bmatrix} \text{ ....(II)}$$

In the body fixed frame, the angular acceleration of the inertia is $I\dot{\omega}_b{}^b$ and the centripetal force is $\omega_b{}^b \times (I\omega_b{}^b)$. Now these two forces are equal to external torque τ . Hence the equation is given by:

$$I\dot{\omega}_b{}^b + \omega_b{}^b \times (I_b\omega_b{}^b) = \tau \text{ ....(III)}$$

$$\begin{bmatrix} \dot{p} \\ \dot{q} \\ \dot{r} \end{bmatrix} = I_b{}^{-1} \begin{bmatrix} \tau_\phi \\ \tau_\theta \\ \tau_\psi \end{bmatrix} - I_b{}^{-1}(\omega_b{}^b \times (I_b\omega_b{}^b)) \text{ ....(IV)}$$

The various parameters used for the quadcopter are given below:

Parameters	Value	Unit
g	9.81	m/s2
m	2	Kg
l	0.2	M
k	1.32E-5	
Ix	1.25E-2	Kgm2
Iy	1.25E-2	Kgm2
Iz	2.5E-2	Kgm2
b	5.17E-7	

Table 1parameters of the Quadcopter

1. **Control Design:** A PID controller for the above Quadcopter ca=n be designed as follows:

Suppose we have four inputs (i.e. three Euler angles and altitude), then the error dynamics equation is given by:

$$\ddot{e}+K_I\int e+K_d\dot{e}+K_pe=0 \text{ ....(V)}$$

Where, e=desired state-real state and K_I , K_d & K_p are gains.

Now the equation for the controller is:

$$\ddot{\zeta}_c=\ddot{\zeta}_{des}+K_I\int e+K_d\dot{e}+K_pe \text{ ....(VI)}$$

This gives the thrust force as:

$$F=m(g+\ddot{\zeta}_c)/(\cos(\theta)\cos(\phi)) \text{ ....(VII)}$$

1.1. Novel Control Law:

The proposed novel control law is based on the behavior of a viscoelastic material. The relationship between the force and displacement of a viscoelastic material can be represented by multielement models. A model consisting of four elements (two spring and two dampers) is shown below. This type of materials finds there use in vibration reduction [9] [10].

The relation between the force and displacement is derived below:

$$F_m(t)=k_1\delta(t)+c_1\dot{\delta}(t)+k_2(\delta(t)-\vartheta(t))$$

................(VIII)

Where, $\delta(t)$ is the resulting displacement of the four-element (FE) viscoelastic material model. $F_m(t)$ is applied force on the model. $\vartheta(t)$ is displacement in second damper

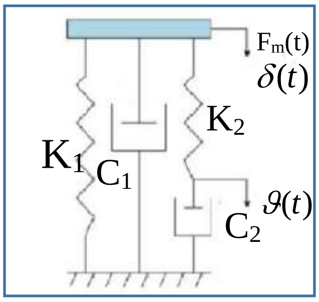


Figure 3Four-element viscoelastic material

as shown in the figure. Now using force balance equation:

$$k_2(\delta(t)-\vartheta(t))=c_2D\vartheta(t) \text{ ....(IX)}$$

Eliminating $\vartheta(t)$ from equation (VIII) we get:

$$F_m(t)=\left(\frac{k_1k_2+(c_2(k_1+k_2)+c_1k_2)D+c_1c_2D^2}{k_2+c_2D}\right)\delta(t) \text{ .(X)}$$

The above equation represents the constitutive equation of a FE viscoelastic model. The novel control law i.e. the four-element control law is based on the above equation. The equation can be rewritten as:

$$F_m(t)=G*\delta(t) \text{ ....(XI)}$$

Where, G is the equivalent transfer function of FE controller. Now the control law for the quadcopter is given as follows:

$$\ddot{\zeta}_c=\ddot{\zeta}_{des}+\left(\frac{k_1k_2+(c_2(k_1+k_2)+c_1k_2)D+c_1c_2D^2}{k_2+c_2D}\right)*e(t)$$

4. RESULT AND DISCUSSION :

The Quadcopter is simulated for the parameters as given in the table. Both the controller is designed as given above. The results are compared for both optimized controllers. The block diagram for the PID/FE controller is given below.

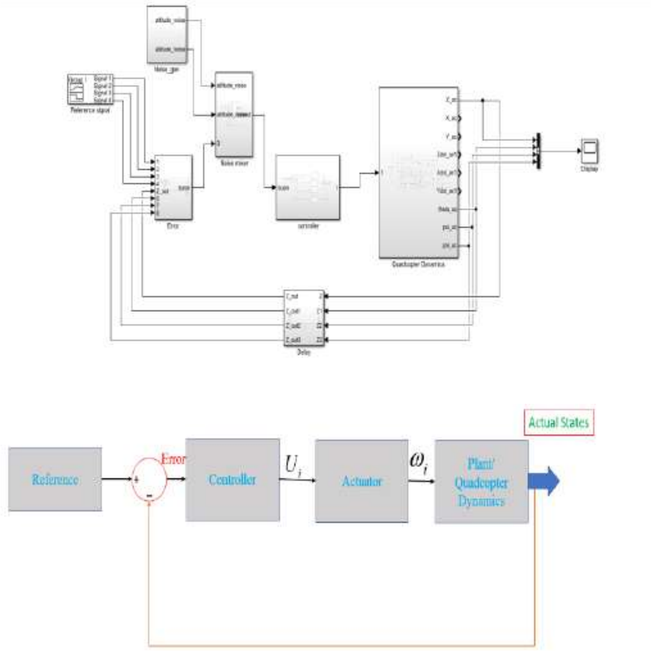


Figure 4Control architecture of Quadcopter FE controller

A basic input signal is given where for the first 1 sec, the altitude increases steadily to 1m then remains stable for 2 sec. After 3 sec it again rises steadily to 2m in 3 sec, then the altitude decreases steadily to 1m in 2 more sec. after this the altitude remain stable until 10 sec. The other inputs are kept constant as shown below.

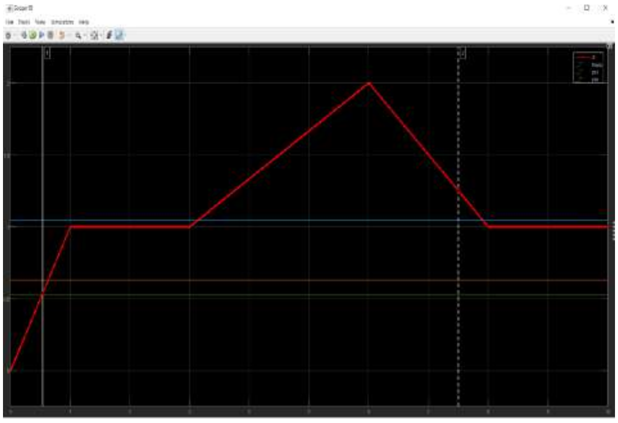


Figure 5Reference signal for both the controller

Both the controllers are optimized by minimizing the mean error and maximum peak. But the two won’t happen simultaneously. Hence, we took the minimum value of the multiplication of the above two.

Altitude variation:

The altitude variation of the quadcopter for the given reference as shown in the figure 5 is given below.

The result for both the controller i.e. PID and FE is shows that in PID controller the error variation is much more compared to the FE controller. The PID controller takes more time to settle and when there is an abrupt change in the input signal the peak overshoot of the PID controller is more. Hence the FE controller is more stable in all situation in the given reference signal. This can be further verified later, when we will see the hysteresis curve for both the controller. Here it is to be noted that for both the controller there is continuous mixing of noise in the input signal, this is why there is a continuous variation in the output state even after settling.

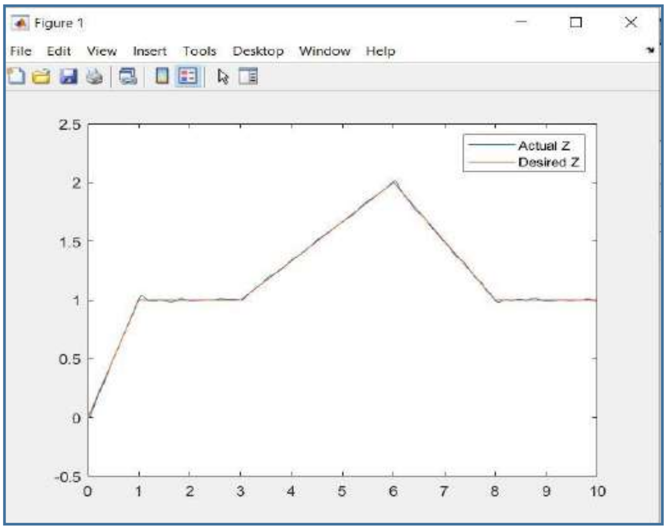


Figure 6PID controller altitude variation

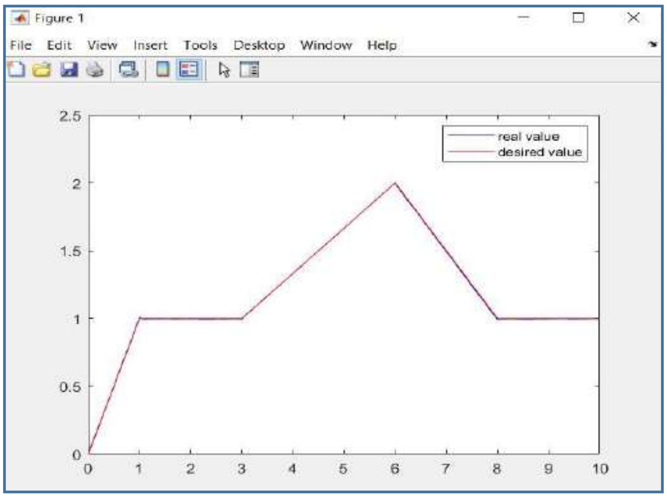


Figure 8FE controller altitude variation

Attitude Variation:

Similarly, the attitude variation is shown for both the controller below. Here as we can see the attitude state is more sensitive to the input noise unlike altitude, Hence the variation in output state is much more compared to the reference signal. But still the FE controller manages to give better result compared to PID controller.

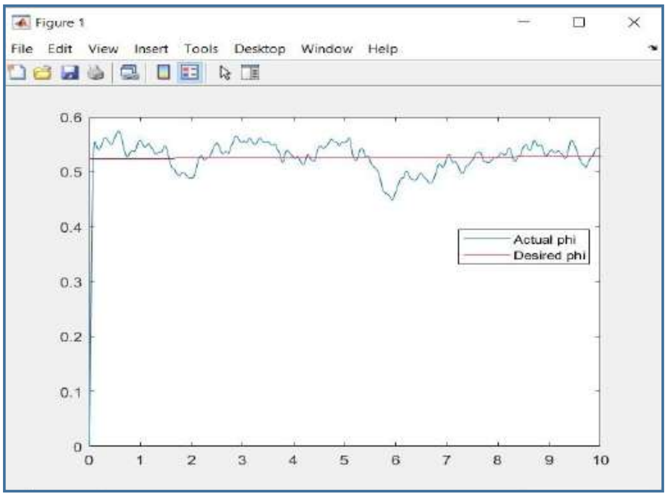


Figure 7Phi variation in PID controller

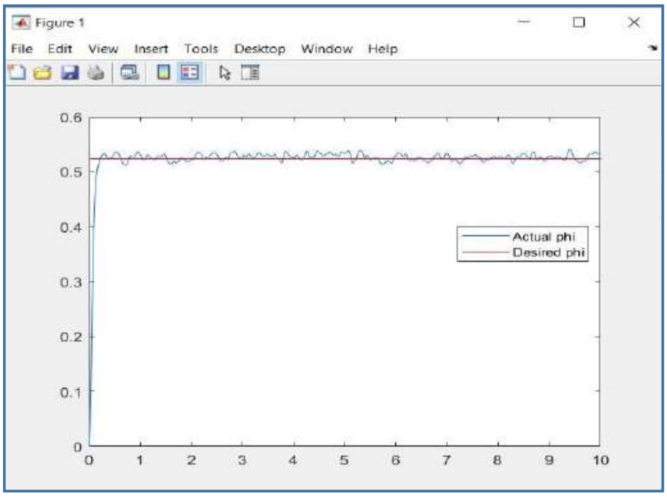


Figure 9Phi Variation in FE controller

It is to be noted that, the above response of both the controller are obtained with continuous noise input (Both for altitude and attitude). The error variation will be consistent due to the continuous noise input which is given below.

Error Variation:

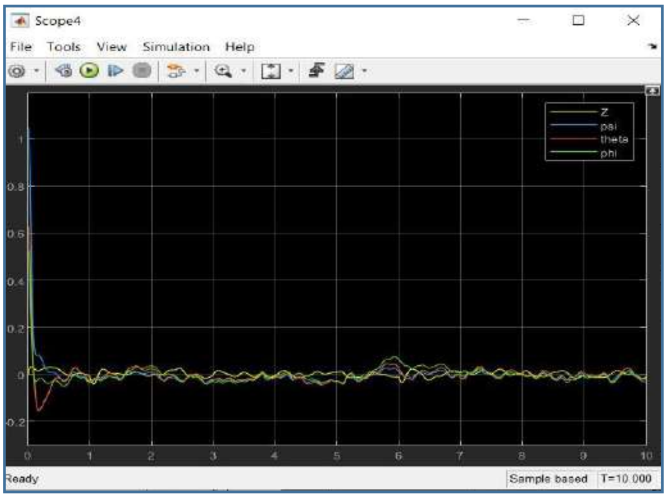


Figure 10Error variation of output states in PID controller

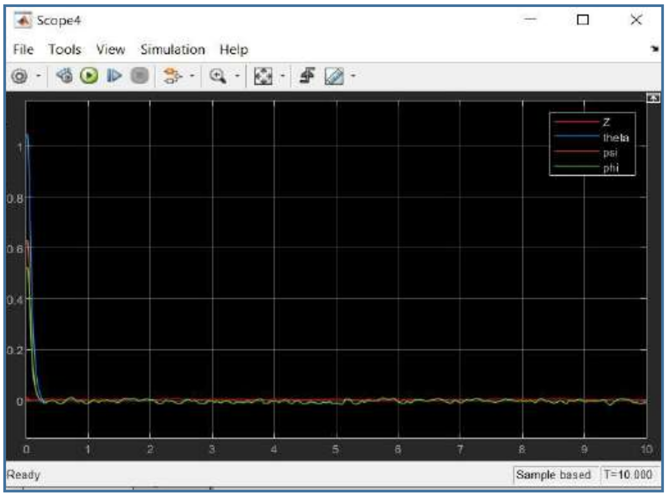


Figure 11Error variation of output states in FE controller

Here, the input force produced in FE controller is way higher than the PID controller. Hence it is saturated to the peak value of PID controller.

Conclusion: In this paper we have presented a dynamic analysis of quadcopter. The quadcopter is controlled by a novel controller which is based on the behavior of four element (FE) viscoelastic material. The numerical simulation was done to show the robustness and accuracy of FE controller. The controller is optimized and compared with PID controllerThe results are found to be satisfactory. The future work is dedicated to applying the controller in more generalist system of Multicopters and the control of manipulator on a drone.

II. References

[1] H. H. S. L. W. a. C. J. T. G. M. Hoffmann, "“Quadrotor helicopter flight dynamics and control: Theory and experiment,”," in *Proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit*, AUG,2007.

[2] H. Huang, G. M. Hoffmann, S. L. Waslander, and C. J. Tomlin, "“Aerodynamics and control of autonomous quadrotor helicopters in aggressive maneuvering,”," in *IEEE International Conference on Robotics and Automation*, may,2009.

[3] A. Tayebi and S. McGilvray, "Attitude stabilization of a four-rotor aerial robot," in *43rd IEEE Conference on Decision and Control, vol. 2, pp. 1216–1221*, 2004.

[4] I. C. Dikmen, A. Arısoy, and H. Temelta¸s,, "“Attitude control of a quadrotor,”," in *4th International Conference on Recent Advances in Space Technologies, pp. 722–727*, 2009.

[5] Z. Zuo, "“Trajectory tracking control design with command-filtered compensation for a quadrotor,”," *IET Control Theory Appl*, vol. 4, pp. pp.2343-2355, 2010.

[6] S. Bouabdallah, A. Noth, and R. Siegwart, "“PID vs LQ control techniques applied to an indoor micro quadrotor,”," in *IEEE/RSJ International Conference on Intelligent Robots and Systems,*, 2004.

[7] T. Madani and A. Benallegue,, "“Backstepping control for a quadrotor helicopter,"," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[8] G. V. Raffo, M. G. Ortega, and F. R. Rubio, "“An integral predictive/nonlinear H1 control structure for a quadrotor helicopter,”," *Automatica,*, vol. 46, pp. 29-39, 2010.

[9] Tukesh Soni,J K DUTT, A S DAS, "Parametric Stability Analysis of Active Magnetic Bearing-Supported Rotor System with a Novel Control Law subject to Periodic Base Motion," *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS*, 2019.

[10] A S Das,JK Dutt, K Ray, "Active Vibration Control of Flexible Rotors on Maneuvering Vehicles," *AIAA,* vol. 48, 2010.

[11] D. Kastelan, M. Konz, J. Rudolph, "Fully Actuated tricopter with Pilot supporting control," in *IFAC papers online*, 2015.

[12] Abhishek, Abhinadan Tripathi, "Six Rotor UAV Helicopter DYnamics and control: Theory and simulation," *International journal of advance research in electrical electronics and instrumentation engineering,* vol. 2, no. 11, 2013.

[13] Kaito Isogai, Ryo Inohara, Hideo Nakano, Hideaki Okazaki, "Modeling and simulation of motion of a Quadcopter," in *International Symposium on Nonlinear Theory and it's Application*, Yugawara,Japan, November 27th-30th, 2016.

Network Routing in SDNs Using Topology Based Multitask Neural Modelling

Sowmya Sanagavarapu
Dept.of Computer Science and Engineering
Anna University, Chennai
Chennai, India
sowmya.ssanagavarapu@gmail.com

Sashank Sridhar
Dept. of Computer Science and Engineering
Anna University, Chennai
Chennai, India
sashank.ssridhar@gmail.com

***Abstract*— Software Defined Networks (SDNs) are adapted for their high programmability and security offered by them for use in enterprise networks. The SDN Controller present in these networks have the function to route the packets in the network to enable communication between the end-point devices connected to the network. In this paper, we created an SDN environment in Mininet using Ryu Controller to collect the communication data to build a topology-based system that could perform dynamic deep neural routing in the network. The neural routing model implemented here is a multitask learning model hosted on top of the SDN Controller that will perform network routing. The trained neural network had given a high-performance accuracy of 99.73% and the proposed system is compared and measured with the existing network routing algorithm in SDNs.**

Keywords*— *Software Defined Networks, SDN routing, Deep neural network routing, Multitask learning

I. Introduction

The widespread use of multi-functional networks requires the optimization of traffic and the distribution in the management of a large number of devices within these networks [1]. The need for intellectualization [2] of the networks to make them inherently secure with mechanisms to defend themselves against attacks and to detect anomalies is required to be performed. Providing activation and customization of networks like the Intent Based Networks (IBNs) [3] along with Network Function Virtualization [4] with high scalability has driven to meet the needs of users and enterprises to adapt these networks for transmission of large amounts of data across the world.

Software Defined Networks (SDNs) have a three-tier architecture[5] : the Application Plane, the Control Plane and the Data Plane. These networks provide flexibility to control traffic in the network by decoupling the control plane from the data plane for simplifying the network operation and management for network innovation. The open and dynamically controlled network environment opens up possibilities to programmatically control the network with intelligent route planning, traffic prediction and classification [6].

SDNs are a new paradigm that has a centralized architecture, compared to the traditional networking architecture, where the SDN Controller [7] present in the Control Layer has the complete knowledge of the network, including the topology of the switches formed by the network, network traffic pattern and handling traffic engineering management function. Hence it performs the function of routing in the system to establish communication across the various hosts in the network. The Controller uses a communication protocol such as OpenFlow [8] that allows it to host the Control Plane and Data Plane by giving the orders for the switches for their activation and to write rules into their Flow tables [9].

SDNs help to create affordable networks with high security since the need for intelligent switches is replaced by the Controller that uses Dynamic Routing Algorithm [10] to perform routing operations in the network. The SDN Controller controls the operations spanned out in the entire network due to its centralized architecture [11] and implements firewalls in the network for hosting an Intrusion Prevention System [12]. Routing of packets in the network is handled by the SDN controller since it is aware of the SDN schema, referring to the network topology and structure. The controller learns the configuration of the switches in the network topology by flooding the Address Resolution Protocol (ARP) packets to the switches to establish the data routes [13].

A disadvantage presented by the ARP flooding is that it takes a significant amount of time during which the switches could be vulnerable to a wide range of attacks such as Denial-Of-Service (DOS) attacks [14] due to the temporary escalated traffic in the network. To avoid this, a neural network model can be trained to dynamically determine routes that a packet can take based on the instantaneous network parameters without increasing the traffic in the network.

Multitask learning [15] is used to learn inter-dependencies between the various traffic parameters in the network for multi-output neural modelling. The multitask model consists of a number of subtasks that could be used to solve multiple parts of the same problem. This can be used to predict the activation of switches to write rules for SDN routing.

In this paper, a multitask model is set up based on the topology of the SDN network under study. Topology [16] of the SDN network helps determine the number and the level of switches present. The inputs of the multitask model are network parameters and each switch in the topology can be modelled as a subtask and the outputs will be used to determine which of the switches should be active for the given inputs. The trained neural model is used to implement the routing to avoid flooding at each instance of running the network. The routes predicted by the neural model are added into the flow table of switches via REST APIs in the SDN.

The rest of the paper is organized as follows. Section II gives the summary of some of the best works done in SDN and Deep

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Neural network routing. The system design of the developed model is given in Section III, followed by the implementation details in Section IV. The analysis of the multitask learning model and the evaluation of integration of the model with the SDN controller are given in Section V and VI. Section VII presents the conclusion of the paper and Section VIII summarizes the proposed future works for the paper.

II. Related Works

The summary of some of the state-of-the-art work done in the field of Deep Neural Network Routing is given below.

Mohammed et al [17] explored machine learning and deep learning techniques for the classification and prediction of traffic in SDNs. They had studied the usage of Convolutional Neural Networks (CNN), Autoencoder networks and (Recurrent Neural Networks) RNN along with (Long Short-Term Memory) LSTMs for the non-linear transformation of data to extract useful features of classification.

Mestres et al [18] worked with knowledge defined networking that uses deep learning techniques to gather knowledge about the network, and exploit that knowledge to control the network using logically centralized control facilities provided by SDNs. The designed system used an artificial network with traffic as input and average delays as output features to prevent potential bottlenecks, packet losses or performance drops.

In our paper, three types of network parameters are extracted: traffic, structure and path parameters. The traffic and structure serve as inputs and the path parameters serve as outputs to our neural model. The number of outputs for the model is determined by the number of switches in the network which is in turn dependent on the topology of the SDN network.

Wu et al [19] developed an Artificial Intelligence Enabled Routing (AIER) mechanism with congestion avoidance in SDNs to alleviate the impact of monitoring periods with dynamic routing. The AIER mechanism adds an Artificial Neural Network (ANN) hosted in the Control plane to select a suitable path to avoid congestion in the network. The dataset for the AIER model is collected from generating all data flows in the network with a congestion flag.

In our paper, we develop a neural network model that determines the routing path based on network parameters at that instant. The outputs of the neural model are translated into routing paths that are used to write rules from the SDN Controller into the flow table of the switches.

Zhang et al [20] developed a model using multitask learning for short term traffic state forecasting using Gated Recurrent Units for intelligent transportation systems. They used residual mappings and extracted informative features of the model to study the impact of the size of the training data on model performance to prevent bottlenecks in the networks due to scaling up of the training dataset.

The multitask learning architecture we have used in predicting routes for our SDN consists of a number of sub-tasks each of them have the activation of a switch present in the network depending on the hosts communicating in the network at that instance.

Zuo et al [21] had proposed traffic engineering with a learning-based network path planning using GEANT topology and grid networks in SDN. They formed a sequence-to-sequence model to learn forwarding paths of nodes in the network to capture their essential sequential features.

In our paper, multitask learning has been implemented to generate subtasks to predict the activation of the switches in the topology-based SDN. To take into account the effect of data size to train the model, the dataset has been collected over a variety of congestion windows and its effects are studied in the results in a clos topology for its high scalability.

Mao et al [22] used a CNN for intelligent routing at real-time for the changing traffic patterns in the SDN. to compute the path combinations with high accuracy. The data used for training the proposed CNN was collected from running each path in combination using conventional routing protocols.

In our implementation, we have extended the ANN to a multitask learning deep neural network where the path prediction is divided into a number of subtasks to predict the path for the activation of switches and the dataset used comprises of the flows collected by simulating traffic with various congestion windows in the SDN.

In summary, the system implemented in our paper uses a Multitask learning deep neural network architecture to predict the routes through the activation of the switches in the constructed Clos topology network for the SDN. The dataset for the training of this model was collected by using the network traffic parameters obtained from simulating varying congestion and traffic flowing in the network to increase the efficiency of the performance of the model. This neural model is integrated with the SDN Controller using the REST APIs provided in the SDN Mininet interface.

III. System Design

This section gives a brief description of the design of the SDN system with the deep neural routing model for network routing.

A. Dataset Description

The dataset that was used to train the Multitask Learning deep neural network comprises three types of parameters [23].

1) Traffic Parameters

These parameters given in Table I contain information about the traffic flow in the network to avoid possible bottlenecks in the network and to increase the efficiency of performance of SDNs.

TABLE I. Traffic Parameters Collected from the iPerf Command

No.	Name	Description
1	Interval	Time Taken between successive packets
2	Ct	TCP connect time
3	Transfer	Amount of data transferred
4	Bandwidth	Bandwisth of the link
5	Write	Total number of socket writes
6	Err	Total number of non-fatal socket write errors
7	Rtry	Number of TCP Retries
8	Cwnd	TCP Congestion Window
9	RTT	Round Trip Time
10	NetPwr	Network Power as Throughput/RTT

2) *Structure Parameters*

These represent the hosts which are involved in the active communication in a particular data exchange between two end-point devices. The communication between hosts could take place between hosts connected to the same switch or any other switch in the SDN.

3) *Path parameters*

The first layer of switches in the network refer to the switches connected directly to the end-point devices or the hosts in the network. The higher layer of switches that are involved in the routing are traced using a Packet Tracer tool such as sflow-rt [24] used on the Mininet [25] interface of the constructed SDN. The path followed by the data packet through the switches is fed into as binary values to train the deep neural network model.

B. *Dataset Preprocessing*

The dataset consisting of the traffic, structure and path parameters are normalized before using them to train the multitask learning model. Normalization [26] is performed to scale the data to a value between 0 and 1. Input values can be of a wide range which decreases the speed of learning as the time taken for the neural model to converge to local minima during gradient descent is high. Normalization ensures that the learning time is reduced and more features are learnt from the data. Min-Max normalization ensures that the inputs are within a range of (0,1).

C. *Multitask Learning*

The implemented Deep Neural architecture is based on multitask learning.

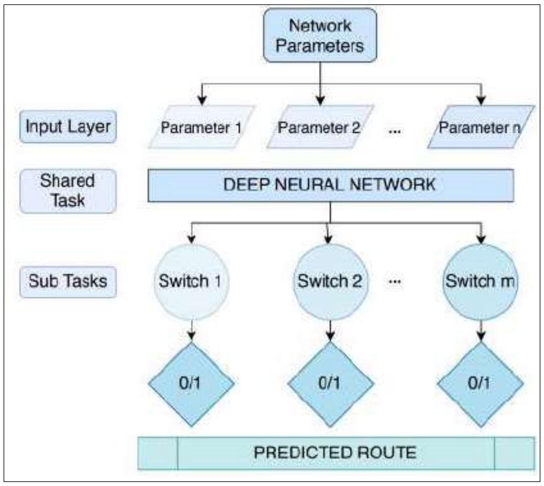


Fig. 1. Multitask Learning Model for Route Prediction

Figure 1 shows the overall design for the multitask learning model. Multitask learning [27] helps the neural network model learn the overall network parameters and decide which switches will be active for the given parameters. If a single task learning model is used, then the neural network would be able to predict if a switch is active or not but the prediction would have been an independent decision without taking into consideration the activation conditions of the other switches. In a multitask learning model, the model can predict which all switches are activated at the given instant of time. The activation of higher level switches depends on which lower level switches are active.

In the given design, the network parameters are given as input. The neural network learns the variations in the network conditions as the shared task. The prediction of which switches will be active for the given network conditions is carried out by the sub tasks. Each sub task is the activation condition of each switch. The number of subtasks in the model is dependent on the topology of the SDN network and the number of switches present in the network.

D. *Modelling of the Multitask Neural Network*

The Multitask Learning Architecture consists of a shared Neural Network and then independent Neural networks that predict the outputs at each switch. The shared neural model is also known as a shared representation.

1) *Forward propagation:*

The network parameters are given as inputs to the shared representation. The forward propagation is given by

$$Z^l = W^l A^{l-1} + b^l \tag{1}$$

Where, W^l and b^l are the weight and bias at layer l and A^{l-1} is the activation at layer $l-1$. When the value of l is 0 i.e., it is the initial layer, then A^{l-1} takes the values of the input X given to the network.

$$A^{l-1} = \begin{cases} X \; for \; l = 0 \\ A^{l-1} \; for \; l > 0 \end{cases}$$

The output of the forward propagation i.e. Z^l is passed through an activation function f^l.

$$A^l = f^l(Z^l) \tag{2}$$

The activated output of layer l is passed as input to the next hidden layer. Once the inputs are propagated through all the hidden layers in the shared representation, the output of the shared representation is given as input to each of the sub tasks. Each subtask has their own set of weights and biases and the output of the shared representation is propagated through these layers. At the final output layer of each subtask, a cost function evaluates the predicted output and the actual output.

When there are i sub tasks, the cost function for each sub task J_i is given by,

$$J_i = cost(\hat{y}_i, y_i) \tag{3}$$

where $\hat{y}_i$ is the predicted output and y_i is the actual output for the subtask i.

The overall cost for the shared representation J is the weighted sum of the costs of each individual sub task.

$$J = \sum \propto_i J_i \tag{4}$$

where $\propto_i$ is the weighting factor that determines the contribution of each sub task to the overall cost of the shared representation. $\propto_i$ is determined during the hyperparameter tuning stage of the neural model.

2) *Backpropagation:*

Backpropagation is done to reduce the overall cost of the neural model by optimizing the weights and biases. In a multitask learning model, backpropagation is applied to each sub task by using the cost function J_i of each sub task i and then the backpropagation is applied to shared representation by using the aggregated cost J.To perform backpropagation the gradients of cost are calculated with respect to weights and biases. For each sub task i, the gradients are calculated for each hidden layer l of the sub task.

$$\frac{\partial J_i}{\partial w_i^l} = \frac{\delta J_i}{\delta z_i^l} A^{l-1} \quad (5)$$

$$\frac{\partial J_i}{\partial b_i^l} = \frac{\delta J_i}{\delta z_i^l} \quad (6)$$

The weights and biases are optimized by using the gradients.

$$W_i^l = W_i^l - \rho \frac{\partial J_i}{\partial w_i^l} \quad (7)$$

$$b_i^l = b_i^l - \rho \frac{\partial J_i}{\partial b_i^l} \quad (8)$$

Where, ρ is the learning rate of the model.

Once the backpropagation is applied to the sub tasks, the loss is propagated through the shared representation. The aggregated cost J is used to calculate the gradients for each hidden layer l of the shared representation. The weights and biases are optimized by using the gradients.

$$\frac{\partial J}{\partial w^l} = \frac{\delta J}{\delta z^l} A^{l-1} \quad (9)$$

$$\frac{\partial J}{\partial b^l} = \frac{\delta J}{\delta z^l} \quad (10)$$

$$W^l = W^l - \rho \frac{\partial J}{\partial w^l} \quad (11)$$

$$b^l = b^l - \rho \frac{\partial J}{\partial b^l} \quad (12)$$

Where, ρ is the learning rate of the model.

E. Configuring REST APIs in the SDN

Representational State Transfer or REST APIs [28] are used to communicate between the SDN Controller and the applications or services that would be running in the network to facilitate automation with network programming. OFCTL_REST API [29] is used for retrieving and updating the switch stats to help debug and update the flow table of the switches for network routing.

IV. System Implementation

This section gives the implementation details of the Deep Neural network along with its integration with the SDN.

A. SDN Setup

The SDN was set up in a Windows 10 with Intel i7 7th Gen Processor with NVidia GeForce 940MX with Oracle VM Virtual Box with 4GB RAM allocation to host the Mininet version 2.2.1 with OpenFlow 1.3 on Ubuntu 18.04. RYU controller 4.34 is established in the SDN as the central coordinator of the network making up the Control Plane with OpenVSwitch 2.14. The network follows a clos topology [30] that uses double-layered switches. The network constructed here consists of 8 hosts connected to 7 switches with a Ryu Controller.

This network can be extended as per the requirement in the constructed SDN. Clos topology is the most common topology used in the cloud networks and is a multi-layer switching network where the number of switching layers determines the structure of the clos network. The clos topology has three stages- the ingress stage, middle stage and egress stage. Each packet entering the ingress stage can be routed through any of the middle stage switches to the relevant egress stage switch. The advantages of using a clos network is that it can be scaled horizontally and also has low latency in accessing data. In case any switch in the middle stage of the clos network fails, then the packets can be routed through any of the other remaining middle stage switches thereby guaranteeing fault tolerance.

B. Dataset collection using the iPerf command

1) Network parameters

The traffic parameters are collected by the iPerf command in the XTerm terminal of the hosts in the SDN. iPerf tool [31] is used to measure the bandwidth and the quality of a network link by implementing a client-server architecture between the source host and destination host which are communicating across the network. The parameters that the iPerf command uses is given in Table II.

TABLE II. IPERF COMMANDS ON THE SERVER AND THE CLIENT SIDE

Server side	
Command	**Description**
-s -server	Starts iPerf in server model and waits for an iPerf client to contact it
Client side	
Command	**Description**
-c -client <Host>	Starts iPerf in client mode and connects with the iPerf server <Host> (IP address or DNS name)
-t time <Time> (default: 10 seconds)	Sets the duration of the connection in seconds
-I -interval	Seconds between periodic bandwidth reports
-b -bandwidth <BW>	Sets the bandwidth for data transfer
-e -enhanced output	Display enhanced output reports
-w -window size <Size>	TCP window size
-f -format	Format to report: Kbits, Mbits, Kbytes, Mbytes
-h -help	Outputs the help text
-v -version	Outputs the version

2) Structure Parameters

The communication between the source and destination host is established by using simple ping commands. The commands for transferring data from the source host are done by using the XTerm terminal [32] for that particular host.

3) Path Parameters

A packet tracer tool such as sflow-rt is used in the SDN to find the path used by the data packets to travel across in the network. This tool captures all the data packets being currency transmitted across the network. The packets using the OpenFlow protocol are traced and the switches that transfer the data packets in the data transfer are recorded in the dataset. During any data transfer between two hosts, the first layer switches connected to the host are always active. The second layer switches, connected only to the first-layer switches, are chosen based on the present state of traffic in the system.

In Figure 2 a), we observe that the second-layer switch S5 is actively used for the data transfer between the hosts connected to the first layer switches. In Figure 2 b), S7 is the second layer switch chosen by the constructed clos topology in the SDN.

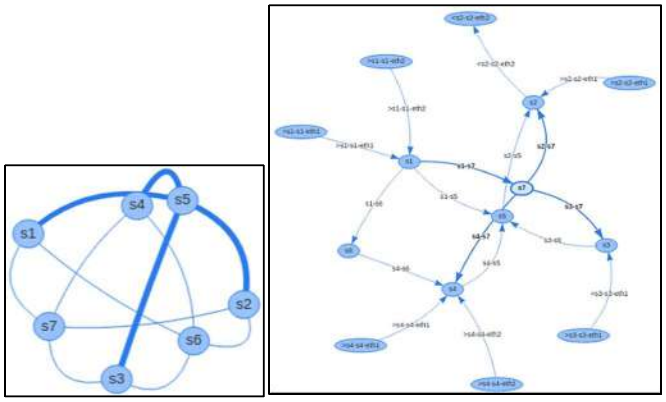


Fig. 2. a) Tracing of the OpenFlow packets using sflow-rt to identify the active Layer-2 switch(S5), b) The switch S7 is chosen as the Layer-2 active switch for data transfer

C. Preprocessing the Dataset

The dataset collected has a total of 1,048,575 samples. The dataset is cleaned to remove null instances. The dataset was split into a training and testing set with a ratio of 80:20. Normalization is done by using Keras' MinMaxScaler [33] for each input field. The minimum value of the field is subtracted from each instance of the field and this is divided by the maximum value in that field.

D. Structure of the Deep Neural model for network routing

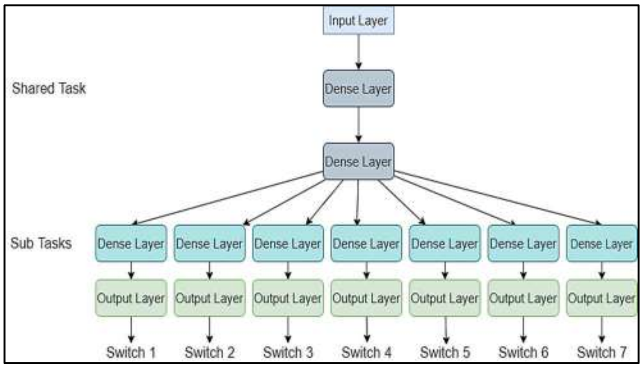


Fig. 3. Implementation of Multitask Learning Neural Model

Figure 3 shows the overall implementation of the multitask learning neural model. The input layer has 23 nodes corresponding to 23 network parameters extracted in the dataset. The shared task of neural tries to identify variation of features within the input dataset. The shared model comprises 1 hidden layer with 16 nodes activated using ReLu Activation function [34]. ReLu increases the speed of convergence as it does not activate for any negative input values. The output of the shared representation is fed to 7 sub tasks. Each sub task represents the activation of a corresponding switch. The number of subtasks is based on the topology and structure of the network chosen. The outputs of the subtasks are passed through the output layer which has one neuron activated by Sigmoid function [34]. Sigmoid function is used as it is a binary classification problem where if the output probability is greater than 0.5, then the switch is activated.

E. Parameters for Training

The parameters of training can be observed in Table III. Adam Optimizer [35] is used. Adam optimizer combines Momentum and Root Mean Square Propagation (RMSP) to control gradient descent such that the step-size can pass local minima but the training stops at the global minima. Binary cross entropy [36] is the loss function used.

TABLE III. TRAINING PARAMETERS OF THE MODEL

Parameter	Value
Optimiser	Adam
Loss Function	Binary Cross Entropy
Batch Size	128

F. Training

The training of the neural model is done on Google Colab with Intel(R) Xeon(R) CPU @ 2.20GHz Processor, 13 GB RAM and 12GB NVIDIA Tesla K80 graphics processor.

TABLE IV. ACCURACY AND LOSS VARIATION WITH NUMBER OF ITERATIONS

Iterations	Accuracy	Change in accuracy (in %)	Loss	Change in loss (in %)
4	97.32	-	6.0	-
8	98.65	1.36	4.67	22.16
12	99.73	0.08	4.0	14.34
16	99.82	0.09	3.2	20

The accuracy of the deep neural model is calculated with increasing number of iterations as shown in Table IV. It is interpreted that by increasing the number of iterations the values of accuracy increase linearly. Loss is calculated using the binary cross entropy function for each epoch. Table IV shows that the loss in the model decreases with increasing the number of iterations of the training dataset.

The training is stopped before the model overfits to learn noise and irrelevant data but undertraining the model will decrease its ability to generalize.

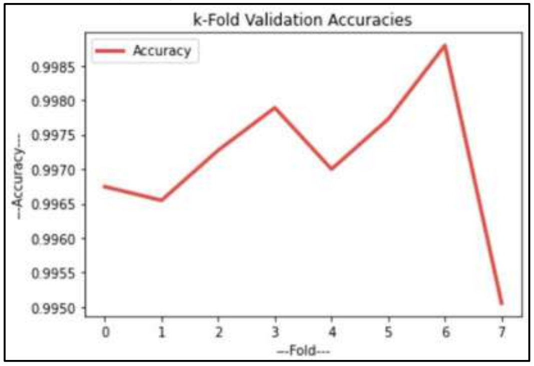


Fig. 4. k-Fold Cross Validation performed with the variation of Accuracy across 8-folds

When the accuracy increases steeply then the model may be overfit and if the number of iterations is too low then the model may be underfit. Based on the change in accuracy and loss, the model is trained for 12 iterations. In order to ensure that the model does not overfit, Stratified k-Fold Cross Validation [37] is performed with the number of folds taken as 8. This is implemented for cross-verification of the developed model to train over the collected dataset before testing. Figure 4 shows the variation of accuracy score for the 8 folds of cross validation. The average accuracy score is found to be 0.9973.

G. *Integration of the Deep Neural model with the SDN Controller*

The SDN Controller uses a standard dynamic routing algorithm for updating the flow tables. This approach incurs overhead due to flooding and over-writing of rules, making it harder and slower for the expansion of the network, decreasing the efficiency of SDN's performance. The integration of the deep neural model with the SDN was done by hosting the trained model on the Controller to predict the path between two hosts for data transfer. The output of the neural model is translated to feed into the OFCTL_REST API that is used to write the rules into the flow tables of the switches to accept the data packets from the corresponding switches or hosts.

V. Evaluation of Multitask Learning Model

This section talks about the results and analysis of the multitask model designed to perform the network routing.

A. *Variation of accuracy on the size of dataset*

Accuracy per batch of dataset is measured with increasing size of the dataset collected for training the deep neural network as seen in Table V. The accuracy is observed to increase with data size being used. The increase in the size of data being used will help the neural model to train with a variety of data that has varying levels of traffic simulations and paths. The loss incurred by the model with increasing dataset size and its variance is given in Table V. The loss decreases with the increase in data size linearly due to higher number of available cases for the model to train and generalize before testing of the model is performed.

TABLE V. Accuracy and Loss Variation with the Dataset Size

Dataset size	**Accuracy**	**Change in accuracy (in %)**	**Loss**	**Change in loss (in %)**
0.25	88.63	-	88.42	-
0.5	93.45	5.43	52.98	40.08
0.75	96.61	3.38	6.39	87.93
1.0	99.73	3.22	3.13	51.01

B. *Variation of True Positive Rate and True Negative Rate with Dataset Size for multilayer switches*

True Positive Rate (TPR) or Sensitivity [38] is defined as the number of times the deep neural model categorizes a positive value correctly. The trained model should be able to identify the switches to be activated in the actual path during network routing.

True Negative Rate (TNR) or Specificity [38] is defined as the number of times the negative values are correctly identified. In this case, the true negative rate should be high enough to predict which switches needn't be active in the network for a particular flow of communication between the hosts.

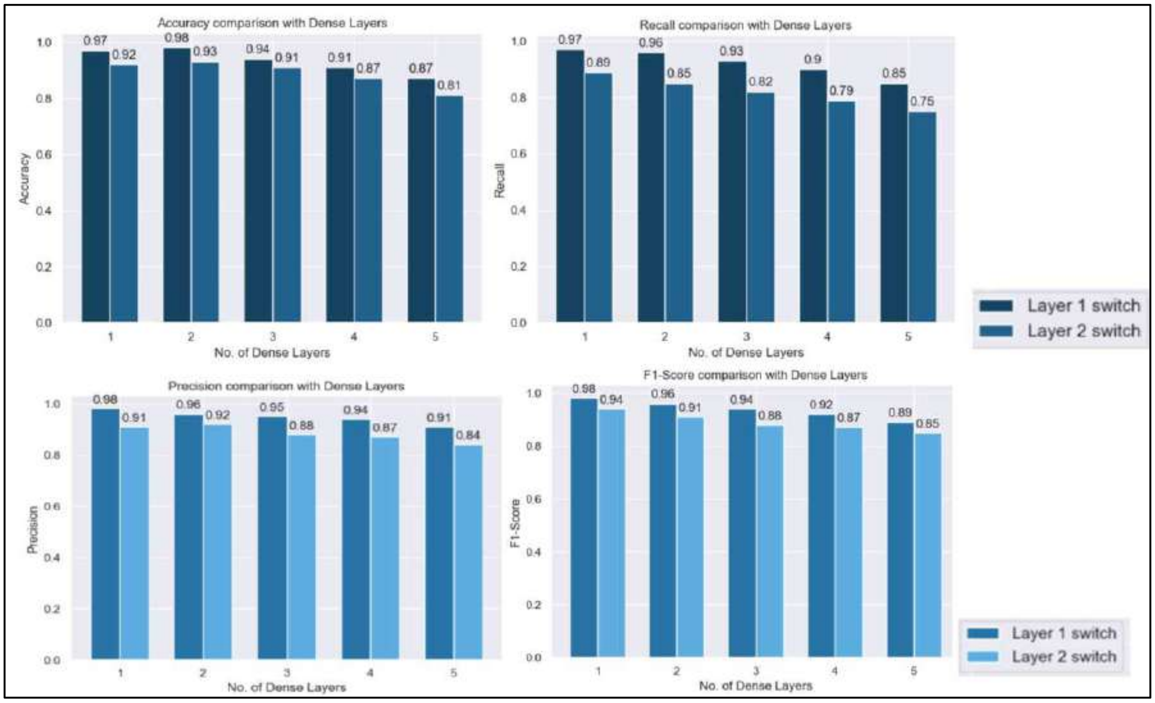


Fig. 5. a) Accuracy on Number of Dense Layers with Layer-1 and Layer-2 switches b) Recall on Number of Dense Layers with Layer-1 and Layer-2 switches c) Precision on Number of Dense Layers with Layer-1 and Layer-2 switches d) F1-Score on Number of Dense Layers with Layer-1 and Layer-2 switches

As it can be observed from Table VI, the ability of the neural model to predict the positive values correctly increases linearly with larger dataset. The prediction of Layer-2 switches varies with the traffic present in the system, which changes dynamically. The size of the dataset should be sufficiently high to prevent the false negatives and false positives in the network to increase the model's performance.

TABLE VI. TPR AND TNR VARIATION WITH THE DATASET SIZE

Dataset size	Layer chosen	TPR	Change in TPR (in %)	TNR	Change in TNR (in %)
0.25	Layer-1	94.14	-	96.73	-
	Layer-2	89.46	-	84.56	-
0.5	Layer-1	95.03	0.94	97.49	0.78
	Layer-2	92.56	3.46	86.37	2.14
0.75	Layer-1	97.21	2.29	97.88	0.4
	Layer-2	94.42	2.0	89.21	3.28
1.0	Layer-1	99.35	2.2	98.46	0.59
	Layer-2	97.65	3.42	90.21	1.12

C. *Accuracy, Recall, Precision and F1-Score on Number of Dense Layers with Layer-1 and Layer-2 switches*

Accuracy [39] gives the measure of how close the predicted values are to the actual value. Recall [39] value gives the proportion of true positive values out of all the positive values predicted by the network. Precision [39] gives the proportion of predicted positive values that are actually positive and its variance with number of dense layers. F1- score [39] combines the precision and recall value of the trained deep neural model using a harmonic mean. This is highly useful in cases such as here where there is an uneven class distribution.

The variation of Accuracy, Precision, Recall and F1-Scores of Layer-1 and Layer-2 switches in the clos network constructed here are plotted against the varying number of dense layers in Figures 5 a), b), c), d). The accuracy indicates how well the model has trained on the dataset.

It is observed from Figures 5 a), b), c) and d) that the accuracy, precision, recall and F1-Score does not seem to increase with increasing the number of dense layers. This is due to the over-fitting of the data by increasing the number of trainable layers in the model. The complexity of the model would also increase leading to poorer performance with increasing number of dense layers.

D. *ROC Curve and AUC Value*

An ROC curve or Receiver Operating Characteristic curve [40] is a graph showing the performance of a classification model at all classification thresholds. The AUC or the area under the ROC curve [40] gives the aggregate measure of performance across all possible classification thresholds. For the developed Deep Neural network model, the AUC area is at 0.987 as seen in Figure 6.

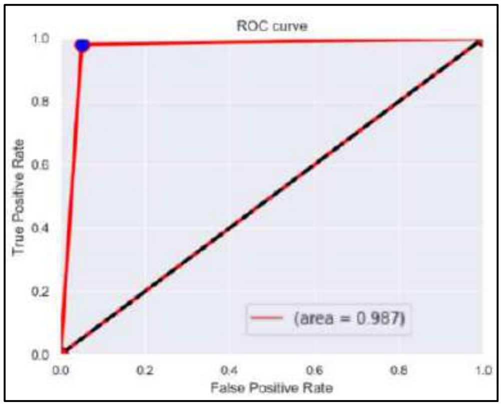


Fig. 6. ROC Curve of the Neural Model

VI. EVALUATION OF INTEGRATION OF NEURAL ROUTING WITH THE SDN CONTROLLER

In this section, the integration of the neural routing model with the SDN Controller is analyzed.

A. *Retry value in network- existing system vs proposed system*

The performance of the SDN system with the Deep Neural network for routing integrated and without the neural network is done by using the number of retry values [42] required when sending the data packets across the network.

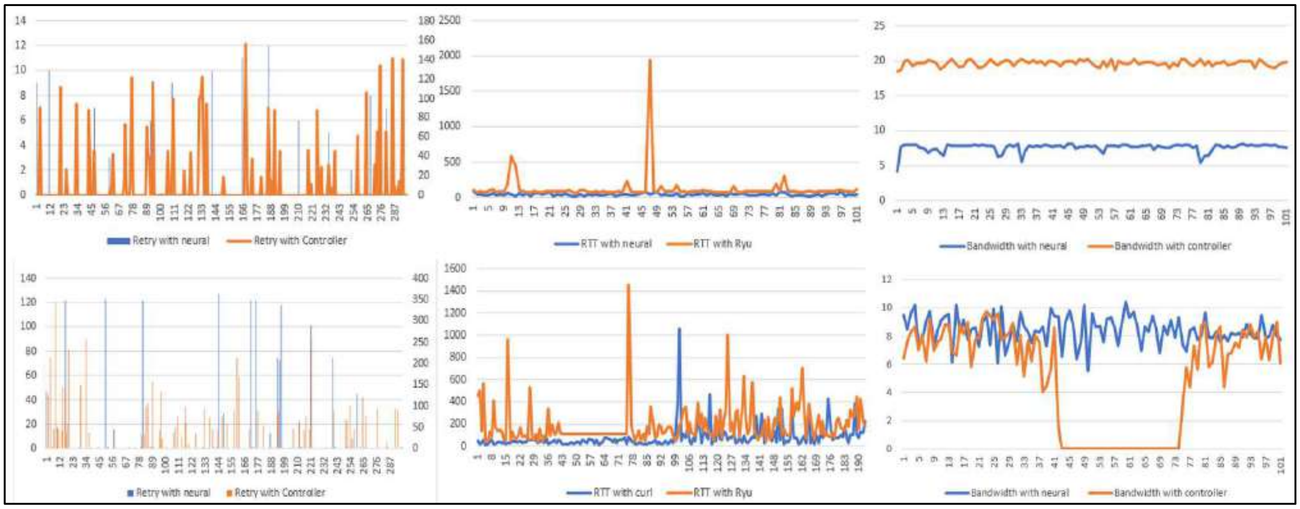


Fig. 7. Comparison of SDN with Ryu Controller and Deep Neural Routing a) Retry, b) Round Trip Time and c) Bandwidth in Layer-1 Switches (above) and d) Retry e) Round Trip Time and f) Bandwidth in the Layer-2 Switches (below)

The Retry value in the neural network integrated system for Layer-1 switches is plotted in Figure 7 a) and for Layer-2 switches in Figure 7 d). It can be seen that the number of retry values for the neural routing model is significantly lower as compared to the traditional SDN routing model for both Layer-1 and Layer-2 switches.

B. *Round Trip Time value in network- existing system vs proposed system*

The Round-Trip Time or RTT [41] is another measure of the performance of the SDN system as it determines the speed of transfer of data packets in the network. The performance of the system with Deep Neural routing model integrated and without neural is observed by the low RTT values in Figure 7 b) and Figure 7 e) for Layer-1 and Layer-2 switches respectively.

C. *Bandwidth Value in network- existing system vs proposed system*

The bandwidth [42] of the network is observed to drop to zero in the existing system as in Figure 7 f) that the proposed system handles better by limiting the overhead in the system especially with higher layers of switches. The low congestion of the data flowing in the network will enable the system to maintain a steady bandwidth as observed in Figures 7 c) and f).

VII. Conclusion

A deep neural network based on multitask learning architecture was implemented using the path, network and traffic parameters collected from the SDN. These parameters were collected dynamically from the network using packet tracer tools deployed on the Application Layer of the SDN. Clos topology containing multilayer switches was constructed in the network where the data flow between the hosts occurred through first and higher layer switches. The constructed neural routing model will dynamically determine the higher-layer switches in the network by analyzing the collected traffic parameters present in the network and activating the switches by writing rules. The activation rules for switches present in the data path during active communication between the hosts are written into the flow table of the switches using OFCTL_REST API. The trained deep neural model was evaluated on various parameters to determine the efficiency of its performance and the model performed switch activations in the network with an accuracy of 99.73%. The performance of the SDN using the dynamic routing model in the existing system is compared with the deep neural routing model using the values of Retry, Round Trip Time and Bandwidth values collected during the data flows when the SDN Controller performed routing on the data with the deep neural model performing the route prediction.

VIII. Future Works

The implemented deep neural routing model using multitask learning architecture is topology based and the dependence of this model on the constructed architecture could be explored by using LSTM architecture for dynamic expansion of the network. The scalability of the network can be extended by the employment of more hosts in the network and across multiple networks in a SDWAN. Further, the prediction of dynamic traffic in the network could be implemented with the route prediction of the data collected from the traffic and congestion forecasts from the future communication between the hosts in the network.

References

[1] M. Rahouti, K. Xiong, and Y. Xin, “Secure Software-Defined Networking Communication Systems for Smart Cities: Current Status, Challenges, and Trends,” IEEE Access, vol. 9, pp. 12083–12113, 2021.

[2] E. Pencheva, I. Atanasov, and I. Asenov, “Toward Network Intellectualization in 6G,” in 2020 XI National Conference with International Participation (ELECTRONICA), pp. 1–4, 2020.

[3] L. Pang, C. Yang, D. Chen, Y. Song, and M. Guizani, “A Survey on Intent-Driven Networks,” IEEE Access, vol. 8, pp. 22862–22873, 2020.

[4] D. M. F. Mattos, P. B. Velloso, and O. C. M. B. Duarte, “An agile and effective network function virtualization infrastructure for the Internet of Things,” Journal of Internet Services and Applications, vol. 10, no. 1, 2019.

[5] J. Xie et al., “A Survey of Machine Learning Techniques Applied to Software Defined Networking (SDN): Research Issues and Challenges,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 393–430, 2019.

[6] Y. Zhao, Y. Li, X. Zhang, G. Geng, W. Zhang, and Y. Sun, “A Survey of Networking Applications Applying the Software Defined Networking Concept Based on Machine Learning,” IEEE Access, vol. 7, pp. 95397–95417, 2019.

[7] P. Sun, Y. Hu, J. Lan, L. Tian, and M. Chen, “TIDE: Time-relevant deep reinforcement learning for routing optimization,” Future Generation Computer Systems, vol. 99, pp. 401–409, 2019.

[8] A. Markoborodov, Y. Skobtsova, and D. Volkanov, “Representation of the OpenFlow Switch Flow Table,” in 2020 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC), pp. 1–7, 2020.

[9] B. Isong, R. Molose, A. M. Abu-Mahfouz, and N. Dladlu, “Comprehensive Review of SDN Controller Placement Strategies,” IEEE Access, vol. 8, pp. 170070–170092, 2020.

[10] Q. Han, S. Cheng, and L. Zeng, “An Intellectual Routing Algorithm based on SDN,” in 2020 IEEE/CIC International Conference on Communications in China (ICCC), pp. 1144–1149, 2020.

[11] A. Prajapati, A. Sakadasariya, and J. Patel, “Software defined network: Future of networking,” in 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1351–1354, 2018.

[12] A. A. Y. R. Fares, F. L. de Caldas Filho, W. F. Giozza, E. D. Canedo, F. L. Lopes de Mendonca, and G. D. Amvame Nze, “DoS Attack Prevention on IPS SDN Networks,” in 2019 Workshop on Communication Networks and Power Systems (WCNPS), pp. 1–7, 2019.

[13] O. Flauzac, C. J. G. Santamaria, F. Nolot, and I. Woungang, “An SDN approach to route massive data flows of sensor networks,” International Journal of Communication Systems, vol. 33, no. 7, pp. 1–14, 2020.

[14] A. Prakash and R. Priyadarshini, “An Intelligent Software defined Network Controller for preventing Distributed Denial of Service Attack,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 585–589, 2018.

[15] S. Sanagavarapu and S. Sridhar, “Dynamic Routing Framework Proposal for SDWAN using Topology-based Multitask Learning,” in 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1–8, 2020.

[16] L. Mamushiane, J. Mwangama, and A. A. Lysko, “Given a SDN Topology, How Many Controllers are Needed and Where Should They Go?,” in 2018 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), pp. 1–6, 2018.

[17] A. R. Mohammed, S. A. Mohammed, and S. Shirmohammadi, “Machine Learning and Deep Learning Based Traffic Classification and Prediction in Software Defined Networking,” in 2019 IEEE International Symposium on Measurements & Networking (M&N), pp. 1–6, 2019.

[18] A. Mestres et al., “Knowledge-Defined Networking,” ACM SIGCOMM Computer Communication Review, vol. 47, no. 3, pp. 2–10, 2017.

[19] Y.-J. Wu, P.-C. Hwang, W.-S. Hwang, and M.-H. Cheng, “Artificial Intelligence Enabled Routing in Software Defined Networking,” Applied Sciences, vol. 10, no. 18, p. 6564, 2020.

[20] K. Zhang, L. Wu, Z. Zhu, and J. Deng, “A Multitask Learning Model for Traffic Flow and Speed Forecasting,” IEEE Access, vol. 8, pp. 80707–80715, 2020.

[21] Y. Zuo, Y. Wu, G. Min, and L. Cui, “Learning-based network path planning for traffic engineering,” Future Generation Computer Systems, vol. 92, pp. 59–67, 2019.

[22] B. Mao, F. Tang, Z. Md. Fadlullah, and N. Kato, “An Intelligent Route Computation Approach Based on Real-Time Deep Learning Strategy for Software Defined Communication Systems,” IEEE Transactions on Emerging Topics in Computing, pp. 1–12, 2019.

[23] S. Sanagavarapu and S. Sridhar, “SDPredictNet-A Topology based SDN Neural Routing Framework with Traffic Prediction Analysis,” in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0264–0272, 2021.

[24] “sFlow-RT,” sflow-rt.com. https://sflow-rt.com/ (accessed Nov. 29, 2020).

[25] M. Hasan, H. Dahshan, E. Abdelwanees, and A. Elmoghazy, “SDN Mininet Emulator Benchmarking and Result Analysis,” in 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 355–360, 2020.

[26] S. Bhanja and A. Das, “Impact of Data Normalization on Deep Neural Network for Time Series Forecasting,” arXiv:1812.05519 [cs, stat], 2019, Accessed: Nov. 29, 2020. [Online]. Available: https://arxiv.org/abs/1812.05519.

[27] J. Zhou, P. Hong, J. Pei, and D. Li, “Multi-Task Deep Learning Based Dynamic Service Function Chains Routing in SDN/NFV-Enabled Networks,” in ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pp. 1–6, 2019.

[28] W. Zhou, L. Li, M. Luo, and W. Chou, “REST API Design Patterns for SDN Northbound API,” in 2014 28th International Conference on Advanced Information Networking and Applications Workshops, pp. 358–365, 2014.

[29] “ryu.app.ofctl_rest — Ryu 4.34 documentation,” ryu.readthedocs.io. https://ryu.readthedocs.io/en/latest/app/ofctl_rest.html (accessed Nov. 29, 2020).

[30] A. Singh et al., “Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google’s Datacenter Network,” Communications of the ACM, vol. 59, no. 9, 2016.

[31] V. Gueant, “iPerf - The TCP, UDP and SCTP network bandwidth measurement tool,” Iperf.fr, 2013. https://iperf.fr/ (accessed Nov. 29, 2020).

[32] D. Kumar and M. Sood, “Software Defined Networks (S.D.N): Experimentation with Mininet Topologies,” Indian Journal of Science and Technology, vol. 9, no. 32, 2016.

[33] “sklearn.preprocessing.MinMaxScaler — scikit-learn 0.22.1 documentation,” Scikit-learn.org, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (accessed Nov. 29, 2020).

[34] B. Ding, H. Qian, and J. Zhou, “Activation functions and their characteristics in deep neural networks,” in 2018 Chinese Control And Decision Conference (CCDC), pp. 1836–1841, 2018.

[35] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in 3rd International Conference on Learning Representations, ICLR 2015, pp. 1–15, 2015.

[36] K. Janocha and W. M. Czarnecki, “On Loss Functions for Deep Neural Networks in Classification,” Schedae Informaticae, vol. 25, pp. 49–59, 2017.

[37] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” in 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 78–83, 2016.

[38] E. Martin et al., “Sensitivity and Specificity,” Encyclopedia of Machine Learning, pp. 901–902, 2011.

[39] V. Deart, V. Mankov, and I. Krasnova, “Development of a Feature Matrix for Classifying Network Traffic in SDN in Real-Time Based on Machine Learning Algorithms,” in 2020 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC), pp. 1–9, 2020.

[40] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, M. Ghogho, and F. El Moussa, “DeepIDS: Deep Learning Approach for Intrusion Detection in Software Defined Networking,” Electronics, vol. 9, no. 9, p. 1533, 2020.

[41] M. Althobyani and X. Wang, “Implementing an SDN based learning switch to measure and evaluate UDP traffic,” Computers & Electrical Engineering, vol. 66, pp. 342–352, 2018.

[42] E. R. Jimson, K. Nisar, and Mohd. H. bin Ahmad Hijazi, “Bandwidth management using software defined network and comparison of the throughput performance with traditional network,” in 2017 International Conference on Computer and Drone Applications (IConDA), pp. 71–76, 2017.

A Light-based Interpretation of Schrodinger's Wave Equation and Heisenberg's Uncertainty Principle with Implications on Quantum Computation

Dr Pravir Malik
Deep Order Technologies, USA
pravir.malik@deepordertechnologies.com

Abstract— **Quantum computation as currently conceived is based on the largely unproven Copenhagen Interpretation of Quantum Mechanics. By viewing light as a multi-layered, symmetrical construct though, it is possible to interpret quantum-level dynamics assumed as fundamental, differently. Hence, looking at Schrodinger's Wave Equation and Heisenberg's Uncertainty Principle from the point of view of light, it becomes possible to understand quantum-level dynamics as an outcome of a multi-layered, symmetry-based model of light. Such a different view of quantum-level dynamics suggests a different way to conceive of quantum computation. As such, Schrodinger's Wave Equation can be viewed as an arbitration to take information from behind the quantum-veil that may exist in antecedent layers of light, and through such arbitration or rate of change of the wave-function, compute it into material existence. Heisenberg's Uncertainty Principle suggests that meta-level function seeking to precipitate or to be arbitrated into material existence may take different form while still fulfilling the intent of the meta-level function. Quantum computation, therefore, can be conceived as a creative as opposed to a solely constructive process. The object of quantum computation in such an interpretation of quantum phenomena is nothing other than to continue to create something new, or to continue to enhance materialization of meta-level function, rather than to simply construct based on regurgitating programming-based instruction.**

Keywords— ***Quantum Computation, Schrodinger's Wave Equation, Heisenberg's Uncertainty Principle, Symmetrical-Model of Light, Planck's Constant, Speed of Light, Copenhagen Interpretation of Quantum Mechanics***

I. INTRODUCTION

The jury is still out on the dynamics that animate the quantum level.

It is assumed, based on the Copenhagen Interpretation of Quantum Mechanics, that a quantum-object will exist in a number of superposed states until measured. The act of measurement will cause the quantum-object to collapse into an observable state, and therefore it will be impossible, given today's science and technology, to know what is happening behind the quantum-veil.

Yet quantum computation assumes knowledge of phenomena behind the quantum-veil, and subsequently that an n-qubit quantum computer can exist in 2^n (2 to the power n) states simultaneously. Further, it is assumed that if these qubits can be entangled, then the combination of the dual properties of superposition and entanglement, will confer quantum computers with extraordinary processing power.

If, however, an alternative light-based view of quantum phenomena were to be adopted (Malik, 2018b), there are different implications for quantum computation (Malik, 2020). Such implications can be teased out by interpreting cornerstone quantum-level equations, such as Schrodinger's Wave Equation and Heisenberg's Uncertainty Principle.

Schrodinger's equation, which seeks to model how a quantum state of a quantum system changes with time, or in other words seeks to model matter as a wave rather than as a particle (Stewart, 2012), is depicted in the following, Equation (1):

$$i\frac{h}{2\pi}\frac{\partial}{\partial x}\psi = \hat{H}\psi$$

Eq. 1: Schrodinger's Wave Equation

ψ depicts a wave form and can be thought of as a probable cloud of possible states. $\hat{H}$ is the Hamiltonian operator, which acts as a focusing function. In its essence what the equation is likely suggesting is that the way a wave form changes over time is equivalent to some expressible state of the possibilities inherent in the cloud of possible states.

Heisenberg's uncertainty principle, that calls out the difficulty in measuring any two properties of a quantum-object

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

at the same time, and in the following, Equation 2, focusing on momentum (p) and positions (x) of a quantum-object, is depicted as:

$$\Delta p \; X \; \Delta x \; \geq \; \frac{h}{4\pi}$$

Eq. 2: Heisenberg's Uncertainty Principle

This paper will first summarize a multi-layered, symmetry-based model of light that will be associated with the "cloud of possible states" (Section II). It will then suggest an interplay (Section III) between Planck's Constant, h, and the speed of light, c, to offer interpretations and some quantum computation implications of the Schrodinger equation (Section IV) and Heisenberg's uncertainty principle (Section V) based on the proposed model of light (Malik, 2018a). The model of light, interpretations of the fundamental quantum-level equations, and some implications for quantum computation will then be summarized (Section VI).

II. LIGHT-BASED MODEL

Light has a significant impact on the experienced nature of reality. Fundamentals such as space, time, and the possibilities of the movement of objects are all tied to the reality of light traveling at speed c, 186,000 miles per second in a vacuum (Einstein, 1995). By extrapolating on the necessity for light to move at a constant speed of c it is possible to construct a multi-layered light-based model (Malik et al 2019; Malik 2020) which can provide significant insight into the nature of quanta, and subsequently on the interpretations of the Schrodinger equation and the Heisenberg principle.

In this model the infinite information conceived were light to travel infinitely fast depicted by the row $c_\infty\colon [Pr, Po, K, H]$ in Equation 3, below, precipitates into material reality, c_U - where light travels at speed c, via intermediate realities where light is envisioned to exist at speeds slower than infinity, but faster than c. These intermediate realities are specified by rows c_K and c_N respectively, such that $c_U < c_N < c_K < c_\infty$. Note that Einstein's Theory of Relativity does not disallow speeds of light greater than c: it is the acceleration to speed c from a slower speed that is not possible (Perkowitz, 2011). Further, spaces with light speeds greater than c should be viewed as conceptual spaces made to vary were light to travel at different speeds, or property spaces, being separate from but influencing physical space as explored by Nobel Physicist Frank Wilczek (Wilczek, 2016).

Referring to the matrix-equation, Equation 3, that follows, precipitation itself takes place via a series of quantization functions. The first quantization (↓) is specified by $(\downarrow\ \mathrm{R}_{\mathrm{C}_\mathrm{K}} = \mathrm{f}(\mathrm{R}_{\mathrm{C}_\infty}))$ and suggests that reality (R) at $\mathrm{c}_\mathrm{K}, \mathrm{R}_{\mathrm{C}_\mathrm{K}}$, is a function (f) of reality at c_∞, $\mathrm{R}_{\mathrm{C}_\infty}$.

The states of all-presence (Pr) formed because light is instantaneously present in whatever volume is being considered, all-power (Po) formed by the ability of light to overpower any other emergence, all-knowledge (K) formed by the fabric of light being able to record any appearance or disappearance of event, and all-harmony (H) formed by everything being connected in the nature of the all-present light, are mathematically transformed into large sets as specified by $\mathrm{c}_\mathrm{K}\colon [\mathrm{S}_{\mathrm{Pr}}, \mathrm{S}_{\mathrm{Po}}, \mathrm{S}_\mathrm{K}, \mathrm{S}_\mathrm{H}]$, where S_{Pr} is the set of all-presence, S_{Po} is the set of all-power, S_K is the set of all-knowledge, and S_H is the set of all-harmony, respectively.

A further quantization takes place via $(\downarrow\ \mathrm{R}_{\mathrm{C}_\mathrm{N}} = \mathrm{f}(\mathrm{R}_{\mathrm{C}_\mathrm{K}}))$, and suggests that reality (R) at $\mathrm{c}_\mathrm{N}, \mathrm{R}_{\mathrm{C}_\mathrm{N}}$, is a function (f) of reality at c_K, $\mathrm{R}_{\mathrm{C}_\mathrm{K}}$.

Hence, elements from each of the four sets combine in unique combinations, specified by $\mathrm{c}_{\mathrm{N}:}\ \mathrm{f}(\mathrm{S}_{\mathrm{Pr}} \ \mathrm{x}\ \mathrm{S}_{\mathrm{Po}} \ \mathrm{x}\ \mathrm{S}_\mathrm{K} \ \mathrm{x}\ \mathrm{S}_\mathrm{H})$, to create a bases for a practically infinite number of unique seeds or functions.

A final quantization, specified by $(\downarrow\ \mathrm{R}_{\mathrm{C}_\mathrm{U}} = \mathrm{f}(\mathrm{R}_{\mathrm{C}_\mathrm{N}}))$, suggests that reality (R) at $\mathrm{c}_\mathrm{U}, \mathrm{R}_{\mathrm{C}_\mathrm{U}}$, is a function (f) of reality at c_N, $\mathrm{R}_{\mathrm{C}_\mathrm{N}}$. This quantization results in material reality, specified by $\mathrm{c}_\mathrm{U}\colon\ [\mathrm{S}, \mathrm{T}, \mathrm{E}, \mathrm{G}]$, where S refers to Space, T refers to Time, E refers to Energy, and G refers to Gravity. Specifically, "Space" – envisioned to be the arena for all the subtle-seeds to exist and subsequently leading to the creation of material possibility; "Time" – ensuring that the possibilities in the seeds are worked out; "Energy" – associated with seeds and their conversion into matter; and "Gravity" – specifying relationships between seed and seed and seeds and seeds. Further, Space, being a repository of archetypes represents light's property of knowledge (K); Time, assuring maturity regardless of opposition represents light's property of power (Po); Energy, allowing seeds to have presence, represents light's property of presence (Pr); and Gravity, allowing relationship between seed and seed, represents light's property of harmony (H).

Hence the multi-layered fourfold light-based model is summarized as (3):

$$\begin{bmatrix} \mathrm{c}_\infty\colon [\mathrm{Pr}, \mathrm{Po}, \mathrm{K}, \mathrm{H}] \\ (\downarrow\ \mathrm{R}_{\mathrm{C}_\mathrm{K}} = \mathrm{f}(\mathrm{R}_{\mathrm{C}_\infty})) \\ \mathrm{c}_\mathrm{K}\colon [\mathrm{S}_{\mathrm{Pr}}, \mathrm{S}_{\mathrm{Po}}, \mathrm{S}_\mathrm{K}, \mathrm{S}_\mathrm{H}] \\ (\downarrow\ \mathrm{R}_{\mathrm{C}_\mathrm{N}} = \mathrm{f}(\mathrm{R}_{\mathrm{C}_\mathrm{K}})) \\ \mathrm{c}_{\mathrm{N}:}\ \mathrm{f}(\mathrm{S}_{\mathrm{Pr}} \ \mathrm{x}\ \mathrm{S}_{\mathrm{Po}} \ \mathrm{x}\ \mathrm{S}_\mathrm{K} \ \mathrm{x}\ \mathrm{S}_\mathrm{H}) \\ (\downarrow\ \mathrm{R}_{\mathrm{C}_\mathrm{U}} = \mathrm{f}(\mathrm{R}_{\mathrm{C}_\mathrm{N}})) \\ \mathrm{c}_\mathrm{U}\colon\ [\mathrm{S}, \mathrm{T}, \mathrm{E}, \mathrm{G}] \end{bmatrix}_{\mathrm{Light}}$$

Eq. 3: Multi-Layered Fourfold Light-Based Model

Note that because space-time-energy-gravity appears to come into existence when light precipitates to speed c, it is fair to assume that the deeper nature and activity at the quantum veil before matter is formed, is of the substance of space-time-energy-gravity. As proposed in The Origins and Possibilities of Genetics (Malik, 2019), space-time-energy-gravity can also be thought of as the script used to write a "law" about any specific emergence, which get aggregated into the overall dynamics of Space, Time, Energy, and Gravity as experienced at the macro level.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

III. The Interplay Between Planck's Constant and Speed of Light

In this model of Light, quanta are perceived as a bridge mechanism that allows subtle information in a conceptual space determined by a faster moving speed of light, to quantize or materialize some of that information, more concretely, in a conceptual space determined by a relatively slower moving speed of light.

This implies that emergence in the layer where light exists at c, has to take place via the device of quanta. Planck's discovery that energy at the subatomic level requires a minimum threshold 'quantum' to express itself perhaps relates to this. Einstein postulated quanta as a fundamental property of light itself, rather than as something that arose in the interaction of light with matter as Planck thought. Note that Planck's treatment of quanta was more as a mathematical convenience that allowed the derivation of an equation that explained the curve of radiation wavelengths at varying temperatures of a heated black body (Isaacson, 2008). Einstein's theory produced a law of the photoelectric effect where the energy of emitted electrons would depend on the frequency of light. Einstein received the Nobel Prize for this discovery (Isaacson, 2008).

Summarizing, if c is the upper limit of the layer associated with c_U, then it makes sense that the lower limit h (Planck's constant) should be inversely proportional to c. Hence, Equation 4:

$$h \propto \frac{1}{c}$$

Eq. 4: Relationship Between Planck's Constant and Speed of Light

This relationship is substantiated by combining two well-known equations: the first is Einstein's photoelectric equation connecting energy with frequency of light as depicted by Equation 5, and the second is the electromagnetic equation connecting speed of light with wavelength and frequency as depicted by Equation 6:

$$E = h\nu$$

Eq. 5: Einstein's Photoelectric Equation

$$c = \nu\lambda$$

Eq. 6: Relationship Between Speed of Light, Wavelength, and Frequency

To yield Equation 7:

$$h = \frac{E\lambda}{c}$$

Eq. 7: Connecting h and c

About h, H.A. Lorentz the Dutch scientist has commented in The Science of Nature (Lorentz, 1925): "We have now advanced so far that this constant not only furnishes the basis for explaining the intensity of radiation and the wavelength for which it represents a maximum, but also for interpreting the quantitative relations existing in several other cases among the many physical quantities it determines. I shall mention a few only, namely the specific heat of solids, the photo-chemical effects of light, the orbits of electrons in the atom, the wavelengths of the lines of the spectrum, the frequency of the Roentgen rays which are produced by the impact of electrons of given velocity, the velocity with which gas molecules can rotate, and also the distances between the particles which make up a crystal. It is no exaggeration to say that in our picture of nature nowadays it is the quantum conditions that hold matter together and prevent it from completely losing its energy by radiation."

Elaborating on the interplay between c and h:

- When light slows down, then the binding factor of matter, h, can change. Hence, at light speeds greater than c, the resulting 'h' would become smaller than h since there is an inverse proportionality between c and h.
- h in this physical realm is such that matter forms in the way that it does. At faster than c speeds of light, this binding factor, h, would be smaller in value, and hence result in a materialization where matter would not be bound in the way it is when light travels at c.
- In Light's state of traveling infinitely fast, 'h' would be zero, implying that the essential properties of Light will pervade existence. As light slows down, to create vast sets of properties anchored in each of the four essential properties of Light as depicted in (3), 'h' must be such that it allows the essential properties to splinter into a vast number of variations of itself. Further, 'h' must be such that it allows these variations to be accessed in any subsequent layer of Light. This may suggest that each element could exist in a relatively materialized 'field-like' form.
- With the second quantization function or mathematical transformation, light would be slower than in the first quantization or transformation, and yet faster than c. The resulting 'h' would hence be such that the material form would be something between being 'field-like' and 'particle-like', the latter being the basis of matter as it begins to manifest in the physical realm where light travels at c.
- Perhaps such an 'h' that accompanies the second transformation will allow form to materialize to be 'wave-like', which would also have the needed characteristic of linking various elements from their individual 'fields', to create a unique though subtle wave-form seed.
- It is such a unique wave-form seed that with the existence of the h corresponding to c, then becomes the basis of the quantum-bridge that allows matter in its known form to arise.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

IV. Interpretation of Schrodinger Wave Equation Based on Model of Light & Implication for Quantum Computation

Restating Schrodinger's equation (1):

$$i\frac{h}{2\pi}\frac{\partial}{\partial x}\psi = \hat{H}\psi \qquad (1)$$

ψ depicts a wave form and can be thought of as a probable cloud of possible states. $\hat{H}$ is the Hamiltonian operator, which is a focusing function, and in its essence what the equation may be suggesting is that the way a wave form changes over time is equivalent to some expressible state of the possibilities inherent in the cloud of possible states.

But in reference to the multi-layered model of Light depicted in (3), the cloud of possible states is another way of saying that behind physical form, "form" is represented in another way than physically. Schrodinger's equation, a cornerstone in quantum theory, alludes to the mystery that accompanies the lifting of the material veil. Different interpretations of the quantum world have arisen to suggest the meaning of potentiality implicit in this equation. As summarized by Kleinman in his book "The Four Faces of the Universe" (Kleinman, 2006), at one end are theories to do with 'hidden variables' that bring into focus our ignorance of the deeper aspects of the system being studied, to David Bohm's notion of the 'implicate order' in which any element contains enfolded within it the totality of the universe (Bohm, 1983), to a 'many-worlds' interpretation asserting that every observation of a quantum system splits the universe into parallel and disconnected worlds.

It is interesting to note that in his lectures on Schrodinger's equation Feynman (Gottlieb, 2013) has stated: "Where did he get that [equation] from? Nowhere. It is not possible to derive it from anything you know. It came out of the mind of Schrödinger".

The Copenhagen Interpretation of Quantum Mechanics in fact does away with the real questions about what reality is altogether, by simply focusing on the observed universe. This orientation perhaps stems from one of its founders, Neils Bohr, who claimed that it was not the purpose of physics to answer questions about the nature of reality (Kleinman, 2006).

Considering Schrodinger's equation therefore, 'i' is a complex number and suggests the interplay of two dimensions, one being real, and one being 'imaginary'. But the 'imaginary' dimension could be thought of as none other than the meta-levels implicit in the mathematical model in this treatise. Further, $\frac{h}{2\pi}$ is in line with the suggestion that h will have to become a fraction of itself as c increases. Hence, the change in the wave function, $\frac{\partial}{\partial x}\psi$, is intimately related to i and $\frac{h}{2\pi}$, and perhaps more fully makes sense when considered in the context of i x $\frac{h}{2\pi}$ x $\frac{\partial}{\partial x}\psi$ - which has to be the case when dealing with the integration of dynamics of multiple levels of light.

Further, the change in the wave function, $\frac{\partial}{\partial x}\psi$, is related to $\hat{H}\psi$, and suggests that there is some system "energy", represented by the Hamiltonian, $\hat{H}$, that when applied to the existing wave, ψ, will indicate how the wave will be expressed going forward.

Therefore, 'form' can be seen to exist as Light, and in this point of view Schrodinger's equation becomes an arbitration to take information from behind the quantum-veil that may exist in antecedent layers of light, and through such arbitration or rate of change, compute them into material existence. Quantum computation, therefore, becomes a predominantly creative as opposed to a solely constructive function. The object of quantum computation in such an interpretation of quantum phenomena is nothing other than to create something new, rather than to construct, based on regurgitating programming-based instruction. That which is created as 'new' is nothing other than enhanced or even unique function that is proposed to exist in antecedent layers of light.

V. Interpretation of Heisenberg's Uncertainty Principle Based on Model of Light & Implication for Quantum Computation

In his book, The Little Book of String Theory, Princeton University's Gubser (Gubser, 2010) describes the effect on approaching absolute zero temperature on molecules. He takes the example of water molecules and relates that one cannot make the water molecules colder than absolute zero, -273.15 Celsius. This is so because there is no more thermal energy to suck out at that temperature. However, quantum uncertainty, the phenomenon which relates the momentum and location of electrons in atoms necessitates that the water molecules will still vibrate. Gubser suggests this by considering Heisenberg's uncertainty relation depicted previously by (2) and reproduced here for convenience:

$$\Delta p\ X\ \Delta x\ \geq \frac{h}{4\pi} \qquad (2)$$

In (2) Δp is the uncertainty in a particle's momentum, Δx is the uncertainty in the particle's location, and h is the Planck's constant. In frozen water crystals it is precisely known where the water molecules are, and therefore Δx is fairly small. This means that Δp has to be considerably larger, and therefore that the water molecules are still vibrating even though they are at absolute zero. This innate vibration, known as 'quantum zero-point' energy, expresses the phenomenon of quantum fluctuations.

The Planck's constant order of magnitude (10^{-34}) though, suggests that the boundary between the layer so created when light travels at 'c' and the antecedent layers of light, manifests as the phenomenon of quantum fluctuations, or the uncertainty relation, or the quantum zero-point energy. In reality this phenomenon is nothing other than an expression of the essential play of unique seeds or functions that is posited as a key formative force behind any organization occurring in the layer so created when light travels at speed c.

In this interpretation the thermal energy describes the essential energy of movement of molecules, while the uncertainty relation suggests the phenomenon of function-precipitation from other layers of Light, "physically" linking layers of light. This also casts a different interpretation on the cosmological phenomenon of quantum vacuum, reinforcing in line with this treatise, that the “emptiness” of space is in fact a veil to infinitely more information seeking precipitation at the material level.

But further, it may also be suggested that the uncertainty principle itself is only valid at the layer so created when light travels at speed c, and that too, because of the finiteness of c. This finiteness as already suggested implies h, which implies that if the position of a particle is going to be observed by shining light on it, the light has to have at least a quantum of energy. But to determine the position of a particle accurately, light of a shorter wavelength would have to be used (Hawking, 1988) which would necessarily have relatively higher energy as compared with higher wavelength light, which in turn would interfere with the velocity and hence momentum of the particle. The uncertainty in measuring the momentum could also therefore be thought of as a consequence of the finiteness of the speed of light, c.

If the speed of light were to approach ∞ miles per second, as suggested in Section III, the quantum would be smaller and the uncertainty in measuring position or momentum would be reduced. At the layer where light is proposed to travel infinitely fast there would therefore be no uncertainty since light would accurately tell both position and momentum definitively.

Hence, the uncertainty principle may be further qualified, as in Equations 8 - 10:

$$\Delta p \; X \; \Delta x \; \geq \frac{h}{4\pi}$$

Eq 8: Uncertainty Principle when Light at 186,000 miles per second

$$@\mathrm{C_Q}\!: \; \Delta p \; X \; \Delta x \; \rightarrow \; 0$$

Eq 9: Uncertainty Principle at when Light >> 186,000 miles per second

$$@\mathrm{C}_{M_3}\!: \; \Delta p \; X \; \Delta x = 0$$

Eq 10: Uncertainty Principle when Light traveling infinitely fast

The notion of position and momentum becoming finite when light travels at c, also may imply that space, time, and quanta are emergent rather than absolute properties, as also suggested in Section II. This is also the conclusion of Arkani-Hamed of the Institute of Advanced Studies in the following thought experiment (Wolchover, 2013):

‘Locality says that particles interact at points in space-time. But suppose you want to inspect space-time very closely. Probing smaller and smaller distance scales requires ever higher energies, but at a certain scale, called the Planck length, the picture gets blurry: So much energy must be concentrated into such a small region that the energy collapses the region into a black hole, making it impossible to inspect. “There’s no way of measuring space and time separations once they are smaller than the Planck length,” said Arkani-Hamed. “So we imagine space-time is a continuous thing, but because it’s impossible to talk sharply about that thing, then that suggests it must not be fundamental — it must be emergent.”

Unitarity says the quantum mechanical probabilities of all possible outcomes of a particle interaction must sum to one. To prove it, one would have to observe the same interaction over and over and count the frequencies of the different outcomes. Doing this to perfect accuracy would require an infinite number of observations using an infinitely large measuring apparatus, but the latter would again cause gravitational collapse into a black hole. In finite regions of the universe unitarity can therefore only be approximately known.’

The notion of emergence of fundamentals such as space, time, and quanta, further reinforce the idea that quantum computation is a fundamentally creative as opposed to a constructive act. What Heisenberg’s uncertainty principle is suggesting is that while a meta-function will fulfil its purpose, it may manifest in a variable process due to the different sets of dynamics existing in the different conceptual layers of layers. If a hypothetical quantum computer were to be constructed in a layer of light closer to light traveling at infinite speed, the uncertainty would be less, as also suggested by the preceding equations. This would be because there would be fewer influences arbitrating possibility.

VI. SUMMARY AND CONCLUSION

The jury is still out on the dynamics that animate the quantum level.

It is assumed, based on the Copenhagen Interpretation of Quantum Mechanics, that a quantum-object will exist in a number of superposed states until measured. The act of measurement will cause the quantum-object to collapse into an observable state, and therefore it will be impossible, given today’s science and technology, to know what is happening behind the quantum-veil.

Yet quantum computation assumes knowledge of phenomena behind the quantum-veil, and subsequently that an n-qubit quantum computer can exist in 2^n (2 to the power n) states simultaneously. Further, it is assumed that if these qubits can be entangled, then the combination of the dual properties of superposition and entanglement, will confer quantum computers with extraordinary processing power.

The alternative light-based model leveraged in this paper is predicated on the infinite information conceived were light to travel infinitely fast depicted by the row c_∞: $[Pr, Po, K, H]$ in the (3). Information from this layer precipitates into material reality, c_U - where light travels at speed c, via intermediate realities where light is envisioned to exist at speeds slower than infinity, but faster than c. These intermediate realities are specified by rows c_K and c_N respectively, such that $c_U < c_N < c_K < c_\infty$.

Such a model gives rise to an alternative view of quantum-level phenomena as suggested by the alternative interpretation of foundational quantum-level equations such as Schrodinger’s wave equation and Heisenberg’s uncertainty principle. Leveraging the alternative interpretation of Schrodinger’s equation, quantum computation can be seen to be a creative act arbitrating information or function from information-rich conceptual spaces created by light travelling at faster than speed

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

c. Heisenberg's uncertainty principle reinforces such a model of light suggesting that meta-level function has leeway in the way at it manifests but will manifest to fulfil the function it is intended to.

Leveraging such interpretations suggests an alternative to view the field of quantum computation, to be creative rather than solely constructive.

References

1. Einstein, A. 1995. Relativity: The Special and General Theory. New York: Broadway Books.

2. Gubser, S. 2010. The Little Book of String Theory. Princeton University Press

3. Hawking, Stephen. 1988. A Brief History of Time. New York: Bantam Books

4. Isaacson, W. 2008. Einstein: His Life and Universe. Simon and Schuster. New York.

5. Kleinman, R. 2006. The Four Faces of the Universe: An Integrated View of the Cosmos. Lotus Press: Twin Lakes.

6. Lorentz, H.A. 1925. The Science of Nature. Vol. 25, p 1008. Springer

7. Malik, P. 2018a. Cosmology of Light. Google Books

8. Malik, P. 2018b. The Emperor's Quantum Computer. Google Books.

9. Malik, P. 2019. The Origin and Possibilities of Genetics. Google Books.

10. Malik, P. Pretorius, L. 2019. An Algorithm for the Emergence of Life Based on a Multi-Layered Symmetry-Based Model of Light. 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). 10.1109/CCWC.2019.8666554

11. Malik, P. 2020. "Light-Based Interpretation of Quanta and its Implications on Quantum Computing," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0719-0726, doi: 10.1109/CCWC47524.2020.9031279.

12. Perkowitz, S. 2011. Slow Light. London: Imperial College Press

13. Stewart, Ian. 2012. In Pursuit of the Unknown. Basic Books. New York.

14. Wilczek, F. 2016. A Beautiful Question: Finding Nature's Deep Design. New York: Penguin Books

15. Wolchover, N. 2013. A Jewel at the Heart of Quantum Physics. Quanta Magazine. https://www.quantamagazine.org/20130917-a-jewel-at-the-heart-of-quantum-physics/

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Evaluation of Moth-Flame Optimization, Genetic and Simulated Annealing tuned PID controller for Steering Control of Autonomous Underwater Vehicle

Sudarshan K. Valluru
Center for Control of Dynamical Systems and Computation
Dept. of Electrical Engineering
Delhi Technological University
Delhi-110042, India
sudarshan_valluru@dce.ac.in

Karan Sehgal
Center for Control of Dynamical Systems and Computation
Dept. of Electrical Engineering
Delhi Technological University
Delhi-110042, India
karansehgal_2k18ee089@dtu.ac.in

Hitesh Thareja
Center for Control of Dynamical Systems and Computation
Dept. of Electrical Engineering
Delhi Technological University
Delhi-110042, India
hiteshthareja_2k18ee084@dtu.ac.in

***Abstract*—This paper describes an optimal bio-inspired PID controller for accurate steering management of the Autonomous Underwater Vehicle system (AUV). To achieve precise control performance, a PID controller is designed, and its gain parameters K_p, K_i, K_d are tuned by applying Simulated Annealing (SA), Genetic Algorithm (GA) and Moth-Flame Optimization Algorithm (MFO). The experimental response corresponding to the unit step and square input waveform for these proposed nature-inspired optimization algorithms were obtained. The response characteristics like overshoot, rise time, settling time and performances index ITAE were calculated and compared. The experimental results show that MFO-PID is highly efficient, followed by GA and SA, respectively.**

Keywords—Moth-flame, GA, SA, AUV, PID, Optimization

I. Introduction

AUV, which stands for an autonomous underwater vehicle, can perform several operations in shallow and deep-sea environments. They have been successfully applied in various fields including military operations, commercial and research purposes etc. Fitted with electronic subsystems, they allow the robot to steer efficiently in harsh surroundings while undergoing the assigned tasks without any human input. This coherent nature of the AUV is due to its six degrees of freedom (DoF). The device's robust interconnection is shown in Fig.1. and the symbolic notations of position and velocity terms of AUV [1] are shown in Table. I. Despite these adroit networks of subsystems, due to the presence of natural and environmental disturbances such as tidal waves and ocean currents, etc., control of such vehicles becomes an arduous task.

For achieving a more potent control, researchers have utilised several intelligent control methods for AUV control [2], [3]. However, controllers like PD, PID have proved to showcase a more straightforward controlling approach. Moreover, such schemes possess more uncomplicated application in implementation from the linear regime's computational point of view [4]. In contrast, however, PID controllers are suffering from tedious computations during the changes in system parameters because of the occurrence of natural perturbations. Many papers are also available in the literature that erudite a controller's application and design, synthesized using a PID control process for the robust steering control of the AUV system [5], [6]. Such controllers are also in demand to control multiple other systems like the trajectory tracking and control of TRMS and Ball beam systems [7], [8].

This paper implements a PID based controller, which is optimally tuned by a bio-inspired meta-heuristic optimization approach referred to as the Moth-Flame Optimization (MFO), Genetic Algorithm (GA) and Simulated Annealing (SA) for the robust control of an AUV. It is found that the performance MFO-PID is better as compared to GA-PID and SA-PID.

This paper is arranged as mathematical modelling of AUV system is explained in section II, followed by section III &IV, which include synthesis of PID based controller applying MFO. Section V gives the experimental responses of the AUV. Eventually, section VI provides the conclusion.

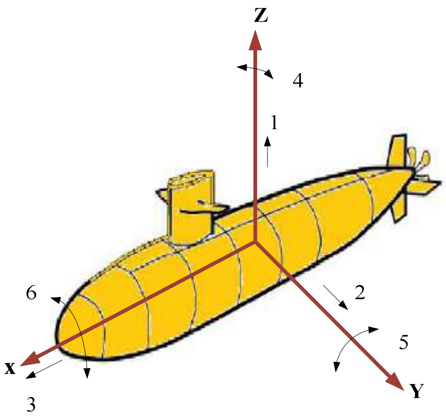


Fig1. Six Degrees of Freedom of an AUV

TABLE I. Notations for AUV Motion

Direction Of Motion	*Moment And Force*	*Earth-Fixed Frame (Position)*	*Body-Fixed Frame(velocity)*
Surge (Motion along $X - axis$)	X	x	u
Sway Motion along $Y - axis$)	Y	y	v
Heave (Motion along $Z{-}axis$)	Z	z	w
Roll (Rotation along $X - axis$)	K	Φ	p
Pitch (Rotation along Y$-axis$)	M	θ	q
Yaw (Rotation along $Z - axis$)	N	Ψ	r

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. AUV MODELLING

To formulate a PID controller for the steering management of an autonomous underwater system, we would first require its general transfer function which gives information regarding the yaw angle, and its relation with the deflection parameter. For this, the mathematical modelling of an AUV [9] is done. This is carried out by considering two different frames of reference, namely earth-fixed and body-fixed frame. System coordinates of this prototype are expounded using three mutually perpendicular axes starting from a random point. North and East correspond to x and y-axis, respectively. Increasing depth conform with the z-axis. The position vector η and velocity vector v can be described by the equations (1) and (2) respectively.

$$v = [u\, v\, w\, p\, q\, r]^T \qquad (1)$$
$$\eta = [x\, y\, z\, \phi\, \theta\, \Psi]^{\mathrm{T}} \qquad (2)$$

In pure steering plane, simplification of these equations is carried out by considering the origin of the body-fixed frame[10] to concur to the center of gravity:

$$m(\dot{v} + u_o r) = \sum Y \qquad (3)$$
$$I_Z \dot{r} = \sum N \qquad (4)$$

Surge speed (u_o), is fixed at 0.75 m/s. Assuming the pitch angle and roll to be small:

$$\Psi = \frac{\sin\phi}{\cos\theta} q + \frac{\cos\phi}{\cos\theta} r \approx r \qquad (5)$$

In matrix form, the equations (1) to (5) are rewritten as (6):

$$\begin{bmatrix} m - Y_{\dot{v}} & -Y_{\dot{r}} & 0 \\ -N_{\dot{v}} & I_{zz} - N_{\dot{r}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{r} \\ \dot{\Psi} \end{bmatrix} + \begin{bmatrix} -Y_v & -Y_r + mv_o & 0 \\ -N_v & -N_r & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} v \\ r \\ \Psi \end{bmatrix} = \begin{bmatrix} Y_{\delta i} \\ N_{\delta i} \\ 0 \end{bmatrix} \delta_r \qquad (6)$$

By substituting the values of vehicle parameter of AUV[11], [12] dimensions, the hydrodynamic coefficient for u_o at 0.75 m/s and by applying the state space approach to get equations (7) to (9)

$$\dot{x}(t) = Cx(t) + Du(t) \qquad (7)$$
$$C = \begin{bmatrix} -0.114 & -0.2647 & 0 \\ 0.0225 & -0.2331 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$
$$D = [0.0211 \text{ -0.0258 } 0]^{\mathrm{T}}$$
$$U = \delta_r \qquad (8)$$

The relation between yaw (ψ) and rudder deflection (δ_r) in terms of the transfer function is acquired as:

$$\frac{\Psi(s)}{\delta_r(s)} = \frac{-0.0258s - 0.0024}{s^3 + 0.3445s^2 + 0.319s} \qquad (9)$$

Now that the system's requisite transfer function has been derived, we can synthesize the proposed compensator.

III. DESIGN OF PID CONTROLLER FOR AUV

PID controller, is a conventional control scheme, is applied extensively for precise control of various systems. Controllers based on PID have been successfully implemented in the design and control of multiple systems[13]–[15]. The PID displays sufficient stability margins, optimum time responses, better system characteristic properties such as low overshoot, and lesser settling time. It consists of proportional, integral and derivative gain parameters which are functions of error between the desired set point and actual system output. The general unity feedback characteristic equation of AUV with PID control laws are written as equations (10), (11) and (12).

$$c(t) = k_p e(t) + k_i D^{-1} e(t) + k_d D e(t) \qquad (10)$$
$$U(s) = \frac{c(s)}{e(s)} = k_p + k_i s^{-1} + k_d s \qquad (11)$$
$$1 + G(s)U(s) = 0 \qquad (12)$$

Now for finding the accurate values of the operational gains(k_p, k_i, k_d), the characteristic equation is optimized based on the Integral Time Absolute Error (ITAE) performance index, i.e., integral of time multiplied by absolute error to minimize the error signal.

$$\text{ITAE} = \int_0^t t|e(t)|dt \qquad (13)$$

PID controller gains are then tuned using MFO, GA and SA to observe their performances for comparison. The block diagram of PID Controller based optimization of AUV using MFO/GA/SA by taking ITAE as the cost function is shown in Fig.2.

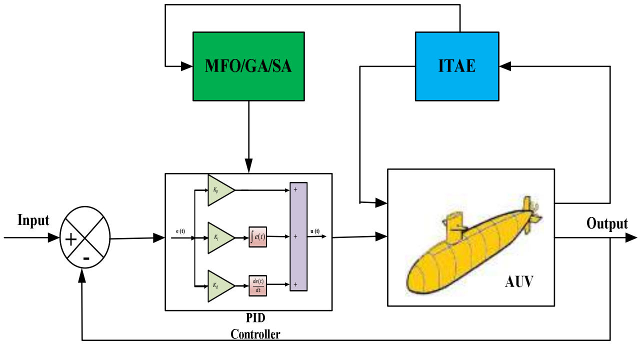


Fig.2. Block Diagram of unity feedback optimal PID control

IV. TUNING OF PID CONTROLLER USING SA, GA AND MFO

PID controllers are commonly used in all dynamical systems, but it requires a monotonous tuning of control actions to avoid sluggishness of the system response. Nature and bio-inspired optimization algorithms can diminish the computational difficulties in the monotonous tuning of PID controllers, thereby the AUV's steering control is with minimum human interventions. Here, the PID controller in AUV's steering loop is tuned by Simulated Annealing, Genetic Algorithm and Moths Flame Optimizer methods.

A. Simulated Annealing Tuned PID Controller

Simulated Annealing is one of the most widely used methods for optimizing control problems of dynamical systems. This algorithmic technique is inspired by the relationship of combinational optimization and quantum and classical or statistical mechanics laws. A simple mechanism of cooling of material is applied in steps till the lowest energy is reached. The state at this lowest energy is the optimized state. Pseudocode for the SA algorithm applied for tuning the gain parameters of the proposed PID controller is given as:

```
Step 1: The ranges for the 3 gain coefficients of a PID control are fixed in the form of an objective function $f(x) = [K_p, K_i, K_d]^t$
Step 2: Set $t_o$ as the temperature at the beginning of the process
Step 3: Take $s_i$ as the initial stage of the system
Step 4: Take $s_{opt}$ as the desired optimized state of the system
Step 5: Initially, assign t as $t_o$ and $s_{opt}$ as $s_i$
Step 6: **for** $t = 1$ to $t_{max}$ **do**
        Assign $s_{opt+1}$ to adjacent/nearest $(s_{opt})$
        Change $\Delta E$ to $(f(s_{opt+1})$-$f(s_{opt}))$
        **if** $min\ (1, e^{-\Delta E/t}) \geq$ random (0,1) **then**
                Update $s_{opt}$ to $s_{opt+1}$
        **end if**
        Assign t as the temperature-schedule(t)
**end for**
Step 7: Output the final solution of the optimized function
```

The values for the 3 gain coefficients Kp, Ki and Kd obtained using SA method are -19.9298, -4.9419 and -41.25 respectively. However, the simulated annealing technique poses a critical drawback while working on minimization problems. For instance, any change in the system values that decrease the cost function 'f ' will be accepted as desired, but sometimes, changes that increase 'f' might also get counted. This happens with a probability p, known as the transition probability. Due to this disadvantage offered by SA, we use GA to stabilize the system.

B. Genetic Algorithm Tuned PID Controller

Genetic algorithm (GA) is a nature-based optimization technique applied to solve computational problems. It is used on a chromosome population where every chromosome represents a solution that has an associated fitness value to it. This value defines how optimal a solution is. Some arbitrarily generated population is initially taken, followed by the selection process, which is fitness based. The next step involves recombination to develop the next generation. For the above step, parent genes are used to obtain child chromosomes. Several iterative processes continue till the stopping criteria is achieved. Pseudocode for the genetic algorithm is:

```
Step 1: Designate the ranges for the control parameters of the PID controller to model an objective function as- $f(x) = [K_p, K_i, K_d]^t$
Step 2: Generation of initial source population of $M$ chromosomes randomly.
Step 3: Calculation of the fitness $F$ using Eq. (14)
Step 4: **for** $i = 1: n$
        Take two chromosomes from the current population.
        Crossover is applied to chromosomes with crossover rate $x$.
        Mutation is applied to the chromosome with mutation rate $m$ to generate a new chromosome.
        Add the above generated chromosome to next generation population.
        Current Population is replaced with next generation population.
    **end for**
Step 5: Finally, output the solution of the optimized global best.
```

One practical step for applying GA is to evaluate the fitness value for all the chromosomes to reduce the error signal $e(s)$. PID controller is applied to minimize this error. As this fitness value is inversely proportional to the value of the performance index, we can define the chromosomes' fitness as (14).

$$F = \frac{1}{Performance\ Index} \tag{14}$$

The tuned optimized parametersK_p, K_i and K_d, using GA are - 17.4092, -1.8497 and -38.4368, respectively. The response obtained using GA shows an overshoot of more than 10%. This result also shows that GA does not have a high speed of convergence that is why an alternative algorithm, designed using evolutionary strategies can be deployed to obtain a faster and more efficient performance of the system. The moth flame optimization algorithm (MFO) is one such technique.

C. Moths Flame Optimization Algorithm tuned PID

MFO is a newly developed nature-based solution finding mechanism made from the algorithms inspired by the population search strategy. This bio-inspired optimization algorithm is very flexible, and can easily be implemented in finding the optimal solution of real-world problems. Some applications of MFO are available in literature ranging from the tuning of controllers such as fuzzy-PID, fuzzy-PI, PID, PI. Moth–flame optimization (MFO) technique was introduced by Mirjalili[16]. It initiates by creating moths arbitrarily inside the solution region by a transverse orientation shown in Fig.3.

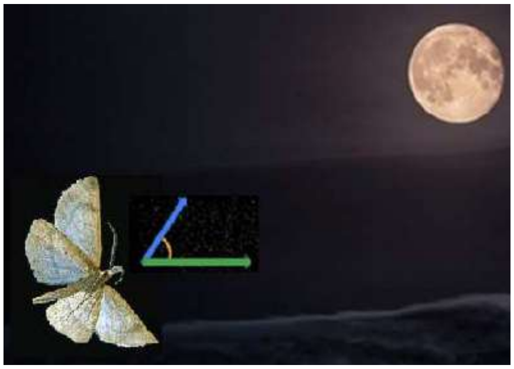


Fig.3 Moth's Transverse Orientation

After this, the fitness function values for every moth are calculated and tagged the most optimum position by flame. Then, depending on the spiral movement function, the moths' positions are modified to attain more acceptable positions sorted by a flame, the new most acceptable positions of individuals are upgraded, and replicating the previous operations.

This process takes place until the total number of iterations have been completed. The MFO algorithm has three main postulates. These are given below:

1.Initializing Population:

It is taken that all moths can fly in all the dimensions. The moth set can be represented as:

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,d} \\ b_{2,1} & \cdots & \cdots & \cdots & b_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{x,1} & b_{x,2} & \cdots & \cdots & b_{x,d} \end{bmatrix} \qquad (15)$$

Where x stands for the number of moths and d gives the number of dimensions in the solution region.

Array to store the values of the fitness function is as follows:

$$OB = \begin{bmatrix} OB_1 \\ OB_2 \\ \vdots \\ OB_x \end{bmatrix} \qquad (16)$$

The given matrix defines the flames in the d-dimensional region, and their fitness value vector follows it.

$$F = \begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & \cdots & F_{1,d} \\ F_{2,1} & \cdots & \cdots & \cdots & F_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{x,1} & F_{x,2} & \cdots & \cdots & F_{x,d} \end{bmatrix} \qquad (17)$$

$$OF = \begin{bmatrix} OF_1 \\ OF_2 \\ \vdots \\ OF_x \end{bmatrix} \qquad (18)$$

The solutions are moths and flames. What distinguishes them is how we treat and update them after every iterative step. Actual searching agents which move throughout the search area are moths. Flames are optimal positions of moths which have been derived until now.

2.Updating Moths' Position:

The optimal global value can be obtained for the optimization problem; this algorithm employs three steps. These are given below:

$$MFO = (I, P, T) \qquad (19)$$

where I describe the function, which gives the first moth population randomly

$I: \phi \rightarrow \{B, OB\}$

P signifies the moths' movement in the search area

$P: B \rightarrow B$

T is the condition for termination

$T: B \rightarrow true, false$

The equation below, explains the function I, which applies random distribution.

$$\mathrm{M(p,q)} = \big(\mathrm{ub(p)} - \mathrm{lb(q)}\big) * \mathrm{rand(\,)} + lb(p) \qquad (20)$$

Where lb and ub represent the lower and the upper bounds, respectively.

Three conditions which should be followed while applying a logarithmic spiral are:

- The starting point of the spiral should begin from the moth.
- The ending point of the spiral should be in the flame position.
- Spiral's range fluctuation is not supposed to surpass the search space.

$$S\big(B_p, F_q\big) = D_p \cdot e^{bt} \cdot \cos(2\Pi t) + F_q \qquad (21)$$

Where D_p represents the region within p^{th} moth and q^{th} flame.

$$Dp = |Fq - Bp| \qquad (22)$$

Bp describes the logarithmic spiral's shape, and t stands for any arbitrary value within $[r, 1]$. The spiral motion guarantees the balance between exploitation and exploration close to the flame. Where r changes from [-1,2] in the whole process of iteration, which is called as the convergence constant. The logarithmic spiral shape, as described above, is shown in Fig.4.

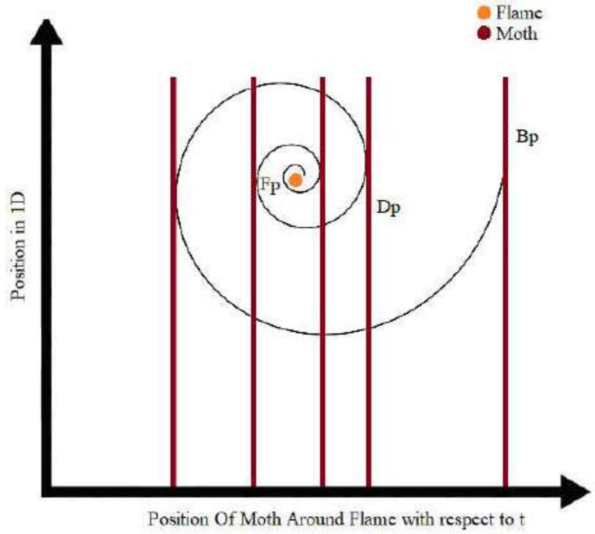


Fig.4. Logarithmic spiral shape

3.Mistakes Update in Flames

Moth positions are updated in n different locations inside the search region, which might reduce the exploitation of the most optimized solutions. Thus, minimizing the flames solves the conflict using an equation given below:

$$flame\ no. = round\left(N - l * \frac{N-1}{T}\right) \quad (23)$$

N represents the total flames
l represents the present iteration
T represents the total iterations
Pseudocode for the MFO algorithm is described as:

```
Step 1: Assign the fixed ranges for each control parameter of the proposed PID controller in the form of a cost function as- $f(x) = [K_p, K_i, K_d]^t$
Step 2: Initialize the Moth-Flame population
Step 3: Initialize position of moth $B$ arbitrarily
Step 4: for $p = 1\ to\ n$ do
            Calculate the value of fitness function F
        end for
Step 5: While iterations<=total iterations do
            Update flame no by using Eq. (23)
            $OB$ = Fitness Func($B$);
            if iteration == 1
            $F$ = sort($B$);
            $OF$ = sort ($OB$);
            else
            $F$ = sort ($B_{t-1}$, $B_t$);
            $OF$ = sort ($B_{t-1}$, $B_t$);
            end
Step 6: for $p = 1:n$
            for $q = 1:d$
            Update the values of $r$ and $t$
            Calculate the value $D$ using Eq. (22)
            Update $S(p,q)$ using Eq. (21)
            end
end
Step 7: Consequently, the precisely tuned solution of the optimized cost function is output.
```

The tuned optimally calculated parametersK_p, K_i and K_d, by using MFO are -15.5343, --0.025 and -38.93, respectively.

V. Simulation and Results

For comparison of the optimization performances of Moth-Flame Optimization (MFO), Genetic Algorithm (GA) and Simulated Annealing (SA) method, the output response of AUV, controlled by tuned PID controllers are observed for unit step and unit square wave input. In the square wave input case, the chosen frequency for simulation was fixed at 15 mHz.

The tuned PID controller's step and square wave responses for MFO-PID, GA-PID and SA-PID controllers for AUV steering control are shown in Fig.5. and Fig.6 respectively.

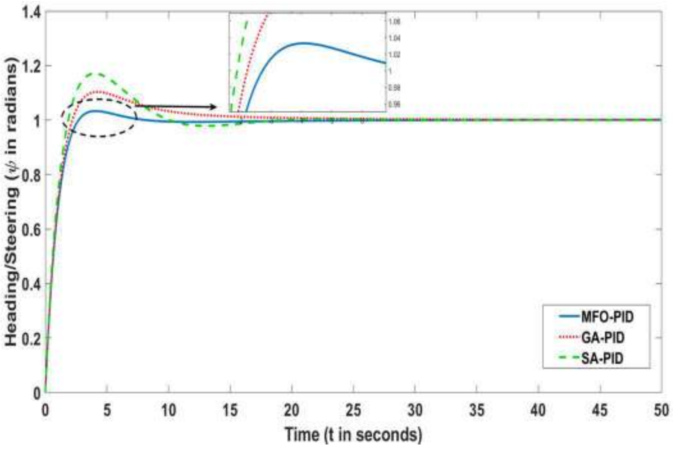


Fig.5. Unit Step Responses by the respective tuned controllers

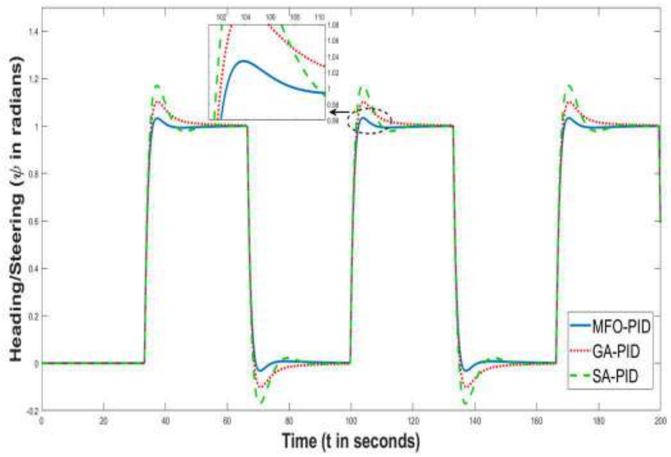


Fig.6. Square Responses by the respective tuned controllers

The precisely optimized solutions of the gain parameters are given in Table II.

TABLE II. GAIN PARAMETERS OF TUNED PID CONTROLLER

Gain Parameters			
PID Parameters	***SA***	***GA***	***MFO***
K_p	-19.9298	-17.4092	-15.5343
K_i	-4.9419	-1.8497	-0.025
K_d	-41.25	-38.4368	-38.93

The PID transient performance characteristics using MFO, GA and SA are given in Table III.

TABLE III. TRANSIENT PERFORMANCE OF TUNED PID CONTROLLER

ITAE			
Tuning Algorithm	***SA***	***GA***	***MFO***
Rise Time (Sec)	1.37	1.56	1.72
%Overshoot	17.1	10.3	3.3
Settling Time (Sec)	14.1	12.4	5.7

The cost comparison amongst MFO, GA and SA are given in Table IV.

TABLE IV. COST COMPARISON

ITAE ERROR			
Tuning Algorithm	***SA***	***GA***	***MFO***
Error(J)	0.6396	0.4371	0.2066

It is observed from Table III that the MFO tuned PID controller has remarkable performance as compared to GA tuned PID and SA tuned PID in terms of lesser overshoot and settling time. It proves that the MFO tuned PID for AUV steering control is optimally and accurately meeting the design objectives. It is also noted from Table IV that the cost and error are drastically reduced in contrast to GA tuned PID and SA tuned PID controllers.

VI. CONCLUSION

To obtain the robust control for an autonomous underwater vehicle, this paper presents the application of a bio-inspired optimization algorithm called Moth-Flame Optimization to optimally tune the gain parameters of a PID based controller for efficient motion stabilization of the AUV system. Tuning of the proposed PID controller is done for an error-based performance index ITAE. Genetic Algorithm and Simulated Annealing Method are also used to compare the system's step and square responses. The response obtained for MFO-PID is clearly, better than GA-PID, followed by SA-PID in case of overshoot, settling time and rise time. The scope of future work can be the application of this algorithm to design more complex controllers for advanced systems.

REFERENCES

[1] M. A. Abkowitz, *Stability and Motion Control of Ocean Vehicles*. The MIT Press, 1969.

[2] J. Lorentz *et al.*, “A fuzzy rule-based algorithm to train perceptrons,” *IEEE J. Ocean. Eng.*, vol. 19, no. 4, pp. 359–367, Nov. 2020, doi: 10.1016/S0165-0114(03)00242-2.

[3] A. J. Healey and D. Lienard, “Multivariable Sliding-Mode Control for Autonomous Driving and Steering of Unmanned Underwater Vehicles,” *IEEE J. Ocean. Eng.*, vol. 18, no. 3, pp. 327–339, 1993, doi: 10.1109/JOE.1993.236372.

[4] Astrom. K.J. and Hagglund.T, *PID controllers: theory design and tuning*. North Carolina, USA: Instrument Society of America, Research Triangle Park, 1995.

[5] S. P. Hou and C. C. Cheah, “Can a simple control scheme work for a formation control of multiple autonomous underwater vehicles?,” *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 5, pp. 1090–1101, 2011, doi: 10.1109/TCST.2010.2076388.

[6] F. Kong, Y. Guo, and W. Lyu, “Dynamics Modeling and Motion Control of an New Unmanned Underwater Vehicle,” *IEEE Access*, vol. 8, pp. 30119–30126, 2020, doi: 10.1109/ACCESS.2020.2972336.

[7] S. . Valluru, M. Singh, Ayush, and A. Dharavath, “Design and Experimental Implementation of Multi-loop LQR, PID, and LQG Controllers for the Trajectory Tracking Control of Twin Rotor MIMO System.,” in *Intelligent Communication, Control and Devices.*, K. A. Choudhury S., Mishra R., Mishra R., Ed. Springer Nature Singapore Pte Ltd, 2019, pp. 599–608.

[8] S. K. Valluru, M. Singh, and S. Singh, “Prototype Design and Analysis of Controllers for One Dimensional Ball and Beam System,” in *1st IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems*, 2016, pp.1–6.

[9] K. P. Valavanis, D. Gracanin, M. Matijasevic, R. Kolluru, and G. A. Demetriou, “Control Architectures for Autonomous Underwater Vehicles,” *IEEE Control Systems Magazine*, vol. 17, no. 6, pp. 48–64, 1997.

[10] J. A. Monroy, E. Campos, and J. A. Torres, “Attitude control of a Micro AUV through an embedded system,” *IEEE Lat. Am. Trans.*, vol. 15, no. 4, pp. 603–612, 2017, doi: 10.1109/TLA.2017.7896344.

[11] B.Jalving, “The NDRE-AUV Flight Control System,” *IEEE J. Ocean. Eng.*, vol. 19, no. 4, pp. 497–501, 2007.

[12] A.Healey and Marco D.B, “Experimental verification of mission planning by autonomous mission execution and data visualization using the NPS AUV II,” in *IEEE Symposium on Autonomous Underwater Vehicle Technology*, 1992, pp. 65–72, doi: 10.1109/AUV.1992.225193.

[13] Sudarshan. K.Valluru, R. Kumar, and R. Kumar, “Design and Implementation of L-PID and IO-PID Controllers for Twin Rotor MIMO System,” in *IEEE International Conference on Power Electronics, Control and Automation (ICPECA)*, 2019, pp. 1–5, doi: 10.1109/icpeca47973.2019.8975542.

[14] S. K.Valluru and M. Singh, “Performance investigations of APSO tuned linear and nonlinear PID controllers for a nonlinear dynamical system,” *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 442–452, 2018, doi: 10.1016/j.jesit.2018.02.001.

[15] S. K. Valluru, M. Singh, M. Singh, and V. Khattar, “Experimental Validation of PID and LQR Control Techniques for Stabilization of Cart Inverted Pendulum System,” in *3rd IEEE International conference on Recent Trends in Electronics, Information and Communication Technology*, 2018, pp. 708–712.

[16] S. Mirjalili, “Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm,” *Knowledge-Based Syst.*, vol. 89, pp. 228–249, 2015, doi: 10.1016/j.knosys.2015.07.006.

How Stability of Hybrid Coupler Characteristic Affects Front-End Isolation of In-Band Full Duplex System

Soheyl Soodmand
dept.electrical & electronic engineering
university of bristol
Bristol, United Kingdom
soheyl.soodmand@bristol.ac.uk

Kevin A. Morris
dept.electrical & electronic engineering
university of bristol
Bristol, United Kingdom
kevin.morris@bristol.ac.uk

Mark A. Beach
dept.electrical & electronic engineering
university of bristol organization
Bristol, United Kingdom
m.a.beach@bristol.ac.uk

***Abstract*— In In-Band Full Duplex (IBFD) transceivers, Electrical Balance Duplexers (EBDs) provide Transmit (TX)-Receive (RX) isolation to implement a form of self-interference (SI) cancellation to facilitates simultaneous transmission and reception from a single antenna. EBD works by coupling transmitter, receiver, antenna, and balancing impedance using a hybrid coupler. The balancing impedance in the EBD needs to be equal as much as possible to the antenna impedance to achieve a high isolation while the antenna impedance variations limit the isolation bandwidth. Hybrid couplers are also not ideal elements, and their S parameters are variable in frequency domain. An antenna with a stable impedance, designed by authors, has been connected in this work to two different hybrid couplers in the EBD stages. One coupler is a commercial coupler and the other one, designed by the authors, has more S parameter stability in the frequency domain than the commercial one. It is shown that by using the designed coupler in the front-end EBD stage more stable isolation bandwidth and 45% better isolation value at ultra-wideband range of 1.5-3.5GHz are obtained, in comparison when using the commercial one.**

Keywords—duplexers, in-band full duplex, self-interference cancellation, couplers, antennas

I. Introduction

In-band full-duplex (IBFD) systems can theoretically double link capacity of Time Division Duplex (TDD) and Frequency Division Duplex (FDD) systems thus allowing simultaneous transmission and reception on the same frequency and reduce wireless latency [1], [2]. Transmitting and receiving on the same time-frequency resource results in strong co-channel Self-Interference (SI), which can be more powerful than the desired receive signal [2]. Any residual SI, due to unsuitable transmit (Tx) to receive (Rx) isolation, will effectively increase the receiver noise floor therefore reduces the capacity of the receive channel [3].

Existing IBFD designs [4], [5] involve various combinations of analog cancellation, digital cancellation, and antenna based suppression to provide the required isolation. Digital cancellation [6] cannot properly prevent SI from overloading the receiver. Analog cancellation [5], [6], can provide significant isolation and prevents receiver overloading in most cases [5]. Antenna based suppression and separation methods can provide significant isolation also, however these designs require additional antennas [6]. Single antenna full duplex systems in [5] use circulators to provide some level of Tx-Rx isolation but these are unattractive options due to their cost, size, and limited bandwidth. New duplexers based on SI Cancellation (SIC) at the receiver have received substantial interest to enable IBFD operation [1], [5]–[8].

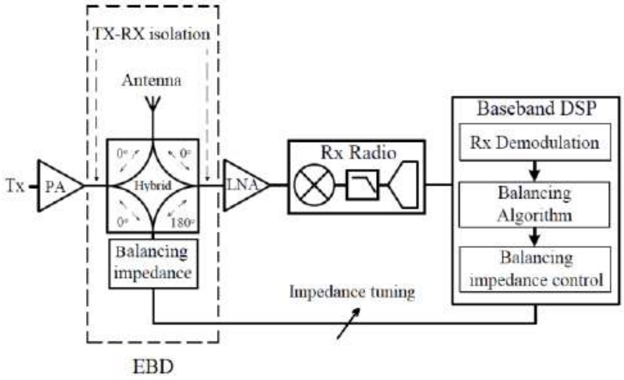


Fig. 1. EBD stage of IBFD system with adaptive balancing impedance

Recent results [3], [7], [9]–[11] have demonstrated that Electrical Balance Duplexers (EBDs), which is of interest as the first stage of passive Radio Frequency (RF) front-end cancellation in IBFD transceivers, implement a form of SIC in order to provide high transmit to receive isolation whilst facilitating simultaneous transmission and reception from a single antenna. EBD could potentially be combined with analog cancellation, digital cancellation, and full duplex MIMO technology [4]. The analog circuit technique of EBD works by coupling the transmitter, receiver and antenna using a four-port hybrid junction, along with a balancing impedance connected to the fourth port, Fig. 1. Using a suitable hybrid coupler, a high transmit-to-receive isolation is expected in the EBD stage when the balancing impedance is equal to the antenna impedance at all frequencies within the aimed bandwidth. However, in practice, the antenna impedance is not an ideal 50 Ω resistor but has a complex impedance having real and imaginary parts. This exhibits variations with respect to frequency so the bandwidth and value of the Tx-Rx isolation will be limited by the impedance mismatching between the antenna and balancing impedance.

Measured real antenna data in [3], [8], and results for a prototype EDB in [12], demonstrate that the variation in antenna impedance significantly reduces the isolation bandwidth. Evaluations that include measured antenna data in the EBD in [7], [9], [11] shows a mean isolation of 35 dB over a 20-MHz bandwidth at 1.9 GHz but poor performance of wider bandwidths, again said because of antenna impedance variations. A Micro- ElectroMechanical Systems (MEMS) implementation of tuneable balancing impedance of the EBD is presented in [13], balancing at 800 MHz and 1900 MHz to provide 43 dB isolation over a 20 MHz bandwidth at each frequency but introduce non-linear distortion into the system. Consequently, to maintain Tx-Rx isolation, the balancing

This research is supported in part by the UK Engineering and Physical Sciences Research Council (EP/K503824/1), and in part by u-blox AG.

978-1-6654-4067-7/21/$31.00 ©2021IEEE

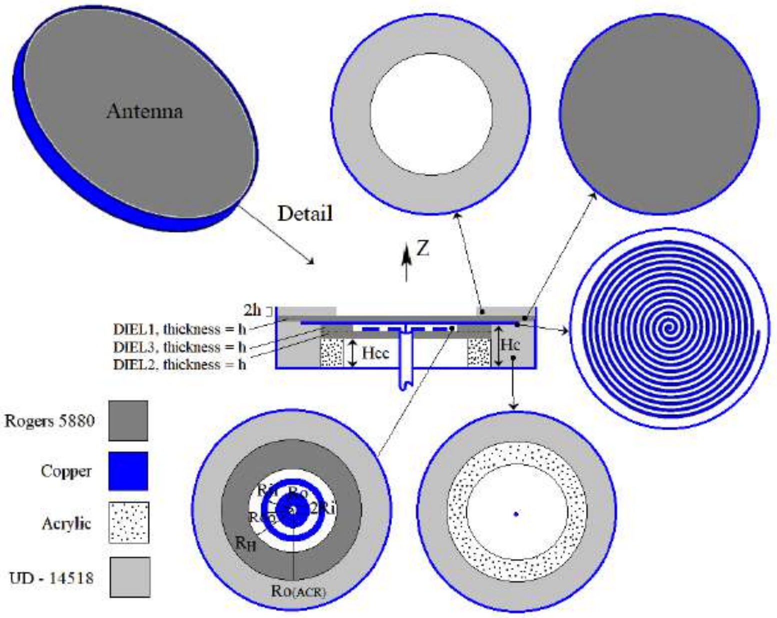


Fig. 2. Designed antenna with stable impedance, Section III.A (whole and cross-section views),

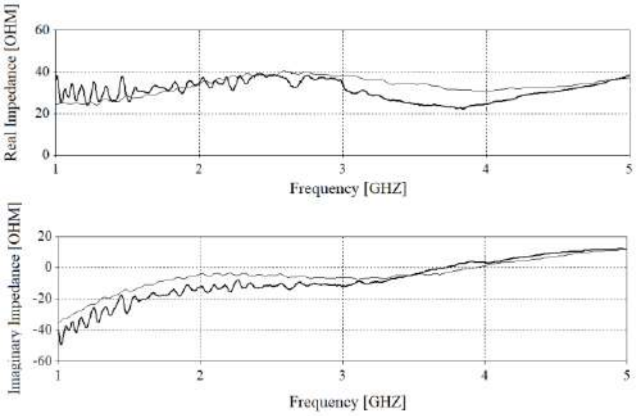


Fig. 3. Measured and simulated impedance of the antenna in Fig. 2.

impedance must be tuneable as it tracks and mimics the antenna impedance as it changes. This requires an adaptive architecture [9], [11], Fig. 1, using a balancing algorithm which extremely limits both the mimic and the isolation bandwidths.

In all above scenarios in the literature, analyses consider the antenna impedance variations as the main factor limiting the SIC when the antenna is connected to a hybrid coupler. The main aim in this work is to investigate how the EBD isolation is affected by stability of coupler S parameters in the frequency domain. In the rest of this paper, in Section II, Tx-Rx gain function of EBD circuitry is briefly reviewed to assess EBD isolation in generic form. Used devices in the experiments of this work (stable impedance antenna, stable S parameter coupler, commercial coupler, and balancing load) are introduced in Section III. Stable impedance antenna and stable S parameter coupler are designed by authors. In Section IV, EBD isolations are measured when the two couplers are connected to the stable impedance antenna in two separate experiments. In comparison when using the commercial coupler in the EBD, it is shown that the EBD isolation bandwidth is increased, and the isolation value is optimized when using the designed coupler with higher S parameter stability. Finally, a conclusion is given in Section V.

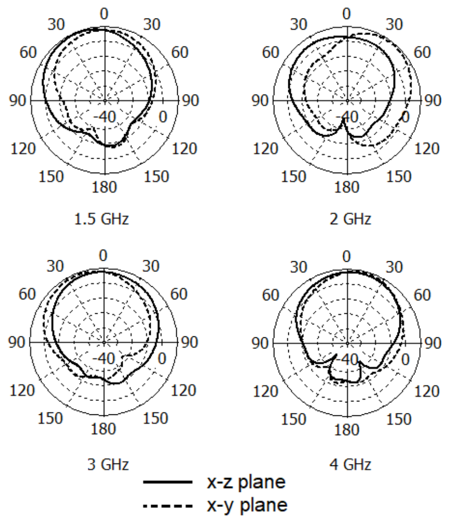


Fig. 4. Measured directivity pattern of the antenna in Fig.2.

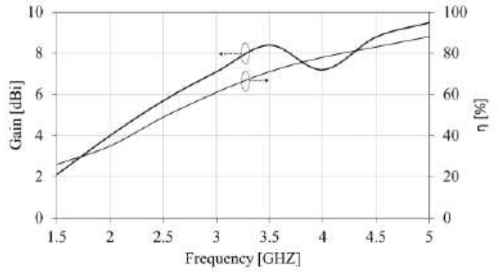


Fig. 5. Measured gain and radiation efficiency of the antenna in Fig. 2.

II. Tx-Rx Gain of Electrical Balance Duplexers

Tx-Rx gain of an symmetrical EBD is given [10] by:

$$G_{Tx-Rx}(\omega) = L\,|\Gamma_{BAL}(\omega) - \Gamma_{ANT}(\omega)|^2. \qquad (1)$$

where Γ_{BAL} and Γ_{ANT} are the complex reflection coefficients of the balancing impedance and antenna impedance,

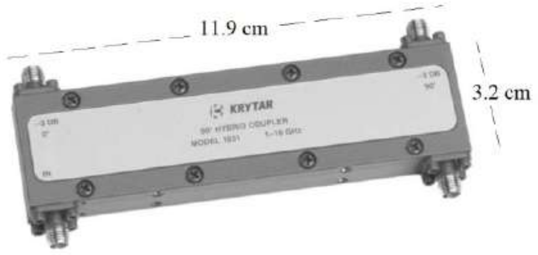


Fig. 6. Commercial 3-dB coupler, Model Krytar1831 [18] (Section III. B)

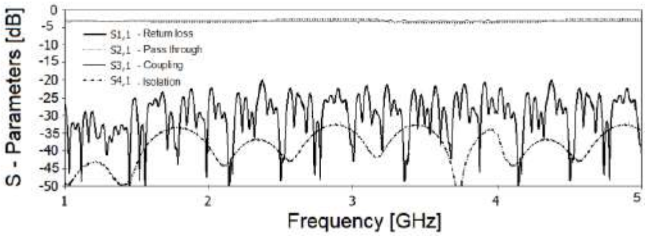


Fig. 7. Measured S parameters of the 3- dB Krytar1831 coupler in Fig. 6 (Section III. B).

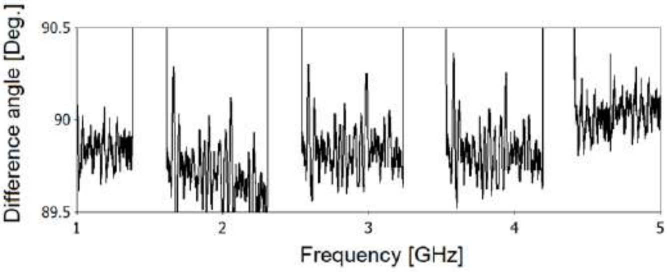


Fig. 8. Measured phase characteristic of the 3-dB Krytar1831 coupler in Fig. 6 (Section III. B), Zoomed view

respectively (see Fig.1), and for an ideal (lossless) hybrid L = ¼. As we can see from (1), the Tx-Rx gain is theoretically zero when the balancing reflection coefficient and antenna reflection coefficient are equal at the carrier frequency. To obtain maximum (theoretically infinite) Tx-Rx isolation over a given frequency band, the balancing impedance must be equal to the antenna impedance at all frequencies within that band such that $\Gamma_{BAL}(\omega) = \Gamma_{ANT}(\omega)$ for $\omega_l<\omega<\omega_h$ where ω_l and ω_h are the lower and upper limits of the band of interest, respectively. Also, coupler characteristics is a determining parameter to calculate the Tx-Rx gain. Results in [10] demonstrate that (1) remains valid for non-ideal circuit.

III. Devices

In this section, four RF elements used in the paper experiments are introduced. First device is an antenna with a stable impedance. More details about the antenna design are given in the authors other papers in [14] and [15]. Also, two couplers are introduced here, first one is a commercial coupler, and the second coupler is designed by the authors. The second coupler has more S parameter stability in the frequency domain than the first one. More details about design procedure of the second coupler are given in author’s separate paper. Also, a wide band 50-ohm RF load is used as the balancing impedance.

A. Antenna with Stable Impedance - Designed

A single arm Archimedean Spiral (AS) is backed by a cavity, Fig. 2, to have an antenna with a unidirectional beam. The AS is formed as a conducting spiral strip arm of width w =2 mm on bottom of a disc shaped Rogers RT 5880 substrate - so called DIEL1- of radius A_C = 62 mm with a dielectric constant of ε_r = 2.2, loss tangent of 0.0009, h=0.79-mm thickness plus a 9 μm copper coating. DIEL1 has no copper on the top. The centerline of the spiral arm is defined by the function of $r = K\,\Phi$, where K is a constant and Φ is the winding angle, ranging from starting angle $\Phi_S = 0.07\pi$ Rad to ending angle $\Phi_E = 28\pi$ Rad. The antenna circumference C is defined by $C = 2\pi\, R_{max}$ with $R_{max} = K\, \Phi_E$ where K = 0.64 mm/rad, R_{max}= 56 mm. The cavity radius, A_C, has 6 mm distance from the arm end to the cavity wall as A_C = 62 mm= 0.31 λ_L whilst λ_L is the wavelength at the lowest design frequency of 1.5 GHz. The distance between the bottom of the cavity and the spiral on the bottom of DIEL1 substrate is considered as H_C = 6.9 mm = 0.035 λ_L. The height of the copper case wall is chosen as H_C + 2h to surround the DIEL1 disc substrate while the case has a 1 mm uniform thickness of copper.

Reflected fields from bottom of the case are attenuated in the designed antenna using electromagnetic absorbers (EMAs) made up of model UD-14518 of ARC Technology to make the antenna impedance more stable in an UWB frequency range. The UD-14518 specified by relative permittivity of ε_r=ε’-jε” ~ 22+3j and relative permeability of μ_r=μ’-jμ”~ 4.5+2j at aimed the bandwidth [16]. As shown in Fig. 2, a ring-shaped Electromagnetic Absorber (EMA) strip - so called EMA1- of optimized width 11mm and with the same height of copper case H_c = 6.9 mm is placed under the antenna arms. Also, second EMA, so called EMA2, is added to above the DIEL1 substrate to more improve impedance stability. EMA2 has an optimized thickness of 2h = 1.6 mm while its outer radius is equal to the cavity diameter of A_C and its inner radius is A_C -17mm = 45mm.

A capacitive impedance matching using two cocentric planar copper rings, Fig. 2, is also used to improve impedance stability. Maximum bandwidth for a stable impedance is achieved when air with dielectric constant of ε_r =1 is considered between the copper rings and the spiral. As seen, the two cocentric planar copper rings are considered on the top surface of a disc shape dielectric material, so called DIEL2, with a distance of h = 0.79 mm between the copper rings and the spiral body. DIEL2 with a thickness of h is made up of the same material of Rogers RT 5880. DIEL2 has a small hole with R_i = 1.7mm radius in the centre for passing the coaxial cable whilst the outer radius of DIEL2 is 41mm. Also, a planar dielectric ring, so called DIEL3, with the same thickness and the same material with DIEL2 is used as a spacer between DIEL2 and DIEL1. DIEL3 has a hole with radius of R_H = 35mm in the centre whilst its outer radius is equal with the outer radius of DIEL2. The internal copper ring has inner radius of R_i=1.7mm and outer radius of R_O= 5.5 mm while the external ring has inner radius of R_{ii} = 8mm and outer radius of R_{OO} =11 mm. To hold the substrates and spiral a cylindrical ring is 3D printed on Acrylic material with a dielectric constant of $\varepsilon_{r(ACR)}$ = 3.5, internal radius of R_{iA} = of 41 mm, outer radius of R_{oA} = 51 mm and height of $H_{CC} = H_C$ -2h = 5.3mm. The antenna is fed by a 50 Ω coaxial cable whereas the inner conductor of the cable is connected to the AS at its starting angle and outer ground conductor of the cable is connected to the inner radius of internal capacitive ring at the top surface of DIEL3.

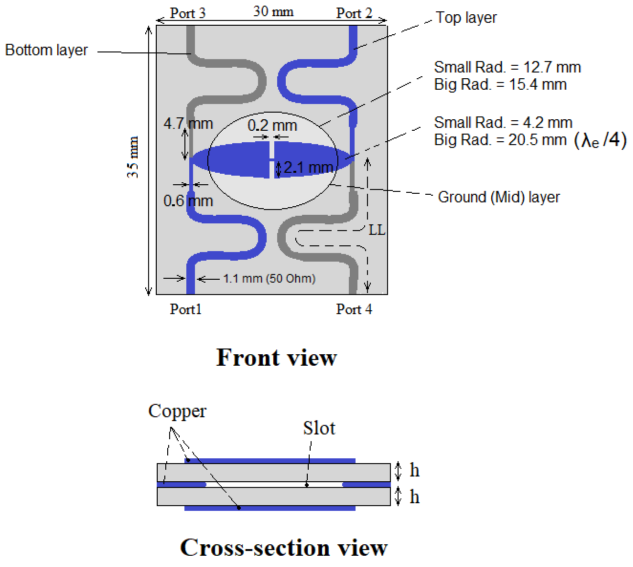


Fig. 9. Designed 3-dB coupler with stable S parameter (Section III. C)

The Finite Integration Technique (FIT) with high meshing in CST MICROWAVE STUDIO [17] software is used for simulations. Measured and simulated antenna impedances with a stable impedance at 1.5 GHz to 3.5 GHz range are shown in Fig. 3, with an average variation of about 15 Ω for the measured imaginary and real impedance. Also, measured directivity pattern of the antenna and its measured gain/efficiency vs frequency plots are shown respectively in Fig. 4 and Fig. 5. The antenna has a measured Axial Ratio (AR) below 3 dB indicating a circular polarization over 1.6 GHz (AR plot is not shown here). These results indicate good impedance stability at 1.5 - 3.5 GHz range for the designed antenna while the antenna also has a suitable radiation performance at this Ultra-wide Band (UWB) bandwidth. The fabricated antenna is shown in Section IV.

B. *Commercial 3-dB Coupler (Krytar Model 1831)*

Krytar coupler model 1831 is an UWB commercially available coupler [18], Fig. 6. Measurement results of the coupler by Vector Network Analyzer (VNA) are given in Fig. 7 and Fig. 8. As seen, this commercial coupler features UWB characteristics with the coupling of 3 ± 1 dB at 1-5 GHz bandwidth. Also, isolation in the order of better than 30 dB and return loss in the order of better than 20 dB are measured. It is seen that the measured S parameters are not very stable in the frequency domain as there are dense variations around 15 dB for isolation and 25 dB for return loss at 1-3.5 GHz range. In Fig. 8, it is observed that the measured phase difference between output ports is dominantly around 90^o over the bandwidth, needed for a suitable quadrature hybrid coupler.

C. *3-dB coupler with Stable S Parameter - Designed*

Proposed miniaturized coupler capable of providing tight coupling over an UWB band, is shown in Fig. 9 where its coupling mechanism is similar to [19] and [20]. The differences for the proposed model concern the shaping factor of the broadside coupled strips, the slot, electrical size and most importantly the operation frequency which is in the ultra-high frequency band. This coupler consists of three conductor copper layers which are interleaved by two dielectric substrates. The top copper layer includes ports 1 and 2. The bottom copper layer is similar to the top layer, but the ports here are ports 3 and 4. Ports 3 and 4 are on opposite sides of the substrate compared to ports 1 and 2. The two layers are coupled via a slot, which is made in the copper layer supporting the top and bottom dielectrics. As seen in Fig. 9, the two microstrip conductors and the slot are of an elliptical shape. The 50 Ω microstrip lines are included to make connections to SMA ports. The structure features double symmetry with respect to both horizontal and vertical plans.

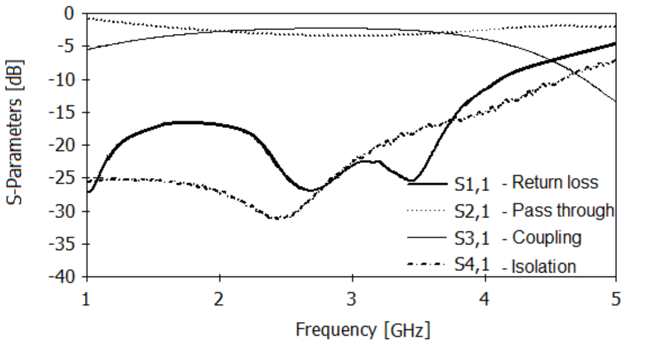


Fig. 10. Fig. 7. Measured S parameters of the designed 3-dB coupler with stable S parameter in Fig. 9 (Section III. C).

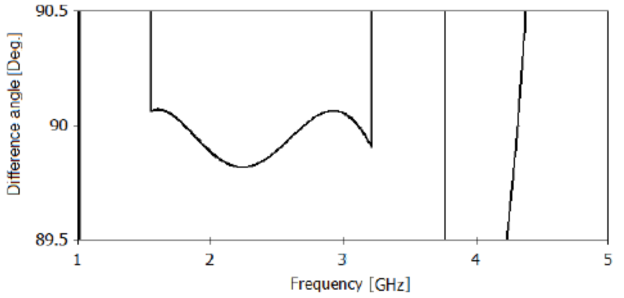


Fig. 11. Measured phase characteristic of the designed 3-dB coupler with stable S parameter in Fig. 9 (Section III. C), Zoomed view

Fig. 12. Wideband 50-ohm RF Load (Section III. D)

Analysis starts in a similar way to the ones described in [21] for the equivalent rectangular shaped microstrips. If characteristic impedance of the microstrip ports of the coupler is $Z_0 = 50\,\Omega$ and the coupling factor is $C_{\text{dB}} = 3$ dB, the values of even mode characteristic impedance Z_{ev} and odd mode characteristic impedance Z_{od} are calculated as 175.5 Ω and 14.2 Ω, respectively.

The validity of the presented design is tested in the 1.5–3.2 GHz frequency band where the centre frequency of operation is 2.35 GHz. A Rogers RO4350B substrate with a dielectric constant of 3.48 and a loss tangent of 0.0037, h = 0.51mm thickness, plus 35- μm-thick conductive copper is used for the coupler development. The elliptical body length is chosen as 20.5 mm ~ λ/6 ~ $\lambda_e/4$ where λ_e is effective wavelength and λ is free space wavelength at the central frequency of 2.35 GHz. The return loss, coupling, and isolation of the designed coupler are first verified by running high mesh FIT in the CST software and the final obtained dimensions are shown in Fig. 9.

Four phase shifters each having length of LL = 33 mm= $0.4\ \lambda_e$ are added to the terminals of elliptical bodies to adjust output phase difference to 90^o, needed for a suitable quadrature hybrid coupler, Fig. 9. To maintain a compact size, phase shifters are formed as curved microstrip lines. Simulations do not show considerable differences in results when the curves angle are not smaller than 75 degree. A combination of impedance matching technique and structural modifications also have been employed to optimize the coupler results. The impedance matching is carried out in a similar way for all four ports by narrowing the width of tracks which connect the ports to the elliptical body as shown in Fig. 9. Also, two narrow slots have been etched on each elliptical body as shown in the same figure. Measurement results by VNA are given in Fig. 10 and Fig. 11. The coupler features measured UWB characteristics with coupling of 3 ± 1 dB at the aimed 1.5-3.2 GHz band. Also, smooth isolation in the order of better than 25 dB and return loss in the order of better than 17 dB is achieved as there are non-dense variations around 10 dB for isolation and 9 dB for return loss at 1-3.5 GHz range. It is seen that the S parameters for the designed coupler have less variation density and more stability in the frequency domain in comparison with the commercial Krytar 1831 in Section III.B. It is observed in Fig. 11 that the measured phase difference between ports 2 and 3 is about 90^o over the target band. The fabricated coupler is shown in Section IV. These results indicate that this compact coupler with 35 mm × 30 mm × 1.1 mm $(0.27\ \lambda\times 23\ \lambda\times 0.009\lambda)$ dimension operates as a backward wave quadrature coupler.

D. *Wideband 50-ohm RF Load*

A wideband commercial 50-ohm RF load (Fig. 12) - which has the nearest commercially available impedance to the antenna impedance - is used in the experiments.

IV. Tx-Rx Isolation in EBD Stage and Comparisons

In this section Tx-Rx isolations of EBD are investigated when the devices introduced in Section III are used and the obtained results are compared together. The designed stable impedance antenna (Section III.A) is connected to the commercial Krytar 1831 hybrid coupler (Section III.B) and stable S parameter coupler (Section III.C) in two separate experiments to from EBD, Fig. 13. For the Krytar 1831 coupler from Fig. 7 it is seen that the measured S parameters are not stable in the frequency domain at 1-3.5 GHz range as there are dense isolation variations around 15 dB and return loss variations around 25 dB while for the designed coupler (Fig. 10) there are less dense variations with the isolation variations around 10 dB and return loss variations around 9 dB. A 50 Ω wideband 50-ohm RF load (Section III.D) is also connected as balancing impedance to both couplers. The experiment setups are shown in Fig. 14.

Average isolation of 20 dB and 29dB are measured by VNA at 1.5GHz-3.5GHz bandwidth respectively for EBD with the Krytar 1831 coupler and EBD with the designed stable S parameter coupler, Fig. 13. While both EBDs use the same antenna and the same 50 Ω RF load, it is seen that when using the designed coupler more stable UWB bandwidth and 45% better isolation value are obtained in comparison when using the Krytar 1831 coupler.

First reason of variations and limitations still seen in the obtained optimized UWB isolation could be the small inherent impedance variations of the designed coupler and the antenna.

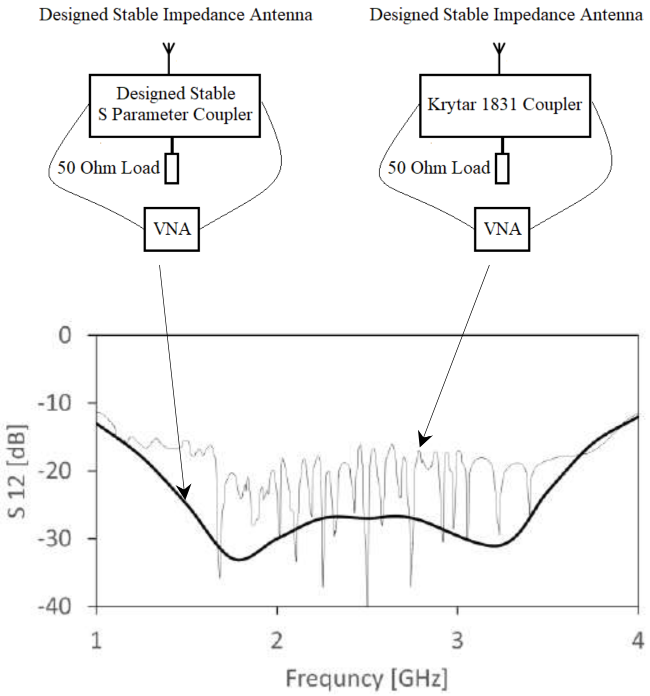


Fig. 13. Measured Tx-Rx EBD isolation when using designed stable impedance antenna with the commercial Krytar 1831 coupler and the designed stable S parameter coupler

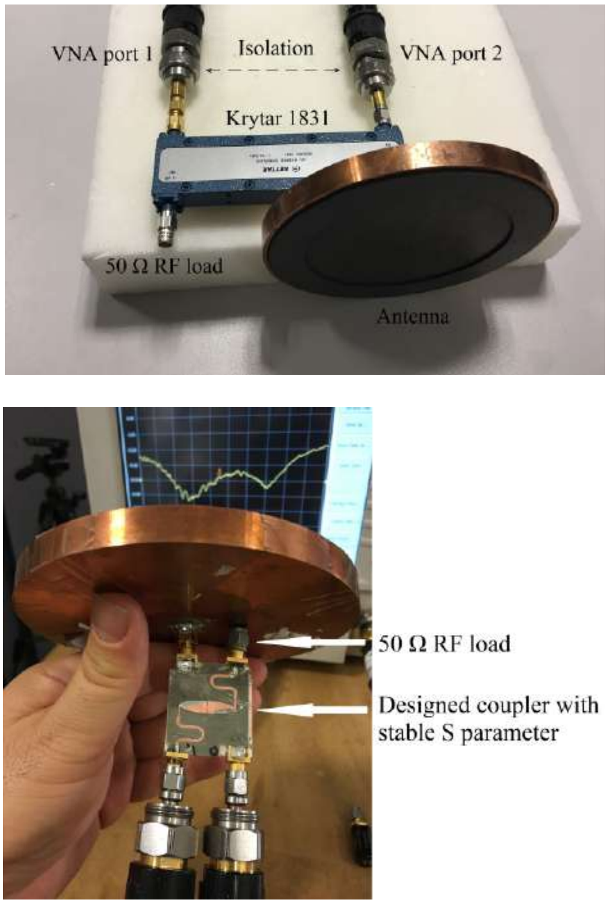


Fig. 14. EBD Setup to measure Tx-Rx isolation when the designed stable impedance antenna is connected to: (Top) Krytar 1831 hybrid coupler (Below) Designed stable S parameter coupler.

The second reason could be impedance mismatch between antenna and 50 Ω balancing load.

V. Conclusion

Electrical Balance Duplexers (EBDs) in In-Band Full Duplex (IBFD) transceivers, isolates transmit and receive signals to implement a form of self-interference cancellation to facilitates simultaneous communication from single antenna. EBD works by coupling transmitter, receiver, antenna, and balancing impedance using a hybrid coupler. The balancing impedance in the EBD needs to be equal as much as possible to the antenna impedance to achieve a high isolation where the antenna impedance variations limit the isolation bandwidth. Hybrid couplers are also not ideal elements, and their S parameters are variable in the frequency domain. An antenna with a stable impedance, designed by authors, has been connected to two different hybrid couplers in the EBD stages. One coupler is commercial Krytar model 1831 and the other one, designed by authors, has more S parameter stability in the frequency domain. It is shown that using hybrid coupler with more stable isolations results to higher isolation bandwidth and better isolation value in the EBD stage at an UWB frequency range from 1.5 GHz to 3.5 GHz.

References

[1] L. Laughlin *et al.*, "Tunable Frequency-Division Duplex RF Front End Using Electrical Balance and Active Cancellation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 12, pp. 5812–5824, 2018.

[2] S. N. Venkatasubramanian, C. Zhang, L. Laughlin, K. Haneda, and M. A. Beach, "Geometry-Based Modeling of Self-Interference Channels for Outdoor Scenarios," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 5, pp. 3297–3307, 2019.

[3] L. Laughlin, M. A. Beach, K. A. Morris, and J. L. Haine, "Optimum single antenna full duplex using hybrid junctions," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 9, pp. 1653–1661, 2014.

[4] E. Aryafar, M. Khojastepour, K. Sundaresan, S. Rangarajan, and M. Chiang, "MIDU: Enabling MIMO full duplex," in *Proc. ACM Int. Conf. Mob. Comput. Netw*, 2012, pp. 257–268.

[5] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *Proc. ACM SIGCOMM,* 2013, pp. 375–386.

[6] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4296–4307, 2012.

[7] L. Laughlin, C. Zhang, M. A. Beach, K. A. Morris, and J. Haine, "A widely tunable full duplex transceiver combining electrical balance isolation and active analog cancellation," in *IEEE 81st Veh. Tech. Conf.*, 2015, vol. 2015, pp. 1–5.

[8] B. Debaillie *et al.*, "Analog/RF solutions enabling compact full-duplex radios," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 9, pp. 1662–1673, 2014.

[9] L. Laughlin, C. Zhang, M. A. Beach, K. A. Morris, and J. L. Haine, "Passive and active electrical balance duplexers," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 1, pp. 94–98, 2016.

[10] S. H. Abdelhalem, P. S. Gudem, and L. E. Larson, "Hybrid transformer-based tunable differential duplexer in a 90-nm CMOS process," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 3, pp. 1316–1326, 2013.

[11] L. Laughlin, M. A. Beach, K. A. Morris, and J. L. Haine, "Electrical balance duplexing for small form factor realization of in-band full duplex," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 102–110, 2015.

[12] B. Van Liempd, J. Craninckx, R. Singh, P. Reynaert, S. Malotaux, and J. R. Long, "A dual-notch +27dBm Tx-power electrical-balance duplexer," in *European Solid-State Circuits Conference*, 2014, pp. 463–466.

[13] C. Zhang, L. Laughlin, M. A. Beach, K. A. Morris, and J. L. Haine, "Micro-electromechanical impedance control for electrical balance duplexing," in *Europ. Wireless Conf.*, 2016, pp. 263–268.

[14] S. Soodmand, M. A. Beach, and K. A. Morris, "Small Antenna with Stable Impedance and Circular Polarization," in *IEEE 15th European Conference on Antennas and Propagation (EuCAP)*, 2021.

[15] S. Soodmand, K. Morris, and M. Beach, "Quantization of Impedance Stability in Frequency Domain," in *IEEE International Electrical Engineering Congress (iEECON2021)*, 2021.

[16] "[Online]. Available." https://www.hitek-ltd.co.uk/index.php/downloads/dl/file/id/9822/product/0/ud_14518_rev_b_8ghz_urethane_1_14mm.pdf.

[17] "CST Studio Suite Electromagnetic Field Simulation Software (2020)." DASSAULT SYSTÈMES, [Online]. Available: https://www.3ds.com/products-services/simulia/products/cst-studio-suite/.

[18] "Krytar 1831 - Datasheet." https://krytar.com/pdf/1831.pdf.

[19] T. Tanaka, K. Tsunoda, and M. Aikawa, "Slot—coupled directional couplers on a both—sided substrate MIC and their applications," *Electronics and Communications in Japan (Part II: Electronics)*, vol. 72, no. 3, 1989.

[20] A. M. Abbosh and M. E. Bialkowski, "Design of compact directional couplers for UWB applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 25, no. 22, pp. 189–194, 2007.

[21] M. F. Wong, V. Fouad Hanna, O. Picon, and H. Baudrand, "Analysis and design of slot-coupled directional couplers between double-sided substrate microstrip lines," in *IEEE MTT-S International Microwave Symposium Digest*, 1991, pp. 2123–2129.

Effects of Noise on Machine Learning Algorithms Using Local Differential Privacy Techniques

Krishna Chaitanya Gadepally
Dept. of Electrical and Electronics Engineering
Birla Institute of Techology and Science, Pilani
Hyderabad, India
krishna.gadepally@gmail.com

Sameer Mangalampalli
Data and Machine Learning Engineering
Comatrix Labs
Hyderabad, India
sameer.m1@comatrixlabs.io

***Abstract*— Noise has been used as a way of protecting privacy of users in public datasets for many decades now. Differential privacy is a new standard to add noise, so that user privacy is protected. When this technique is applied for a single end user data, it's called local differential privacy. In this study, we evaluate the effects of adding noise to generate randomized responses on machine learning models. We generate randomized responses using Gaussian, Laplacian noise on singular end user data as well as correlated end user data. Finally, we provide results that we have observed on a few data sets for various machine learning use cases.**

Keywords—Differential Privacy, Randomization, Privacy

I. Introduction

Data is being collected in large amounts from a web browser to purchasing of data for applying machine learning algorithms to engage with customers. Machine learning algorithms work by studying a lot of data and updating their parameters to encode the relationships between features and the output in that data. If Machine learning is used to solve important tasks, like making a cancer diagnosis model, credit card eligibility, customer engagement, customer lifetime value, then features need to be developed using personally private information leading to invasion of privacy. In a study by PWC more than 83% of participants did not want to share their due to fear of privacy loss. [1]

In the 1990s, the American state of Massachusetts made medical data publicly available which included minimal demographic information: birth date, zip code and gender, in addition to the diagnosis. A computer scientist Latanya Sweeney purchased a voter list with 54,000 names. Sweeney was able to link the demographic information in the medical records released by the government to the demographic information in the voter database and in many cases, she was able to match an unlabelled medical record to the patient's name using the voter list she had bought earlier. She tried this data-matching technique for the then-governor William Weld. Just six people in the city of Cambridge shared his birthday. Out of the six, only three of them were men. Weld was the only one who lived in the correct zip code. Cynthia Dwork came up with a new way to define privacy as a quantifiable measure, also giving a formal guarantee that information is not leaked, called "differential privacy". [2]

"Differential privacy" describes a promise, made by a data holder, or curator, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available." [3]

Differentially private algorithms are able to answer a large number of aggregates, statistical queries approximately, so that the approximate answers can draw roughly the same conclusions as if the original data. Local differential privacy is a model of differential privacy with the added restriction that approximates personal responses of an individual such that we do not reveal too much about the user's personal data. Some of the methods in Differential privacy methods are:

Randomization
Laplace Method
Exponential Mechanism

In this paper we explore methods for local differential privacy by adding noise to generate randomized responses for features in Machine learning models.

II. DEFINITIONS

Our idea is to explore methods adding noise and generate randomized responses to be used in features. Here are few ways noise can be expressed mathematically:

1. Noise limitations expressed mathematically:

Gaussian noise

x: Gaussian random variable
The probability density function, p of a Gaussian random function, x is:

$$p(x) = \frac{1}{\sigma\left(\sqrt{2\pi}\right)} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean and σ is the standard deviation.

Laplacian mechanism

The Laplace Distribution (centred at 0) with scale b is the distribution with probability density function:

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

$$L(x|b) = \frac{1}{2b} e^{-\left(\frac{|x|}{b}\right)}$$

where b is a scale parameter.

2. Pearson Correlation

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where:
r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-sample

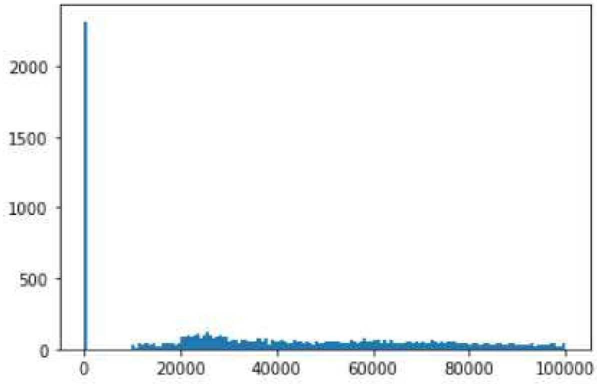
III. METHODOLOGY

A sample dataset with a selected subset of features is shown below. ‘Engaged’ is the output of the dataset.

Engaged	Income	Number of Open Complaints	Number of Policies	Total Claim Amount
0	71941	0	2	198.23
1	21604	0	1	379.20

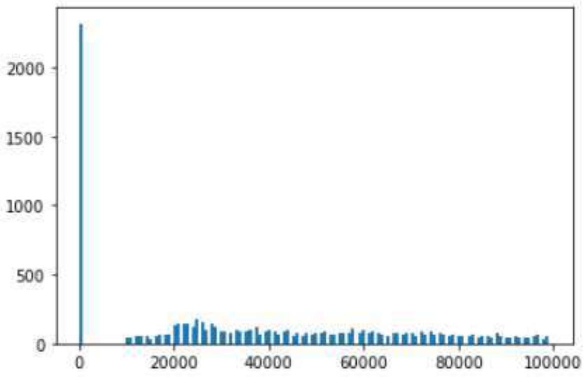
Before noise was added to Income:

Shown below is a histogram plot of ‘Income’ before noise addition to ‘Income’.

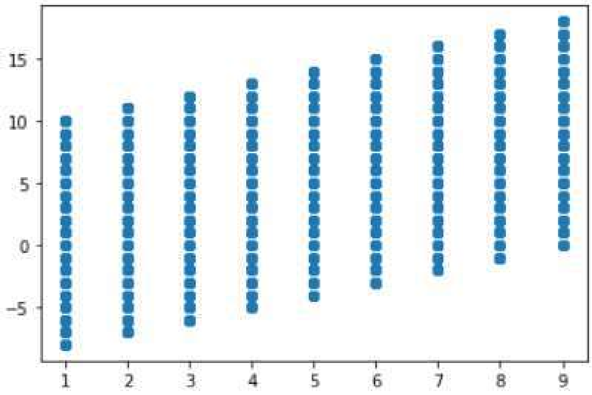


After noise is added to Income:

Shown below is a histogram plot of ‘Income’ after noise addition to ‘Income’.



Shown below is a scatter plot which shows how ‘Number of Policies’ values’ change as noise is added.
The horizontal axis corresponds to ‘Number of Policies’ variable before noise addition.
The vertical axis corresponds to ‘Number of Policies’ variable after noise addition.



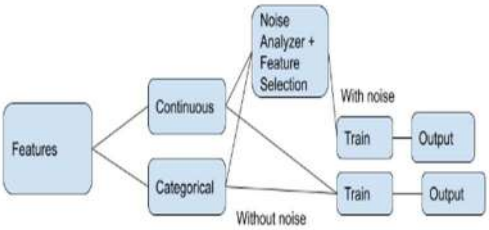
Pseudo Code	
Input	f(x) = {f1, f2, ..., fn}: features of x G(x): Gaussian noise function
Function	Repeat for all features f in {f1, f2, ..., fn}: Take a random feature column, fx ∈ f For each M in {Logistic Regression, Random Forest Classifier}: if fx is categorical: call pseudoCodefunction1 else call pseudoCodefunction2

Pseudo Code -- Function 1	
Input	Example data x1, x2, …, xn in X Random sample percentage for selecting values to add noise
Function	For x ∈ X:

	Take random sample x with sampling probability L/N Compute For each value compute the noise g(x): Clip noise g(x) to g(x)’ Add Noise x’ = x + g(x)’ within a bound Fx’ append x’ where Fx’ is the new noise feature

Pseudo Code -- Function 2	
Input	Input: Example data x1, x2, ..., xn Random sample percentage for selecting values to add noise.
Pre-process	If Categorical introduce a new random value based on data: Method 1: Increase a max bound Method 2: Introduce a categorical value in range Bounds changed and gaussian function updated
Function	For x ∈ X: Take random sample x with sampling probability L/N Compute For each value compute the noise g(x): Add Noise x’ = x + g(x) within a bound Fx’ append x’ where Fx’ is the new noise feature

IV. RESULTS



The results were determined using the above workflow for ML training. The Noise Analyzer and Feature Selection module were used to add noise to the features. This was done using Python and scikit-learn in Jupyter notebooks. Finally, the ML models were trained with noise and without noise as expressed in the flow.

The IBM-Watson Customer Marketing Value data set was selected. 'Engaged' is the output column. 'Yes' and 'No' in 'Response' feature were converted into 1 and 0 respectively and made the output column. Categorical variables were factorized and added to the dataset.

Eleven features were chosen for randomization, namely, ‘Customer Lifetime Value’, ‘Education’, ‘Gender’, ‘Income’, ‘Location Code’, ‘Marital Status’, ‘Months Since Last Claim’, ‘Months Since Policy Inception’, ‘Number of Open Complaints’, ‘Number of Policies’ and ‘Sales Channel’ from the data set.

Model: Logistic Regression

1. Before Gaussian noise and randomized responses were added to the features, a logistic regression model was trained on the data set. The training accuracy was 86.41%.

Feature chosen	Trained on noisy dataset; Tested on original dataset	Trained on original dataset; tested on noisy dataset
Customer Lifetime Value	86.41%	86.39%
Education	86.40%	86.32%
Gender	86.40%	86.40%
Income	86.39%	86.45%
Location Code	86.40%	86.30%
Marital Status	86.40%	86.41%
Months Since Last Claim	86.36%	86.19%
Months Since Policy Inception	86.42%	86.26%
Number of Open Complaints	86.40%	86.22%
Number of Policies	86.35%	86.11%
Sales Channel	86.45%	86.38%

Model: Random Forest Classifier

2. Before Gaussian noise and randomized responses were added to the features, a random forest classifier was trained on the data set. The training accuracy was 100%.

Feature chosen	Trained on noisy dataset; tested on original dataset	Trained on original dataset; tested on noisy dataset
Customer Lifetime Value	100%	100%
Education	100%	99.26%
Gender	100%	100%
Income	100%	100%
Location Code	100%	99.37%
Marital Status	100%	99.75%

Months Since Last Claim	100%	98.9%
Months Since Policy Inception	99.99%	97.89%
Number of Open Complaints	100%	99.98%
Number of Policies	99.99%	100%
Sales Channel	100%	99.77%

3. In contrast to the previous results, here three features are considered instead of one for randomization.

'Customer Lifetime Value', 'Income' and 'Marital Status' are chosen for randomization.

Model	Trained on noisy dataset; tested on original dataset	Trained on original dataset; tested on noisy dataset
KNN	100%	100%
SVM	85.68%	85.68%

'Months Since Last Claim', 'Months Since Policy Inception', 'Number of Open Complaints' are chosen for randomization.

Model	Trained on noisy dataset; tested on original dataset	Trained on original dataset; tested on noisy dataset
KNN	96.66%	94.70%
SVM	85.68%	85.68%

4. 'LocationCode' was found to have the highest Pearson Correlation value, which meant that 'LocationCode' was the feature which had the closest to a linear relationship with the output 'Engaged'.

Model	Trained and tested on original dataset	Trained on noisy dataset; tested on original dataset	Trained on original dataset; tested on noisy dataset
KNN	100%	99.9%	99.68%
SVM	85.68%	85.68%	85.68%
Logistic Regression	86.41%	86.38%	86.30%
RF	100%	100%	99.42%

V. CONCLUSION

From our experiments we can safely conclude machine learning applications will not be affected by differential privacy preserving mechanisms and in fact can become allies. If proper care is taken to add noise to the data, we have seen the models are almost alike with noise and without noise. We could easily extend the methodology to build features using SQL join, distributed joins as well as advanced data aggregation methods. There are several real world machine learning applications like targeting customers, marketing funnels, segmentation, churn prediction in ecommerce, communication, retail as well as health care applications where differential privacy and local differential privacy techniques could be applied in order to protect the privacy of the end user data. Personalization while preserving end user privacy is something that we all want and we hope we were able to present a strong case for it in our study.

VI. REFERENCES

[1] PwC. 2021. *Consumers Trust Your Tech Even Less Than You Think*.

[2] Dwork, C., 2021. *Differential Privacy*.

[3] Dwork, C. and Roth, A., 2014. *The algorithmic foundations of differential privacy*.

[4] Privacytools.seas.harvard.edu. 2021. *Differential Privacy*.

[5] Geng, Q. and Viswanath, P., 2021. *Optimal Noise Adding Mechanisms for Approximate Differential Privacy*.

[6] Mironov, I., Talwar, K. and Zhang, L., 2021. *Renyi Differential Privacy of the Sampled Gaussian Mechanism*.

[7] Erlingsson, U., Pihur, V. and Korolova, A., 2021. *RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response*.

[8] Agrawal, D., Aggarwal, C. 2001. *On the Design and Quantification of Privacy Preserving Data Mining Algorithms*.

[9] Chawla, S., Dwork, C., McSherry, F., Smith, A., Wee, H., 2005. *Toward Privacy in Public Databases*.

Spectrum Segmentation Techniques for Edge-RAN Decoding in Telemetry-Based IoT Networks

Michael Schadhauser*, Joerg Robert*, Albert Heuberger* and Bernd Edler†
*Lehrstuhl für Informationstechnik mit dem Schwerpunkt Kommunikationselektronik (LIKE)
*†Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany
†International Audio Laboratories Erlangen (AudioLabs), a joint research institute between the Friedrich-Alexander
Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute of Integrated Circuits (IIS), 91058 Erlangen, Germany
Email: {michael.schadhauser, joerg.robert, albert.heuberger}@fau.de, bernd.edler@audiolabs-erlangen.de

***Abstract*—The possible fields of application for small sensor nodes are tremendous and still growing fast. Concepts like the *Internet of Things* (IoT), *Smart City* or *Industry 4.0* adopt wireless sensor networks for environmental interaction or metering purposes. As they commonly operate in license-exempt frequency bands, telemetry transmissions of sensors are subject to strong interferences and possible shadowing. Especially in the scope of Low Power Wide Area (LPWA) communications, this scenario results in high computational effort and complexity for the receiver side to perceive the signals of interest. Therefore, this paper investigates means to an adequate segmentation of receive spectra for a partial spectrum exchange between base stations of telemetry-based IoT sensor networks. The distinct interchange of in-phase and quadrature (IQ) data could facilitate stream combining techniques to mask out interferences amongst other approaches. This shall improve decoding rates even under severe operation conditions and simultaneously limit the required data volume. We refer to this approach of a reception network as Edge-RAN (Random Access Network). To cope with the high data rates and still enable a base station collaboration, especially in wirelessly connected receiver mesh networks, different filter bank techniques and block transforms are examined, to divide telemetry spectra into distinct frequency sub-channels. Operational constraints for the spectral decomposition are given and different filter methodologies are introduced. Finally, suitable metrics are established. These metrics shall assess the performance of the presented spectrum segmentation schemes for the purpose of a selective partial interchange between sensor network receivers.**

***Index Terms*—Wireless sensor network, IoT, spectrum segmentation, Low Power Wide Area (LPWA) communications, long range telemetry, Edge-RAN decoding**

I. Introduction

Over the last recent years, the *Internet of Things* (IoT) is one of the most driving forces for developments and research interest in the area of wireless sensor networks. The conception comprises the communication between numerous different sensor nodes with the aim of environmental interaction or data aggregation, for example in infrastructure monitoring [1]. For many applications, the mobility of nodes must be preserved, putting harsh constraints on weight and energy consumption when powered by battery. The layout of these nodes is therefore targeted on an extremely energy efficient minimalist design. This in turn shifts complexity towards the receiver side for detection and decoding, which can be considered a general characteristic of Low Power Wide Area (LPWA) sensor networks [2]. Often, license-exempt frequency bands are utilized, inevitably exposing signals to strong interferences. Given also the nodes' restricted transmit power, the signal-to-noise-ratio (SNR) at receiving stations can be quite unfavorable, particularly for high path losses in distances of several kilometers and severe shadowing by obstacles like trees or buildings. While a station may detect the presence of a burst, the signal strength might not be sufficient for an overall successful decoding. In Figure 1, such a typical scenario is depicted with several receiving stations spanning a network and continuously monitoring the spectrum for transmissions of sensors within range. Especially for highly mobile transmitter nodes, for example attached on bats for the purpose of wildlife tracking [3], that are frequently changing their position relative to the reception grid, the receive spectra can differ notably between stations. At this point, the idea arises to further improve the system decoding rate via an interchange of inphase and quadrature (IQ) data among different base stations. This approach is then in compliance with the schemes of *Cooperative Multi-Point* (CoMP) [4] or *Distributed Base Station Cooperation* [5] discussed in the context of the fifth mobile communication standards (*5G*). Within the scope of telemetry-based IoT sensor networks, we refer to this conception as *Edge-RAN* (*Random Access Network*) decoding, in allusion to *Centralized-RAN* (also *Cloud Radio Access Network*, C-RAN) in 5G. Contrary to the central approach of C-RAN, a distributed demand-based pre-processing of IQ streams within the edge of the receive network among the collaborating base stations is performed. By providing

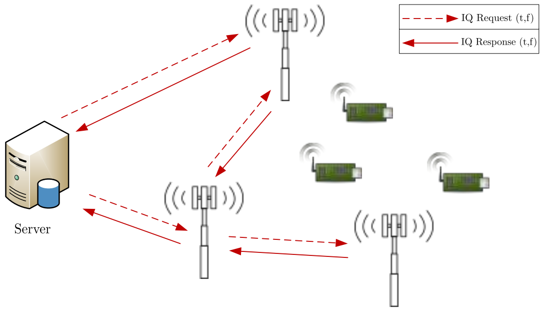


Fig. 1. Telemetry reception network utilizing Edge-RAN decoding for an overall improved decoding rate in a dynamic IoT sensor network.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

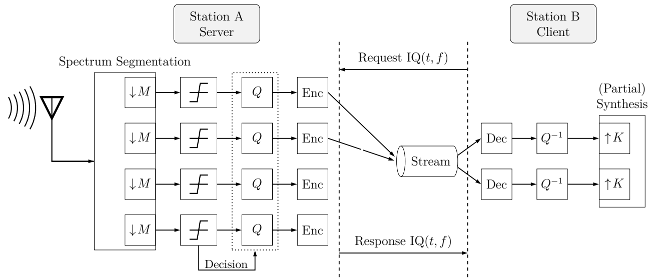


Fig. 2. Proposed telemetry-dedicated IQ compression scheme.

several spectra synchronized in time and obtained by various receivers, the combination of those streams or a fragmentary exchange between them, may compensate for distinct sample sections, that would otherwise be affected by interferences or suppressed by shadowing. A server (*cf.* Figure 1) might then request a specific fraction of manipulated interference-cleared spectral recordings, in order to provide the resulting streams to several connected users for an individual final decoding. A wirelessly linked and rate restricted base station system, however, renders this collaboration quite challenging. While we do not address stream combining techniques here, we focus on means to provide the desired IQ interchange over a rate-limited channel and thus facilitate those subsequent algorithms. We propose to divide the receive spectra into sub-channels, that exploit the small bandwidth of telemetry signals. This shall enable a partial spectral exchange to cope with the rate limitations of links in an IoT base station network. In Figure 2, the overall IQ compression scheme is shown. It comprises the frequency segmentation as its key feature, followed by in-band signal detection providing a decision about the presence of user signals to a content-adapted quantization (Q) stage and later encoding (Enc) of requested sub-bands. The decoding side may only reconstruct a fraction of the entire spectrum.

The paper is structured as follows. At first, state of the art IQ compression schemes are addressed in Section II. Section III introduces spectrum segmentation for a distinct exchange and combination of sub-streams in telemetry-based IoT networks. Constraints on possible approaches are given and suitable filter bank techniques and block transforms are illustrated. In Section IV, the metrics of coding gain, basis restriction and complexity measures are established and the performance of presented filtering procedures, regarding their eligibility for telemetry data, is discussed. Section V finally summarizes the findings and concludes this paper.

II. State of the Art IQ Compression

As has been stated in Section I, IQ compression is especially tackled in the scope of 5G. To reduce the rate requirements within the backhaul, Guo et al. [6] proposed a scheme that directly acts on time domain data. A stream is divided into groups, that are processed via *block floating point encoding*. Vosoughi et al. [7] apply a similar method, yet operate on difference symbols and an adapted Arithmetic coder. A lossy extension is given by *least significant bit removal*. While these approaches are simple and applicable to generic signals of any shape, they are reported to achieve only a minor compression factor of about 3.5 [7]. Samardzija et al. [8] exploit the specific signal characteristics in mobile communications, applying a deliberate downsampling, block scaling and non-linear quantization. Nieman et al. [9] follow the same idea. They remove the signal's cyclic prefix, perform decimation and a spectrum-adapted quantization via noise shaping. Contrary to those time domain approaches, Grieger et al. [10] conduct the compression in frequency domain. Via a Fast Fourier Transform (FFT), a subset of carriers is quantized and Huffman encoded. Although these proposals promise higher compression rates, they are specifically designed for orthogonal frequency division multiplex (OFDM) signals. The algorithms make use of particular signal properties and require a-priori knowledge. Hence, they are not directly transferable to streams in case of IoT networks. Here, a generic compression scheme is required, easily adaptable to unknown narrowband transmissions, that are heavily fluctuating over time. Another approach that operates on transform coefficients is the scheme denoted as *Compressive Sampling* or *Compressed Sensing* [11]. This conception is based upon the assumption, that a signal's representation can be sparse, thus exhibiting only few non-zero coefficients within the frequency domain. The Nyquist sampling theorem presumes a rather fully occupied spectrum. In this way, it is deemed to be too pessimistic. These considerations give rise to a quasi sub-Nyquist rate reduction, implicitly compressing any input signal in the process of sampling. While in [12] and [13] corresponding telecommunication compression frameworks, referred to as *Random Demodulator* and *Modulated Wideband Carrier*, have been presented, a large-scale adoption is still impeded by the system's susceptibility to noise aggregation. The inverse transform comprises the extensive search for a sparse solution of an under-determined system. However, an IoT spectrum is inherently exposed to rather harsh noise (*cf.* Figure 3). Furthermore, possible short time interferences counteract the sparsity assumption, leaving a successful reconstruction impossible [14], [15].

Existing generic compression approaches can not provide sufficient performance, whereas schemes proposed in the context of 5G are based upon OFDM-related signal characteristics. The conception of *Compressive Sampling* suffers from noise, which is the prevailing process in IoT spectra. To overcome these limitations, the scheme in Figure 2 shall be adopted. The division of spectra in distinct sub-bands is a crucial component. Therefore, this paper investigates proper means of spectrum segmentation, in order to exploit the narrow bandwidth characteristic of low rate telemetry transmissions. This shall be a building block for an efficient compression of reception spectra in IoT sensor networks.

III. Band Segmentation Approach for the Interchange of Telemetry Spectra

The interchange of spectra among telemetry base stations shall exploit the inherent macro diversity of the reception

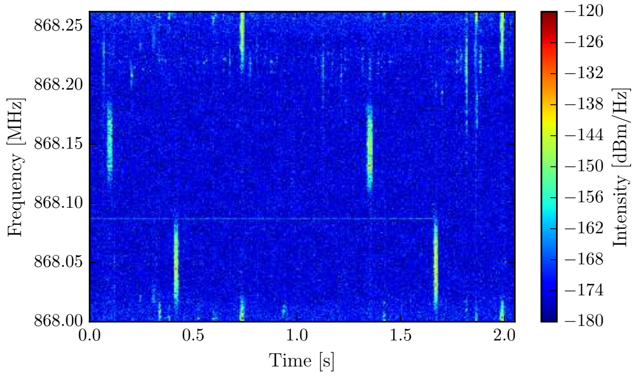


Fig. 3. Spectogram of a typical telemetry stream in the SRD frequency band.

network. IQ streams of different stations might be combined to mask out interferences or compensate for notches in the channel transfer characteristic, as observed by one receiver but not the other. The spectral segmentation might also limit the overall noise bandwidth and thus increase the relative in-band SNR. This can alleviate decoding by concealing possible strong disruptive frequency components nearby, that would otherwise dominate the signal as seen by the receiver. In a practical setup, this network must meet the requirements concerning ease of installation, maintenance and cost. Wired connections and high speed fiber-based solutions are therefore not a viable option. Wireless approaches like WiFi or radio link systems are both flexible and economical, but are limited in rate. This issue gets apparent, when considering an actual large scale telemetry network as presented in [1]. Here, a common exposed receiver records the frequency-scattered signals being broadcast by a distributed network of numerous sensor nodes for the purpose of infrastructure monitoring. Scanning the complete *Short Range Devices* (SRD) band from $863\,\mathrm{MHz}$ to $870\,\mathrm{MHz}$ and gathering complex samples with 24 bit resolution in a configuration with up to four antennas, easily results in a necessary rate of $1.34\,\mathrm{Gbit/s}$. Even with the latest standards of WiFi, this high rate can not be provided. The rate problem appears even more acute for a fully meshed receiver network or in a hierarchical network structure that aggregates several streams in the backbone. As elaborated before, in order to alleviate this challenge, we propose a partial exchange by means of a spectrum segmentation.

Figure 3 exemplarily depicts an excerpt of a typical actually recorded telemetry spectrum with approximately $2\,\mathrm{MHz}$ bandwidth centered around the frequency of $868.13\,\mathrm{MHz}$, acquired by the reception system presented in [1]. As this range corresponds to the license free SRD radio bands (equivalent to the frequency bands around $915\,\mathrm{MHz}$ in the United States), numerous transmissions are present. Besides persistent or temporary carriers, also momentary bursts are perceivable. One has to note, that common to all of the signals shown, is their comparatively low bandwidth in relation to the sampling range. Often rates of only a few kHz or even less are sufficient for IoT use cases with low overall data throughput. An example is the *LoRa*[1] protocol with a typical bandwidth of $125\,\mathrm{kHz}$, gaining increased research interest in the scope of low power and long range sensor applications. This is in contrast to mobile communication with a constantly high channel occupancy. Besides, only a minor fraction of the stream is actually populated with user data, while the rest is dominated by noise. It is this characteristic of momentary narrowband transmissions that gives rise to the segmentation of telemetry spectra into a frequency-time grid. With knowledge about the transmit hopping pattern, a base station for which the connected decoder indicated a failed decoding, might then state a request for a specific frequency band and time span to all receivers within the network (*cf.* Figure 3). Rather than an exchange of the complete spectrum, a partial and selective streaming of data with minimal rate can be performed, still adhering to the limited rate of each wireless link within the wireless receiver mesh network. Also, by this separation, each of the resulting frequency bands within the addressable frequency-time grid can be compressed individually depending on its actual content (presence of signals or prevailing channel noise).

This raises the question which spectral segmentation method at server side is advantageous for the purpose of partial telemetry IQ exchange and related Edge-RAN decoding at the client side. Considering the context of spectrum compression in sensor network applications, corresponding constraints have to be imposed on potential techniques for partitioning:

- *Handling of complex values*: As the sample stream is composed of IQ samples, the algorithm must be suitable for complex data. Furthermore, the phase of the representing complex rotator must be preserved, as actual payload information might be encoded into this parameter by modulation.
- *Critical sampling*: Considering an overall compression, critical sampling is a favorable feature for the filtering procedure. This implies that the total number of samples over all resulting filter sub-channels remains constant compared to the input [16]. Even moderate oversampling structures would increase the data volume. This would force a possible subsequent compression scheme to settle for the loss that the preceding analysis stage provoked.
- *Aliasing cancellation* (AC): Adherence to critical sampling denotes an operation at the Nyquist rate. This inevitably leads to Aliasing, that has to be canceled when recomposing the sub-channels or at least suppressed as best as possible, to not corrupt the payload signal within the frequency range of interest. One refers to those methods providing perfect reconstruction (PR) or near perfect reconstruction (NPR), respectively [16].
- *Relation to time-frequency pattern*: For a selective interchange of partial spectra, a relation to the actual physical parameters of frequency and time must be given for the obtained general transform coefficients. This would for example also preclude the previously mentioned methodology of *compressive sampling* [11], often cited

[1] *https://lora-alliance.org/ last accessed 10 June 2020*

in relation to compression in wireless sensor networks, but not abiding by this restraint.

- *Scalability*: A partitioning technique must support various numbers of sub-bands, thus implementing a scalable sub-channel bandwidth to properly adopt to the signal characteristics.

In the following, we briefly introduce the spectral segmentation algorithms that, at least partially, adhere to the given constraints. Basically, the approaches can be grouped into two fundamental categories.

Filter Bank Techniques

Spectral decomposition based on filter bank structures can be considered as the more classical approach. As there is no priority to different sub-bands in the application at hand, a uniform filter bank with equal spectral spacing and bandwidth can be targeted. It is thereby common to design a single *prototype* filter with a relative bandwidth according to the desired number of sub-channels. This prototype is then modulated to form the overall filter bank structure [16]. The *Discrete Fourier Transform* (DFT) [17] is among the most utilized approaches and will be considered for the evaluation. Besides, the so-called *Modified Discrete Fourier Transform* (MDFT) proposed by Fliege et al. [18] fulfills the stated constraints above and will be issued in the following analysis. In [19], Viholainen et al. presented the *Exponentially Modulated Filter Bank* (EMFB) as an alternative complex modulated critically sampled filter approach. It will be considered as the third filter bank representative. It is characterized by an odd channel stacking, as opposed to the DFT and MDFT with even stacking.

Block Transforms

While for the filter bank techniques the common sub-filter characteristic is designed in frequency domain, the block transforms operate in the equivalent time domain. Here, the prototype is considered a *window* function, weighting the input sample sequence organized in successive blocks [20]. Detached from the domain of filtering and referring to the notation of transforms, the sub-channel filters are also identified as *basis vectors*. Requirements for Aliasing cancellation and perfect reconstruction for the case of a full assembly of the spectrum are expressed in terms of those windows and summarized in the phrase of *Time Domain Aliasing Cancellation* (TDAC) [21]. As representatives for block transforms, the *Modulated Complex Lapped Transform* (MCLT), *Extended Complex Modulated Lapped Transform* (ECLT) [20] and the *Karhunen-Loève Transform* (KLT) [22] are considered. While the latter can be deemed critically sampled, the MCLT and ECLT do actually state PR structures with inherent two-fold oversampling. However, they provide exponential modulation, processing of complex streams and efficient implementations. Therefore, they are also incorporated into analysis.

Selection of proper PR and NPR prototypes

Due to the various filter structures and modulation approaches, there are different constraints for Aliasing cancellation and (near-)perfect reconstruction. These constraints are mapped to certain restrictions on the prototypes. The only prototype that allows for PR in the case of a DFT is rectangular in time [23]. With a length of L samples, the corresponding number of channels M is equal to L and the DFT structure is thereby fully determined. A popular solution for the prototype of the MCLT is a sine window [24]

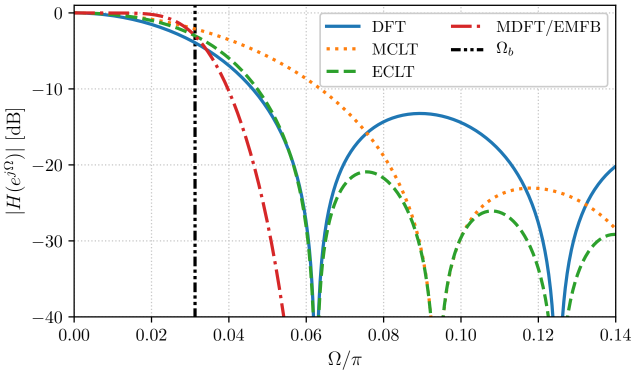


Fig. 4. Prototype transfer functions. Sub-channel border is indicated by Ω_b.

$$p[n] = -\sin\left[\frac{\pi}{2M}(n+0.5)\right], \qquad n = 0,\ldots,2M-1 \quad (1)$$

where the length L of the window $p[n]$ is restricted to $2M$ samples. The ECLT can be considered a generalization of MCLTs, extending the basis function to arbitrary length $L = 2KM$, with K a positive integer. The filter retrieval process for PR includes a highly non-linear and rather sensitive complex optimization, for which a global optimum is not guaranteed [25]. For $K = 2$, however, a possible closed-form solution is given by [24]

$$p[n] = 0.5\cos\left[\frac{\pi}{2M}(n+0.5)\right] - \frac{1}{2\sqrt{2}}. \quad n = 0,\ldots,4M-1 \quad (2)$$

With the MDFT and EMFB as classical filter bank approaches, their prototype is designed utilizing the *Quadrature Mirror Filter* (QMF) property [24]. The filter structure assures for a direct Aliasing cancellation of adjacent sub-channels, while other aliasing components are supposed to be sufficiently suppressed by the filter's stopband attenuation [18]. Various filter design methods for this purpose do exist, whereas we resort to the iterative optimization process that was proposed by Doblinger et al. in [26], with $L = 16M$ and a desired stopband attenuation of $100\,\mathrm{dB}$. The KLT can be considered a special case of a varying block transform, depending on the actual signal itself. Rather than a fixed set of basis vectors or equivalently sub-channel filters, the basis vectors are dynamically computed as the eigenvectors of the estimated autocorrelation matrix within the current block. The transform is thus constantly changing, trying to decorrelate its output coefficients, so no fixed filter structure can be stated in advance. Figure 4 illustrates the one-sided transfer functions $\left|H(e^{j\Omega})\right|$ of the discussed prototypes in decibel scale for a sub-channel number of $M = 32$. All curves have been normalized for DC gain, where again no prototype can be given for the KLT due to its non-static structure. The entity Ω_b represents the channel border. Distinctive differences in the filters' frequency responses are evident. While the prototypes of the MCLT and ECLT can provide perfect reconstruction, their progression in frequency exhibits rather large side lobes. In contrast, the common MDFT and EMFB prototype features a much smoother progression within the passband and a harsh decay in the stopband. Yet, these approaches can generally only allow for a near perfect reconstruction. The perfect reconstruction constraint is thus not bounded to the classical filter optimization goals. It is therefore of scientific interest to investigate the feasibility of the presented transforms for spectrum segmentation in a telemetry-dedicated IQ compression scheme.

IV. Filter Bank Evaluation

A fundamental relation of source coding is the so-called rate-distortion function. Without restriction to a specific signal source distribution, it can be expressed by [27]

$$R(D) \cong 0.5 \log_2 \left(\frac{\varepsilon^2 \sigma^2}{D} \right), \tag{3}$$

where ε^2 is a factor depending on the actual signal distribution and σ^2 is the corresponding variance. Equation (3) thus relates the minimal rate $R(D)$ needed to preserve the signal information under the maximal allowed distortion metric D. Regarding (3) with a direct dependence to σ^2, in the context of a sub-band coding system for IQ compression, it is favorable for a segmentation to concentrate stream information in as few sub-bands as possible. This in turn gives room for a harsh compression and less rate for the low variance channels, while only a minor number of channels need higher rates.

A. Coding Gain

As this feature is also desired in the context of audio and video applications [28], [29], one can adopt metrics commonly known in multimedia processing for the purpose of IQ segmentation evaluation. A transform's ability to focus the signal energy in just a few sub-bands is therefore desirable, denoted as the *energy compaction* characteristic. This can be measured by the metric of *transform coding gain* G_{TC} [28]

$$G_{TC} = \frac{\frac{1}{M} \sum_{i=0}^{M-1} \sigma_i^2}{\left(\prod_{i=0}^{M-1} \sigma_i^2 \right)^{\frac{1}{M}}}. \tag{4}$$

The coding gain relates the arithmetic mean of the transform sub-channel variances σ_i^2 to the geometric mean. A more irregular variance distribution yields higher gains and indicates a better energy compaction. Figure 5 depicts the metric of coding gain in decibel scale for the presented transforms. The signal shown in Figure 3 serves as an input and the gain is calculated along (4) in dependence of the sub-channel number M.

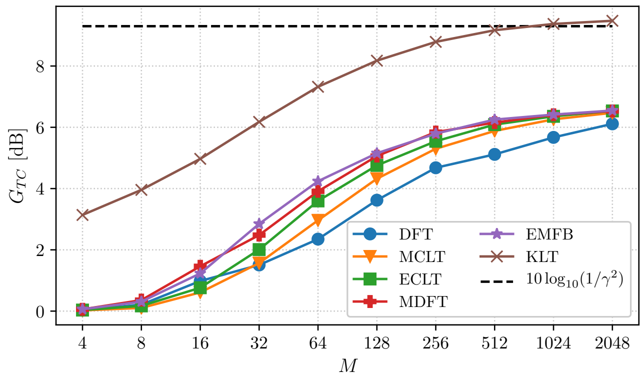


Fig. 5. Coding gain in dependence of the sub-channel number M.

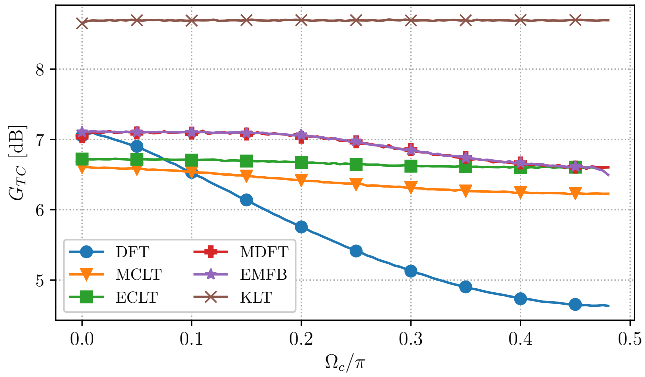


Fig. 6. Coding gain in dependence of carrier frequency relative to the frequency resolution grid.

For all segmentation techniques the coding gain increases with an increasing number of sub-channels M. The relative channel bandwidth is reduced, approaching the narrowband signal bandwidth, such that a separation of distinct telemetry transmissions becomes feasible. Among the fixed filter structures, the EMFB and MDFT indicate the best compaction compared to the DFT and the presented lapped transforms. This advantage of the EMFB over the MDFT might yet be related to the channel stacking, so also to the relative position of signal carriers within the test signal to the transform grid. The KLT, however, shows an overall superior performance with an offset of about 3 dB to the EMFB curve. With its adaptivity to the current signal autocorrelation, it decorrelates the channel coefficients. It maximizes the coding gain and gives an upper bound for a given block length, under the assumption of a prevailing quasi-Gaussian overall source [30]. For a very high sub-band number M, all curves show a saturation behavior. The KLT reaches the maximal possible gain for the spectrum in Figure 3 determined by the the spectral flatness measure γ^2 [28], the inverse of (4) for $M \to \infty$, as indicated by the black dashed line. A slight exceedance is seen, due to computation inaccuracies and spectral smoothing via windowing (a *von-Hann* window was used). Figure 6 shows the coding gain evaluations in dependence of a varying normalized carrier frequency Ω_c.

The position of a carrier in frequency domain, relative to the transform grid, that equally subdivides the spectrum, is decisive for the containment of energy within the corresponding sub-channel. Here, $M = 32$ is set and an unmodulated carrier $e^{j\Omega_c t}$ superimposed by complex Gaussian noise $n(t) \sim \mathcal{CN}(0, \sigma_n^2)$ is simulated for a SNR of 5 dB. The carrier frequency Ω_c is swept from the relative location 0.0 coinciding with the sub-channel center frequency up to 0.5 being positioned right between the actual transform grid. Again, the KLT exhibits the best performance, unaffected by the carrier position. It is followed by the MDFT and EMFB, whose curves almost perfectly coincide. For all fixed filter bank structures, a drop in gain is notable when the deviation from the transform grid increases. While this loss is rather moderate for the filter bank techniques as well as for the lapped transforms MCLT and ECLT, especially the DFT suffers from this effect. The finding for the coding gain simulations gets apparent, when considering the prototype characteristics in Figure 4. The blue curve for the DFT not only exhibits the most decreasing progress in the passband region left to the sub-channel border $\Omega_b = \frac{\pi}{M}$, but also shows the highest out-of-band ripples. This not only makes it susceptible to signal carriers off the ideal transform grid as shown in Figure 6, but also leads to an undesired spreading of signal energy across neighboring channels, notable by the inferior coding gain in Figure 5. The prototype for the MDFT was the same as for the EMFB and shows a flat passband progression, but then drastically drops before the channel border Ω_b, confirming the behavior for both coding gain results presented.

B. Basis Restriction Error

When only a partial subset of the whole IQ stream is exchanged, also only a fraction of the spectrum's energy is contained. A metric that reflects the ability of a segmentation technique to preserve the overall energy under elimination of certain sub-channels is the *basis restriction error* [29] or equivalently the *compaction efficiency* [22]

$$\eta_E(l) = \frac{\sum_{r=0}^{l-1} \sigma_r^2}{\sum_{r=0}^{M-1} \sigma_r^2}. \tag{5}$$

Here, only a subset of $l = 0, \ldots, M-1$ sub-channels are used to re-transform the spectral coefficients to the time domain reconstruction signal. The sub-channel variances σ_r^2 are thereby sorted in decreasing order and all corresponding sub-band coefficients addressed by an index $r \geq l$ are set to 0. Evaluating (5) for the input spectrum in Figure 3 yields the result shown in Figure 7.

Again, $M = 32$ channels have been simulated and only the first l sub-bands, exhibiting the highest variance, have been used for reconstruction. Larger values of $\eta_E(l)$ closer to 1 indicate better efficiency. As before, the KLT expectedly outperforms the other techniques. Given the configuration of prototypes as in Section III, the EMFB shows also here a superior result among the fixed structure approaches, revealing a better channel segmentation than the ECLT and MCLT following. Interpreting the basis restriction in (5) in the sense of Aliasing distortion, the EMFB also exhibits a slight advantage over the MDFT, with its internal downsampling factor being half of that of the MDFT [19].

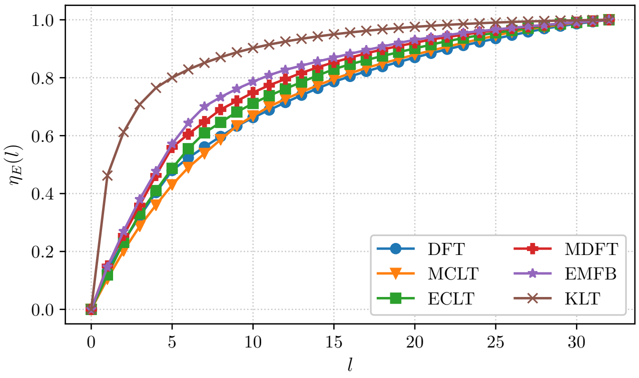


Fig. 7. Basis restriction error for the presented segmentation techniques.

C. Measures of Complexity

Besides the characteristics of PR or NPR and the energy compaction for later compression or partial spectrum transfer, the aspects of design complexity, computational effort, overhead and signal delay are decisive for practical usage. For the transforms as described in Section III, these factors are listed in Table I. The implementation aspect of complexity for the evaluated techniques is addressed in terms of the number of multiplications $\mu(M)$ and additions $\alpha(M)$ (real valued) for the calculation of the next M complex sub-channel samples in the process of segmentation. While there are various computationally efficient approaches in literature, we rely on the *Cooley-Tukey* algorithm [31] in the case of the DFT. For the MCLT and ECLT the setup presented in [32] is employed, based upon lattice structures and butterfly operations. The polyphase and crossover scheme applied in [33] serves as a reference for the MDFT, while the EMFB is assumed to be implemented by a fast lapped transform structure for a pair of cosine/sine modulated filter banks [32]. The KLT is considered a complex matrix multiplication. In order to get also a visual impression of the transforms' complexity, Figure 8 depicts the computational effort in terms of multiplications $\mu(M)$, in dependence of the number of sub-channel samples M to be calculated. When comparing $\mu(M)$ and $\alpha(M)$, an asymptotic effort can be noticed for all filter bank approaches, with the MDFT and EMFB showing a slightly higher cost due to the larger prototype length used (*cf.* Section III). The KLT, however, reveals a quadratic increase of $\mu(M)$ for increasing M, which is manifested by the harshly rising progression in logarithmic scale, seen in Figure 8. It appears to be feasible for minor sub-channel numbers M. Though, for higher and more practical values of M, its complexity is not manageable any more. Furthermore, the calculation of the eigenvectors and the related matrix inversion are not even taken into account here. While it may be the quasi-optimal transform with regard to energy compaction and basis restriction, its computational

TABLE I
IMPLEMENTATION RELATED ASPECTS OF FILTER TECHNIQUES FOR IQ STREAM SEGMENTATION. THE COMPUTATIONAL EFFORT IS COMPARED IN TERMS OF THE NUMBER OF MULTIPLICATIONS $\mu(M)$ AND ADDITIONS $\alpha(M)$ TO COMPUTE THE NEXT M COMPLEX SUB-CHANNEL SAMPLES.

Transform	$\mu\,(M)$	$\alpha\,(M)$	Overhead	Design Complexity	$d\,[T_s]$
DFT	$2M\log_2 M$	$3M\log_2 M$	none	none	$M-1$
MCLT	$M\,(\log_2 M+5)$	$M\,(3\log_2 M+5)$	none	none	$2M-1$
ECLT	$M\,(\log_2 M+7)$	$M\,(3\log_2 M+7)$	little/moderate	none/extensive	$4M-1$
MDFT	$M\,(\log_2 M+30)+4$	$M\,(3\log_2 M+34)-4$	moderate	extensive	$16M+\frac{M}{2}-1$
EMFB	$M\,(\log_2 M+19)$	$M\,(3\log_2 M+19)$	moderate	extensive	$16M-1$
KLT	$4M^2$	$4M^2-2M$	extensive	none	$M-1$

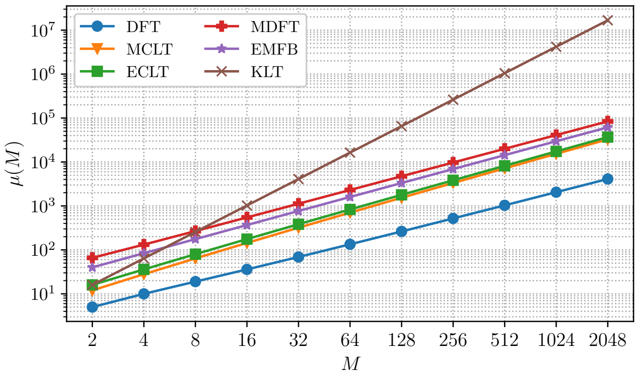


Fig. 8. Quantity of multiplication operations $\mu(M)$, dependent on the number M of new sub-channel samples to be obtained in each transform.

complexity renders it rather impractical for the scenario of an Edge-RAN telemetry reception network. The complexity of the remaining transforms correlates with their respective channel selectivity, where the EMFB structure seems to offer a reasonable balance between those aspects. The qualitative overhead given in Table I also supports these findings, as it is quite extensive for the KLT having to transmit the eigenvectors for each single stream segment, while the filter bank and lapped transforms only rely on a single transmission of prototype coefficients. Under the presented configuration, the MCLT, ECLT and DFT use fixed windows deemed known to the remote site. Therefore, they do not contributing to the overhead. Nevertheless, it has to be noted again, that the ECLT, MCLT introduce two-fold oversampling, reducing the effective comparable compression ratio. The prototype design complexity coheres to the energy compaction ability, showing higher effort for ECLT, EMFB and MDFT, but in exchange also offering a more desirable channel separation performance. The last column represents the overall transient delay $d\,[T_s]$ in terms of sampling periods T_s. It is of interest for a proper layout of buffer storage in the process of stream synchronization. Considering the block transforms as filter banks, their corresponding delay conforms to the filter order $L-1$ [34], with L the prototype length, respectively. The MDFT and EMFB exhibit also here a higher cost. Again, this is related to the longer prototype length and their PR structure. Nevertheless, these schemes bring the benefit of improved frequency selectivity, which is deemed to be of more practical relevance.

V. CONCLUSIONS

This paper investigated means for spectrum decomposition, taking advantage of the narrowband characteristic of telemetry transmissions. A partial IQ exchange between base stations in an Edge-RAN IoT network shall exploit the inherent macro diversity. Partial spectrum combining may mask interferers and shadowing effects seen at various stations and improve the overall decoding success rate. The selective transmission of sub-bands shall also limit the overall data volume, as only a minor spectrum portion is actually populated with low-rate telemetry signals. Utilization of this characteristic is inevitably required for a multi-station collaboration over wireless rate-limited links. With the spectrum division as the most essential building block, this publication gave a holistic view on spectrum decomposition methods. Thereby, the operation conditions and spectral features in narrowband IoT sensor networks have been accounted for, not having been issued by literature so far. The eligibility of the presented approaches has been supported by both, theoretical analysis and simulation on actually recorded real-word data. Constraints, such as critical sampling and Aliasing containment have been given, upon which eligible filter and block transform approaches have been presented. Among those, the Karhunen-Loève Transform (KLT) was found optimal regarding the employed metrics of coding gain and basis restriction error. Despite its performance, the usage of the KLT for spectrum decomposition can be considered critical. The prerequisite of scalability for larger sub-channel numbers M can not be satisfied, because the computational effort and related overhead drastically increases to a non bearable extent. For the given prototype configuration, the Exponentially Modulated Filter Bank (EMFB) performed best among the static approaches, yet at the cost of a higher design and computational effort. While it might be sub-optimal, the fixed structure allows for reduced overhead and shows a robust behavior against deviations of carrier frequencies relative to the transformation grid. In general, it transpires that the transfer function of the prototype is decisive with regard to the metrics of coding gain and basis restriction. With the Modulated Complex Lapped Transform (MCLT) and Discrete Fourier Transform (DFT) being restricted in their prototype length L, that meet the preassigned perfect reconstruction (PR) conditions, they exhibit a notable decay in passband. Furthermore, they and can not provide sufficient stopband attenuation for an adequate energy containment and minimal spectral leakage. The Modified Discrete Fourier Transform (MDFT), Exponentially Modulated Filter Bank (EMFB) and Extended Complex Lapped Transform (ECLT) do not suffer from this limitation. They can be implemented efficiently, yet the design process is deemed simpler and more stable for the case of MDFT and EMFB. While they involve more computational complexity, these near-perfectly reconstructing approaches appear superior over the perfect reconstruction transfoms with respect to selectivity. This aspect is deemed most important for a partial sub-band exchange. Given their high selectivity capability while still exhibiting feasible computational complexity and overhead, the EMFB and MDFT are deemed the best choices for band segmentation within a spectral sub-band compression system for telemetry IQ data.

Acknowledgment

This work has been supported in part by the German Research Foundation (DFG) under grant no. FOR 1508, operating under the name *"BATS"*.

References

[1] J. Robert, H. Lieske, A. Heuberger, J. Bernhard, and G. Kilian, "Large area experimental telemetry network for infrastructure monitoring applications," in *Smart SysTech 2015; European Conference on Smart Objects, Systems and Technologies; Proceedings of*, Jul. 2015, pp. 1–6.

[2] X. Xiong, K. Zheng, R. Xu, W. Xiang, and P. Chatzimisios, "Low power wide area machine-to-machine networks: key techniques and prototype," *Communications Magazine, IEEE*, vol. 53, no. 9, pp. 64–71, Sep. 2015.

[3] M. Schadhauser, J. Robert, and A. Heuberger, "Concept for an adaptive low power wide area (lpwa) bat communication network," in *Smart SysTech 2016; European Conference on Smart Objects, Systems and Technologies*, Jul. 2016, pp. 1–9.

[4] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in lte-advanced systems," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 44–50, Nov. 2012.

[5] Y. Hadisusanto, L. Thiele, and V. Jungnickel, "Distributed base station cooperation via block-diagonalization and dual-decomposition," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, Nov. 2008, pp. 1–5.

[6] B. Guo, W. Cao, A. Tao, and D. Samardzija, "Lte/lte-a signal compression on the cpri interface," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 117–133, Sep. 2013.

[7] A. Vosoughi, M. Wu, and J. Cavallaro, "Baseband signal compression in wireless base stations," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, 2012, pp. 4505–4511.

[8] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3216–3225, 2012.

[9] K. Nieman and B. Evans, "Time-domain compression of complex-baseband lte signals for cloud radio access networks," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, 2013, pp. 1198–1201.

[10] M. Grieger, S. Boob, and G. Fettweis, "Large scale field trial results on frequency domain compression for uplink joint detection," in *2012 IEEE Globecom Workshops*, Dec 2012, pp. 1128–1133.

[11] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[12] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 520–544, Jan 2010.

[13] M. Mishali and Y. C. Eldar, "Xampling: Analog data compression," in *2010 Data Compression Conference*, March 2010, pp. 366–375.

[14] ——, "Sub-nyquist sampling," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 98–124, Nov 2011.

[15] M. A. Davenport, J. N. Laska, J. R. Treichler, and R. G. Baraniuk, "The pros and cons of compressive sensing for wideband signal acquisition: Noise folding versus dynamic range," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4628–4642, Sept 2012.

[16] P. Vaidyanathan, *Multirate Systems And Filter Banks*, ser. Electrical engineering. Electronic and digital design. Dorling Kindersley, 1993.

[17] A. Oppenheim, R. Schafer, and J. Buck, *Discrete-time Signal Processing*, ser. Prentice Hall international editions. Prentice Hall, 1999.

[18] T. Karp and N. J. Fliege, "Modified dft filter banks with perfect reconstruction," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 11, pp. 1404–1414, Nov 1999.

[19] A. Viholainen and M. Renfors, "Alternative subband signal structures for complex modulated filter banks with perfect reconstruction," *2004 IEEE International Symposium on Circuits and Systems*, vol. 3, pp. III–525, 2004.

[20] H. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 3, Mar 1999, pp. 1421–1424.

[21] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, Oct 1986.

[22] A. Akansu and R. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*, ser. Electronics & Electrical. Academic Press, 2001.

[23] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, ser. Prentice Hall Signal Processing Series, A. V. Oppenheim, Ed. Prentice-Hall, Inc., 3 1983.

[24] H. Malvar, *Signal Processing with Lapped Transforms*, ser. Artech House communications library. Artech House, 1992.

[25] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 969–978, June 1990.

[26] G. Doblinger, "A fast design method for perfect-reconstruction uniform cosine-modulated filter banks," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6693–6697, Dec 2012.

[27] R. M. Gray, *Source Coding Theory*, ser. Kluwer international series in engineering and computer science: Communications and information theory. Springer US, 1990.

[28] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall Professional Technical Reference, 1990.

[29] A. Jain, *Fundamentals of Digital Image Processing*, ser. Prentice-Hall information and system sciences series. Prentice Hall, 1989.

[30] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, Sep 2001.

[31] J. Cooley and J. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[32] A. Viholainen, J. Alhava, and M. Renfors, "Efficient implementation of complex modulated filter banks using cosine and sine modulated filter banks," *EURASIP journal on advances in signal processing*, vol. 2006, Jan. 2006.

[33] T. Karp and N. J. Fliege, "Computationally efficient realization of mdft filter banks," in *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, Sep. 1996, pp. 1–4.

[34] J. Alhava, A. Viholainen, and M. Renfors, "Efficient implementation of complex exponentially-modulated filter banks," in *2003 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 4, May 2003, pp. IV–IV.

Dynamic Hand Gesture Pattern Recognition Using Probabilistic Neural Network

Debasish Bal
Dept. of Electrical and Electronic Engineering
International Islamic University Chittagong
Chittagong-4318, Bangladesh
debasishopu@gmail.com

Asif Mohammed Arfi
Dept. of Computer Science and Engineering
University of Yamanashi
Takeda 4-4-37, Kofu, Yamanashi, Japan
arfieee@hotmail.com

Sujoy Dey
Dept. of Computer Science and Engineering
University of Science and Technology Chittagong
Chittagong-4202, Bangladesh
sujoy.ustc2011@gmail.com

***Abstract*—Flex Sensor and Gyroscope based hand gloves are being used widely to detect gesture-based sign language patterns to provide aid for speech-impaired people. However, detecting dynamic gestures is not an easy task due to the variability and redundant implementation precautions. This article approaches a dynamic hand gesture recognition system by using Probabilistic Neural Network (PNN) deliberately focused on practical usage of dynamic gestures on day-to-day life. In this experiment, 10 gesture patterns have generated by using a data glove. For each gesture, 360 training input vectors are generated to train the PNN model, and the output has provided via a speaker.**

Keywords—flex sensor, gyroscope, dynamic gesture, PNN

I. Introduction

There are approximately 360 million people around the world who are deaf and mute [1]. Due to speech and hearing impairment, these people have to communicate by using sign language. By using a convenient Man-Machine Interface (MMI), it is possible to generate the record and validate such sign language gestures to mediate a communication system between mute and unmute people. Kinect sensor-based works are proposed in previous years to design gesture recognition prototype [2]–[6]. Kinect is a depth sensor that has designed by Microsoft. However, these systems are not portable and cheap. User needs stationary Kinect sensor on board to work with it. Authors have used Convolutional Neural Network (CNN) to detect static gestures from the user on [7], [8]. Static sensor-based work also has done using Time-of-flight (ToF) camera [9] and Dynamic Time Wrapping algorithm [10]. As these works are based on the camera sensor, they are highly dependent on the image quality, lighting condition, camera angle and many more factors. There is much approach to detecting various versions of sign language as well. Authors approached to detect Vietnamese, Indian and American Sign Language [11], [12]. Most of these approaches are not capable of detecting dynamic gestures.

This proposed system will approach a dynamic gesture recognition system to eliminate these problems at a much affordable solution as well as portability. This system also consists of a speaker module to generate Bangla speech as output for each corresponding gesture.

II. Methodology

A. Hand Gloves Design

The hand glove has designed using flex sensors attached with the fingers.

- Three flex sensors have used for gesture recognition.
- These three sensors have mounted in index, middle and ring finger.
- An MPU-6050 which is a three-axis and accelerometer and three-axis gyroscope has used. This device uses I2C protocol for communication and has 16-bit built-in ADC channel for high accuracy.
- A Raspberry Pi 3B has used for interfacing the flex sensors and the MPU-6050.
- An MCP-3008 ADC has used to interface the flex sensors with Raspberry Pi as Pi does not have inbuilt analog pins.

B. Mathematical Model

The mathematical models are used to process raw data from sensors and using them for gesture recognition purpose. The first 20 reading of sensors at the time *t* for generating a training set, can be expressed as:

$$\{x_{g,k,1}^{n}(t),\ x_{g,k,2}^{n}(t), x_{g,k,3}^{n}(t), \dots x_{g,k,20}^{n}(t)\} \qquad (1)$$

Where *x* is the signal value, *n* is the user index, *k* is the replicate index, and *t* is the time.

A replicate of a gesture by a person can be expressed as:

$$\Pi_{n,g,k} = \{x_{g,k,1}^{n}(i.dt), x_{g,k,2}^{n}(i.dt), x_{g,k,3}^{n}(i.dt) \dots x_{g,k,20}^{n}(i.dt) : i \in \{0, 1, \dots D_{n,g,k}\}\} \qquad (2)$$

Where *dt* is the interval between sampling of sensors data and $D_{n,g,k}$ is the last reading from sensor by n^{th} user for g^{th} gesture.

C. Overall Framework

The proposed gesture recognition system consists of three stages of processing. The framework consists of three main procedures where the value of gyroscope is preprocessed first. In the next stage, the raw data from flex is retrieved. After gathering these data, the sensors data have trained and tested by PNN further to analyze the recorded gesture. The PNN has specifically chosen to solve the classification problem for each dynamic gestures. The overall system has demonstrated in Fig 1.

D. Reading MPU-6050

The MPU-6050 is consists of a three-axis gyroscope and accelerometer integrated on a single chip. This sensor is also called 6 Degree of Freedom (DoF) device because of its six outputs. It measures the rotational velocity of the rate of change of the angular position over time in X, Y, and Z-axis.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Micro-Electro-Mechanical Systems (MEMS) technology has used inside this sensor.

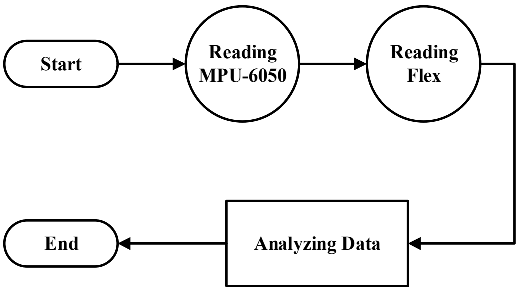


Fig. 1. Overall demonstration of the system

The outputs of the gyroscope are measured in degrees per second. The unit for measuring acceleration is meter per second squared. To read the sensor, we have used the wire library after that the by resetting the power management register; the sensor was reset. The register address is as follows:

TABLE I. MPU-6050 REGISTER ADDRESS

Register	B7	B6	B5	B4	B3	B2	B1	B0
0x6B	DEVICE_RESET	SLEEP	CYCLE		TEMP_DIS	CLKSEL[2:0]		

As the datasheet says, the data of each axis is stockpiled in two registers. The addresses of these registers are shown in Table 2.

TABLE II. MPU-6050 ACCELEROMETER DATA REGISTERS

Register	B7	B6	B5	B4	B3	B2	B1	B0
0x3B	ACCEL_XOUT[15:8]							
0x3C	ACCEL_XOUT[7:0]							
0x3D	ACCEL_YOUT[15:8]							
0x3E	ACCEL_YOUT[7:0]							
0x3F	ACCEL_ZOUT[15:8]							
0x40	ACCEL_ZOUT[7:0]							

To read all the registers value correctly, at first the value of X, Y and Z-axis for first six register are requested. After that, we read the second register, perform 2's complement and combine them appropriately to generate the actual values. Here, the roll φ and pitch θ angles are calculated for normalized accelerometer reading G_p.

$$\frac{G_p}{\left\|G_p\right\|} = \begin{pmatrix} -sin\theta \\ cos\theta sin\varphi \\ cos\theta cos\varphi \end{pmatrix} \Rightarrow \frac{1}{\sqrt{G_{px}^2 + G_{py}^2 + G_{pz}^2}} \begin{pmatrix} G_{px} \\ G_{py} \\ G_{pz} \end{pmatrix} = \begin{pmatrix} -sin\theta \\ cos\theta sin\varphi \\ cos\theta cos\varphi \end{pmatrix} \qquad (3)$$

$$tan\varphi_{xyz} = \left(\frac{G_{py}}{G_{pz}}\right) \qquad (4)$$

$$tan\theta_{xyz} = \frac{-G_{px}}{\sqrt{G_{px}^2 + G_{pz}^2}} \qquad (5)$$

Similarly, the register values of Gyroscope are:

TABLE III. MPU-6050 GYROSCOPE DATA REGISTERS

Register	B7	B6	B5	B4	B3	B2	B1	B0
0x43	GYRO_XOUT[15:8]							
0x44	GYRO _XOUT[7:0]							
0x45	GYRO _YOUT[15:8]							
0x46	GYRO _YOUT[7:0]							
0x47	GYRO _ZOUT[15:8]							
0x48	GYRO _ZOUT[7:0]							

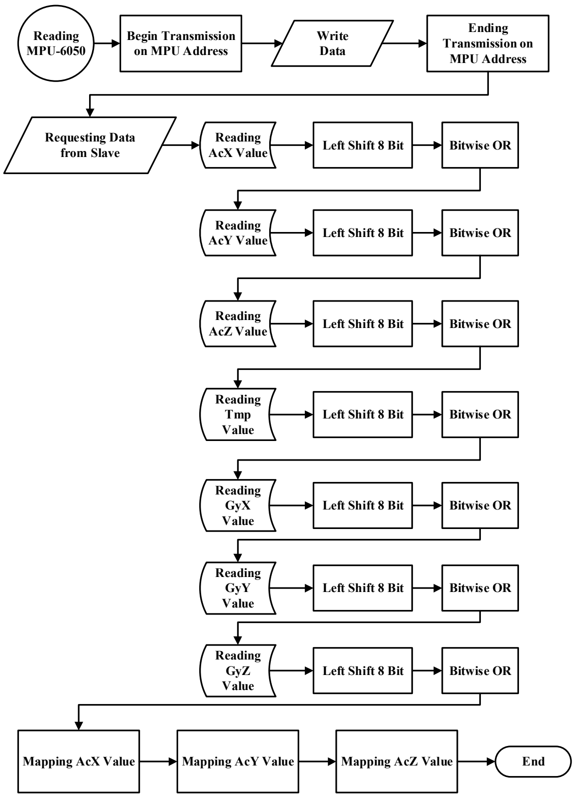


Fig. 2. MPU-6050 Data Reading Procedural

E. *Reading Flex Sensors*

Flex sensors are ideal for measuring the amount of deflection. Generally, these sensors can be capacitive, fiber optic or conductive ink-based.

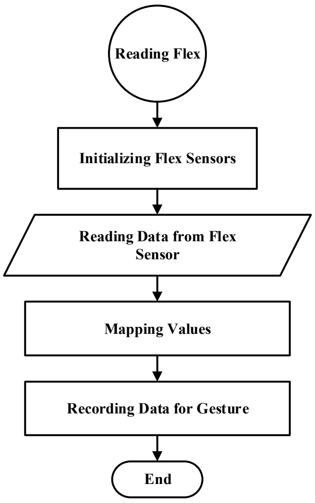

Reading Flex
Initializing Flex Sensors
Reading Data from Flex Sensor
Mapping Values
Recording Data for Gesture
End

Fig. 3. Flex Sensor Data Reading Procedural

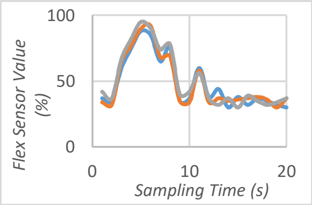
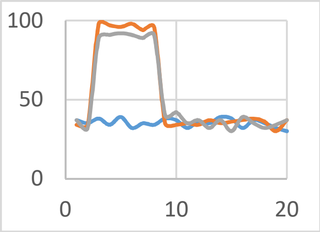
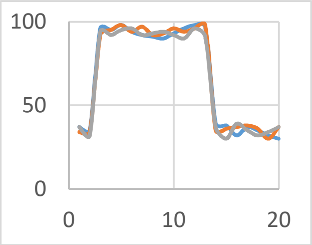
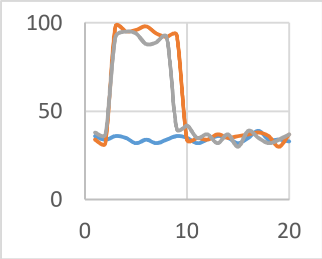
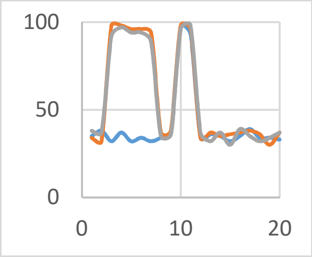
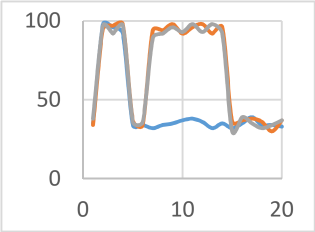
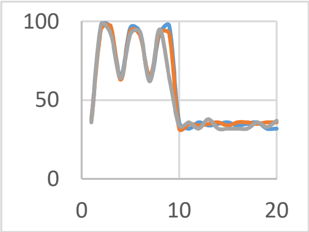
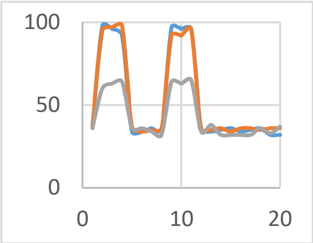
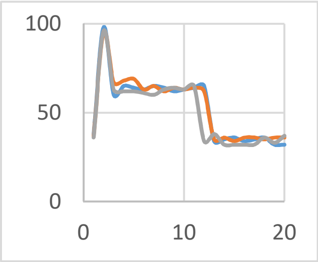
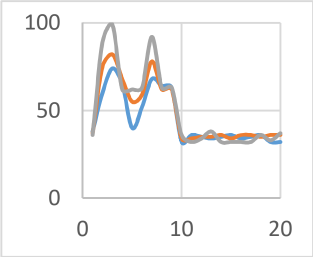
We have used conductive ink-based flex sensors here. Flex sensor works as a variable pin resistor. The output from a flex sensor can be determined from the following equation.

$$V_{out} = V_{in}\left(\frac{R_1}{R_1 + R_2}\right) \tag{6}$$

F. Building Gestures

Gestures have a changeable property, which changes dynamically every time we execute a particular gesture. Therefore, it is necessary to record a gesture for more than once to identify it correctly. The gesture properties are plotted below.

We noted the parameters by recording the values for flex sensors and gyroscope from 88 volunteers for 20 times. We have trained each gesture for 1760 times using a Probabilistic Neural Network algorithm. The algorithm implemented in TensorFlow 1.15 with Python 3.5 GPU support on Nvidia 720m card. The learning rate was 0.1 at 180 epochs.

No	Bangla Speech	Flex Sensor Data
1	আসসালামুয়ালাইকুম (Greetings!)	
2	আমার নাম অপু (My name is Opu.)	
3	আমি চট্টগ্রাম থা□ি । (I live in Chittagong.)	
4	আপনার নাম □ি ? (What's your name?)	
5	আমি ভালো আছি। (I'm fine.)	
6	আপনি □ে মন আছেন? (How are you?)	
7	এই দি□ে আসুন। (Please come here.)	
8	এখানে বসুন। (Please Sit down.)	
9	আমি এ□ দম ভালো নেই। (I am not well.)	
10	বিদায়। (Goodbye!)	

III. Result

To evaluate the performance, the system was tasted with 23 disabled people. Every person has repeated each gesture for 20 times. Therefore, each gesture was tested for 460 times. In the case of gestures 7 and 10, the gesture recognition algorithm was having a problem to detect them accurately. The values of gyroscope are considered at this position to resurrect the detection issue, which was successful for both gestures.

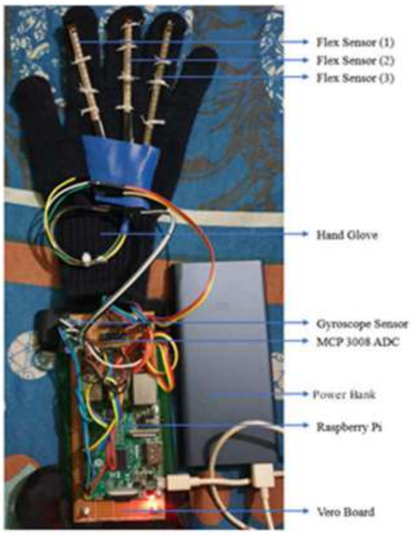


Fig. 4. System Implementation

The Gyroscope data has plotted in Fig 5 to discriminate between the gesture 7 and 10.

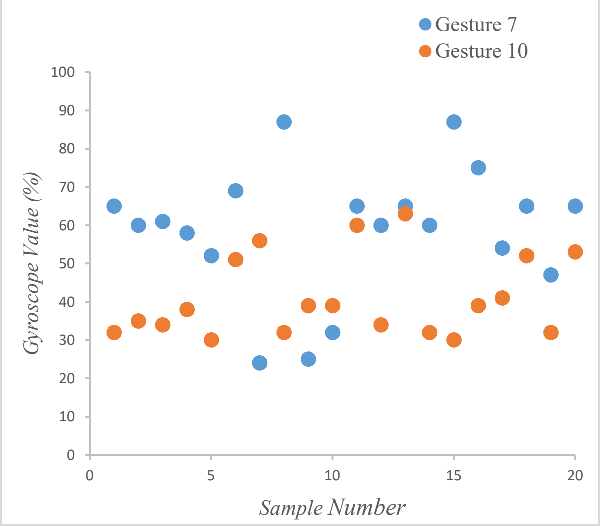


Fig. 5. Gyroscope Datapoints for Gesture 7 and 10

The system has performed overall 92.7% accuracy. The performance can be improved by adding more training datasets.

IV. Conclusion

Our proposed hand gloves have made using three flex sensors. Primarily, that limits our capability to track finger gesture for the remaining two fingers. However, due to this purpose, the device's overall cost is less, which is a tremendous advantage from a consumer perspective. On the other hand, the dynamic gesture that has been used here is entirely chosen by us. As these are not universally recognized patterns, the user may need to adapt to the patterns on the first hand. But the output generated using dynamic gestures are considerably faster than static gesture patterns like American Sign Language (ASL).

Mainly we have proposed a method to recognize dynamic gesture patterns using flex sensor and gyroscope which can be used not only in speech but also in drone control, haptics, augmented reality etc. We have converted the data generated from gloves into a two-dimension matrix where in x-axis sampling time and in y-axis flex sensor's resistances are plot. Thus, it will give us a visual idea about the data. This proposed method helped to identify gestures from different user even though the user response wasn't the same as the training data because we have used a form of pattern recognition.

References

[1] J. Banks, L. Hayes, L. Smith, and K. Millsap, "Deafness and Hearing Loss," in *Encyclopedia of Diversity in Education*, 2013.

[2] L. Liu, X. Wu, L. Wu, and T. Guo, "Static Human Gesture grading based on Kinect," in *2012 5th International Congress on Image and Signal Processing, CISP 2012*, 2012, pp. 1390–1393.

[3] H. Li, L. Yang, X. Wu, S. Xu, and Y. Wang, "Static hand gesture recognition based on HOG with kinect," in *Proceedings of the 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2012*, 2012, vol. 1, pp. 271–273.

[4] T. Q. Vinh and N. T. Tri, "Hand gesture recognition based on depth image using kinect sensor," in *Proceedings of 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science, NICS 2015*, 2015, pp. 34–39.

[5] S. Bhattacharya, B. Czejdo, and N. Perez, "Gesture classification with machine learning using Kinect sensor data," in *Proceedings - 2012 3rd International Conference on Emerging Applications of Information Technology, EAIT 2012*, 2012, pp. 348–351.

[6] S. Bessa Carneiro, E. D. F. M. De Santos, T. M. A. De Barbosa, J. O. Ferreira, S. G. S. Alcalá, and A. F. Da Rocha, "Static gestures recognition for Brazilian Sign Language with kinect sensor," in *Proceedings of IEEE Sensors*, 2017.

[7] X. Guo, W. Xu, W. Q. Tang, and C. Wen, "Research on optimization of static gesture recognition based on convolution neural network," in *Proceedings - 2019 4th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2019*, 2019, pp. 398–400.

[8] X. Wang, Z. Chen, X. Wang, Q. Zhao, and B. Liang, "A comprehensive evaluation of moving static gesture recognition with convolutional networks," in *Proceedings of 2018 3rd Asia-Pacific Conference on Intelligent Robot Systems, ACIRS 2018*, 2018, pp. 7–11.

[9] G. Simion and C. D. Caleanu, "Multi-stage 3D segmentation for ToF based gesture recognition system," in *2014 11th International Symposium on Electronics and Telecommunications, ISETC 2014 - Conference Proceedings*, 2015.

[10] G. Plouffe and A. M. Cretu, "Static and dynamic hand gesture recognition in depth data using dynamic time warping," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 2, pp. 305–316, Feb. 2016.

[11] E. Abraham, A. Nayak, and A. Iqbal, "Real-Time Translation of Indian Sign Language using LSTM," in *2019 Global Conference for Advancement in Technology, GCAT 2019*, 2019.

[12] L. T. Phi, H. D. Nguyen, T. T. Q. Bui, and T. T. Vu, "A glove-based gesture recognition system for Vietnamese sign language," in *ICCAS 2015 - 2015 15th International Conference on Control, Automation and Systems, Proceedings*, 2015, pp. 1555–1559.

Privacy Enhanced Energy Prediction in Smart Building using Federated Learning

Sai Venketesh Dasari
IIIT
Bangalore, India
suryasai.venkatesh@iiitb.org

Kaushal Mittal
IIIT
Bangalore, India
kaushal.mittal@iiitb.org

Sasirekha GVK
IIIT
Bangalore, India
sasirekha@iiitb.ac.in

Jyotsna Bapat
IIIT
Bangalore, India
jbapat@iiitb.ac.in

Debabrata Das
IIIT
Bangalore, India
ddas@iiitb.ac.in

***Abstract*— Prediction of energy consumption is useful in energy budgeting of smart grids, expenditure budgeting by consumers, and comfort management of smart buildings. Essentially, building management systems of smart buildings need to manage energy, efficiently. In order to do this, energy prediction plays an important role. The energy consumption, in general, is predicted using machine learning algorithms. However, machine learning algorithms demand massive amounts of data for performing well. Acquisition of this data from data owners can lead to privacy breaches. Federated learning is a framework of distributed systems which can mitigate privacy breaches to certain extent. Federated learning has as such been developed for mobile edge devices like vehicles, phones etc. In this paper, a novel application of federated learning framework focused on the smart building energy prediction scenario is presented. The architectural details on how the federated learning framework is applied is presented. Also, the performance of this prediction model as compared to centralized method of machine learning is discussed. Deep neural networks are used with the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) dataset to realize and evaluate this architecture.**

Keywords—Smart building, Federated learning, Deep Neural Network, Energy prediction.

I. Introduction

Smart buildings comprise of lighting, surveillance, Heating, Ventilation and Air Conditioning (HVAC), water supply etc. subsystems along with the inhouse kitchen appliances. These subsystems are automatically controlled to provide user comfort energy efficiently. Prediction of energy consumption of such smart buildings is crucial because of the involved energy efficiency related decision making. The predicted energy can help the smart buildings in optimizing their energy consumption. In other words, by using energy prediction, the building management system can automatically minimize the expenditure towards energy, without compromising on the user comfort [1]. Energy predictions can also help energy budgeting by smart grids. Peak demands in certain periods affect the stability of the grid. Hence, the smart grid technology, needs to predict and schedule the load in real-time, and thus achieve the energy demand-supply balance [2] & [3]. The prediction techniques used need to forecast the energy consumption accurately for a specific time interval, under varying environmental conditions, using the past data.

With the advances in the artificial intelligence and data analytics, several Machine Learning (ML) based building energy consumption prediction techniques have been discussed in literature. For example, [4] proposes a novel hybrid prediction approach based on the evolutionary deep learning method combining genetic algorithm with long short-term memory. [5] discusses ML approaches like artificial neural network, Support Vector Machines (SVMs), Gaussian-based regressions, and clustering applied in forecasting and improving building energy performance. [6] presents a deep learning-based model to predict cooling energy consumption of a building based on outdoor weather conditions. A composite model of power prediction based on fuzzy C-means clustering and grey wolf optimized back propagation neural network is proposed in [7]. However, it has been realized that, in order for any ML algorithm to perform well, it is crucial to train the model with large amounts of data.

The data required to train the ML algorithms could be sensitive and be a threat to the data owners' privacy. The Internet of Things (IoT) devices and applications are usually deployed in our homes and workplaces, while the ML algorithms run remotely on cloud. This can result in unethical usage of the data shared with the cloud applications. Sharing of data even within an organization has issues because of privacy, competition, policies etc. These privacy concerns get aggravated further, when cloud services are adopted and inter-organizational data needs to be handled. The commercial value of that data plays an important role in these privacy breaches.

In order to tackle the serious privacy breaches while using cloud-based ML applications, privacy-sensitive data need to be stored and processed securely, privately, and within the ethical regulations. Electronic health/medical records, location information, browsing history, photographs are few examples of such private data. General Data Protection Regulation (GDPR) enforced by the European Union on May 25, 2018, aims to protect users' personal privacy, and provide data security [8]. GDPR suggests techniques like data minimization, anonymization, pseudonymization etc. for privacy protection.

978-1-6654 -4067-7/21/$31.00 ©2021 IEEE

Adhering to these regulations helps in building trust towards the data handlers.

In addition to the regulations, privacy enhanced ML techniques play an important role in preventing privacy and ethical breaches. For Eg., homomorphic encryption, differential privacy, multi-party computation, and federated learning are few such techniques. Such privacy enhanced ML techniques are being researched actively. In [9], a hybrid approach for breaking down large, complex Deep Neural Networks (DNNs) for cooperative, privacy-preserving analytics using a concept of Siamese fine-tuning is discussed. [10] describes how non-linear SVMs can be practically used for image classification. Classification is applied on data encrypted with a Somewhat Homomorphic Encryption (SHE) scheme. [11] compares the efficiency and privacy of two emerging solutions to privacy-preserving ML, namely local differential privacy and federated learning. Differential privacy is a distributed data collection strategy, where each client adds digital noise to data locally before submitting to the server. Federated learning, on the other hand, is for multi-parties to train privacy preserving ML models in the cloud services.

Federated Learning (FL) collaboration method of privacy enhancement was first coined by Google [12] & [13]. It is a privacy-preserving distributed machine learning framework, where the model reaches the data, instead of data reaching the model. The main idea in this framework is to build ML models based on datasets that are located at the data owners' device, thus preventing data leakage. The advances in the edge processing are synergizing such distributed data frameworks.

In this paper, the application of federated learning framework towards development of privacy enhanced energy prediction architecture for smart buildings is presented. A DNN model applied to the ASHRAE dataset is used to realize this framework. The framework is implemented in Python using PySyft and PyTorch libraries [14]. The accuracy and timing performances of this framework are compared with the centralized approach of ML. Section II describes the scenario & dataset along with the FL Framework. Section III discusses the data partitioning and feature engineering. Section IV shows the design, implementation, and evaluation of the framework. Finally, the conclusions are presented.

II. Scenario & Dataset

The architecture is of a horizontal federated-learning system, where all participants have the same data structure. The participants collaboratively train their machine-learning model with the help of a cloud server. A typical assumption is that the participants are honest, whereas the server though honest is curious. Therefore, no leakage of information from any participants to the server is allowed.

Fig.1 shows the scenario where the N different buildings defined as N data owners $\{B_1, . . . B_N\}$. Each of the data owners train a local machine-learning model using their respective datasets $\{D_1, . . .D_N\}$ [15].

Training process of such a system usually constitutes of the following steps. **Initialization:** The model owner at the cloud decides upon the architecture of the DNN model and identifies the participants by selecting random buildings. Buildings get the weather data from the Weather Forecaster. Model owner sends a copy of its model to buildings.

Step 1: At each building the data owner uses the local data to train its local model and compute gradients. The masked gradients are sent to the Trusted Third Party (TTP) server.

Step 2: The TTP server performs secure aggregation without learning information about any participant. The aggregated model is sent to the model owner for updating the global model.

Step 3: The model owner sends back the updated global model to the buildings' data owners.

Step 4: Data owners update their respective local models.

The random selection of data owners (buildings in this case) improves the model as discussed in [16]. Inclusion of all the large number of workers (data owners) in a training epoch, can create a situation wherein, changes done by one data owner will be cancelled out by the other. In other words, the model parameters from a data owner can become insignificant. Due to this situation, important changes recommended by one of the buildings will not be captured in the global model. This issue can be solved by using a small number of random data owners for each global epoch. However, if the number of workers is too small, then the model may take a very long time to converge. For aggregation performed in step 2, [16] suggests taking a weighted average of the received model. The weights are proportional to the amount of data contributed towards training by that data owner. Let n_k is the amount of data possessed by the k^{th} data owner and n be the total amount of data on which the model has been trained. Then, the weights at the model owner at $t+1$ th global epoch are taken as the weighted average of the weights from data owners as in (1).

$$w_{t+1} \leftarrow \sum_{k=1}^{N} \frac{n_k}{n} w_{t+1}^{k} \qquad (1)$$

Step 1 to 4 are repeated till convergence of the loss function. The objective of the local model is to minimize the loss function ϵ defined as in (2). The Root Mean Square Log Error (RMSLE) of federated learning is computed as an average of RMLSE of local models.

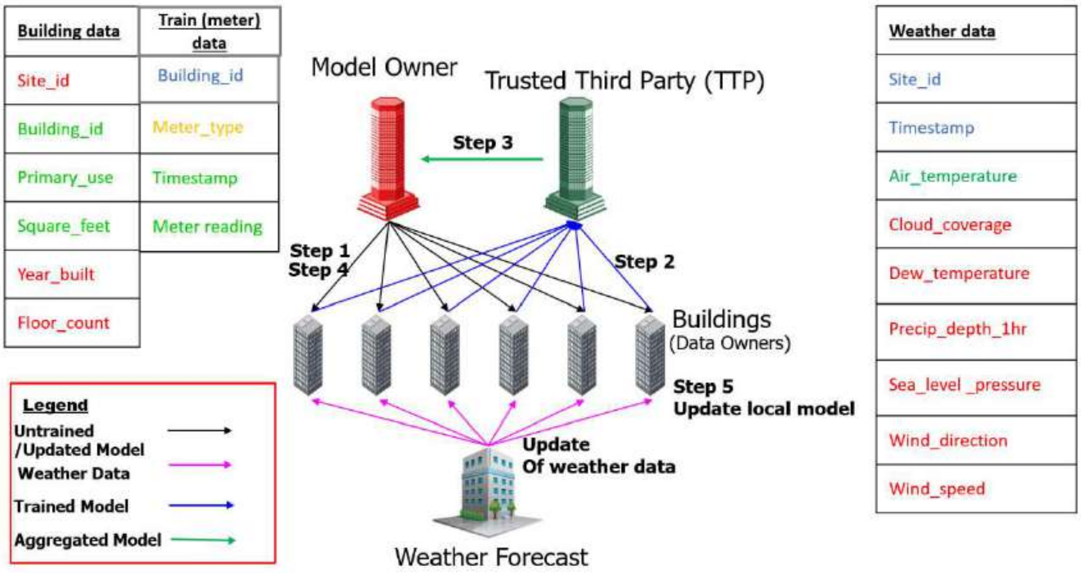
$$\epsilon_{FL} = \frac{1}{N} \sum_{k=1}^{N} \epsilon_k \qquad (2)$$

At each local model the RMSLE is given by (3).

$$\epsilon_k = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(log(p_i + 1) - log(a_i + 1)\right)^2} \qquad (3)$$

Where, m = number of datapoint instances per batch. p_i = predicted meter reading for i^{th} row, a_i = actual meter reading for i^{th} row [17].

The dataset considered is from ASHRAE [17]. It contains the building data, meter data and the weather data provided as a part of the ASHRAE great energy predictor III competition. The competition is to build models, to accurately predict the energy consumption or meter reading. The meters are of four different types: namely, electricity, chilled water, hot water, and steam meters. The data features are listed in Fig. 1.



Building data
Site_id
Building_id
Primary_use
Square_feet
Year_built
Floor_count

Train (meter) data
Building_id
Meter_type
Timestamp
Meter reading

Weather data
Site_id
Timestamp
Air_temperature
Cloud_coverage
Dew_temperature
Precip_depth_1hr
Sea_level _pressure
Wind_direction
Wind_speed

Fig. 1: Scenario of federated learning for smart buildings and dataset features

III. FEATURE ENGINEERING

A. Dataset description:

The ASHRAE dataset is vast, comprising data from more than 1000 buildings over a span of 3 years and of the following 3 types.

1) Train: Size of train data (20216100, 4).

Train data consists of a) Building_id - Refers to the identity (ID) of the building out of 1448 buildings, corresponding to the row, b) Meter - (0-Electricity, 1- Chilled Water, 2 - Steam, 3- Hot Water). c) Timestamp - It is the time corresponding to the meter reading. d) Meter Reading - The target variable after scaling it to with log function. The loss function is RMSLE on the meter reading.

2) Building: Size of building data (1449, 6)

Building Data contains a) Site_id - It is an ID pointing to the site at which the building is present, b) Building_id - A unique ID representing each building. c) Primary_use - it refers to the main purpose of the building, for what this building is being used for. For eg., education. d) Square_feet -It is the gross area of the building. e) Year_built -Year building was opened. f) Floor_count - Number of floors of the building.

3) Weather: Size of weather data (139773, 9)

Weather Data: a) Site_id: Unique Id representing the site of measurements. b) Timestamp: Time at which the reading was taken. c) Air_temperature: provided in degree Celsius. d) Cloud_coverage: Portion of the sky covered in clouds, in oktas e) Dew_temperature: provided in degree Celsius f) Precip_depth_1_hr: It refers to the amount of rainfall/precipitation occurred in one hour. It is measured in mm. g) Sea_level_pressure: refers to air pressure h) Wind_direction: The direction of wind flow in degrees (0-360). i) Wind_speed: It is measured in meters per second.

B. Preprocessing

Preprocessing involving cleaning, transforming, and reducing is based on the observations, which are discussed further in this sub-section.

Observation 1: Not every building has every type of meter. Majority of the buildings have electricity meters (meter 1). This non-uniform presence of meters can be seen in Fig. 2 which shows a plot of the meters present in each building.

Observation 2: A periodic trend in the meter reading across the day, week and yearly is noticed. These trends are expected as the energy consumption primarily depends on the weather patterns across the year as well as day.

Observation 3: Primary_use feature contains categorical data. Only five dominant categories, namely education, office, entertainment public services and lodging are seen and the number of samples in other categories are negligible.

Observation 4: The air temperature variation is highly correlated to the meter reading as expected.

To capture the trends which were observed in observation 2, timestamp column is split into 4 new columns: Hour, Day, Weekend and Month. 1 represents weekend, that is, when the day is Saturday or Sunday else it is 0. The tables are merged with features shown in red of Fig. 1, i.e., Site_id, Building_id and Timestamp. The following features are dropped after this join. 1. Floor_count: Most of the Buildings (80%) do not have floor count or have very low floor count. Year_built - Again most of the buildings do not have Year_built. No correlation between Year_built and power consumption has been observed. Next, Precip_depth_1_hr, Sea_level_pressure, Wind_direction, Wind_speed, Cloud_coverage, Dew_temperature and Site_id are the other features dropped as these do not show significant correlation with the meter reading.

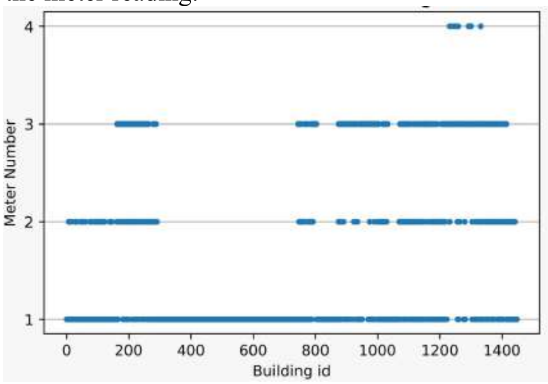


Fig. 2: Building_id versus meters available

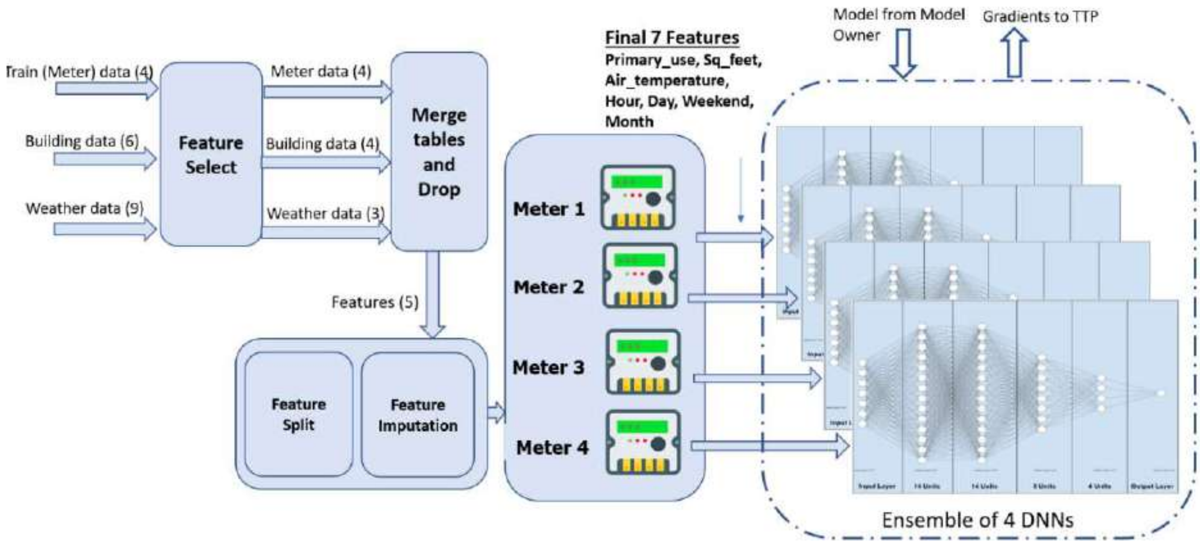


Fig. 3: Feature engineering and DNN model at each Data owner

The rows with meter reading 0 are dropped, as the hourly power consumption is not expected to be 0 generally, and 0-meter reading also does not contribute towards training. The missing values in a column are imputed with the median of the corresponding column. Since the buildings may not have all the four types of meters, an ensemble of 4 models has been created, each for predicting for meter. The data was thus split based on meter type. The shortlisted features are shown in green in Fig. 1. With the timestamp split into Hour, Day, Weekend, Month, along with Air_temperature, Square_ft, Primary_use are the 7 features fed to the DNN.

DNN comprises of a fully Connected Feed Forward Neural Network with Rectified Linear Unit (ReLU) as the activation function. The architecture of the model is arrived at progressively, with number of layers being added and a wider layer provided at the start to overcome underfitting.

The predicted meter reading with the Timestamp, Building_id and Site_id can be used in the visualization and further usage of the energy prediction data, which is beyond the scope of this paper. Fig. 3 details the processes discussed in this section.

IV. Implementation & Evaluation

The Jupyter notebook-based Python scripting using PySyft and PyTorch libraries is used to develop and evaluate the centralized and federated learning models. Both the models are trained in a CPU provided by Google Colab basic/free version.

A. *Centralized learning:*

In order to improve the training speed, the number of input variables (as mentioned in feature engineering) are reduced. The data normalization also helps the speed up the training process. The convergence of the model with different learning rates is observed experimentally. Results for the Learning rate = 0.001 and Batch size (m) = 512 are as seen in Fig. 4. The results presented here are for all the 1448 buildings' data considered meter wise. The training and validation tests were carried out epoch wise till convergence. The number of epochs required to converge, and the final accuracy depends on the distribution of data. Meters 2, 3 & 4 have data containing higher number of 0 values as compared to meter 1. Hence, higher number of epochs are required for convergence in case of meters 2, 3, 4. The reason for meter readings being 0 could be for instance, hot water not turned-on during summer months, resulting in 0 in meter 4 readings.

B. *Federated Learning:*

The FL framework uses the architecture as shown in Fig.1 and the ML model described in section III. 3 local epochs are considered for each global epoch fixed after experimentation. In order to delve deeper into the analysis, only meter 1 is considered and reduced sizes of building data set are considered.

1) *Accuracy:*

Fig.5 shows the RMSLE versus global epochs of federated learning for different sizes of datasets (10, 25, 50, 100) buildings for meter 1 computed using (2). It is observed that the performance does not vary drastically with increasing global epochs. In Fig. 6, the variation of RMSLE for increasing building numbers for both centralized and federated learning are seen. The centralized learning improves with increasing data size as expected. There is no such improvement in case of federated learning. This can be explained that, as the number of buildings increases and the changes due to one building gets canceled out by the other buildings. But this very fact is very useful because it prevents the global model from overfitting. The performance of the centralized model improves as expected.

The parameters that can be varied for experimentation at ML model level are the learning rate, batch size and depth of the model. In FL framework, the number of local epochs for one communication round, can be varied to study the performance. Better aggregation techniques are also recommended.

2) *Timing:*

Fig. 7 shows average time (in seconds) taken by the models to train for one epoch with different amounts of data. It can be observed that there is a massive difference between time taken by centralized model and federated learning model, approximately 70-100 times. To understand the timing requirements, a parallel algorithm has been implemented where in the PySyft calls of federated learning are avoided. Multiple ML models are run for each building data simultaneously. For realizing this model, the complete dataset is split according to data corresponding to each building and is stored in an array. Similarly, an array is made to store models corresponding to each building. The models are then trained with respective datasets using pure PyTorch. The training process is the same as centralized and operations are performed on simple tensors, unlike PySyft where pointers are used for operation on SyftTensors. Also, the weights for each model can be directly updated here. In comparison, each parameter in PySyft is a pointer to the value which needs to be updated [18].

The timing of this learning implementation is very close to that of centralized learning. Parallel learning has an overhead compared to centralized training as one needs to load and save multiple models and datasets. This overhead is trivial as compared to the delays due to inter process communications between the data owners and model owner in the FL framework, using PySyft. This inter process communication in PySyft is by means of sending or sharing tensors between workers (processes). It should be noted that all models are simulated in the same system. Overhead also increases as the number of buildings increases as expected.

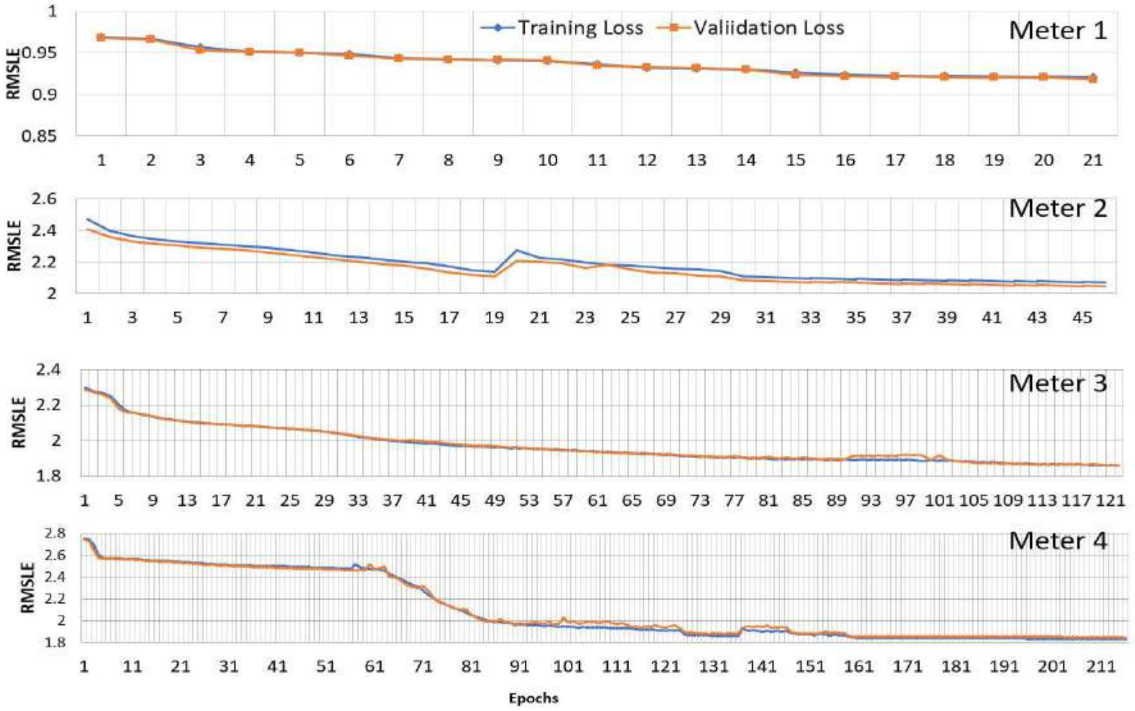


Fig. 4: RMSLE versus epochs of the DNN for different meters

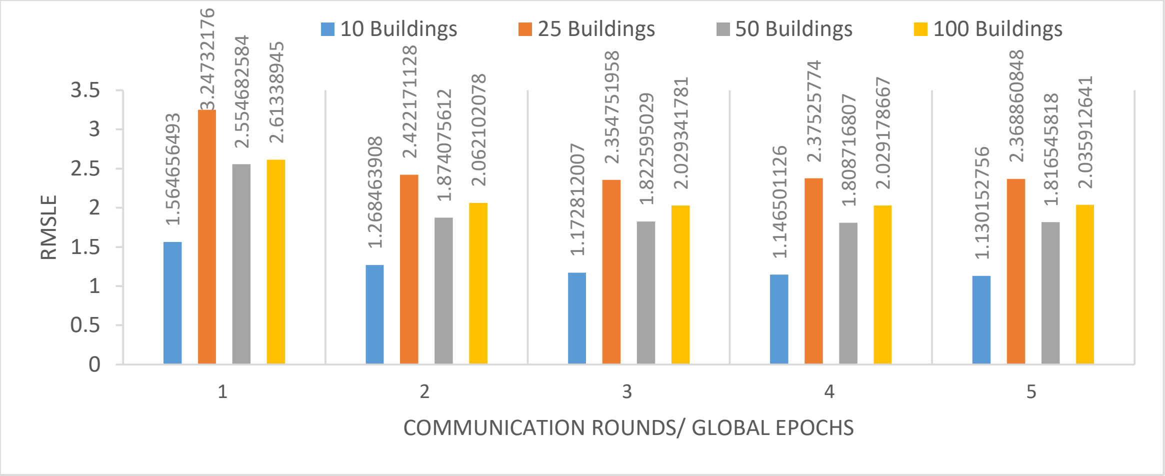


Fig. 5: RMSLE versus communication rounds of federated learning

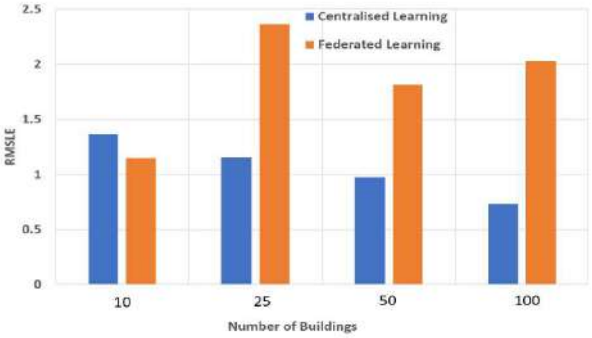


Fig. 6: RMSLE with increasing number of buildings centralized versus federated learning

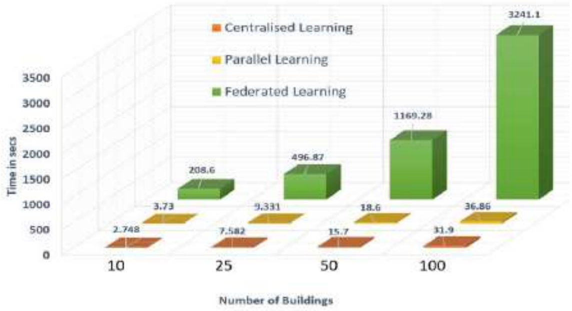


Fig. 7: Time taken with different number of building data for centralized, parallel, and federated learning.

V. Conclusions

The federated learning framework was applied successfully to the smart building energy prediction scenario. This enhances the privacy of the smart building energy prediction system as the data owners do not share the data with the model owners. Introduction of the TTP further helps in secure aggregation. The design details of how the federated learning framework is applied is presented along with the performance of this prediction model as compared to centralized method of machine learning. Further work can include more efficient implementation using better aggregation techniques than weighted average. Homomorphic encryption or differential privacy can be applied to enhance to prevent data and model leakages.

In centralized learning the local data is available at the cloud and can be de-anonymized. For example, by matching Site_id & timestamp of building data with the weather data, Building_id attribute can be narrowed down to very small number. The corresponding meter readings give an indication of occupancy/non-occupancy, aiding the burglary plans. In the case of federated learning data is not visible to anyone expect its owner, which will eliminate the risk of leaking sensitive information.

Acknowledgement

The authors are thankful to Ministry of Electronics (MeitY) and IT, Government of India and Cognizant Technologies Ltd. for supporting this work. under the project titled COPAS.

References

[1] I. Sülo, S. R. Keskin, G. Dogan and T. Brown, "Energy Efficient Smart Buildings: LSTM Neural Networks for Time Series Prediction," 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey, 2019, pp. 18-22, doi: 10.1109/Deep-ML.2019.00012.

[2] A. Almalaq, J. Hao, J. J. Zhang and F. Wang, "Parallel building: a complex system approach for smart building energy management," in IEEE/CAA Journal of Automatica Sinica, vol. 6, no. 6, pp. 1452-1461, November 2019, doi: 10.1109/JAS.2019.1911768.

[3] Khaled Alzaareer and Claude Ziad El-Bayeh , " Steps towards smart energy self-sufficient buildings", Article, Newsletter, IEEE Smartgrid(2020) https://smartgrid.ieee.org/newsletters/april-2020/steps-toward-smart-energy-self-sufficient-buildings

[4] A. Almalaq and J. J. Zhang, "Evolutionary Deep Learning-Based Energy Consumption Prediction for Buildings," in *IEEE Access*, vol. 7, pp. 1520-1531, 2019, doi: 10.1109/ACCESS.2018.2887023.

[5] Seyedzadeh, S., Rahimian, F., Glesk, I. et al., " Machine learning for estimation of building energy consumption and performance: a review," in Eng. 6, 5 (2018). https://doi.org/10.1186/s40327-018-0064-7.

[6] Amasyali, Kadir & El-Gohary, "Deep Learning for Building Energy Consumption Prediction", Nora1,2 1 University of Illinois at Urbana-Champaign, United States (2017).

[7] Ying Tian, Junqi Yu, Anjun Zhao,"Predictive model of energy consumption for office building by using improved GWO-BP," in Energy Reports,Volume 6, 2020, Pages 620-627, ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2020.03.003.

[8] [website], "General Data Protection Regulations," https://gdpr-info.eu/, [Accessed on 09.02.2021].

[9] S. A. Osia et al., "A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics," in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4505-4518, May 2020, doi: 10.1109/JIOT.2020.2967734.

[10] Barnett, A. et al. "Image Classification using non-linear Support Vector Machines on Encrypted Data." *IACR Cryptol. ePrint Arch.* 2017 (2017): 857.

[11] H. Zheng, H. Hu and Z. Han, "Preserving User Privacy for Machine Learning: Local Differential Privacy or Federated Machine Learning?," in *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 5-14, 1 July-Aug. 2020, doi: 10.1109/MIS.2020.3010335.

[12] M. Aledhari, R. Razzak, R. M. Parizi and F. Saeed, "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," in IEEE Access, vol. 8, pp. 140699-140725, 2020, doi: 10.1109/ACCESS.2020.3013541.

[13] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," in IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, May 2020, doi: 10.1109/MSP.2020.2975749.

[14] Andre Macedo Farias, "Private AI — Federated Learning with PySyft and PyTorch," Avaliable at: https://towardsdatascience.com/private-ai-federated-learning-with-pysyft-and-pytorch-954a9e4a4d4e , [Accessed on 8.02.2021].

[15] Yang, Qiang; Liu, Yang; Chen, Tianjian; Tong, Yongxin, " Federated Machine Learning," in ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19. doi:10.1145/3298981.

[16] McMahan, H., Eider Moore, D. Ramage, S. Hampson and Blaise Agüera y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data." *AISTATS* (2017).

[17] Clayton et.al.,"The ASHRAE Great Energy Predictor III competition: overview and results., Science and Technology for the Built Environment, 26:10, 1427-1447, (2020), DOI: 10.1080/23744731.2020.1795514 , arXiv:2007.06933.

[18] Théo Ryffel et.al.,"A generic framework for privacy preserving deep learning", Available: https://arxiv.org/pdf/1811.04017.pdf, [Accessed on 8.02.2021].

Virtex 7 FPGA Implementation of 256 Bit Key AES Algorithm with Key Schedule and Sub Bytes Block Optimization

Mahendrakumar Gunasekaran
Xilinx India Pvt Ltd, Hyderabad
gmah@xilinx.com

Kumar Rahul
Xilinx India Pvt Ltd, Hyderabad
kumarr@xilinx.com

Santosh Yachareni
Xilinx India Pvt Ltd, Hyderabad
santoshy@xilinx.com

***Abstract*— Hardware Security plays a major role in most of the applications which include net banking, e-commerce, military, satellite, wireless communications, electronic gadgets, digital image processing, etc. Cryptography is associated with the process of converting ordinary plain text into unintelligible text and vice versa. There are three types of cryptographic techniques; Symmetric key cryptography, Hash functions and Public key cryptography. Symmetric key algorithms namely Advanced Encryption Standard (AES), and Data Encryption Standard use the same key for encryption and decryption. It is much faster, easy to implement and requires less processing power. The proposed 256-bit AES algorithm is highly optimized in Key schedule and Sub bytes blocks, for Area and Power. The optimization has been done by reusing the S-box block. We are optimizing the algorithm with a new approach where internal operations are 32-bit operations, as compared to 128-bit operations. The proposed implementation helps in re-using the same hardware in a pipelined fashion which results in an area reduction by 72% using slice registers, 62% using slice LUT's and 61% using LUT-FF Pairs. This in turn results in a power reduction by 78% in a FPGA implementation. The throughput (Mbps) of the proposed implementation using Virtex-7 (xc7vx485tffg1157) FPGA improved by 10%.**

Keywords*— *AES *(Advanced Encryption Standard),* ***FPGA*** *(field programmable gate array),* ***LUT*** *(Look up table),* ***Mbps*** *(megabit per second),* ***sub*** *(sub bytes),* ***shift*** *(shift rows),* ***mix*** *(mix column),* ***add*** *(add round key).*

I. Introduction

Cryptography is associated with the process of converting ordinary plain text into unintelligible text and vice versa. There are three types of cryptographic technique namely - Symmetric key cryptography, Hash functions and Public key cryptography.

Symmetric key algorithms namely Advanced Encryption Standard (AES), Data Encryption Standard use the same key for encryption and decryption. It is much faster, easy to implement and requires less processing power.

In this paper, we have discussed about implementation of area, power and performance-based architecture of 256-bit key AES algorithm. Also, we discussed about the PPA comparison of the conventional and proposed based implementation in FPGA.

II. Architecture of AES Algorithm

A. Architecture of AES algorithm

AES algorithm implementation is done using four operations namely Sub Bytes, Shift Rows, Mix Columns and Add Round Key. Fig. 1 shows the architecture of 256-bit AES algorithm. In total there are 14 rounds of operation for encryption and 14 rounds for decryption. The ciphertext after encryption will be transmitted across the channel. The receiver side will decrypt the message using same key which is used in encryption.

In 256-bit AES algorithm, the key size is 256 bits, but all the data size is 128 bits. Data include message to be encrypted, cipher text and the decrypted message.

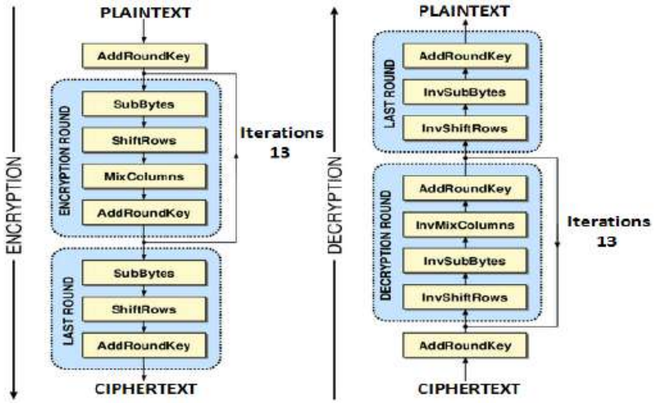
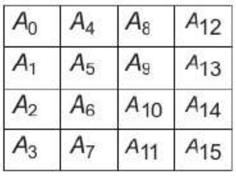


Fig. 1 Architecture of 256 AES Algorithm

Fig. 2 explains the internal data structure of 128-bit data. The 128-bit data is used as 4x4 matrix, where each elements of the matrix is of 8 bits. Since all the four operations are performed on columns basis, we convert the 128-bit data in 4x4 matrix with each element being 8 bits.



A_0	A_4	A_8	A_{12}
A_1	A_5	A_9	A_{13}
A_2	A_6	A_{10}	A_{14}
A_3	A_7	A_{11}	A_{15}

Fig. 2 Data Structure of 128-bit Message

III. Implementation of AES Algorithm

This paper explains about the implementation of both conventional and proposed architecture of 256 AES algorithm. This paper also compares the Power, Performance and Area number in FPGA implementation.

A. Conventional Implementation of 256 bit AES Algorithm

256-bit AES encryption block is implemented in 14 rounds. Each round consists of Add Round Key, Sub Bytes, Shift Rows, Mix column. Round 0 consists of only Add round Key operation as shown in Fig. 3. Round 14 consists of Sub Bytes, Shift Rows and Add Round Key operations, which need 3 clock cycles as shown in Fig. 3. Rounds 1 to 13 consists of all the four operations as shown in Fig. 13. We do a distinct operation in each clock cycle. Hence once the hardware has been implemented for Add Round Key, Sub Bytes, Shift Rows, Mix column, the same hardware can be used for all the 14 rounds [7] [10]. None of the four operations shares the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

same clock cycle. Fig. 3 shows the sequence of round operation with specific sequence of 4 operations to complete the AES encryption. AES algorithm is serial process, i.e. output of first round is the input to the second round. Hence, we can use the same hardware for each round.

Round 0	01 cycle
Round 1 to 13	52 cycles
Round 14	03 cycles
Total	**56 cycles**

Table: 1 Cycles required in Each Round

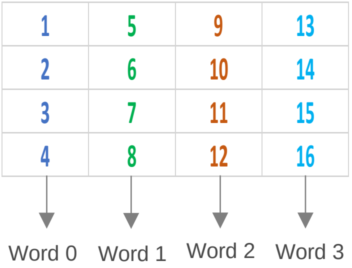
		Cycle
Round 0	Add Round Key	1
	Sub bytes	2
	Shift Rows	3
	Mix column	4
Round 1	Add Round Key	5
	Sub bytes	6
	Shift Rows	7
	Mix column	8
Round 2	Add Round Key	9
	Sub bytes	10
	Shift Rows	11
	Mix column	12
Round 3	Add Round Key	13
	.	.
	.	.
	.	.
	Sub bytes	54
	Shift Rows	55
Round 14	Add Round Key	56

Fig. 3 Structure of conventional implementation

Fig. 3 shows the structure of conventional implementation of 256-bit key AES algorithm.

B. Data Structure

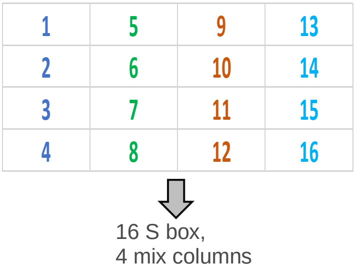
Fig. 4 shows the data structure of 128-bit matrix. Each column consists of 4 elements of 8 bits each, so in total we have 32 bits per word.



1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

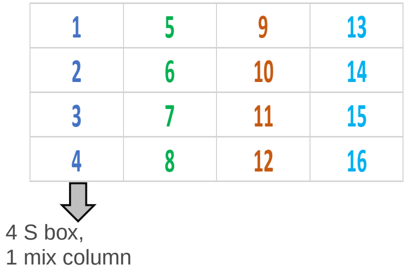
Fig. 4 Word Format of 128-bit Data

Fig. 5 shows the number of S-box required and Mix Column required to implement conventional AES algorithm.



1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

Fig. 5 S-box required for conventional Method



1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

Fig. 6 S box required for Proposed Method

Fig. 6 shows the proposed 32-bit AES implementation. We are doing operations per word (32-bits) in each cycle. Number of blocks required for conventional (128 bit) and proposed (32-bit) implementation are as follows

1) *S box - 16, 4 per clock cycle*
2) *Mix column block - 4, 1 per clock cycle*

In the proposed implementation, we are using the same S-box hardware for both encryption and decryption. Affine transform is the only difference between S-box (encryption) and inverse S-box (decryption). All the other logic (for encryption and decryption) is same for the S-box and inverse S-box. Hence, we are reusing every logic other than the Affine transform for encryption and decryption. Fig. 7 shows the mux selection between S-box and inverse S-box [1]. While doing AES encryption S-box path is chosen and while doing AES decryption inverse S-box path is chosen [1].

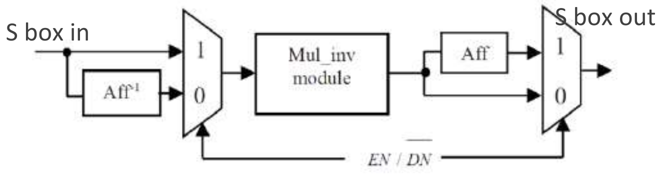


Fig. 7 Combined structure of S box and Inverse S box

C. Proposed 32 bit operations Implementation

In proposed 32-bit operation method, we are reusing S-box and Mix Column blocks. In proposed design "Mix Column" and "Add Round Key" together we called Mix block.

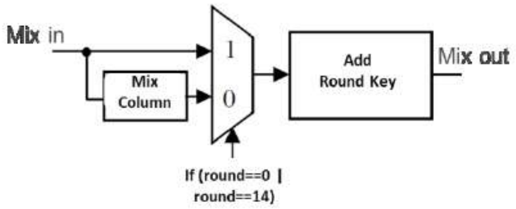
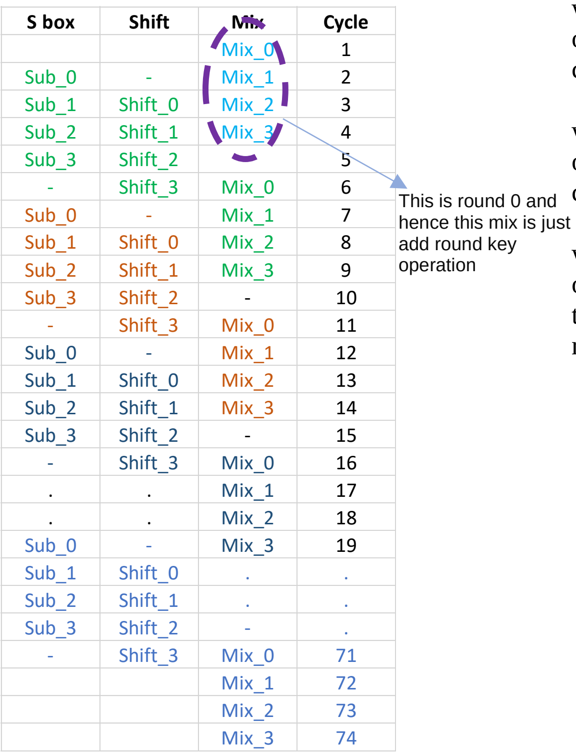


Fig. 8 Combined structure of Mix Column and Add Round Key - Mix

Fig. 9 shows the pipeline structure of the proposed design, where each color represents different round as follows,

Mix – round 0
Mix – round 1
Mix – round 2
Mix – round 3
Mix – round 14



S box	Shift	Mix	Cycle
		Mix_0	1
Sub_0	-	Mix_1	2
Sub_1	Shift_0	Mix_2	3
Sub_2	Shift_1	Mix_3	4
Sub_3	Shift_2		5
-	Shift_3	Mix_0	6
Sub_0	-	Mix_1	7
Sub_1	Shift_0	Mix_2	8
Sub_2	Shift_1	Mix_3	9
Sub_3	Shift_2	-	10
-	Shift_3	Mix_0	11
Sub_0	-	Mix_1	12
Sub_1	Shift_0	Mix_2	13
Sub_2	Shift_1	Mix_3	14
Sub_3	Shift_2	-	15
-	Shift_3	Mix_0	16
.	.	Mix_1	17
.	.	Mix_2	18
Sub_0	-	Mix_3	19
Sub_1	Shift_0	.	.
Sub_2	Shift_1	.	.
Sub_3	Shift_2	-	.
-	Shift_3	Mix_0	71
		Mix_1	72
		Mix_2	73
		Mix_3	74

This is round 0 and hence this mix is just add round key operation

Fig. 9 Pipelined structure of mix operation

From Fig. 9, each word having a size of 32 bits. In **cycle 1**, we are doing Mix operation of word 0 (mix_0). We can denote this as cycle1[round0(mix_0)].

Hence this 32-bit word is available to undergo 32-bit Sub operation. Hence in **cycle 2** we are doing sub operation of word 0 (sub_0) and Mix operation of word 1 (mix_1). We have valid input for "sub" block word 0 in clock cycle 2, and hence we don't need to wait for all the 4 words "mix" block operation to complete. We can denote this as cycle2[round1(sub_0), round0(mix_1)].

In clock **cycle 3**, we are doing sub operation for word 1 (sub_1), shift operation of word 0 (shift_0) and mix operation of word 2 (mix_2), and. We can denote this as cycle3[round1 (sub_1, shift_0), round0(mix_2)].

In clock **cycle 4**, we are doing sub operation for word 2 (sub_2) and shift operation for word 1 (shift_1) and mix operation for word 3 (mix_3). We can denote this as cycle4[round1(sub_2, shift_1), round0(mix_3)].

Since all 128-bit (4 words) round 0 mix operation completed, we don't have mix operation in cycle 5. In clock **cycle 5**, we are doing sub operation for word 3 (sub_3) and shift operation for word 2 (shift_2). We can denote this as cycle5[round1(sub_3, shift_2)].

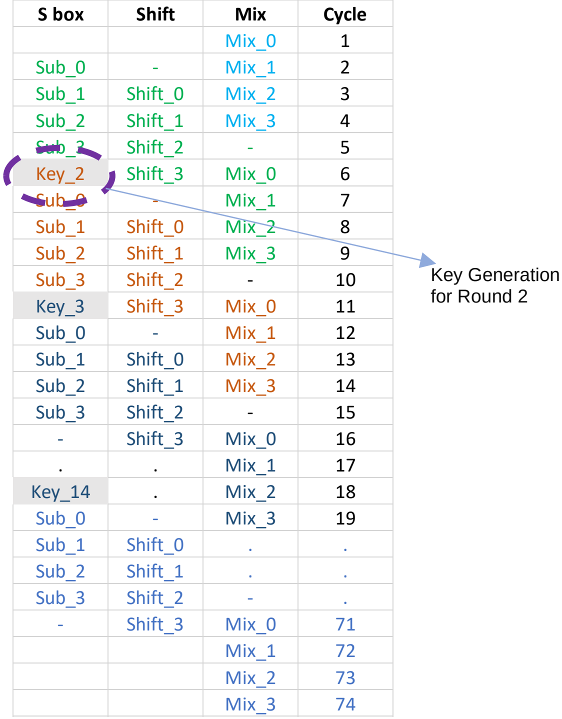
Since all 128-bit (4 words) round 0 sub operation completed, we don't have sub operation in cycle 6. In clock **cycle 6**, we are doing round 1 shift operation of word 3 (shift_3) and mix operation of word 0 (mix_0) and. Since we already have the last byte value from sub_3, we are using that for mix_0. In this way we don't need to wait extra one cycle of shift operation to start the mix operation. We can denote this as cycle6[round1(shift_3, mix_0)].

In clock **cycle 7**, we are doing sub operation for word 0 (sub_0) and mix operation for word 1 (mix_1). We can denote this as cycle7[round2(sub_0), round1(mix_1)].

In clock **cycle 8**, we are doing sub operation for word 1 (sub_1) shift operation for word 0 (shift_0) and mix operation for word 2 (mix_2). We can denote this as cycle8[round2(sub_1, shift_0), round1(mix_2)].

In clock **cycle 9**, we are doing sub operation for word 2 (sub_2), shift operation for word 1 (shift_1) and mix operation for word 3 (mix_3). We can denote this as cycle9[round2 (sub_2, shift_1), round1 (mix_3)].

In clock **cycle 10**, we are doing sub operation for word 3 (sub_3) shift operation for word 2 (shift_2). The same order of execution repeats for all 14 rounds. We can denote this as cycle10[round2(sub_3, shift_2)]. The same sequence repeats for all 14 rounds.



S box	Shift	Mix	Cycle
		Mix_0	1
Sub_0	-	Mix_1	2
Sub_1	Shift_0	Mix_2	3
Sub_2	Shift_1	Mix_3	4
Sub_3	Shift_2	-	5
Key_2	Shift_3	Mix_0	6
Sub_0	-	Mix_1	7
Sub_1	Shift_0	Mix_2	8
Sub_2	Shift_1	Mix_3	9
Sub_3	Shift_2	-	10
Key_3	Shift_3	Mix_0	11
Sub_0	-	Mix_1	12
Sub_1	Shift_0	Mix_2	13
Sub_2	Shift_1	Mix_3	14
Sub_3	Shift_2	-	15
-	Shift_3	Mix_0	16
.	.	Mix_1	17
Key_14	.	Mix_2	18
Sub_0	-	Mix_3	19
Sub_1	Shift_0	.	.
Sub_2	Shift_1	.	.
Sub_3	Shift_2	-	.
-	Shift_3	Mix_0	71
		Mix_1	72
		Mix_2	73
		Mix_3	74

Key Generation for Round 2

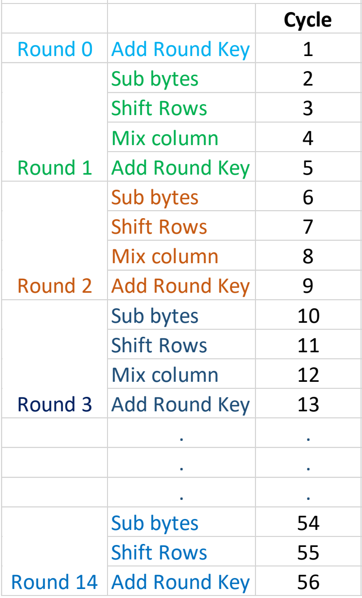
Fig. 10 Pipelined structure of Key gen block

As shown in Fig. 10, in cycle 6 we are using sub bytes S box for key generation block, because of this we don't need extra S box for key generation block. We are generating 128-bit key for every 5 cycles, so that it requires only 4 S box in one cycle. In conventional method, we need 8 S box for key generation block. 128-bit key generated in cycle 6 will be used in cycle 14 mix operation. Similarly, key generated in cycle 11 used in mix operation of cycle 19.

As shown in Fig. 10, in cycle 9 we have valid output for round 1 (cycle 5 to 9), so we need 5 cycles to perform round 1. Total we need 74 clock cycles to complete AES encryption.

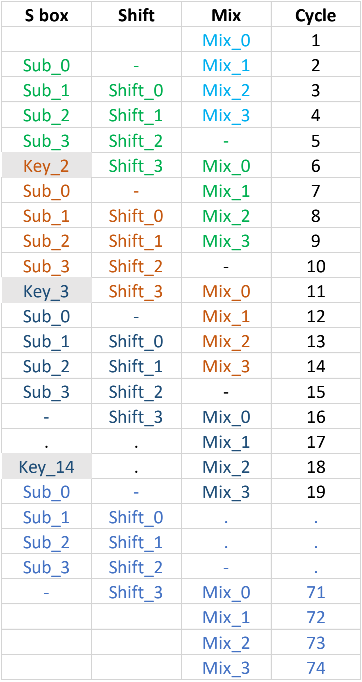
Round 0	04 cycles
Round 1 to 14	70 cycles
Total	**74 cycles**

Table: 2 Cycles required in each round



		Cycle
Round 0	Add Round Key	1
	Sub bytes	2
	Shift Rows	3
	Mix column	4
Round 1	Add Round Key	5
	Sub bytes	6
	Shift Rows	7
	Mix column	8
Round 2	Add Round Key	9
	Sub bytes	10
	Shift Rows	11
	Mix column	12
Round 3	Add Round Key	13
	.	.
	.	.
	.	.
	Sub bytes	54
	Shift Rows	55
Round 14	Add Round Key	56

Fig. 11 Conventional Method



S box	Shift	Mix	Cycle
		Mix_0	1
Sub_0	-	Mix_1	2
Sub_1	Shift_0	Mix_2	3
Sub_2	Shift_1	Mix_3	4
Sub_3	Shift_2	-	5
Key_2	Shift_3	Mix_0	6
Sub_0	-	Mix_1	7
Sub_1	Shift_0	Mix_2	8
Sub_2	Shift_1	Mix_3	9
Sub_3	Shift_2	-	10
Key_3	Shift_3	Mix_0	11
Sub_0	-	Mix_1	12
Sub_1	Shift_0	Mix_2	13
Sub_2	Shift_1	Mix_3	14
Sub_3	Shift_2	-	15
-	Shift_3	Mix_0	16
.	.	Mix_1	17
Key_14	.	Mix_2	18
Sub_0	-	Mix_3	19
Sub_1	Shift_0	.	.
Sub_2	Shift_1	.	.
Sub_3	Shift_2	-	.
-	Shift_3	Mix_0	71
		Mix_1	72
		Mix_2	73
		Mix_3	74

Fig. 12 Pipelined structure of proposed method

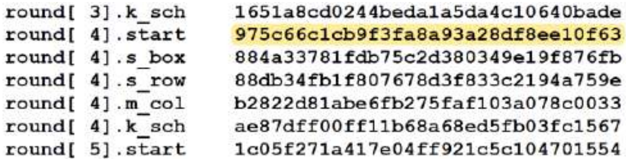
Fig. 11 and Fig. 12 shows the pipelined structure comparison between the implementation of conventional and proposed method.

D. Results and Comparison

Fig. 13 Simulated waveform of standard 256-bit key example

Fig. 13 shows the VCS simulation of 256-bit key AES encryption and decryption. The cipher text is matching with standard 256 AES algorithm results. Also verified all the internal sub bytes, shift rows, mix columns and add round key output with standard test case for 256 key AES implementation.

Fig. 14 Output of round 4 internal operations



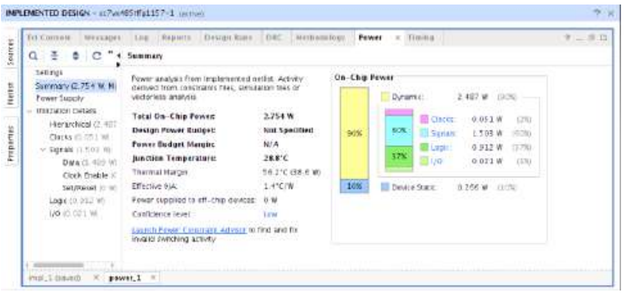
```
round[ 3].k_sch     1651a8cd0244beda1a5da4c10640bade
round[ 4].start     975c66c1cb9f3fa8a93a28df8ee10f63
round[ 4].s_box     884a33781fdb75c2d380349e19f876fb
round[ 4].s_row     88db34fb1f807678d3f833c2194a759e
round[ 4].m_col     b2822d81abe6fb275faf103a078c0033
round[ 4].k_sch     ae87dff00ff11b68a68ed5fb03fc1567
round[ 5].start     1c05f271a417e04ff921c5c104701554
```

Fig. 15 Reference example from AES standard

Fig. 15 shows the round 4 internal operation outputs from AES standard example document “Federal Information Processing Standards Publication 197 November 26, 2001 Announcing the ADVANCED ENCRYPTION STANDARD (AES)”

Fig. 16 Simulated waveform of random inputs

Fig. 16 shows the VCS simulation of 256-bit key AES encryption and decryption with random message and 256-bit key. We are seeing plain text is matching with message.



IMPLEMENTED DESIGN - xc7vx485tffg1157-1 (active)

Power analysis from implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

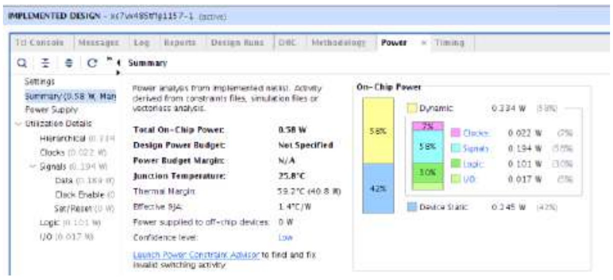
Total On-Chip Power:	2.754 W
Design Power Budget:	Not Specified
Power Budget Margin:	N/A
Junction Temperature:	28.8°C
Thermal Margin:	56.2°C (38.6 W)
Effective θJA:	1.4°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Low

Launch Power Constraint Advisor to find and fix invalid switching activity

On-Chip Power

Dynamic:	2.487 W	(90%)
Clocks:	0.051 W	(2%)
Signals:	1.503 W	(60%)
Logic:	0.912 W	(37%)
I/O:	0.021 W	(1%)
Device Static:	0.266 W	(10%)

Fig. 17 Conventional implementation on-chip power



IMPLEMENTED DESIGN - xc7vx485tffg1157-1 (active)

Power analysis from implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

Total On-Chip Power:	0.58 W
Design Power Budget:	Not Specified
Power Budget Margin:	N/A
Junction Temperature:	25.8°C
Thermal Margin:	59.2°C (40.8 W)
Effective θJA:	1.4°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Low

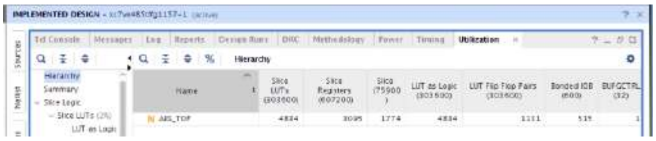
Launch Power Constraint Advisor to find and fix invalid switching activity

On-Chip Power

Dynamic:	0.334 W	(58%)
Clocks:	0.022 W	(7%)
Signals:	0.194 W	(58%)
Logic:	0.101 W	(30%)
I/O:	0.017 W	(5%)
Device Static:	0.245 W	(42%)

Fig. 18 Proposed implementation on-chip power

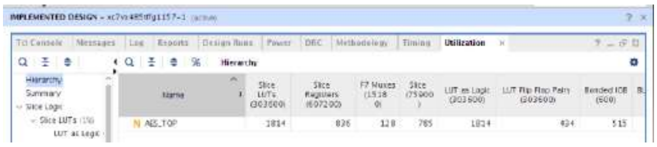
Fig. 17 & Fig. 18 shows the on-chip power in Vivado implementation for conventional and proposed methods.



IMPLEMENTED DESIGN - xc7vx485tffg1157-1 (active)

Name	Slice LUTs (303600)	Slice Registers (607200)	Slice (75900)	LUT as Logic (303600)	LUT Flip Flop Pairs (303600)	Bonded IOB (600)	BUFGCTRL (32)
AES_TOP	4834	3095	1774	4834	1111	515	1

Fig. 19 Conventional implementation Area utilization



IMPLEMENTED DESIGN - xc7vx485tffg1157-1 (active)

Name	Slice LUTs (303600)	Slice Registers (607200)	F7 Muxes (151800)	Slice (75900)	LUT as Logic (303600)	LUT Flip Flop Pairs (303600)	Bonded IOB (600)
AES_TOP	1814	836	128	785	1814	434	515

Fig. 20 Proposed implementation Area utilization

Fig. 19 & Fig. 20 shows the area utilization in Vivado implementation for conventional and proposed methods.

Block	Instance	Conventional	Proposed
Sub bytes	S box	16	4
Mix Column	Mix	4	1
Key Gen	S box	8	0

Table:3 Subblock utilization Comparison

Table:3 shows the comparison for modules used between final implementation of conventional and proposed 256-bit key AES algorithm.

Table: 3 shows that area reduced 4 times in sub bytes and mix column operations in conventional vs proposed methods. S box usage came down to 0 for Key Gen block, since we are reusing same S box of sub bytes.

	FPGA		
Method	Conventioi	Proposed	% Savings
Total cycles for encryption	56	74	
Frequency (MHz)	109	161	
Throughput (Mbps)	249	278	10.53
On chip power (W)	2.75	0.58	78.91
Slice LUT	4834	1814	62.47
Slice Register	3095	836	72.99
LUT Flip Flop Pairs	1131	434	61.63

Table: 4 PPA comparison Conventional vs Proposed

Table: 4 shows the PPA number comparison for conventional and proposed methods in FPGA implementation [6] [8].

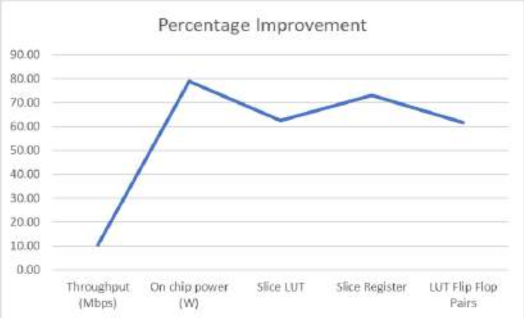


Fig. 23 PPA comparison Conventional vs Proposed

	FPGA		
Method	Proposed	[12]	[13]
Slice LUT	1814	3959	15376
Slice Register	836	1124	5356
LUT Flip Flop Pairs	434	973	2309

Table: 5 Area Comparison for Proposed vs Existing Methods

Table: 5 shows area utilization comparison for proposed vs existing methods [12] [13].

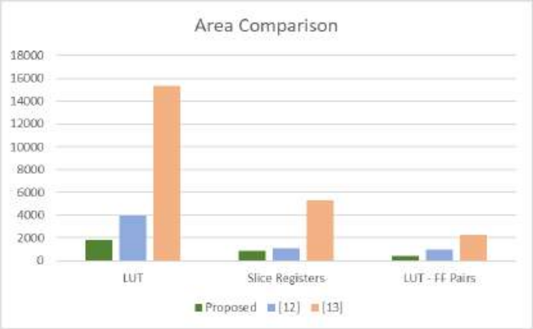


Fig. 24 Area comparison Proposed vs Existing Methods

Fig. 24 shows area utilization comparison in chart for proposed vs existing methods [12] [13].

IV. Conclusion

In this paper, we have compared the PPA numbers of conventional and proposed method in Virtex-7 (xc7vx485tffg1157) FPGA implementation of a 256-bit Key AES algorithm. The proposed implementation has an area reduction by 72% using slice registers, 62% using slice LUT's and 61% using LUT-FF Pairs. This results in a power reduction by 78%. The throughput (Mbps) of the proposed implementation improved by 10%. We proposed reusing the 32-bit Sub bytes and 32-bit Mix column blocks for 128-bit data, reusing the S-box for Sub Bytes and Key Schedule operations, reusing the same hardware for both encryption and decryption. The proposed method is generic and can be used for 128, 196 and 256-bit Key size. The proposed method is generic and can be used with word size operation of 16, 32, 64 bits.

V. References

[1] M. Rajeswara Rao, Dr.R.K.Sharma, SVE Department, NIT Kurushetra "FPGA Implementation of combined S box and Inv S box of AES" 2017 4th International conference on signal processing and integrated networks (SPIN).

[2] Nalini C. Iyer ; Deepa ; P.V. Anandmohan ; D.V. Poornaiah "Mix/InvMixColumn decomposition and resource sharing in AES".

[3] Xinmiao Zhang, Student Member, IEEE, and Keshab K. Parhi, Fellow, "High Speed VLSI architectures for the AES Algorithm", IEEE. VOL.12. No.9. September 2004

[4] Shrivathsa Bhargav, larry Chen, abhinandan Majumdar, Shiva Ramudit "128 bit AES Decryption", CSEE 4840 – Embedded system Design spring 2008, Columbia University.

[5] Atul M. Borkar ; R. V. Kshirsagar ; M. V. Vyawahare "FPGA implementation of AES algorithm".

[6] Announcing the ADVANCED ENCRYPTION STANDARD (AES), November 26 2001.

[7] Yulin Zhang ; Xinggang Wang; "Pipelined implementation of AES encryption based on FPGA" 2010 IEEE International Conference on Information Theory and Information Security.

[8] Yuwen Zhu ; Hongqi Zhang ; Yibao Bao ; "Study of the AES Realization Method on the Reconfigurable Hardware" 2013 International Conference on Computer Sciences and Applications.

[9] Tsung-Fu Lin ; Chih-Pin Su ; Chih-Tsun Huang ; Cheng-Wen Wu; "A high-throughput low-cost AES cipher chip" Proceedings. IEEE Asia-Pacific Conference on ASIC.

[10] C. Sivakumar ; A. Velmurugan ; "High Speed VLSI Design CCMP AES Cipher for WLAN (IEEE 802.11i)" 2007 International Conference on Signal Processing, Communications and Networking.

[11] Vatchara Saicheur ; Krerk Piromsopa ; "An implementation of AES-128 and AES-512 on Apple mobile processor" 2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)

[12] S.P Guruprasad ; B.S Chandrasekar ; "An evaluation framework for security algorithms performance realization on FPGA" 2018 IEEE

International Conference on Current Trends in Advanced Computing (ICCTAC)

[13] N. S. Sai Srinivas ; Md. Akramuddin; “FPGA based hardware implementation of AES Rijndael algorithm for Encryption and Decryption” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).

[14] P. S. Abhijith ; Mallika Srivastava ; Aparna Mishra ; Manish Goswami ; B. R. Singh ; “High performance hardware implementation of AES using minimal resources” 2013 International Conference on Intelligent Systems and Signal Processing (ISSP).

[15] Wei Wang ; Jie Chen ; Fei Xu ; “An implementation of AES algorithm Based on FPGA” 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery

[16] Ashwini M. Deshpande ; Mangesh S. Deshpande ; Devendra N. Kayatanavar; “FPGA implementation of AES encryption and decryption” 2009 International Conference on Control, Automation, Communication and Energy Conservation

Comparison and Analysis of Different Feeding Techniques for MIMO Antenna's UWB Applications with Defected Ground Structure.

MOHAN K N
Dept. of ECE
KLEF Deemed to be University,
Guntur, AP, India
narasimmamohan@gmail.com

K. HIMAJA REDDY
Dept. of ECE
KLEF Deemed to be University,
Guntur, AP, India

N.YASASWINI
Dept. of ECE
KLEF Deemed to be University,
Guntur, AP, India

***Abstract* -- This paper demonstrates the working of UWB MIMO antenna with individual and common feedline ports for excitation. Etched on a highly reliable dielectric, the functioning of designed antenna at different working states of radiating patch elements is evaluated and presented.**

Furthermore, the detailed study and analysis of the generated responses of antenna with common feeding system is elucidated. The variations of the antenna characteristics, geometrical modifications and design parameters are discussed.

Keywords -- Ultra-wideband (UWB), Wilkinson Power Divider (WPD), Multi-Input-Multi-Output (MIMO).

I. INTRODUCTION

With increasing technological advances, constant development in radio wave technologies is rightly acknowledged. In Wireless Communications, the designing of antenna which works under adverse conditions is necessary. Microstrip patch antenna is considered as simple and basic working antenna which paves path to design and fabricate antennas that have high frequency communication channels [6] .

Ultrawide band can be achieved through microstrip antenna with certain geometrical modifications. MIMO antenna uses more than one radiating element to offer high and multiple data transmissions [10]. Power dividers, Directional Couplers and Microwave filters can be used to enhance or isolate desired bandwidth [5] [7].

UWB with MIMO antenna is complex yet efficient combination to tune and obtain user demanded frequency applications in Academical, Industrial, Research and Realtime Implementations. This paper demonstrates the working of two proposed antennas along with brief discussion on their results, design parameters and challenges.

II. METHODOLOGY

A. DESIGN OF UWB MIMO ANTENNA

A MIMO antenna for UWB frequency range is proposed. Dimensions for circular patch antenna at 2.4GHz is calculated. Two circular radiating elements with better isolation is placed on Rogers RT/duroid 5870 ™ dielectric with relative permeability of 2.33 and loss tangent of 0.0012. This is further excited to resonate in ultra-wide band by using partial ground plane of height 20mm. The radius of the patch is calculated by the dimensions of rectangular patch antenna which gives 21.86mm at 2.4GHz.

With the substrate dimensions of 70mm x 140mm x 3.6mm and insertion of slots on the circular patches, the proposed design exhibits power transmission less than -10dB in UWB frequency range. The study and analysis of antenna parameters like radiation pattern, scattering, co and cross polarization, current distribution is presented. At maximum current and minimum voltage point on the surface, the antenna is modified such that the antenna parameters are changed.

Proposed UWB MIMO antenna consists of two radiating patch elements. Hence, two different feeding systems are used for antenna excitation. The working of antenna is now based on excitation magnitude of antenna ports 1 and 2.

Three port states can be derived from above condition. Excitation of antenna can be done with both the ports processing or any one of the ports being terminated. Evaluation of these possibilities are simulated in High Frequency Structure Simulator (HFSS).

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

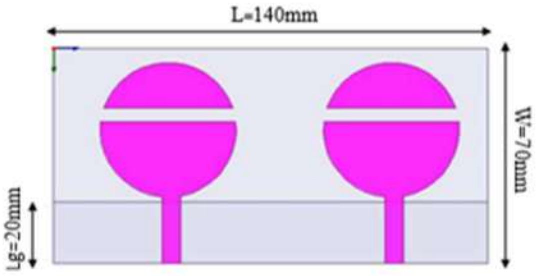


Fig. 1. Design 1 – UWB MIMO Antenna

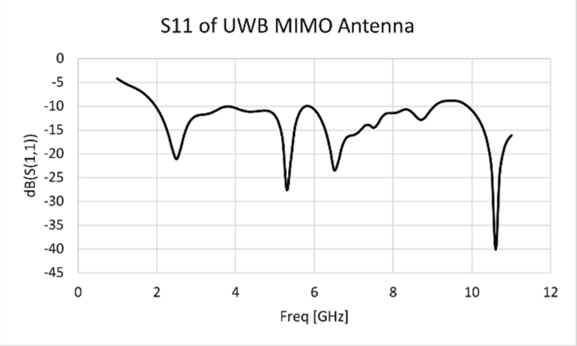


Fig.2. S-parameter (S11) of UWB MIMO Antenna

All the application frequencies like Bluetooth, WLAN, WiMAX, C-Band, X-Band achieved Return loss less than -10dB.

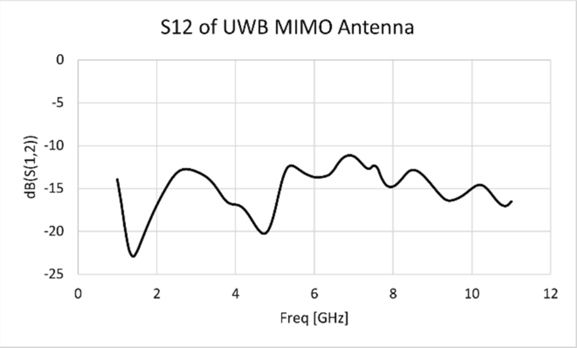


Fig. 3. S-parameter (S12) of UWB MIMO Antenna

S12 represents the power transmitted from port 1 to port 2. The Return loss obtained is less than -10dB.

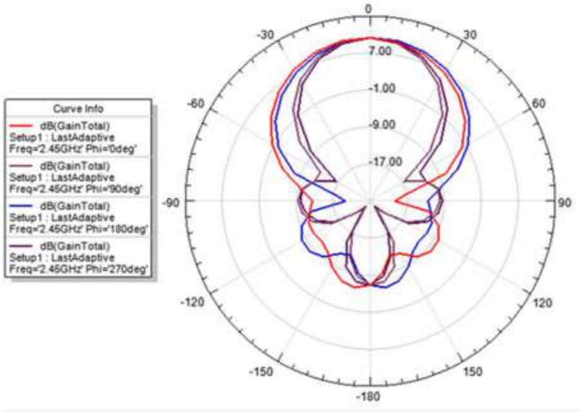


Fig.4. Radiation Pattern of UWB MIMO Antenna

Gain achieved for the proposed antenna is nearly 7dB.

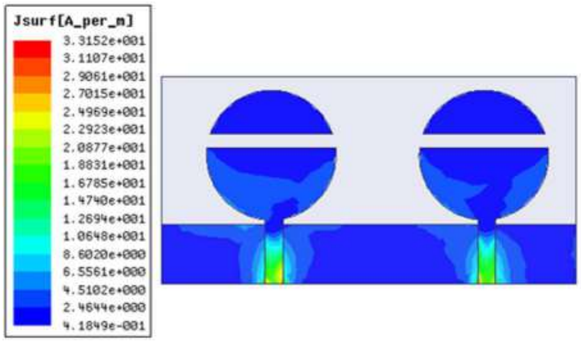


Fig. 5. Current Distribution of UWB MIMO Antenna

The surface current distribution of an antenna determines the electric current transmission all over the surface. This helps in identifying and modifying the antenna geometry to obtain desired outputs. Each color signifies the magnitude of current distribution at the specific area.

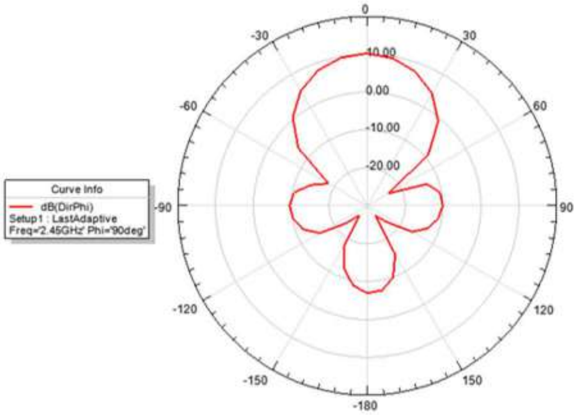


Fig.6. Co-Polarization of UWB MIMO Antenna

The Co-polarization of UWB MIMO antenna is presented above. The orthogonal radiation of obtained co-polarization is the cross-polarization of the proposed antenna. It is helpful to minimize the interference of radiating waves.

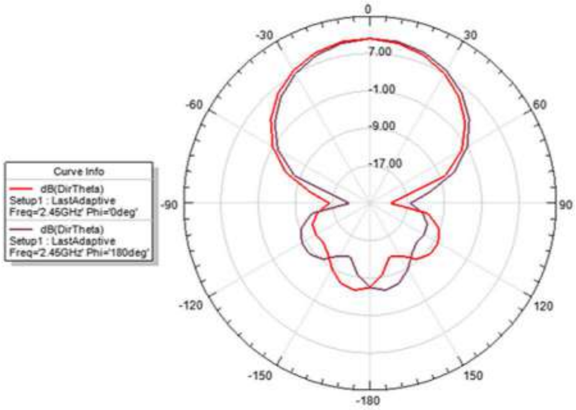


Fig.7. Cross-Polarization of UWB MIMO Antenna.

B. EXCITATION MAGNITUDES OF UWB MIMO ANTENNA

State 1: Excitation Magnitude (1,1).

Here, both the sources have equal magnitude. Both the ports are in ON state. Figure 2 is the Return loss obtained in this state. Figure 4 represents the radiation pattern with gain obtained.

State 2: Excitation Magnitude (0,1). Termination of Port 1.

In this state, the magnitude of port 1 is 0 (OFF state) and the magnitude of port 2 is 1 (ON state). The S-Parameters S11 and S12 when single source is excited is shown below.

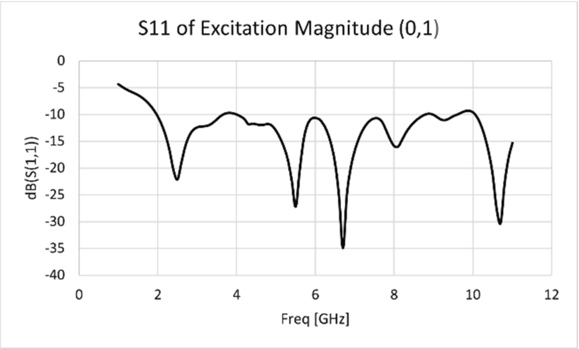


Fig.8. S11 of Excitation Magnitude (0,1)

Return loss when Port 1 is terminated.

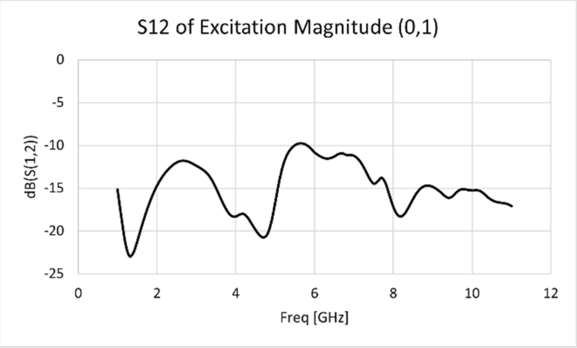


Fig.9. S12 of Excitation Magnitude (0,1)

Power transmission from port 1 to port 2 when port 2 is processed.

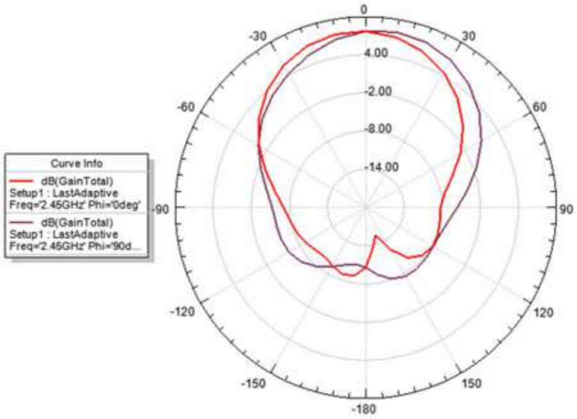


Fig.10. Radiation pattern for Excitation Magnitude (0,1)

State 3: Excitation Magnitude (1,0). Termination of port 2.

In this state, the magnitude of port 1 is 1 (ON state) and the magnitude of port 2 is 0 (OFF state). The S- Parameters S11 and S12 are shown below.

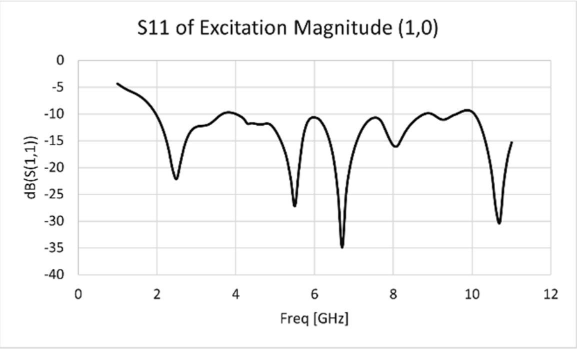


Fig.11. S11 of Excitation Magnitude (1,0)

Return loss when port 2 is terminated and port 1 is processed.

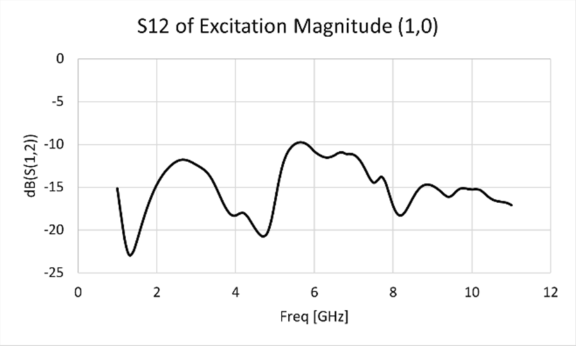


Fig.12. S12 of Excitation Magnitude (1,0)

Power transmission from port 1 to port 2 when port 2 is in OFF state.

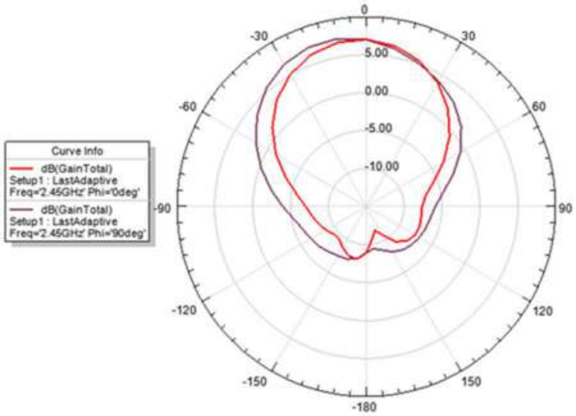


Fig.13. Radiation Pattern for Excitation Magnitude (1,0)

From the above S-Parameters for different excitation states, we can observe the change in Return loss when either of the port is terminated. The significant change in gain is illustrated by below table.

Table I. Maximum Gain at π= 0˚ and π=90˚ for excitation mechanisms

State	1	2	3
Excitation Magnitude **1-Active State** **0-Inactive State**	(1,1)	(0,1)	(1,0)
Port status	Port1 **ON** Port2 **ON**	Port1 **OFF** Port2 **ON**	Port1 **ON** Port2 **OFF**
Gain (dB)	**7dB** (Fig.4)	**4dB** (Fig.10)	**5dB** (Fig.13)
Return loss at 2.4GHz	**19.1433 dB**	**-20.3635 dB**	**-20.3635 dB**

C. DESIGN OF WPD BASED UWB MIMO ANTENNA

Design 1 is limited to classical representation of attaining UWB with MIMO antenna. Design 2 is proposed to avoid the challenges like individual feeding system and high Return loss. The proposed antenna with MIMO acts as type of array for common excitation. Two resistors are placed near the input ports to provide better isolation. The WPD is designed with proper positional and dimensional optimizations. This approach for required application is evaluated and commended with proper study of antenna result parameters.

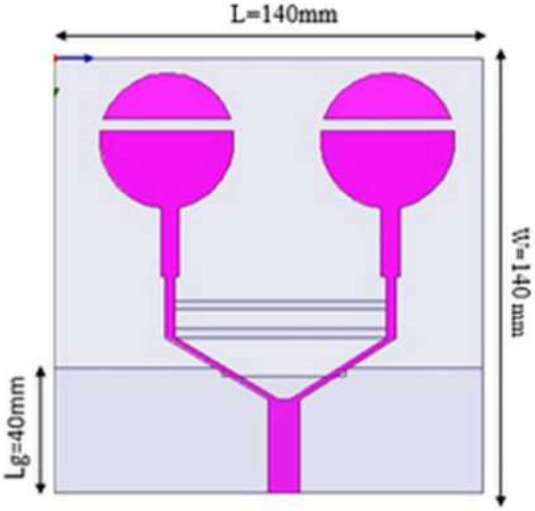


Fig.14. Design of Wilkinson Power Divider Based UWB MIMO Antenna

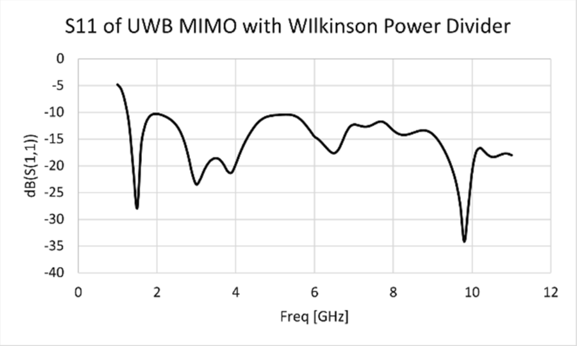


Fig.15. S11 of Wilkinson Power Divider Based UWB MIMO Antenna

Bandwidth of Ultrawide band is less than -10dB for proposed antenna.

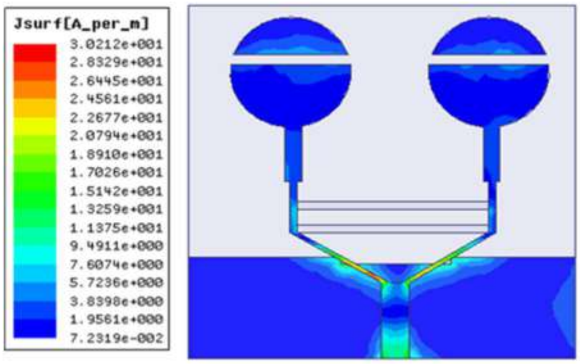


Fig.16. Current Distribution of Wilkinson Power Divider Based UWB MIMO Antenna

With slots etched on both patch and ground plane, the current distribution is presented. Here, the resistors placed are active elements which provide isolation for both input ports.

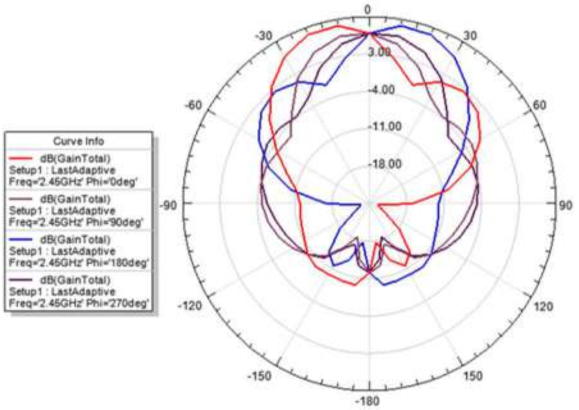


Fig.17. Radiation Pattern of Wilkinson Power Divider Based UWB MIMO Antenna

Evaluated at π= 0˚,90˚,180˚,270˚ and 360˚, the gain obtained for the proposed antenna is nearly 3dB.

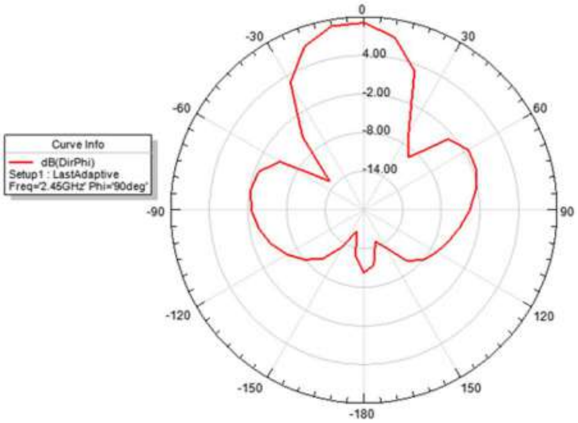


Fig.18. Co-Polarization of Wilkinson Power Divider Based UWB MIMO Antenna

Co-polarization and Cross-polarization of an antenna indicates the radiation of E-fields in its actual and orthogonal direction. Both the polarizations are indicated.

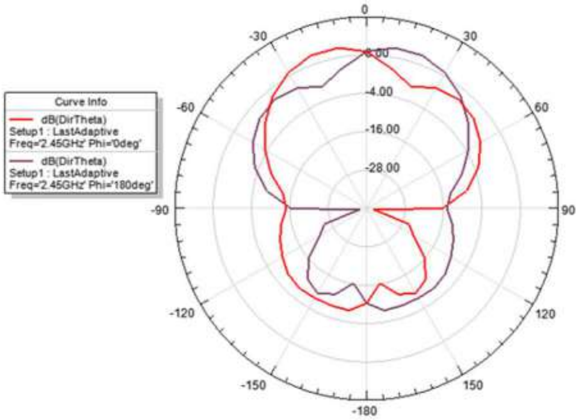


Fig.19. Cross-Polarization of Wilkinson Power Based UWB MIMO Antenna

III. RESULTS

Proposed antennas are designed and simulated in HFSS software. With solution frequency of 2.4GHz, the results are evaluated in terms of Return loss, Bandwidth, Gain and Current distribution. Below is the table which represents the antenna parameters and their responses for both design 1 and design 2.

Table II. Comparison of antenna parameters at 2.4GHz.

Parameters	**Design 1 (Fig.1)**	**Design 2 (Fig.14)**
S11	**-19.64dB** at **2.4GHz** (Bluetooth) **-10.9 dB** at **3.5GHz** (WiMAX) **-10.4dB** at **4GHz** (C-Band) **-21.5dB** at **5.5GHz** (WLAN) **-12.4dB** at **7.7GHz** (X-Band)	**-11.65dB** at **2.4GHz** (Bluetooth) **-18.5 dB** at **3.5GHz** (WiMAX) **-19.6dB** at **4GHz** (C-Band) **-10.6dB** at **5.5GHz** (WLAN) **-11.7dB** at **7.7GHz** (X-Band)
Bandwidth (dB)	**3.1-10.6GHz**. <-10dB **9-9.9GHz** <-8dB	**3.1-10.6GHz** <-10dB

Gain (deg)	**7dB** at **2.4GHz**
	Ɵ=0˚ to 180˚
	Π=90˚
		3dB at **2.4GHz**
		Ɵ=0˚ to 180˚
		Π=90˚

Current Distribution (in W)	1W for each port	2W

The proposed work is compared to previous related works [23] [24] [25], their important parameters are compared and analysed in Table III.

Table III. Comparison of proposed work with existing works.

Related works	**S11**	**Bandwidth (dB)**	**Gain (dB)**
[23]	<-15dB	**3.1-11.2GHz** Band rejection at **3.1-3.8GHz** (Wi-Fi) **5.2-5.8GHz** (WLAN)	Ranges from **2.5-5.5 dB** at given bandwidth
[24]	<-5dB	**2.5-10.6GHz** Band rejection at **5.2-5.8GHz** (WLAN) **7.92-8.395 GHz** (Uplink X-band) **7.252-7.75 GHz** (Downlink X-band)	**14dB** at **3.8GHz** **15dB** at **5.6GHz** **18dB** at **10GHz**
[25]	<-10dB	**3.4-11.7GHz**	**5dB** at **4.1GHz** **5dB** at **5GHz** **5dB** at **9.5GHz**
Proposed work	<-10dB	**2.4-10.6GHz**	**3dB** at **2.4GHz**

IV. CONCLUSION

In this paper, a UWB MIMO antenna is designed and simulated. Antenna excitation for multiple feed sources is analyzed. To avoid individual feeding systems, Wilkinson power divider is integrated with the MIMO system. This increases the Return loss of UWB frequency range throughout its bandwidth. The wireless application frequencies like Bluetooth, WLAN, WiMAX, C-Band, X-Band show decent Return loss in both the cases. Gain and current distributional changes for each antenna is briefed.

The results obtained for both individual fed and Wilkinson fed antenna is discussed. It is concluded that MIMO system with multiple source feeding involves the antenna radiation at different port stages. Using Wilkinson power divider, the gain and overall bandwidth Return loss has been amplified. Moreover, it provides better isolation which helps in achieving desired output.

V. REFERENCES

[1] Islam Md Rafiqul, Waheeb S. A, Sarah Rafiq, M.S Yasmin, M.H. Habaebi, "A 2x2 MIMO Patch Antenna for Multiband Applications," 2017 *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)",* Vol.5, No.4, pp. 383-389.

[2] Karunakar Patchala, Y. Raja Rao, A.M. Prasad, "Triple band notch compact MIMO antenna with defected ground structure and split ring resonator for wideband applications ",*Heliyon*, Vol 6, Issue 1, 2020.

[3] S.Oudayacoumar, M.Amudhan, 2013, "A Compact Hexagonal Structured Dual Band MIMO Antenna for Fixed WiMAX Application", *International Journal of Engineering Research and Technology (IJERT)*, 2013, Vol. 2, Issue 8.

[4] Kamble V.D., Jadhav M.R, "Design of MIMO Antenna for WLAN and WiMAX Application", ICATSA Springer, cham. 2017.

[5] Pozar D M, 1997, Microwave Engineering (Wiley).

[6] Constantine A Balanis 2005 Antenna Theory: Analysis and Design.pdf (Hoboken, NJ: John Willey).

[7] Soufian Lakrit et al., "Design and Analysis of Integrated Wilkinson Power Divider Fed Conformal High gain UWB Array Antenna with Band Rejection Characteristics for WLAN Applications", *Journal of Circuits, Systems and Computers.* 2020.

[8] M. Irshad Khan, M. I. Khattak, S. U. Rahman, A. B. Qazi, A. A. Telba, and A. Sebak, "Design and Investigation of Modern UWB-MIMO Antenna with Optimized Isolation," *Micromachines*, vol. 11, no. 4, p. 432, Apr. 2020.

[9] R. Li, Z. Mo, H. Sun, X. Sun, and G. Du, "A Low-Profile and High-isolated MIMO Antenna for 5G Mobile Terminal," *Micromachines*, vol. 11, no. 4, p. 360, Mar. 2020.

[10] Wael A.E. Ali, Ahmed A. Ibrahim, "A compact double-sided MIMO antenna with an improved isolation for UWB applications", *AEU - International Journal of Electronics and Communications*, Vol 82, 2017, Pages 7-13.

[11] Sanjeev Kumar, Ravi Kumar, Rajesh Kumar Vishwakarma, Kunal Srivastava ,"An improved compact MIMO antenna for wireless applications with band-notched characteristics", *AEU - International Journal of Electronics and Communications*, Vol 90, 2018, Pages 20-29.

[12] Wenhua Chen, Manos M. Tentzeris, Yuan Yao, Yan Zhang, Li Yang, "MIMO Antenna Design and Channel Modeling", International Journal of Antennas and Propagation, vol. 2012, 2 pages.

[13] Santanu Mondal, Kaushik Mandal & Partha Pratim Sarkar (2018) Design of MIMO Antenna for Ultra-Wideband Applications, *IETE Journal of Research*, Vol 64, Issue 4, 497-502.

[14] Rhea Nath, Promod Singh, "MIMO Antenna for UWB Applications", *International Journal of Engineering Trends and Technology – IJETT*, Vol 53, 2017.

[15] Agrawal, Tanvi, & Srivastava, Shweta. (2017). Compact MIMO Antenna for Multiband Mobile Applications. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, 16(2), 542-552.

[16] Banothu Y.V.N.R.Swamy, P.Siddaiah et al., "Design of a Compact 2x2 Multi Band MIMO Antenna for Wireless Applications", International Journal of Recent Technology and Engineering – IJRTE, Vol. 7, Issue 6S2, 2019.

[17] J. Ren, W. Hu, Y. Yin and R. Fan, "Compact Printed MIMO Antenna for UWB Applications," in *IEEE Antennas and Wireless Propagation Letters*, vol. 13, pp. 1517-1520, 2014.

[18] B. Ramesh, M. Ramesh, and V. R. Lakshmi, "Ultra-wideband hexagonal MIMO antenna with Defected Ground Structure (DGS)," *2015 13th International Conference on Electromagnetic Interference and Compatibility (INCEMIC)*, Visakhapatnam, 2015, pp. 109-111.

[19] Srivastava, G., Kanuijia, B., & Paulus, R. (2017),"UWB MIMO antenna with common radiator." *International Journal of Microwave and Wireless Technologies*, 9(3), 573-580.

[20] Charles MacWright Thomas, Huda A. Majid, Zuhairiah Zainal Abidin, Samsul Haimi Dahlan, Mohamad Kamal A. Rahim, Raimi Dewssan " A Study on V-shaped Microstrip Patch MIMO Antenna", *Indonesian Journal of Electrical Engineering and Computer Science* Vol. 5, No. 3, March 2017, pp. 606-611.

[21] S. R. Thummaluru, M. Ameen and R. K. Chaudhary, "Four-Port MIMO Cognitive Radio System for Midband 5G Applications," in *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 8, pp. 5634-5645, Aug. 2019.

[22] W. Jiang, Y. Cui, B. Liu, W. Hu and Y. Xi, "A Dual-Band MIMO Antenna with Enhanced Isolation for 5G Smartphone Applications," in *IEEE Access*, vol. 7, pp. 112554-112563, 2019.

[23] A. Mchbal, N. Amar Touhami, H. Elftouh, A. Dkiouak, "Mutual Coupling Reduction Using a Protruded Ground Branch Structure in a Compact UWB Owl-Shaped MIMO Antenna", International Journal of Antennas and Propagation, vol. 2018, Article ID 4598527, 10 pages, 2018.

[24] Rao T V, Sudhakar A, Raju K P, "Novel Technique of MIMO Antenna Design for UWB Applications Using Defective Ground Structures'', Journal of Science and Industrial Research (JSIR), Vol.77(1), January 2018.

[25] Noor M. Awad, Mohamed K.Abdelazeez, Multislot Microstrip antenna for ultra-wide band applications, Journal of King Saud University - Engineering Sciences, Volume 30,Issue 1,2018,Pages 38-45.

Explainable Predictions of Industrial Emissions

Sudarshan S. Chawathe
School of Computing and Information Science & Climate Change Institute
University of Maine
Orono, Maine 04469-5711, USA
chaw@eip10.org

***Abstract*—Predictive emission monitoring systems for gas turbines are important in the power generation industry. A key task in this context these systems is the prediction of flue gas emissions using process and environmental measurements that are easier to obtain. This paper presents methods for such predictions with an emphasis on explainability. A notable result is that despite the potential restrictions imposed by this emphasis, the numerical accuracy compares very favorably with prior work that uses models that are more difficult to explain.**

***Index Terms*—Predictive Emission Monitoring Systems; Exhaust Emissions Prediction; Gas Turbines; CO; NOx; Machine Learning.**

I. Introduction

Monitoring emissions from industrial systems, such as flue gas emissions from turbines, and in particular emissions of CO (carbon monoxide) and NOx (nitrogen oxides), are of growing environmental and regulatory concern and, consequently, there is a growing need to measure and estimate them accurately. While direct measurements are an obvious and effective method, such measurements often come with high associated costs in equipment, maintenance, and operations. There is therefore significant interest in estimating or predicting these emissions using other, more easily measured variables, such as temperatures and pressures in different parts of a turbine.

This paper studies such predictions of CO and NOx emissions experimentally, using a recently published dataset from prior work. A notable feature of this work is the emphasis on explainable predictions in addition to prediction accuracy.

The main contributions are:

- A detailed exposition of a valuable dataset from prior work, making it more accessible to further research.
- An experimental study of prediction accuracy of explainable methods.
- A detailed examination of some concrete predictors produced by such methods.
- A quantitative investigation of the merit of the attributes in this dataset for prediction.

The main results are:

- Contrary to what may reasonably be expected, limiting prediction models to those that are easily explainable does not result in a penalty in prediction accuracy.
- Simple and well understood explainable methods, such as M5P in particular, provide accuracies that are higher than those reported by more opaque models in prior work on the same dataset.
- Measures of attribute merit based on the *Relief* algorithms are better suited to this data than are measures based on correlation.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Outline: The domain and dataset motivating this work is described in Section II. Explainable prediction methods are outlined in Section III, followed by an experimental study of their prediction accuracy in Section IV. A few representative concrete predictors are examined in Section V. Attribute merit and selection is studied in Section VI. Related work is described by Section VII and Section VIII provides a summary.

II. Dataset

The work in this paper is based on a recently published dataset of turbine flue gas emissions due to prior work [1], [2]. The data span the years 2011 through 2015 and there are 36733 records in total. Each record includes the year in which it was collected along with the values of the attributes summarized by Table I. The first three columns of that table, which list the abbreviations, descriptions, and units of the attributes, are from the original dataset documentation while the remaining columns, which provide a brief description of the distributions of each attribute's values, are computed from the base dataset. The attributes in bold font, **CO** and **NOx**, are the dependent variables and the others are the independent variables.

Fig. 1 summarizes the attribute-value distributions using histograms. While the distributions of some attributes, such as ambient temperature (AT) and ambient humidity (AH) are unsurprising, some of the other attributes, such as compressor discharge pressure (CDP) and gas turbine exhaust pressure (GTEP), have interesting multimodal distributions. Others, such as turbine after temperature (TAT) and turbine inlet temperature (TIT) are characterized by highly skewed distributions and this observation may also be confirmed by noting the high magnitudes of the related statistics in Table I.

TABLE I
ATTRIBUTES IN THE TURBINE FLUE GAS EMISSIONS DATASET [1] ALONG WITH SOME COMPUTED STATISTICS.

Attr.	Description	Unit	Min	Mean	Variance	Skewness	Kurtosis	Max
AFDP	Air filter difference pressure	mbar	2.09	3.93	0.60	0.38	0.22	7.61
AH	Ambient humidity	%	24.08	77.87	209.13	-0.63	-0.27	100.20
AP	Ambient pressure	mbar	985.85	1013.07	41.77	0.19	0.44	1036.60
AT	Ambient temperature	C	-6.23	17.71	55.46	-0.04	-0.83	37.10
CDP	Compressor discharge pressure	mbar	9.85	12.06	1.19	0.24	-0.63	15.16
CO	**Carbon monoxide**	mg/m^3	0.00	2.37	5.12	4.84	49.08	44.10
GTEP	Gas turbine exhaust pressure	mbar	17.70	25.56	17.61	0.33	-0.65	40.72
NOx	**Nitrogen oxides**	mg/m^3	25.90	65.29	136.38	1.03	2.04	119.91
TAT	Turbine after temperature	C	511.04	546.16	46.82	-1.76	2.02	550.61
TEY	Turbine energy yield	MWH	100.02	133.51	243.94	0.12	-0.50	179.50
TIT	Turbine inlet temperature	C	1000.80	1081.43	307.52	-0.89	-0.05	1100.90

III. PREDICTING CO AND NOX

The primary task in the context of the dataset of Section II is predicting the values of the CO and NOx attributes given the values of some or all of the other attributes. Prior work has addressed this task using the connectionist approach and, in particular, using extreme learning machines (ELMs) [1]. In contrast, the focus of this work is on the use of simpler and more explainable machine learning methods. In addition to permitting human oversight and fine tuning, these models also exhibit very competitive accuracy, as summarized in Section IV. In particular, the mean absolute errors in prediction of several of the options listed below are lower than those reported in prior work on the same dataset.

A baseline is established by the 0-R (ZerR) method which predicts the overall mean value of the dependent attribute regardless of the values of the other attributes in an instance. The rule-based DTab method uses Decision Tables [3]. The LinR method uses linear regression with the Akaike information criterion [4]. The related SLin method uses simple linear regression, picking the attribute that gives the lowest squared error. The M5P method uses the *M5'* model tree algorithm [5], [6]. The REPT method builds a regression tree, using information gain, which is pruned using backfitting and reduced-error pruning. The RanT method builds a regression tree by randomly selecting k attributes at each node, using the implementation's default value of $k = 1 + \lfloor \log_2(m) \rfloor$, were m is the number of independent attributes. In this work, m is 9 (10 for one set of experiments) giving $k = 4$. Finally, RanF uses the popular Random Forest ensemble learning method [7].

IV. EXPERIMENTAL EVALUATION

The first set of results is summarized by Figs. 2 and 3, which depict the mean absolute error in the predictions of CO and NOx, respectively, for each calendar year of records in the dataset and for each of the methods noted earlier. In this set of experiments, the dataset was segmented by year and each method was evaluated on each segment using conventional 10-fold cross-validation. While there are a few easily discernible trends across dependent variables, years, and methods, one particularly notable one is the high accuracy of the M5P method. In this and other experiments (unless noted otherwise) the dependent attributes (CO and NOx) are predicted using only the independent attributes. That is, when studying the accuracy in predicting one dependent attribute, the other is excluded from the dataset completely during both training and testing phases.

The above experiments do not distinguish records within each segment by times of observations. Indeed, the dataset includes no timestamps beyond the year of observation, making such considerations difficult to address. To characterize the practical requirement that predictions are most useful when made forward in time, the next set of experiments makes predictions for each calendar year models trained on only the data from the previous calendar year. (The very first year, 2011, is included in training but excluded for predictions, and conversely for the very last year, 2015.) The results are summarized by Figs. 4 and 5. As may be expected, the greater restriction on the training data in comparison with that used by the earlier experiments maps to greater errors in prediction in general. However, as the figures illustrate, some of the methods are more resilient than others in this regard. In particular, the M5P method continues to exhibit the lowest or close to lowest errors in prediction for both CO and NOx.

The next set of experiments studies the effect of the availability of measured values of one of the dependent attributes when predicting the other. That is, unlike in all the other experiments reported here, this set of experiments retains the values of the CO attribute when training and testing a model for predicting NOx, and vice versa. Apart from gaining a better understanding of the data in general, this setup may also have practical ramifications in situations where one of the dependent variables is easier to measure (or to measure more frequently) than the other. The results are summarized by Figs. 6 and 7 using the same conventions as earlier figures. While there are small differences across methods and years, a general observation here is that the availability of the other dependent attribute's value does not significantly improve

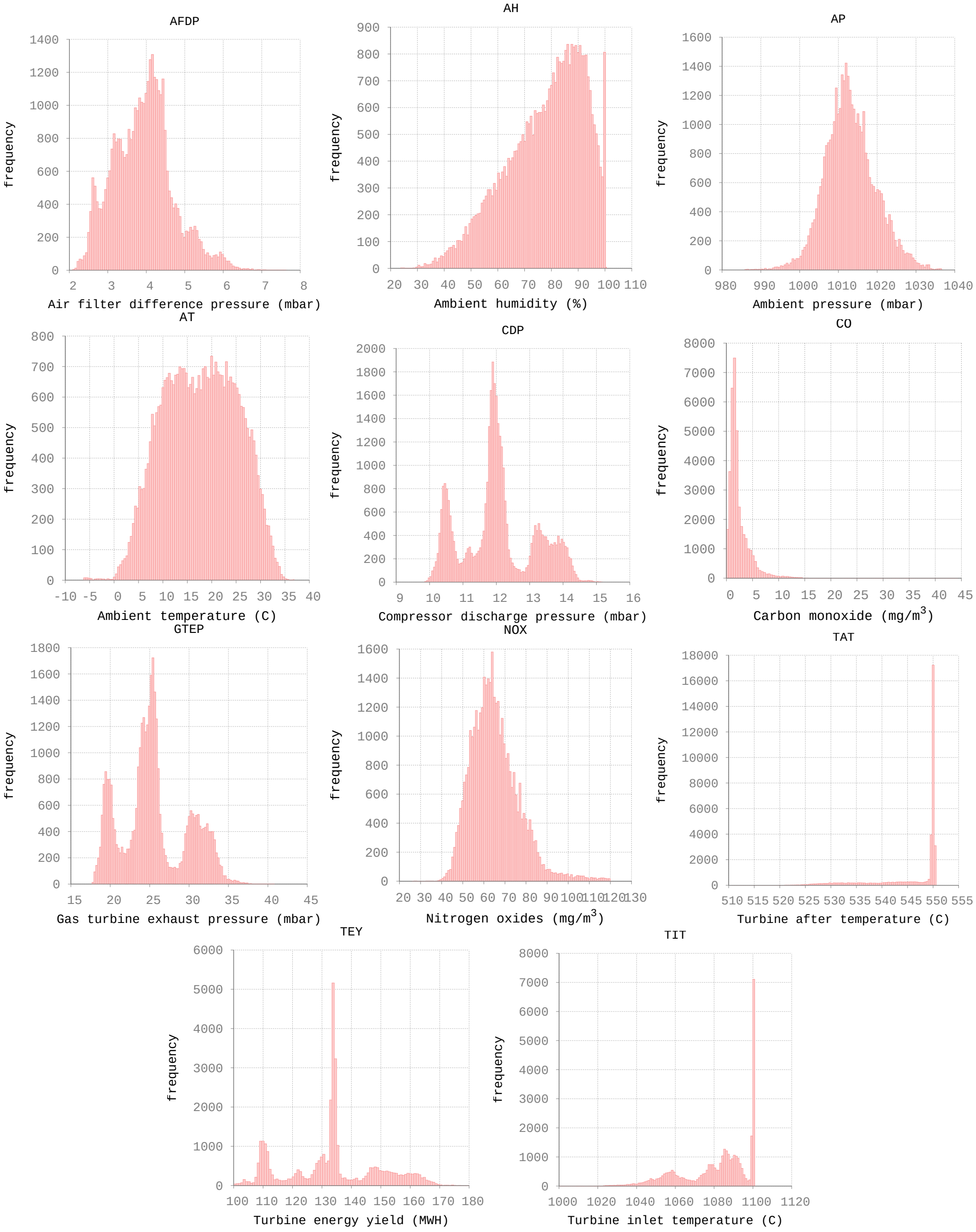


Fig. 1. Histograms for the attributes (Table I) in the experimental dataset [1].

prediction accuracy.

V. Examining Predictors

As noted earlier, this work emphasizes the use of explainable predictions and predictors in addition to the usual quality metrics studied in the previous section. It is interesting to note that one of the more consistently accurate predictors for this dataset is also one that produces prediction models that are scrutable to humans: The M5P ($M5'$) predictor. The implementation underlying the experiments reported in Section IV allows interactive examination of the tree produced by M5P as well as similar artifacts from other methods. A static version of a portion of such a tree appears in Fig. 11. In M5P, the leaves represent linear regression models while the interior nodes serve to guide the prediction to the best linear regression model by using the values of other attributes for branching. A representative sample of the linear regression models at the leaves appears in Fig. 10. Although this particular model is a rather large tree, it is nevertheless intuitively easy for a human to comprehend.

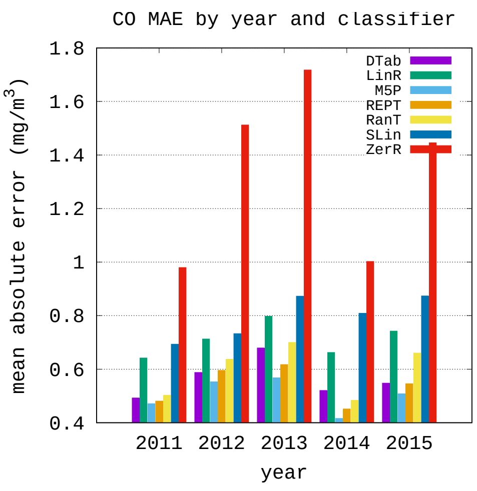


Fig. 2. Mean Absolute Error in CO prediction by classifier and year.

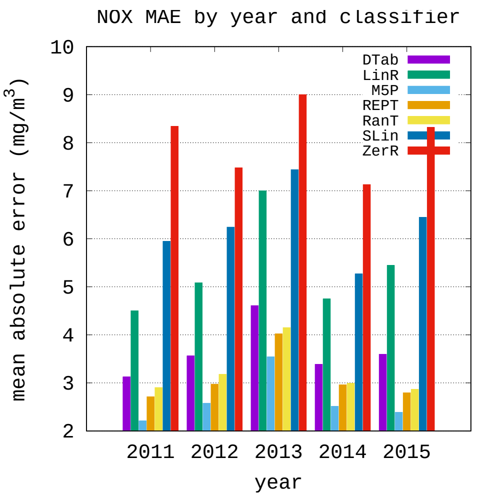


Fig. 3. Mean Absolute Error in NOx prediction by classifier and year.

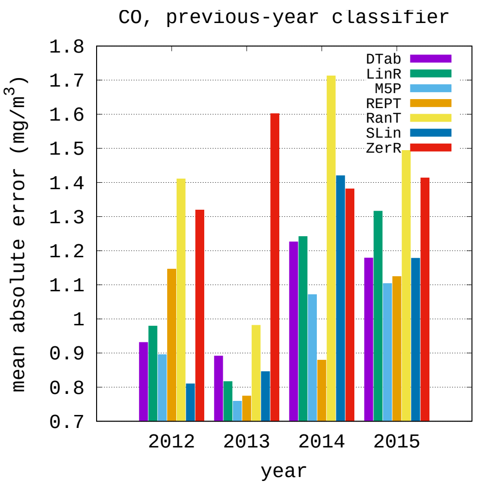


Fig. 4. Mean Absolute Error in CO prediction by previous-year classifier and year. (Year 2011 is used for training models for year 2012 but cannot be used for testing given this restriction.)

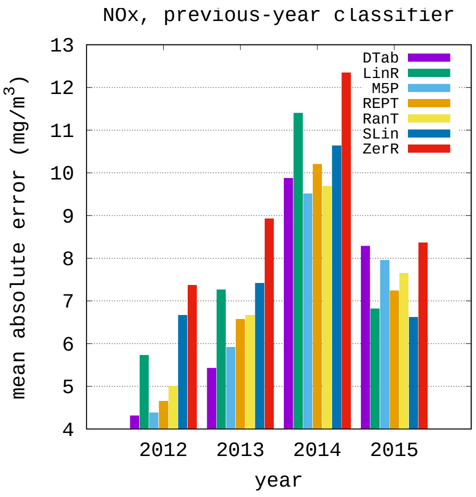


Fig. 5. Mean Absolute Error in NOx prediction by previous-year classifier and year. See also Fig. 4.

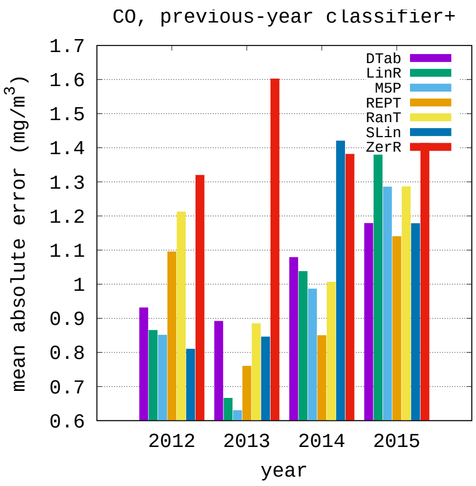


Fig. 6. Mean Absolute Error in CO prediction by previous-year classifier and year, using the other dependent attribute as well.

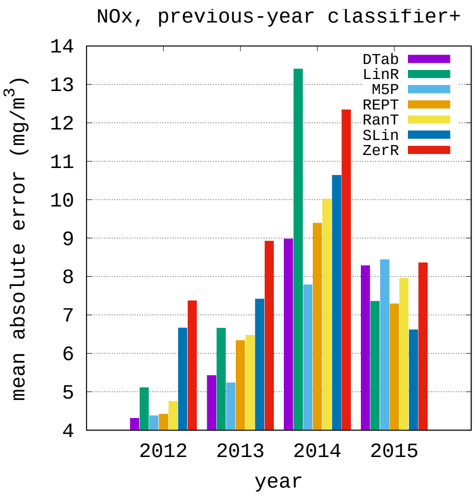


Fig. 7. Mean Absolute Error in NOx prediction by previous-year classifier and year, using the other dependent attribute as well.

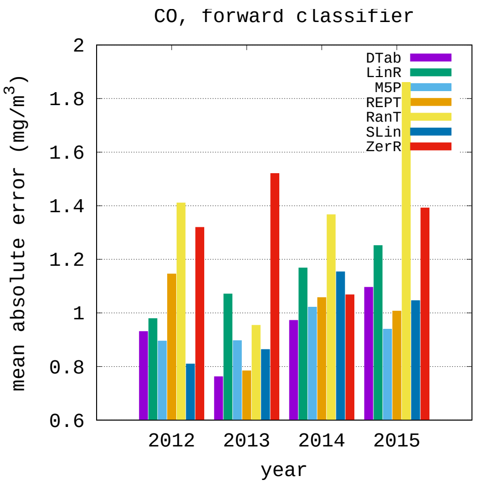


Fig. 8. Mean Absolute Error in CO prediction by forward classifier (using only historical data for each year) and year, using the other dependent attribute as well.

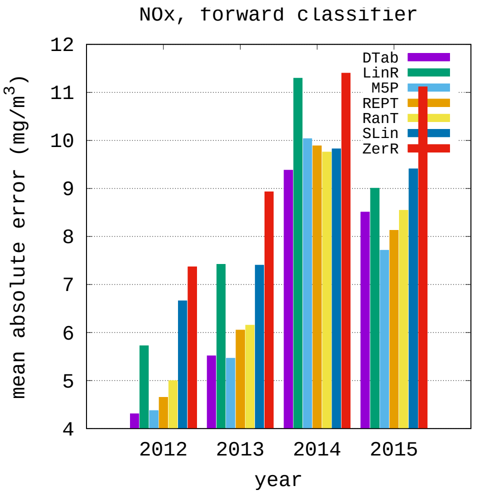


Fig. 9. Mean Absolute Error in NOx prediction by forward classifier (using only historical data for each year) and year, using the other dependent attribute as well.

Another one of the more accurate methods from the previous section, Decision Tables (DTab) also produces models that are suitable for human interpretation, although their tabular form is different from the hierarchical nature of the tree-based models such as M5P or RanT. An excerpt of the decision table produced for predicting NOx values using the 2015 data appears in Fig. 12.

VI. Attribute Merit

Measures of attribute merit (for regression) are frequently used to trim the set of attributes used for prediction. Although the set of attributes in the dataset and underlying domain in the present case is quite small, it is nonetheless useful to examine attribute merit to determine if some attributes may be omitted without significant negative impact on prediction accuracy.

```
LM num: 1                         LM num: 4
NOX =                             NOX =
1.3178 * yr                       -1.7587 * yr
- 0.6062 * AT                     - 0.7118 * AT
- 0.0008 * AP                     + 0.037 * AP
- 0.0643 * AH                     - 0.1764 * AH
+ 3.0378 * AFDP                   + 0.7175 * AFDP
- 1.7196 * GTEP                   + 0.6021 * GTEP
+ 0.0953 * TIT                    + 0.1463 * TIT
- 0.2839 * TAT                    - 0.2638 * TAT
- 0.3692 * TEY                    - 2.1183 * TEY
- 2.6659 * CDP                    - 2.7387 * CDP
- 2383.2419                       + 3866.7144

LM num: 2                         LM num: 5
NOX =                             NOX =
6.2128 * yr                       -1.7587 * yr
- 0.9424 * AT                     - 0.7118 * AT
- 0.0008 * AP                     + 0.037 * AP
- 0.3477 * AH                     - 0.1764 * AH
+ 13.9495 * AFDP                  + 0.7175 * AFDP
- 1.2765 * GTEP                   + 0.6021 * GTEP
- 0.3251 * TIT                    + 0.2328 * TIT
+ 0.6484 * TAT                    - 0.2638 * TAT
- 3.2606 * CDP                    - 2.1466 * TEY
- 12345.0094                      - 2.7387 * CDP
                                          + 3781.1312
LM num: 3
NOX =                             LM num: 6
-0.6452 * yr              NOX =
- 0.6763 * AT                     -1.7587 * yr
- 0.0008 * AP                     - 0.7118 * AT
- 0.1955 * AH                     + 0.1778 * AP
- 0.6063 * AFDP                   - 0.1764 * AH
- 0.2197 * GTEP                   + 0.7175 * AFDP
- 0.6088 * TIT                    + 0.6021 * GTEP
+ 1.3572 * TAT                    + 0.1655 * TIT
- 6.1882 * CDP                    - 0.2638 * TAT
+ 1391.65                         - 1.6279 * TEY
                                          - 2.7387 * CDP
                                          + 3647.8532
```

Fig. 10. Six linear models (among several) used by the M5' regression model excerpted by Fig. 11.

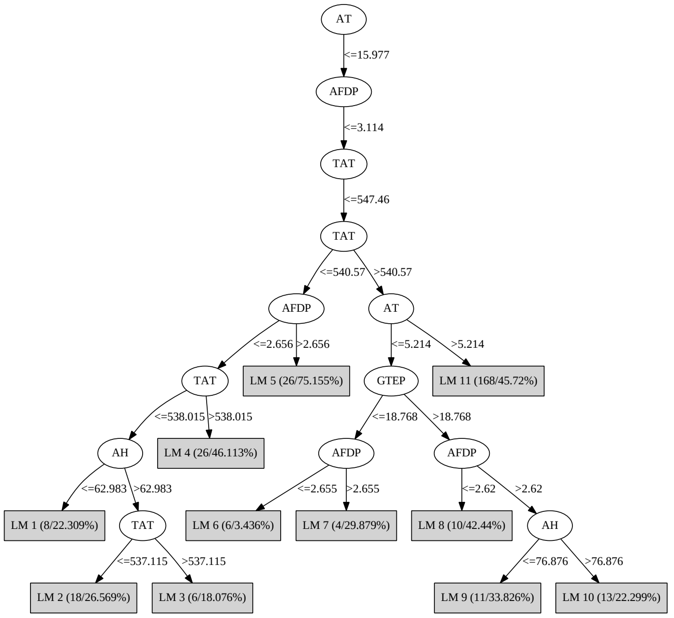


Fig. 11. A portion of the M5' classifier related to the linear models of Fig. 10.

```
AT                  AFDP               GTEP               TIT            TAT                TEY                NOX
(-inf--1.90102]     (3.19204-3.74436] (26.9052-29.207]   (1090.89-inf) (546.653-inf)      (155.656-163.604] 84.334
(15.4341-19.76788]  (4.849-5.40132]    (29.207-31.5088]   (1090.89-inf) (546.653-inf)      (147.708-155.656] 58.179
(6.76654-11.10032]  (4.29668-4.849]    (29.207-31.5088]   (1090.89-inf) (542.696-546.653] (155.656-163.604] 72.824
(11.10032-15.4341]  (3.74436-4.29668] (33.8106-36.1124] (1090.89-inf) (538.739-542.696] (155.656-163.604] 57.752
(24.10166-28.43544] (4.29668-4.849]    (29.207-31.5088]   (1090.89-inf) (546.653-inf)      (147.708-155.656] 55.083
(19.76788-24.10166] (4.29668-4.849]    (29.207-31.5088]   (1090.89-inf) (546.653-inf)      (147.708-155.656] 60.624
(15.4341-19.76788]  (4.849-5.40132]    (31.5088-33.8106] (1090.89-inf) (538.739-542.696] (155.656-163.604] 64.833
(6.76654-11.10032]  (3.74436-4.29668] (29.207-31.5088]   (1090.89-inf) (542.696-546.653] (155.656-163.604] 90.0885
(24.10166-28.43544] (3.74436-4.29668] (33.8106-36.1124] (1090.89-inf) (542.696-546.653] (147.708-155.656] 50.505
(19.76788-24.10166] (3.74436-4.29668] (33.8106-36.1124] (1090.89-inf) (542.696-546.653] (147.708-155.656] 62.6495
...
```

Fig. 12. An excerpt of regression using a decision table.

Omitting an attribute may remove the need for hardware and other costs associated with its measurement.

Figs. 13 and 14 summarize the results of experiments quantifying attribute merit using the correlation metric. The results are a bit counter-intuitive as they allot low scores to several attributes that are intimately associated with the turbine operations (e.g., TIT, TAT, GTEP) and that may be reasonably assumed to have a significant impact on emissions.

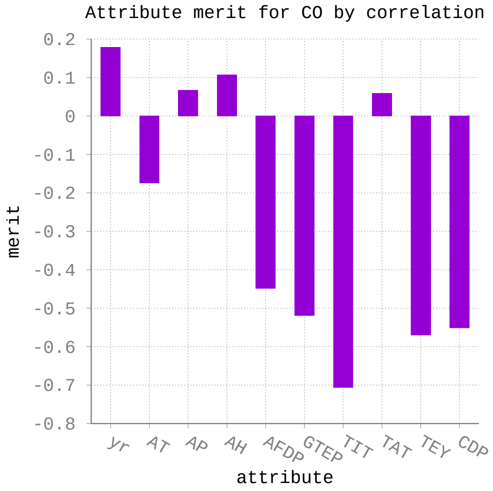


Fig. 13. Attribute merit using the correlation metric for CO prediction using combined data for all years 2011–2015.

Figs. 15 and 16 summarize similar results using the *RReliefF* metric of an attribute, which metric evaluates, for each instance (sampled), the value of that attribute for the nearest instances of the same and different classes [8]–[10]. A notable feature of this and related *Relief*-based attribute merit metrics is that they use contextual information to better estimate the merit of attributes in the presence of strong dependencies. With this metric, in contrast to the correlation metric above, some of the turbine sensor attributes have high scores, as does the air temperature (AT) attribute. These observations invite further study using a pruned set of attributes for predictions, and such work is ongoing.

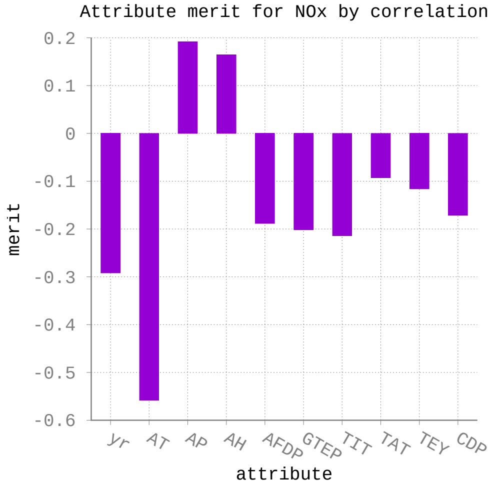


Fig. 14. Attribute merit using the correlation metric for NOx prediction using combined data for all years 2011–2015.

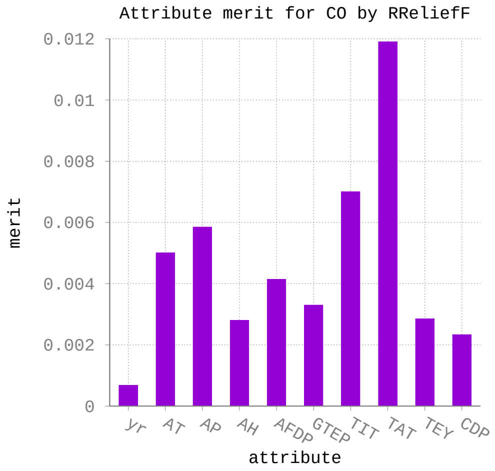


Fig. 15. Attribute merit using the *ReliefF* metric for CO prediction using combined data for all years 2011–2015.

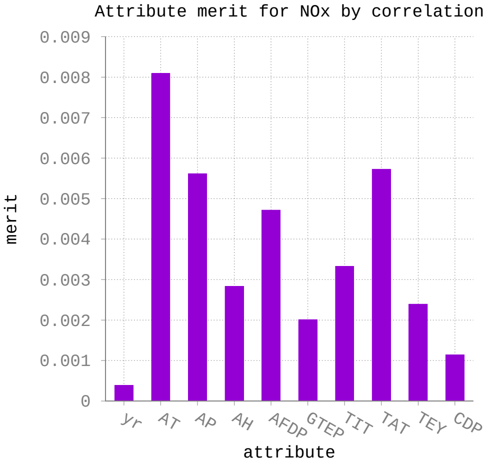


Fig. 16. Attribute merit using the *ReliefF* metric for NOx prediction using combined data for all years 2011–2015.

VII. Related Work

This work was motivated and enabled by prior work using the same dataset that used extreme learning machines (ELMs) for prediction and that provided the dataset as well [1]. This work focuses on explainable results and initially anticipated that emphasis to translate to a modest penalty in prediction accuracy. However, contrary to this expectation, the accuracy of the explainable models here is comparable and typically higher than the accuracies reported earlier, which is a very encouraging result. Although not included in this paper, some experiments on attribute selection with this data also use principal components analysis (PCA), and share some characteristics with work on dimensionality reduction [11].

The methods of this work are part of a larger class of data driven methods that have been successfully applied in diverse domains, such as financial indexes [12]. The evaluation of attribute merits and subsequent selection of a small subset of attributes may be viewed as an optimization problem with regression accuracy as merit and weighted costs of attributes (modeling physical-world costs of acquiring those data) as cost. It may then be studied in the context of the very large and diverse body of work on optimization methods and applications [13], [14].

VIII. Conclusion

Predicting CO and NOx levels in turbine flue emissions is a task of growing importance given the increasing need for such monitoring. While direct measurements provide the most reliable values, they are cumbersome and expensive. Augmenting such direct observations with predictions from more easily and inexpensively measured attributes, such as temperature and pressure values in the ambient environment and at key locations in the turbine, is a valuable way of reducing the cost and complexity of monitoring.

This paper has reported on an experimental study of the accuracy with which such predictions may be made, using a recently published dataset from prior work. In contrast to some earlier work, a notable feature of this work is the emphasis on explainable predictions. While one may reasonably expect this emphasis (restriction) to imply a penalty in prediction accuracy, the study reveals quite the opposite: These predictors provide higher accuracies than those reported in prior work on the same data. Ongoing work is studying the use of attribute merit metrics to trim the set of attributes used for prediction. Evaluations on other datasets are also planned.

Acknowledgments

This work was supported in part by the US National Science Foundation with grant 1848747 and was made possible by the dataset due to prior work [1].

References

[1] Heysem Kaya, Pinar Tüfekci, and Erdinç Uzun, "Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, pp. 4783–4796, 2019.

[2] Tamara Grujic Supuk, Ana Kuzmanic Skelin, and Maja Cic, "Design, development and testing of a low-cost sEMG system and its use in recording muscle activity in human gait," *Sensors*, vol. 14, no. 5, pp. 8235–8258, 2014. http://www.mdpi.com/1424-8220/14/5/8235

[3] Ron Kohavi, "The power of decision tablesa," in *Proceedings of the 8th European Conference on Machine Learning (ECML)*, Apr. 1995, pp. 174–189.

[4] Hirotogu Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*. New York, NY: Springer New York, 1998, pp. 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15

[5] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," in *Poster papers of the 9th European Conference on Machine Learning*. Springer, 1997.

[6] Ross J. Quinlan, "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*. Singapore: World Scientific, 1992, pp. 343–348.

[7] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] Marko Robnik-Sikonja and Igor Kononenko, "An adaptation of relief for attribute estimation in regression," in *Fourteenth International Conference on Machine Learning*, Douglas H. Fisher, Ed. Morgan Kaufmann, 1997, pp. 296–304.

[9] Igor Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *European Conference on Machine Learning*, Francesco Bergadano and Luc De Raedt, Eds. Springer, 1994, pp. 171–182.

[10] Kenji Kira and Larry A. Rendell, "A practical approach to feature selection," in *Ninth International Workshop on Machine Learning*, Derek H. Sleeman and Peter Edwards, Eds. Morgan Kaufmann, 1992, pp. 249–256.

[11] A. Alhowaide, I. Alsmadi, and J. Tang, "PCA, random-forest and pearson correlation for dimensionality reduction in ioT IDS," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Sep. 2020, pp. 1–6.

[12] R. Kyung and M. Kye, "Study on the CBOE volatility data forecast using statistical and computational simulations," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Sep. 2020, pp. 1–5.

[13] Mykel J. Kochenderfer and Tim A. Wheeler, *Algorithms for Optimization*. Cambridge, Massachusetts: MIT Press, 2019.

[14] S. Ray, S. K. Bishnu, A. Chatterjee, and M. Gangopadhyay, "Resonant frequency optimization of cylindrical liquid antenna using particle swarm optimization algorithm," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Sep. 2020, pp. 1–3.

Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis: An Ensemble-based Prediction Model

Ahmed Ibrahim
University of Waterloo
Computer Science Department
Waterloo, Ontario, Canada
a24ibrah@uwaterloo.ca

***Abstract*—Data collected from social media such as tweets, posts, and blogs can assist in an early indication of market sentiment in the financial field. This has frequently been conducted on Twitter data in particular. Using data mining techniques, opinion mining, machine learning, natural language processing (NLP), and knowledge management, the underlying public mood states and sentiment can be uncovered. As cryptocurrencies play an increasingly significant role in global economies, there is an evident relationship between Twitter sentiment and future price fluctuations in Bitcoin. This paper assesses Tweets' collection, manipulation, and interpretation to predict early market movements of cryptocurrency. More specifically, sentiment analysis and text mining methods, including Logistic Regressions, Binary Classified Vector Prediction, Support Vector Mechanism, and Naïve Bayes, were considered. Each model was evaluated on their ability to predict public mood states as measured by 'tweets' from Twitter during the era of covid-19. An XGBoost-Composite ensemble model is constructed, which achieved higher performance than the state-of-the-art prediction models.**

Keywords*—Bitcoin, Sentiment Analysis, Market Movement, prediction Models, Ensemble Modeling, Validation Measures.

I. Introduction

Cryptocurrencies, such as Bitcoin, Ethereum, and Litecoin, are an alternative class of digital assets primarily used as a medium of exchange [1]-[5]. Public key cryptography and blockchain technology are utilized to facilitate decentralized peer-to-peer transactions. Bitcoin, created in 2009, is widely regarded as the world's first cryptocurrency. Following Bitcoin's success, numerous other cryptocurrencies, dubbed 'altcoins' have been developed. The rise of Bitcoin and altcoins has produced a deluge of data on social media platforms, blogs, forums, and countless other online mediums. There have been quite a few researchers trying to predict Bitcoin prices' behavior based on its emotions on social media platforms, such as Twitter, using various machine learning algorithms [6]-[8]. Researchers have been known to get some significant prediction results. However, very few focus on using ensemble modeling to achieve better prediction results.

XGBoost is an ensemble classifier that provides benefits such as no need for normalized data, scalability to larger data sets, and rule-based behavior that is easier for people to interpret. Thus, this paper aims to propose a Composite Ensemble Prediction Model (CEPM) using the notion of sentiment analysis. The CEPM framework is comprised of five stages, 1) text preprocessing, 2) Sentiment Scoring, 3) individual XGBoost classifications, 4) composite ensemble aggregation, and 5) model validation. In stage1, various preprocessing steps are performed, including word quantization, text stemming, and stop word removal. The second stage includes converting tweet text into a sentiment score as a representative of its emotion. Such a task is suited to VADER, a lexicon and rule-based sentiment analysis tool that can deal with the syntax usually used on social media. In the third stage, various instances of the XGBoost classifiers are used. The ensemble modeling is designed to maximize the model performance by utilizing a stacking of ensembles using a majority vote of XGBoost ensembles. Finally, the composite ensemble model is validated using accuracy, recall, precision, and F-scores quality measure. Experimental analysis on Twitter datasets collected during the era of COVID-19 shows that the CEPM model outperforms the individual models. It can be effectively used as an efficient (Bitcoin) BTC predictor to forecast the early market movement of Bitcoin even after the COVID-19 pandemic.

The rest of this paper is organized as follows: Section 2 provided a literature review. In section 3, the text preprocessing is discussed. Vader scoring is presented in section 3. Section 4 presents the adopted classifiers. In section 5, the proposed staking ensemble is introduced. Experimental results and analysis are discussed in section 6. Finally, section 7 concludes the paper and highlights future directions.

II. Literature Review

Several attempts have been made that uses sentiment analysis to predict the early market movement of cryptocurrencies using tweets sentiment [9]-[17]. In [9], authors compared the causality of tweet sentiments, tweet volume, and buyers' ratio to sellers on Twitter with the price returns and daily trading volumes of cryptocurrencies. It has been speculated that sentiments expressed on Twitter could help in predicting cryptocurrency price changes. Li et al. [10] have attempted to demonstrate this concept by training an Extreme Gradient Boosting Regression tree model (XGBoost) with Twitter sentiments to predict ZClassic price changes. The research in [10] provided the KryptoOracle to predict the Bitcoin price for the next minute

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

using current and historical data from Twitter sentiments and Bitcoin closing prices. XGBoost, a regression tree model, was used because of its performance, speed, and retraining simplicity. In [13], Jain et al. attempted to predict the prices of Bitcoin and Litecoin two hours in advance based on the sentiments expressed in current tweets. They wanted to investigate if social factors could predict the prices of cryptocurrencies. So they used a Multiple Linear Regression (MLR) model to predict a bihourly average price from the number of positive, neutral, and negative tweets accumulated every two hours. Authors in [14] compared the significance of different preprocessing techniques for tweets' sentiment analysis. They used four different machine learning algorithms to classify tweets, and they tested 16 different preprocessing methods. Based on their results, it was recommended to use lemmatization, replacing repeated punctuation, replacing contractions, or removing numbers. The research work in [17] attempted to characterize Twitter users who use controversial terms when mentioning COVID-19 on Twitter and trained various machine learning algorithms for classifying such users. The machine learning algorithms trained on these attributes included Logistic Regression, Random Forest, Support Vector Machine, Stochastic Gradient Descent, Multi-Layer Perceptron, and XGBoost. Random Forest had the highest AUC-ROC score out of all algorithms when trained on the baseline, demographic, and geolocation data.

III. TEXT DATA PREPROCESSING METHODS

To categorize a large data set as Twitter, the data be appropriately cleaned to save computational time and increase the data manipulation's overall accuracy. In heavily text-based datasets, stemming and stop word analysis are crucial in the proper analysis [19]-[22].

A. *Text Stemming*

Stemming is a pre-processing method utilized in text mining, natural language processing, and information retrieval applications. It is an effective approach to reduce grammatical and word conjunctions to essentially extract the root form or “stem” to improve searching by automatic sorting of word endings at the time of indexing and searching. Since certain words have similar semantic meanings but different word forms, stemming allows for a reduction in the number of distinct terms in a document and increases the number of retrieved documents. The decrease in overall variability of the text, thus shortening the final output processing time for an Information Retrieval System. In stemming, converting a word to its stem assumes each is semantically related, leaving separate words with different meanings. Two main errors occur with stemming: Over-stemming and under-stemming. In over-stemming, words with different stems are stemmed from the wrong root (false positive), and under-stemming is when words that should be stemmed to a specific root are not (false negative). Porter's stemming is an example of a truncating method that removes suffixes or prefixes of a word. It consists of five steps, where within each step, rules are applied until a condition is met. The suffix is removed if the condition is completed and the subsequent step is performed. The result at the end of the 5th step is the resulting stem. The rules follow the syntax: Porter Stemming usually provides a much better output compared to other stemmers, has less stemming error rates, and also the Porter Snowball stemmer framework is independent of the language being used. A drawback of using the porter stemming algorithm is that the stems produced are not always real words, and the five steps in the algorithm make it a time-consuming process.

B. *Stop Word Removal*

Stop words in documents occur frequently but are effectively insignificant as they are used to join words in sentences. These words do not contribute to the context, and due to their frequency, they hinder information comprehension. Therefore, they are removed because they increase the amount of text in data slowing down information retrieval effectiveness in text mining. Stop words include words like "and", "are", "because" etc.

IV. SENTIMENT ANALYSIS USING VADER SCORING

To categorize tweets, the words must be assigned a positive or negative relative to cryptocurrency markets. A predefined value was assigned to the tweet's specific words to predict cryptocurrencies' probability of increasing or decreasing based on tweet sentiment. These words were cross-referenced with programming libraries containing lexicons of words that were assigned positive and negative values. The text used in early market predictors for cryptocurrencies using tweets was weighted positively or negatively based on these predetermined values.

VADER is a lexicon and rule-based sentiment analysis tool that can handle words, abbreviations, slang, emoticons, and emojis commonly found in social media [23]. It is typically much faster than machine learning algorithms as it requires no training [23][24]. For each body of text, it produces a vector of sentiment scores with negative, neutral, positive, and compound polarities. The negative, neutral and positive polarities are normalized to be between 0 and 1. The compound polarity can be thought of as an aggregate measure of all the other sentiments, normalized to be between -1 (negative) and 1 (positive). VADER was introduced by C.J. Hutto and Eric Gilbert [24]. They found that it performed better than most other sentiment analysis tools and even surpassed some human judges.

V. FORECASTING MODELS

Several machine learning algorithms can be used to drill down the data to analyze how Twitter can be an early market indicator for cryptocurrency prices. Historical research indicates that the most commonly used Twitter Sentiment analysis tools include Vector Support Machines and Naïve Bayes to categorize the data into positive or negative reflections for cryptocurrencies in the market.

A. *Support Vector Machines (SVM)*

SVM is a supervised machine learning algorithm that can be used for classification. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features), with the value of each feature is the value of a particular coordinate. For a binary categorization problem, the classification is performed by finding the hyperplane that differentiates the two classes, effectively separating the data

.

using an n-dimensional plane. In cryptocurrency Tweet predictors, the support vectors are positively or negatively valued words in tweets. SVM can be used to clearly and accurately predict an optimal threshold for positive or negative sentiment towards a cryptocurrency given a given tweet. In cases where there is no optimal solution utilizing a simple one-dimensional line, and data points have substantial outliers, the data needs to be graphed in a higher dimension. It is possible to create an n-dimensional hyperplane by transforming the data set that utilizes the same maximum distance characteristics as a two-dimensional hyperplane. By using kernel functions, mapping the data in higher dimensions is possible. For SVM, kernel functions can be represented in 3D space. These functions take low dimensional input space and transform them into a higher-dimensional space, therefore converting a non-separable problem to a separable problem. As the Logistic regressors do not optimize mislabeled data, we use SVM to minimize the classification error rather than solely rely on Naïve Bayes' likelihood. Therefore, the support vector machines model is chosen for the classification of mislabeled data. Using the hyperplane solution and mapping in the 3D plane, misplaced data can be encompassed in the proper classification. For applications in Twitter and cryptocurrencies, any Tweet related to cryptocurrencies is weighted with positive and negative values, then a hyperplane is placed to separate the data points. Once an initial hyperplane or line is determined and separates the data from each other, the ideal placement is determined. Maximizing the distances between the nearest data point and hyperplane determines the optimal solution. Once this best separating hyperplane is found, all data points added to the data set will be classified based on their position relative to the hyperplane.

B. *Naïve Bayes Classifier*

The Naïve Bayes classification is a simple model to apply to text mining. Naïve Bayes is practical as its assumptions include a feature vector and dependant variable Y. The optimal classification is determined through the maximum likelihood of the given function: While sentiment can have either a positive and negative meaning, for the sake of simplicity in this paper, a simple binary classification is used for Naïve Bayes classification. Thus or large amounts of data with a short 140-character document such as tweets, conditional-based probability can easily be used. There is little opportunity for varying thoughts in tweets about sentiments. This is based on the feature vector, words that are determined to be a positive or negative sentiment. Specifically, the frequency of these texts is collected for this specified model. In NB, targeted positive and negative words can be thought of as cues that direct each document being classified. Any words that appear multiple times with an insignificant or words that cannot be determined under any class can be removed from the documents to cleanse the data to ensure that probability calculations are more accurate. To account for negation, further manipulation of data with the addition of specific text to tag words with a negated meaning can then be counted as cues towards positive or negative sentiments accurately. The words are randomly grouped to determine the document's sentimental value, and each word's frequency is counted. Regardless of the word's position in a document, the words are placed to decrease frequency. This is based on the assumption that the word's position in the text does not affect how it is depicted in a document. The binary variable of each word is counted to determine the sentiment of the document. In this example, the tweet determines whether it is positive or negative or if the Bitcoin price increases or decreases. The maximum likelihood function is used using the prior class's probability with the likelihood that the document is given the class. Since we assume that the documents are independent of each other and do not affect the class, the maximum likelihood function becomes simpler to solve. Though the independence assumption is usually a constraint to using this model, tweets from individuals are unrelated to each other; thus, the independence assumption favorably works with the model.

VI. The Proposed Composite Ensemble Prediction model (CEPM)

We built a composite of the Extreme Gradient Boosting (XGBoost) using a majority vote over multiple cross-validations iterations. This composite is used to achieve a better overall prediction accuracy than baseline classifiers and individual boosting algorithms. XGBoost is a novel machine-learning algorithm that improves the gradient boosting decision tree (GBDT) and can be used for classification and regression problems [18]. XGBoost is a boosting-tree approach that integrates many weak classifiers to form a robust classifier. It uses the CART, classification, and regression tree model.

The CEPM framework is comprised of five stages, 1) text preprocessing, 2) Sentiment Scoring, 3) individual XGBoost classifications, 4) composite ensemble aggregation, and 5) model validation. In the initial stages, various preprocessing steps are performed, including text stemming and stop word removal. The second stage includes converting tweet text into a sentiment score using VADER. The VADER sentiment analysis algorithm was used to assign each tweet a compound sentiment score based on how positive, negative or neutral their words were. The final sentiment score is factored in the number of Twitter followers, likes, and retweets associated with each tweet. The closing price of Bitcoin, the final sentiment score, and the moving average of the last 100 data points were four input variables for our machine learning models. In the third stage, various instances of the XGBoost classifiers are used. The ensemble modeling is designed to maximize the model performance by utilizing a stacking of ensembles using a majority vote of XGBoost ensembles. In this paper, a 10-fold cross-validation method was employed. The dataset was divided into ten parts, 9 of which were taken in turn as the training set, one as the test set. The average value of the ten results was used as the evaluation value of the algorithm performance. Meanwhile, the experiment repeated the above process ten times, and ten evaluation values were obtained for each model, and their mean values and corresponding 95% confidence intervals were counted. The CEPM ensemble model is then validated using various quality measures.

VII. Experimental Analysis and Results

A. *Evaluation Metrics*

It was found that a confusion matrix is the most commonly used measure to determine the quality of the methods used in predicting the real-value cryptocurrency trading strategies. The confusion matrix provides a visual performance assessment of a classification algorithm as a matrix, which is then used to determine the quality of the results given the classification problem. For example, a confusion matrix can analyze models for understanding sentiments toward bitcoin in Tweets. Based on the words used in association with the term "bitcoin," each tweet is assigned to a negative or positive category. Positive tweets are indicators of upward movements in the bitcoin price. The most popular metrics used to evaluate the results presented in a confusion matrix include accuracy, precision, recall, and F-score. Each metric gives a value that can communicate whether the model is a good model or not [25]-[29] .

Accuracy is computed by determining the percentage of observations that were labeled correctly. This measure has been used as evidence to support the quality of some models used to predict bitcoin pricing. However, accuracy is not the most reliable metric since accuracy provides misleading results as the classes are not balanced, as is the bitcoin market. Accuracy is given as a percentage. The closer this value is to 100%, the better the model's predictive ability is. Precision measures the ratio of correct positive inputs. Recall, also known as the sensitivity, measures the ratio of the items present in the correctly identified input. These metrics focus on the true positives, making their results more reliable. If Precision is a higher ratio, it represents a robust predictive ability by the model. Lastly, the F-score; takes the weighted average of precision and recall, taking both false positives and false negatives. This metric is beneficial in evaluating cases with uneven class distributions [30]-[34].

B. *Experimental Datasets*

We used the Twitter dataset from [35][36]. The preprocessing steps over time with BTC's closing prices are computed per minute. As tweets are created much more frequently than once a minute, we aggregated all tweets' scores into a per-minute. The CEPM ensemble model is validated using accuracy, recall, precision, and F-scores quality measure. It can be shown from Table 1 that the XGBoost ensemble has the highest Precision, recall, and F-score as compared to Logistic regression, SVM, NB, and a single XGBoost. We have assessed the proposed CEPM model's performance using "accuracy" as another quality measure, as shown in Fig.1. It can be shown from Table 2 that the CEPM model has achieved an improvement of up to 21%, 16%, 18%, and 22%, in the Precision, Recall, F-score, and accuracy, respectively, as compared to the LR, SVM, NB, and XGBoost. In Fig.2, it can be illustrated that the proposed CEPM takes more iterations measured by the increase in the computational time as compared to other algorithms. Further adjustment to the number of iterations is considered future work to balance the trade-off between accuracy improvement and time overhead.

TABLE I. PRECISION, RECALL, F-SCORE (COVID-19 TWEETS)

	Precision	Recall	F-score
LR	0.6743	0.4532	0.54207141
SVM	0.64843	0.5543	0.59768152
NB	0.665732	0.65421	0.65992071
(XGBoost)	0.78953	0.809532	0.7994059
CEPM	0.8926	0.883474	0.88801355

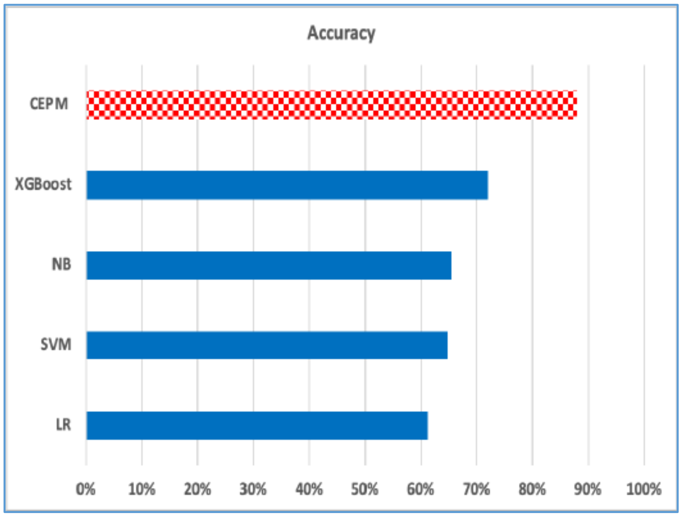


Fig. 1. Accuracy (COVID-19 Tweets)

TABLE II. % OF IMPROVEMENT IN PRECISION, RECALL, F-SCORE, ACCURACY (COVID-19 TWEETS)

Precision	Recall	F-score	Accuracy
21%	16%	18%	22%

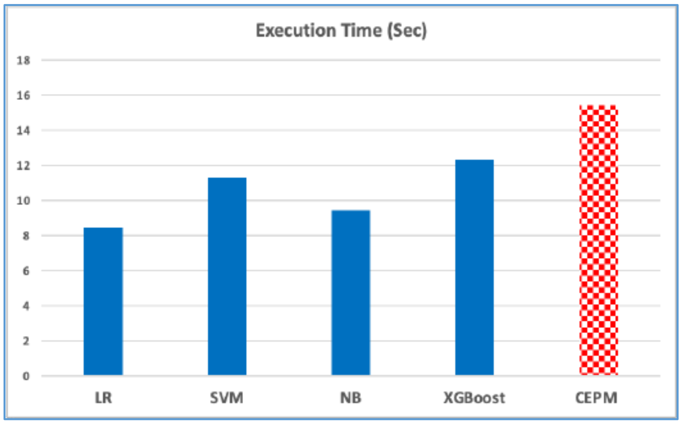


Fig. 2. Execution Time (COVID-19 Tweets)

VIII. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have developed a composite aggregate of the well-known XGBoost classifier to predict the BTC early market movement better. Experimental results show that the proposed CEPM outperforms other state-of-the-art techniques using Twitter datasets collected during the Era of COVID-19. The proposed model can be further adopted to forecast the BTC market even after the COV-19 pandemic to assess individuals and firms trading in future investments. Future research directions would include adjusting the number of incremental iterations of each XGBoost and incorporating various sentiment scoring schemes compared to VADER.

References

[1] Tan, X., & Kashef, R. (2019, December). Predicting the closing price of cryptocurrencies: a comparative study. In *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems* (pp. 1-5).

[2] Ibrahim, A., Kashef, R., Li, M., Valencia, E., & Huang, E. (2020). Bitcoin Network Mechanics: Forecasting the BTC Closing Price Using Vector Auto-Regression Models Based on Endogenous and Exogenous Feature Variables. *Journal of Risk and Financial Management*, *13*(9), 189.

[3] Ibrahim, A., Kashef, R., & Corrigan, L. Predicting market movement direction for bitcoin: A comparison of time series modeling methods. *Computers & Electrical Engineering*, *89*, 106905.

[4] Tobin, T., & Kashef, R. (2020, June). Efficient Prediction of Gold Prices Using Hybrid Deep Learning. In *International Conference on Image Analysis and Recognition* (pp. 118-129). Springer, Cham.

[5] A. F. Ibrahim, L. Corrigan and R. Kashef, "Predicting the Demand in Bitcoin Using Data Charts: A Convolutional Neural Networks Prediction Model," 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), London, ON, Canada, 2020, pp. 1-4, doi: 10.1109/CCECE47787.2020.9255711.

[6] Kashef, R., & Kamel, M. S. (2010). Cooperative clustering. *Pattern Recognition*, *43*(6), 2315-2329.

[7] Kashef, R. (2008). Cooperative clustering model and its applications.

[8] Rasha Kashef, A boosted SVM classifier trained by incremental learning and decremental unlearning approach, Expert Systems with Applications, 2020, 114154, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.114154.

[9] O. Kraaijeveld and J. D. Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," *Journal of International Financial Markets, Institutions, and Money*, p. 101188, Mar. 2020.

[10] T. R. Li, A. S. Chamrajnagar, X. R. Fong, N. R. Rizik, and F. Fu, "Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model", *Frontiers in Physics*, vol. 7, Oct. 2019.

[11] S. Mohapatra, N. Ahmed, and P. Alencar, "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments", arXiv:2003.04967 [cs.CL], Feb. 2020.

[12] C. Kaplan, C. Aslan, and A. Bulbul, "Cryptocurrency Word-of-Mouth Analysis via Twitter", ResearchGate, 2018. [Online]. Available: https://www.researchgate.net/publication/327988035_Cryptocurrency_Word-of-Mouth_Analysis_viaTwitter

[13] A. Jain, S. Tripathi, H. D. Dwivedi and P. Saxena, "Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis", 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, 2018, pp. 1-7, doi: 10.1109/IC3.2018.8530659.

[14] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," *Expert Systems with Applications*, vol. 110, no. Complete, pp. 298–310, Nov. 2018.

[15] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *Journal of Computational Science*, vol. 36, p. 101003, Sep. 2019, doi: 10.1016/j.jocs.2019.05.009

[16] A. Rosen, "Tweeting Made Easier," *Twitter*, 07-Nov-2017. [Online]. Available: https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html. [Accessed: 24-Jul-2020].

[17] H. Lyu, L. Chen, Y. Wang and J. Luo, "Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19," in *IEEE Transactions on Big Data*, doi: 10.1109/TBDATA.2020.2996401.

[18] Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020, September). Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. In *Healthcare* (Vol. 8, No. 3, p. 247). Multidisciplinary Digital Publishing Institute.

[19] Kashef, R., & Kamel, M. S. (2009). Enhanced bisecting k-means clustering using intermediate cooperation. *Pattern Recognition*, *42*(11), 2557-2569.

[20] Kashef, R., & Kamel, M. (2006, November). Distributed cooperative hard-fuzzy document clustering. In *Proceedings of the Annual Scientific Conference of the LORNET Research Network*.

[21] Kashef, R., & Kamel, M. S. (2007, October). Hard-fuzzy clustering: a cooperative approach. In *2007 IEEE International Conference on Systems, Man and Cybernetics* (pp. 425-430). IEEE.

[22] Yeh, T. Y., & Kashef, R. (2020). Trust-Based Collaborative Filtering Recommendation Systems on the Blockchain. *Advances in Internet of Things*, *10*(4), 37-56.

[23] C. J. Hutto, "VADER-Sentiment-Analysis," *GitHub*. [Online]. Available: https://github.com/cjhutto/vaderSentiment. [Accessed: 24-Jul-2020].

[24] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," presented at the Eighth International AAAI Conference on Weblogs and Social Media, May 2014, Accessed: Jul. 24, 2020. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

[25] G. Hass, P. Simon and R. Kashef, "Business Applications for Current Developments in Big Data Clustering: An Overview," 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore, 2020, pp. 195-199, doi: 10.1109/IEEM45057.2020.9309941.

[26] Close, L. and Kashef, R. (2020) Combining Artificial Immune System and Clustering Analysis: A Stock Market Anomaly Detection Model. Journal of Intelligent Learning Systems and Applications, 12, 83-108. doi: 10.4236/jilsa.2020.124005.

[27] Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. Applied Sciences, 10(21), 7748.

[28] Kashef, R. F. (2018, January). Ensemble-Based Anomaly Detetction using Cooperative Learning. In *KDD 2017 Workshop on Anomaly Detection in Finance* (pp. 43-55). PMLR.

[29] M. Ebrahimian and R. Kashef, "Efficient Detection of Shilling's Attacks in Collaborative Filtering Recommendation Systems Using Deep Learning Models," *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore, Singapore, 2020, pp. 460-464, doi: 10.1109/IEEM45057.2020.9309965.

[30] Ebrahimian M, Kashef R. Detecting Shilling Attacks Using Hybrid Deep Learning Models. *Symmetry*. 2020; 12(11):1805. https://doi.org/10.3390/sym12111805

[31] Kashef, R. (2020). Enhancing the Role of Large-Scale Recommendation Systems in the IoT Context. *IEEE Access*, *8*, 178248-178257.

[32] Nawara, D., & Kashef, R. (2020, September). IoT-based Recommendation Systems–An Overview. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-7). IEEE.

[33] Kashef, R., & Niranjan, A. (2017, December). Handling Large-Scale Data Using Two-Tier Hierarchical Super-Peer P2P Network. In *Proceedings of the International Conference on Big Data and Internet of Thing* (pp. 52-56).

[34] Li M, Kashef R, Ibrahim A. Multi-Level Clustering-Based Outlier's Detection (MCOD) Using Self-Organizing Maps. *Big Data and Cognitive Computing*. 2020; 4(4):24. https://doi.org/10.3390/bdcc4040024

[35] Pano, T., & Kashef, R. (2020, September). A Corpus of BTC Tweets in the Era of COVID-19. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-4). IEEE.

[36] Pano, T., & Kashef, R. (2020). A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing*, *4*(4), 33.

Expert Finding In Scholarly Data: An Overview

Abrar A. Almuhanna
Department of Computer Science
College of Computer Science and Engineering
Taibah University
Medina, Saudi Arabia
Abrar_ali_m@hotmail.com

Wael M. S. Yafooz
Department of Computer Science
College of Computer Science and Engineering
Taibah University
Medina, Saudi Arabia
waelmohammed@hotmail.com, wyafooz@taibahu.edu.sa

***Abstract*—In the era of digital transformation, when scholarly literature is rapidly growing, there are hundreds of papers published online daily in different fields, especially in the academic field. The huge volume of research papers published makes it difficult to find an expert/scholar to collaborate within a specific research area. This considers one of the most challenging factors in academia. Many researchers have proposed several methods to rank the authors based on their expertise in a specific area, focusing on co-citation, using keywords and their principal areas of research. The significant relationship or collaborations between scholars are credible. This paper explores the existing methods and tools in related studies to obtain author expertise among the scholarly network in determining the expertise in a specific area. Also, the semantic relations in the heterogeneous networks are presented with an overview of building a visual scholarly network that is beneficial in many pieces of research in data mining and related areas.**

Keywords— Expert Finding, Co-colleaperation, Semantic Similarity, Expert Scholar.

I. Introduction

Presently, the global collaboration between researchers expands that researchers and universities encourage knowledge and resource exchange. Researchers now disseminate their knowledge and publications throughout the world. Besides, digital technology has the power to bridge the distance and promote cross-disciplinary and cross-border collaborations inspiring scholarly networks enhancing the social activity experience at conferences creating a better professional network [1].

From 1945 till 1970 had been the age of rapid growth amongst scholars. Print publishing was customary in the years between 1971 and 1995. Subsequently, researchers experienced the digital age between the years 1996 to 2004. Since 2005, the era of open access and publications took 10 degrees shift. Thus, it is evident that the Internet had been the major contributor to the rise of scholarly community networks [2].

Besides, scholarly communities have revolutionized the method of information publication and research sharing. Many academic communities that provide guidance have emerged influencing the structure of the research community. The website ResearchGate was designed to facilitate access to scholarly research in increasing the collaboration between researchers. People from around the world get registered on this to retain an instant online presence. It allows researchers to maintain their online presence and provides them with a convenient platform to exhibit the research, before publication. Researchers frequently use such platforms to review the research with the scholarly community, in creating professional partnerships. Google Scholar that supports new, inexperienced scientists to obtain assistance and advice from experienced professionals in the field is the best example. Despite this easy access, there are shortcomings in accessibility, usability, support from a range of contributions, open infrastructure, and most importantly, community building in scholarly communication and publishing [3].

According to [4], 265 young researchers use the Internet for informal communication and social networking to create and publish their work. Hence, comprehending the academic and scholarly networks is essential in handling different aspects of ongoing research. The scholarly articles create a professional community where researchers build social networking and connections through old collaborations while creating new links with relevant work. A scholarly network allows improved personal visibility through online profiles of the publications, areas of expertise, and through related content sharing. In the scholarly sphere, a community is a cluster of researchers having similar perspectives on the areas of science. The literature is filled with identifying the communities within a network through structured texts. This assists in various aspects, not being limited to, finding the trend analysis, searching for bibliographic recommendations, and looking for experts in a specific area [5].

However, the academic knowledge contains a massive sum of research work and academic expert information. Every domain needs experts who research the relevant field to explore new dimensions in the specific field. An expert is a master of a particular domain and possesses in-depth knowledge of the related area. Experts have extraordinary subject knowledge in the area of expertise due to the number of years of experience and research work done over a while. In the past decade, obtaining assistance from an academic research expert has gained some attention in simplifying complex tasks. Researchers have developed expert finding systems based on Machine Learning and Natural Language Processing. They had worked with scholarly networks and had collected datasets from large organizational and academic databases to discover the expertise of an individual [6].

Searching for a resource-person is to know the expertise of someone, who excels in a particular area. In some instances, authors are from similar areas of expertise but differ in the specialized field. There are many papers written on computer science. Identifying an author specialized in computer science is crucial because, the author can be a specialist in Artificial Intelligence, Networks, Web Development, or Data Science. Similarly, it is applied to all other domains of science, technology, other humanities groups, and where research work is applied [7].

Despite the advancements in the research domain, this field is still progressing. The researchers are trying to make it more efficient, less time-consuming, and uncomplicated. Most of the studies that focus on automating this process by using the similarity of the keywords [8], the heterogeneous semantic similarity of texts [9], and collecting information about experts through citation networks [10]; containing a group of authors

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

with similar interests can be from different perspectives as presented in Fig.1.

This study presents an overview of studies to scout the author's expertise in the academic network. Also, exploring the heterogeneous information networks is presented through this. Searching for a potential collaborator or an expert within the network is one of the most challenging factors in academia. Additionally, an overview of the process of building a scholarly network is presented.

The rest of the paper is organized as follows. In Section II, we review some related works on the Expert Finding in Scholarly Data categorize them into four sections. While in Section III, we present a discussion on the limitations of the existing methods. The final Section IV concludes the paper with a summary.

II. EXPERT FINDING IN SCHOLARLY DATA

Recurrent Scholarly datasets make the complexity of the researchers improves gradually. This causes a growing interest in obtaining author expertise through an automated process since various researchers with different areas of expertise gather in conducting research projects.

In fact, in many instances, the researcher's expertise recognition is of vital importance. This can be conducted through qualitative attributes or evaluation since there is no known or established ground to assess the data and its needs.

A. Scholarly Data Analysis

There are several approaches under multi-type correlations amongst diverse academic entities within scholarly, big data environments. These industries and academic societies have attracted attention towards scholarly big data. There had been numerous studies conducted to explore international academic collaborations. The collaborator can identify through two aspects as behavioral and attitudinal perspectives [11].

Xia et al. [3], have presented a comprehensive survey of Big Scholarly Data (BSD). They have introduced the complete background of the BSD and have presented an overview of the relevant technologies used in data management. They start with scholarly data collection that has an acquisition framework and information extraction. Besides, they have performed various analyses on popular datasets like statistical analysis, scholarly network analysis, scholarly text mining, etc. They have also divided the scholarly network into five categories. That includes Co-author Networks, Citation Networks, Co-citation Networks, Bibliographic Coupling Networks, and Co-Word Networks. The Scholarly Text Mining is based on Textual Pattern Analysis and Topical Analysis.

Kwiek [12], has investigated the behavioural aspect, for the International Research Collaboration (IRC), and for the attitudinal aspect, the International Research Orientation (IRO) was studied. The dataset used for this study has collected through 17,211 survey forms filled by university scholars from 11 European countries. Additionally, the research has addressed an important question "What makes some European academics more prone to collaborating with international colleagues in research than others?" The results of the experiment found that for the academic-discipline, individuals in hard academic fields consider their research less internationally concerned (low IRO), while they are intensely collaborating with international communities (high IRC). The opposite situation was observed among individuals in soft academic fields. Whiles the results for the academic-generation factor in all 11 European countries for international collaboration, was the lowest among the younger generation of scientists (joined academia since 2001).

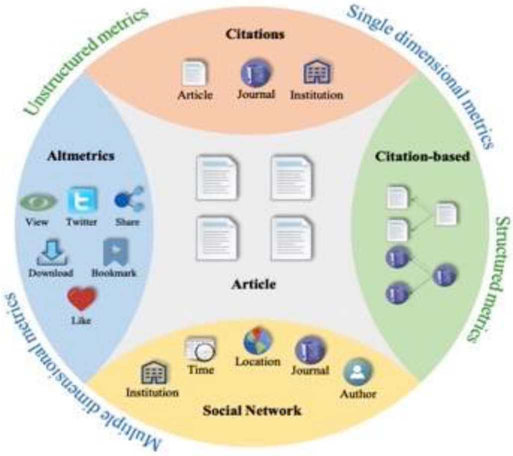


Fig. 1. Types of citation networks [10].

Kyvik et al. [13], examined the degree of scientific collaboration in research for four Norwegian research universities in order to find the effectiveness of research collaboration on the quantity and the quality of research outcomes. Accordingly, it mainly investigates whether the respondent conducts most of their research alone, in a formal group under the university administration, or within an international network. Further, they employed three independent logistic regression models (OLS-regression) to find effective and relevant factors on the research performance of individuals. As a result, the strongest influencer on the number of publications of a researcher is their membership in an international research network.

Liu et al. [14], presented a novel technique to detect scholarly communities that will identify it using unstructured data. For this purpose, they have used three different types of datasets from the real-world. The nodes and edges present between the nodes are basically used to show researchers and the collaborations between those researchers respectively. The proposed technique has three main steps with multiple sub-steps in them. The technique is based upon Reference Entity Recognition, Researcher Relatedness Quantification, and Community Clustering.

Zhou et al. [15], proposed the multi-dimensional network analysis method (AIMN), aims to address the activeness of researches and the popularity of the articles within a social context. It is a computational method that describes multi-type correlations amongst basic academic entities including three relations, Researcher-Researcher, Researcher-Article, and Article-Article. Moreover, they used an improved algorithm known as the Random Walk with Restart model (RWR) in order to calculate the similarity relevance of two nodes. Their results found that an overall performance average consists of between 101-140 nodes, and the performance of AIMN outperformed three other methods compared; MVCWalker, CCRec and basic RWR without AIMN.

Batagelj et al. [16], explored the bibliographic networks and introduces the concept of time in it by using temporal quantities. Also, they have used the authorship network in order to get the author with the largest number of works, and the citation network to know the most cited paper. Thus, by implementing the multiplication of networks, where the derived network can reveal more properties such as temporal co-occurrence, temporal co-authorship, and temporal citations between the journals. Moreover, presented a calculation of fractional productivity of an author in order to calculate the group productivity of authors. Whereas, the properties calculated from the citation network for peer review show that an author can have a smaller number of citations over a large number of years and vice versa.

Batagelj et al. [17], focused to present peer review as an emerging research field by determining the relationships between authors, journals, and their fields. They have taken the time frame of 1950-2016 to determine the relationships, includes citations, recommendations, and collaborations over the specified time period. In this way, by determining authors, their publications, journals, and collaborations at different stages, they made a path to trace evolution over time. Authors have found out three stages of this emerging field. These include the first one that is before 1982 which was the most influential one, then from 1983 to 2002 which is based on biomedical journals, and then the last one from 2003-2016 that contained more specialized journals.

B. Expert Recommendation Systems

Online research search engines and forums are used by many researchers locating scientific articles, in exploring experts to work with, and discovering the venue of publishing, However, they do not receive the optimal outcomes in the initial phases due to the keyword-based searches and lack of interest towards the data. Several recommendation approaches have been proposed in ranking the authors on their expertise in a specific area, some focusing on co-citation, and others on the keywords.

Xiancheng Li et al. [18], proposed a network approach in expertise retrieval. They studied credit allocation for co-authors based on credit allocation algorithms on author-papers using MeSH connections. They employ the HeteAlloc credit allocation algorithm based on path similarities. Here the main task is to allocate credit for each MeSH term in a paper, to correspond to the entire publication history of authors to find out their expertise. However, the unweighted version of the HeteAlloc algorithm is employed since the MEDLINE dataset mostly carries publications with numerous MeSh terms. Compared with the DHA and BL models, this method proved more effective in classifying scholarly expertise when served in different areas.

West et al. [19], extended the capabilities of the Eigenfactor Recommendation algorithm (EFrec) score, proposing two methods as author-based [19], and citation-based [20]. The author-based Eigenfactor was applied in the social science research network (SSRN), which can successfully analyze the author's contribution. Hence, the listing of the authors is quantitative and is based on the number of citations (Fig.2). The results for the author ranking reflect that the first 20% of the authors are responsible for the overall 7.8% Eigenfactor score in author rating. The Institution ranking results show that the US is responsible for 70% Eigenfactor score in the social science research network.

On the other hand, the citation-based method based on finding the most relevant papers for a given article (seed paper). The algorithm creates possible orders for the relevant papers for different users by utilizing the hierarchical arrangement of scientific knowledge. The overall results of an online comparison between EFrec and other recommendation methods co-citation and co-download by using the A-B testing environment indicate that in click-through rates (CTR) the co-citation has 0.26% while EFrec 0.24%, but EFrec comprises a wide range of recommendation coverage than co-citation.

Sriramoju [21], has proposed a framework based on the principle of heat diffusion algorithm that takes three different matrices and conducts web mining. Thus, by bringing the requested web pages and web users into various relationships and compute ranked outcomes. A prototype application was constructed where the findings are primarily related to the consistency of ranked results. The analytical results show that the proposed framework for heat diffusion is successful in seeking the best expert search results.

Javadi et al. [22], proposed the Expert Finding System (EFS), which takes as input the user name and keywords that indicate expertise in that field. Hence, based on that a certain number of ranked experts are the output of the model (Fig.3). Consequently, it provides a list of expert's names in order to collaborate with an expert in a specific field, with an accuracy of 71.50%. However, the high cost and complexity of the calculation of the resemblance of the keywords was a serious problem in the execution. Thus, for saving efficiency, two measurements were estimated and saved for the existing 38,531,031 keyword pairs for the similarity measure period of two years (2012 and 2013).

Zevio et al. [23], proposed a model for finding the experts in the scholarly articles which take the input of data, graph mining, and semantic annotation in order to access the semantic concepts in any scholarly article. Also, they used bi-pattern mining with a new substitute is presented in which the enumerations are reduced by the restriction of component patterns. As a result of precision 0.28 and recall 0.46 by selecting 12 patterns that contained IE field.

Zhou et al. [24], developed a methodology for the extraction of knowledge and expert findings from a numerous article. Compared to the other industry-oriented models, the novelty of the work lies in the use of a hybrid approach that employs not only graph mining (graph abstraction) but also text mining (machine learning algorithms) presenting more comprehensive output. Where the authors, title, and keywords were parsed from the raw data file using CERMINE parser.

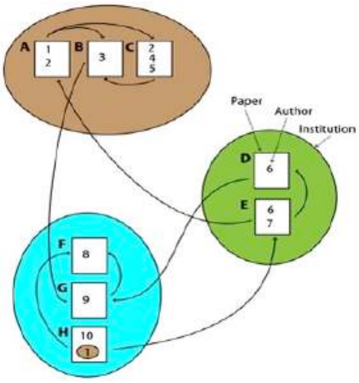


Fig. 2. Journals, authors, institutions citation network [19].

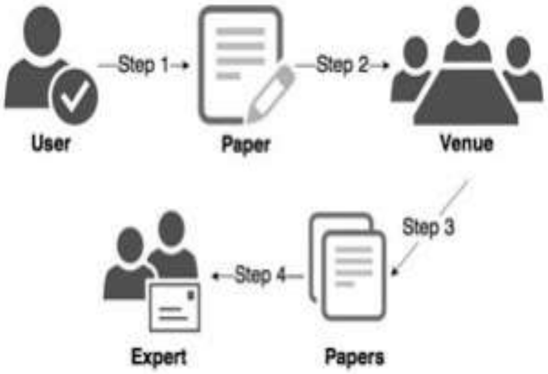


Fig. 3. Expert Finding System model [22].

Liang et al. [25], presented five different models in order to determine a group of experts rather than individual experts, thus by giving a query about a specific topic. Three types of variables were introduced which play a vital role in evaluation: Queries (Q), Documents (D), and Group (G), where the five models are, GDQ, GQD, QDG, QGD, and DGQ. For the evaluation process, the retrieval matrices MAP, p@k, and NDCG@k were used. The results indicate that the QDG model was estimated to be the least performing model among all.

Vergoulis et al. [26], proposed a novel approach known as VeTo, for expanding the set of experts based on the scholarly knowledge graphs. VeTo utilized the latent patterns between the topics and their published venues for the extraction of field specialists. According to the aggregate similarities based on Author-Paper-Topic (APT) and Author-Paper-Venue (APV) meta paths. As result, VeTo-APT and VeTo-APV in both VLDB and SIGMOD conferences achieved better performance, thus by combining both meta paths VeTo will be capable to identify some unique good results.

C. Similarity in HINs

The advancement in Information Communication Technologies makes the nature of the networks heterogeneous, increasing the complexity of the network. Thus, every node carries a different relationship with the other node. Nevertheless, it has a great capability for knowledge discovery due to its enriched semantic.

Kong et al. [27], presented a simple way to classify heterogeneous networks by taking an account of each relationship between the nodes with the use of meta paths and graphs, thus aims to give class labels to these nodes collectively using graphs. The bibliographic graph network (Fig.4) is the used experiment example which has five different types of nodes including author, paper, conference, term, subject, venue. And thus, the heterogeneous networks were classified using graphs, just by using a dependence tree by applying the Breadth-First Search algorithm and send the longer paths into the path set and iterate the process.

Hoang et al. [28], highlighted the importance of establishing a hybrid recommender system that enables scientists to collaborate and find relevant materials. Also, they used acquaintances-based collaboration such as friendship and co-authorship. Research cooperation factored in scientific conferences, co-authors, and collaboration period. In addition, in evaluating the research cooperation metric, the time is included to value co-authors with recent publications more than the ones with outdated publications. The research similarity, on the other hand, describes the success rate for a pair of researchers who have never co-published a paper. This criterion is the more critical one, since unlike the research cooperation that is a quantitative review of the history already known to authors, the research similarity reveals information previously unknown.

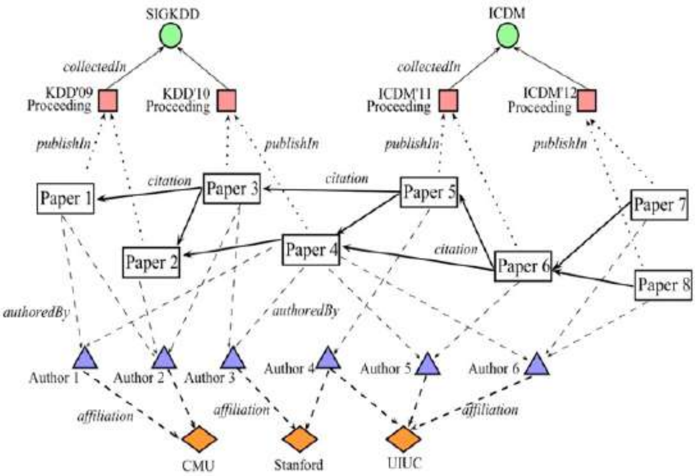


Fig. 4. The bibliographic graph network [27].

D. Semantic Relations

In this era of digitalization, extracting useful or required data from any soft PDF article is a challenging task. A lot of recent studies have been focused on this topic as this can provide useful insights and important data for computation.

Al-Zaidy et al. [29], proposed an approach to extracting data from any given number of documents, which they have incorporated iterative learner and a pattern learning external syntactic in order to extract the semantic relations between the entities. the knowledge base gained by the process in Fig.5, is then fed to a pairing sheet which is usually hypernym-hyponym. As a result, compared to the existing web IE tools, the proposed system established an improvement of the accuracy over 23% to scholarly documents.

Chen et al. [30], proposes a novel technique Weighted Heterogeneous Information Network Containing Semantic Linking (WHIN-CSL) for citation recommendations in three phases. Thus, by the abstract-abstract similarity among any of the two papers, then computing the citing linking hence select the topmost similar papers as semantic linking. Then, build undirected weighted HIN with semantic linking. As result, WHIN-CSL outperforms five baselines including PV-DBOW, PW, PWFC, MMRQ, and BM25, with a 6.9-8.4% better performance percentage for the ANN dataset and 14.3-20.4% better performance percentage for the DBLP dataset.

E. Scholarly Data Visualization

It requires a sort of time to organize and to analyze the scholarly article being published online qualitatively and quantitively. Therefore, in order to organize scholarly networks, the visualization technique is employed to organize the data by the use of statistics and other mapping methods.

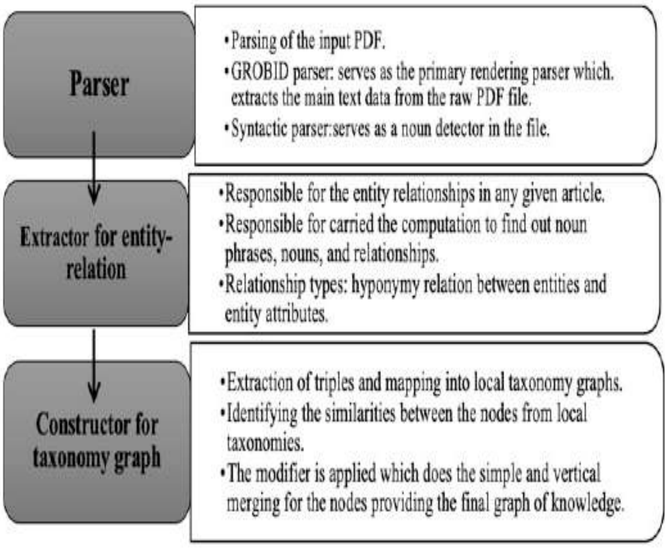

Parser
•Parsing of the input PDF.
•GROBID parser: serves as the primary rendering parser which. extracts the main text data from the raw PDF file.
•Syntactic parser:serves as a noun detector in the file.
Extractor for entity-relation
•Responsible for the entity relationships in any given article.
•Responsible for carried the computation to find out noun phrases, nouns, and relationships.
•Relationship types: hyponymy relation between entities and entity attributes.
Constructor for taxonomy graph
•Extraction of triples and mapping into local taxonomy graphs.
•Identifying the similarities between the nodes from local taxonomies.
•The modifier is applied which does the simple and vertical merging for the nodes providing the final graph of knowledge.

Fig. 5. The proposed model workflow [29].

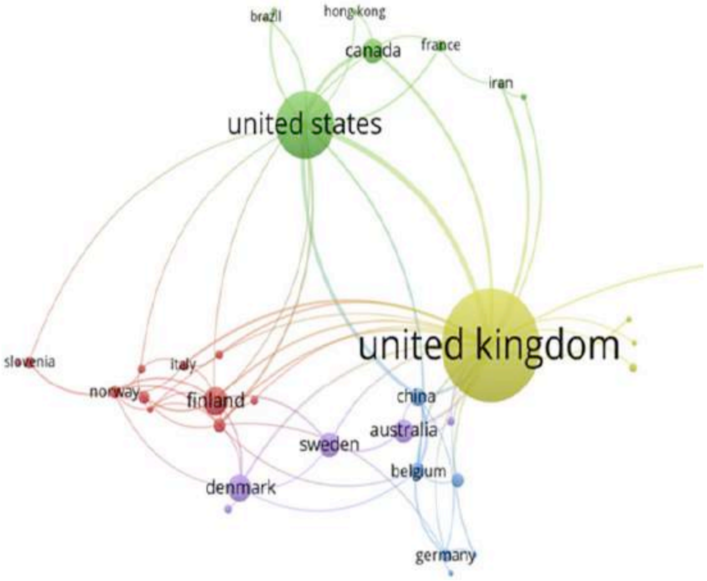


Fig. 6. Visualization of Country co-authorship network [31].

Mokhtari et al. [31], presented a bibliometric visualization and overview of the Journal of Documentation (JDoc) from its initiation from 1945-2018. Also, VOSviewer software utilized for the visualization process and implementation of various techniques such as co-occurrence, co-authorship, and co-citations (Fig.6). Also, the result was represented as ranked tables and graphs, mainly focus on the communicative and influence networks of several scientific actors, and productivity in JDoc.

While Tang et al. [32], tested the International Journal of Fuzzy Systems (IJFS) for the bibliometric data to provide optimization for the evolution and overall development. While incorporates the trends of citation, publication, citation sources, and the details of the paper which are highly cited from the IJFS journal.

Isenberg et al. [33], concentrate on the analysis and visualization of the keywords in three conferences VAST, IEEE InfoVis, and Vis/SciVis. with a total number of 2629 keywords in 1097 papers from 2004–2013. Further, the top 10 keywords are considered for analysis the similarity that occurs between the conferences. The result illustrates that the communities have the similarity in research orientation and the concerns which are of common interest. Additionally, for the 15 most frequent keywords Tableau was used in order to calculate the linear trend lines.

Liu et al. [34], proposed a new system using mining techniques named Web of Scholars. Also, the model uses methods such as searching, mining, and other complex network visualization corresponding to scholars from the field of computer science hence, provides accurate recommendations by the use of querying semantically. Where it relies mainly on the knowledge graph, the model can effectively be used as an interoperable tool which can provide the in-depth analysis.

III. DISCUSSION

This study reviews some existing methods in finding an expert scholar. Initially, it provides an overview by analyzing scholarly data and exploring the importance of international academic collaboration. Next, it provides an overview of how to discover individuals by processing the content of published research papers of a specific field. Next, a scholarly recommendation is obtained using an algorithm based on a network of heterogeneous information. Finally, the visualization techniques in academic networks are discussed. Table I presents a review of the literature study with some of its limitations. The tools and methods of links to the scholars in academia are low. The objective is to develop an interactive visualization network that assists postgraduate students and expert scholars in finding professionals and experts in specific research areas based on expertise.

IV. CONCLUSION

Due to the expanding availability of scholarly datasets, the increase in research problems, and the complexity in research, there has been a growing interest in obtaining the author expertise through an automated process since many researchers from different areas of expertise are required to work on that research projects. While a very common problem that faces the expert finding methods is the qualitative analysis or evaluation as there is no known or established ground to assess that data and it needs several additional measures. This study reviews some existing methods in finding an expert scholar. This also presents an overview of the importance of analyzing scholarly data, finding the semantic similarity in heterogeneous networks, and the visualization techniques in academic networks, that reveal important Implicit information. Additionally, this discusses the approaches, the factors to work on, and some of their limitations. There have been studies conducted by many authors to identify the best author amongst the scholarly community. Nevertheless, the problem of finding the best scholar has not been properly addressed yet. Thus, it is evident that a more systematic and theoretical analysis is required in that matter.

TABLE I. REVIEW OF THE LITERATURE STUDY

Approach / Technique	Objective	Limitation	Works-on	Author (s)
Inferential statistics and a multivariate model approach	•International academic collaboration • Focus on academic research internationalization using cluster of academic discipline	• No clear overview of the intensity of IRC and IRO • Quantification of international co-authorship was done in a self-report base, rather than examine true hard calculations using available academic databases	Co-Author	[3,12]
Analysis using OLS-regression	Find the effectiveness of research collaboration on the quantity and the quality of research outcome	The used of regression models with weak outcome	Co-Author	[13]
Node Embedding	• Discoviring from community clustering: authors mentioned in nearby lines in academic articles belong to a similar research area • Researcher Name Disambiguation (RND)	Their assumption for the authors mentioned in nearby lines belong to a similar research area may fall in some area	Co-Citation, Research Content	[14]
The Random Walk with Restart Model	Building AIMN multidimensional academic network with an improved RWR-based algorithm for recommendations	• Lacks of a justified semantic perspective into the scholarly datasets for extracting insight into the academic collaboration • An adequately detailed discussion on the adjustment plan of the coefficient in the employed equations is missing	Co-Citation, Co-Author, Research Content, Academic Activity	[15]
Network Multiplication	Exploring the bibliographic networks by introducing temporal quantity	No account of the properties of nodes and links thus may affect calculations and computational speed	Co-Citation, Co-Author,Co-Words networks	[16,17]
Credit Allocation	Analyses the degree of each co-author's contribution to collaborative work and assigns expertise level according to it	Lack of ground truth is a big drawback in the validation of this model	Co-Author	[18]
Eigenfactor Recommendation algorithm (EFrec)	EFrec algorithm: citation-based method which based on finding the most relevant papers for a given article (seed paper)	Relaying only on the citation without any semantic information is time-consuming to construct	Co-Citation, Co-Author	[19,20]
Heat diffusion-based ranking	Web mining method to rank the experts from the web search engine query	Lack of ground truth is a big drawback in the validation of this model	Co-occurrence, Co-Words	[21]
Similarity between keywords	Expert Finding System (EFS) using the bibliographic details of author papers, venue, and published articles	High cost and complex-ity of calculation the keywords from more than one perspective in just a period of two years	Co-occurrence, Shared neighbors	[22]
Probabilistic language modeling techniques (Bayes' Theorem)	Five different models in order to determine and ranking a group of experts	Needs more advanced ways to extract individual expert for more efficient identification of an expert given a query topic	binary, graded, and number	[24,25]
Graph-based	Using Graph Mining and semantic annotation for expert finding	The techniques use only the recent patterns which may greatly reduce the accuracy of the collaboration results	Co-Citation, Co-Author	[23]
	VeTo, for expanding the set of experts based on the scholarly knowledge graphs	Their method outperforms all other tested methods, yet its superiority over the baseline method is neglectable especially regarding the obtained results for the SIGMOD	Co-venues, Co-publications topics	[26]
	Classify heterogeneous networks for each relationship between the nodes with the use of meta-paths and graphs	No training or any experiment results	Co-venues, Co-Author	[27]
	Build a collaborators recommender system (CRS) depending on knowledge and acquaintance-based collaborations	Only observes the researchers who either have collaborated in publication in the past or are connected through one or more colleges who have co-authored	Research Content, Academic Activity, Co-Author	[28]
	Build the Web of Scholars system using mining techniques	With regard to scholars' composition, there are no clear approaches to describe the rela tionship in the knowledge graph	Co-Citation, Co-Citation	[34]
VOSviewer	Bibliometric visualization and overview of the Journal of Documentation (JDoc)	• No similarity comparison with the global co-keywords • This study excludes the papers which are not part of the selective database thus under- estimate the impact of studied journals	Influential countries, institutes, authors, author keywords, global keywords	[31]
	Visualization of International Journal of Fuzzy Systems (IJFS) for the bibliometric data	Keyword "fuzzy" is the core of the Journal and the topics being published in IJFS. Thus, it should be excluded where more than 50% of the authors keywords have it	Global keyword, author keyword analysis	[32]

Ward's method and a Squared Euclidean distance metric	Analysis and visualization of the keywords in three conferences VAST, IEEE InfoVis, and Vis/SciVis	Having one shared vocabulary among the three conferences instead of separate three, would enable the authors to broaden their re- search scope and to generalize their findings to a more comprehensive domain of literature	Co-word	[33]

REFERENCES

[1] H. Meishar-Tal and E. Pieterse, "Why do academics use academic social networking sites?" *International Review of Research in Open and Distributed Learning*, vol. 18, no. 1, pp. 1–22, 2017.

[2] J.-C. Gue´don, B. Kramer, M. Laakso, B. Schmidt, E. Sˇimukovicˇ, J. Hansen, R. Kiley, A. Kitson, W. van der Stelt, K. Markram *et al.*, "Future of scholarly publishing and scholarly communication: report of the expert group to the european commission," 2019.

[3] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[4] R. Costas and M. R. Ferreira, "A comparison of the citing, publishing, and tweeting activity of scholars on web of science," in Evaluative Informetrics: The Art of Metrics-Based Research Assessment. Springer, 2020, pp. 261–285.

[5] R. Gonc¸alves and C. F. Dorneles, "Automated expertise retrieval: a taxonomy-based survey and open issues," ACM Computing Surveys (CSUR), vol. 52, no. 5, pp. 1–30, 2019.

[6] S. Lin, W. Hong, D. Wang, and T. Li, "A survey on expert finding techniques," Journal of Intelligent Information Systems, vol. 49, no.2, pp. 255–279, 2017.

[7] M. J. Caley, R. A. O'Leary, R. Fisher, S. Low-Choy, S. Johnson, and K. Mengersen, "What is an expert? a systems perspective on expertise," Ecology and evolution, vol. 4, no. 3, pp. 231–242, 2014.

[8] Q. Shen, "Topic discovery and future trend prediction in scholarly networks."

[9] H. A. M. Hassan, G. Sansonetti, F. Gasparetti, A. Micarelli, and J.Beel,"Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation?" in RecSys (Late-Breaking Results), 2019, pp. 6–10.

[10] X. Bai, H. Liu, F. Zhang, Z. Ning, X. Kong, I. Lee, and F. Xia, "An overview on evaluating and predicting scholarly article impact," Information, vol. 8, no. 3, p. 73, 2017.

[11] J. Czaputowicz and A. Wojciuk, International Relations in Poland: 25 Years After the Transition to Democracy. Springer, 2017.

[12] M. Kwiek, "International research collaboration and international re- search orientation: Comparative findings about european academics," Journal of Studies in International Education, vol. 22, no. 2, pp. 136– 160, 2018.

[13] S. Kyvik and I. Reymert, "Research collaboration in groups and networks: differences across academic fields," Scientometrics, vol. 113, no. 2, pp. 951–967, 2017.

[14] M. Liu, Y. Chen, B. Lang, L. Zhang, and H. Niu, "Identifying scholarly communities from unstructured texts," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2018, pp. 75–89.

[15] X.Zhou,W.Liang,I.Kevin,K.Wang,R.Huang,andQ.Jin,"Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data," *IEEE Transactions on Emerging Topics in Computing*, 2018.

[16] V. Batagelj and D. Maltseva, "Temporal bibliographic networks," *Jour- nal of Informetrics*, vol. 14, no. 1, p. 101006, 2020.

[17] V. Batagelj, A. Ferligoj, and F. Squazzoni, "The emergence of a field: a network analysis of research on peer review," *Scientometrics*, vol. 113, no. 1, pp. 503–532, 2017.

[18] X.Li,L.Verginer,M.Riccaboni,andP.Panzarasa,"Anetworkapproach to expertise retrieval based on path similarity and credit allocation," *arXiv preprint arXiv:2009.13958*, 2020.

[19] J. D. West, M. C. Jensen, R. J. Dandrea, G. J. Gordon, and C. T. Bergstrom, "Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 4, pp. 787–801, 2013.

[20] J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A recommendation system based on hierarchical clustering of an article-level citation network," *IEEE Transactions on Big Data*, vol. 2, no. 2, pp. 113–123, 2016.

[21] S. B. Sriramoju, "A framework for keyword based query and response system for web based expert search," *International Journal of Science and Research" Index Copernicus Value*, vol. 78, 2015.

[22] S. Javadi, R. Safa, M. Azizi, and S. A. Mirroshandel, "A recommenda- tion system for finding experts in online scientific communities," *Journal of AI and Data Mining*, vol. 8, no. 4, pp. 573–584, 2020.

[23] S. Zevio, G. Santini, H. Soldano, H. Zargayouna, and T. Charnois, "A combination of semantic annotation and graph mining for expert finding in scholarly data."

[24] S. Zevio, "Knowledge discovery and enrichment from scholarly data forexpert finding," inThe 21st International Conference on Knowledge En-gineering and Knowledge Management (EKAW) Doctoral Consortium,2018.

[25] S. Liang and M. de Rijke, "Formal language models for finding groups of experts," *Information Processing & Management*, vol. 52, no. 4, pp. 529–549, 2016.

[26] T. Vergoulis, S. Chatzopoulos, T. Dalamagas, and C. Tryfonopoulos, "Veto: Expert set expansion in academia," in *International Conference on Theory and Practice of Digital Libraries*, 2020, pp. 48–61.

[27] X. Kong and P. S. Yu, *Graph Classification in Heterogeneous Networks*, 2018, pp. 958–965.

[28] D. T. Hoang, N. T. Nguyen, V. C. Tran, and D. Hwang, "Research col- laboration model in academic social networks," *Enterprise Information Systems*, vol. 13, no. 7-8, pp. 1023–1045, 2019.

[29] R. A. Al-Zaidy and C. L. Giles, "Extracting semantic relations for scholarly knowledge base construction," in *2018 IEEE 12th international conference on semantic computing (ICSC)*. IEEE, 2018, pp. 56–63.

[30] J. Chen, Y. Liu, S. Zhao, and Y. Zhang, "Citation recommendation based on weighted heterogeneous information network containing semantic linking," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 31–36.

[31] H. Mokhtari, S. Barkhan, D. Haseli, and M. K. Saberi, "A bibliometric analysis and visualization of the journal of documentation: 1945–2018," *Journal of Documentation*, 2020.

[32] M. Tang, H. Liao, and S.-F. Su, "A bibliometric overview and visualization of the international journal of fuzzy systems between 2007 and 2017," *International Journal of Fuzzy Systems*, vol. 20, no. 5, pp. 1403–1422, 2018.

[33] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Mo¨ller, "Visual- ization according to research paper keywords," in *Posters at the IEEE Conference on Visualization (VIS)*, 2014.

[34] J. Liu, J. Ren, W. Zheng, L. Chi, I. Lee, and F. Xia, "Web of scholars: A scholar knowledge graph," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2153–2156.

Applications of Cryptography in Database: A Review

Huanhao Xu
Management and Systems
New York University, SPS
New York, USA
hx798@nyu.edu

Kutub Thakur
Professional Security Studies
New Jersey City University
NJ 07305, USA
kthakur@njcu.edu

Abu S. Kamruzzaman
Seidenberg School of CSIS
Pace University
NY 10570, USA
ak91252p@pace.edu

Md Liakat Ali
Department of Computer Science & Physics
Rider University
NJ 08648, USA
mdali@rider.edu

***Abstract*—Cryptography is the foundation and core of network security. Privacy protection, sensitive information is particularly important, so whether it is system development or app development, as long as there is network communication. As data plays very critical steps for IOT devices it's very crucial that cryptography applied to the databases as well. A lot of information needs to be encrypted to prevent interception and tampering. Many people use cryptography on a daily basis, not everyone is aware of it. This paper investigates the applications of cryptography in the context of databases and offers suggestions to enhance security and privacy.**

Keywords— security, privacy, public key, encryption

I. Introduction

With the popularity of blockchain and cryptocurrencies, more and more people are adopting the foundation slot of blockchain-cryptography. Because cryptography is one of the keys to the safe operation of the blockchain. There exist three categories of cryptographic algorithms known as symmetric cryptography, asymmetric cryptography (public key), and cryptographic hashing. These types are part of most encryption systems, and we'll discuss each of the next and to find where the strengths and weaknesses are. In addition to this, we need to look at what kinds of risks and attacks on cryptography. Then we can talk about the application of cryptography and database encryption techniques.

Cryptography has been a science for many years, and it has been recorded that first cryptography inscriptions were carved around 1900 BC in the tomb of an ancient Egyptian nobleman, Khnumhotep II. The inscription used some strange sacred carving symbols instead of more common symbols. It is not intended to hide information, perhaps to manifest prestige by changing forms.

In 100 B.C., at the height of the Roman Empire, Caesar used encryption to send secret messages to front-line troops. The "Caesar Code" is probably the most mentioned historical password in the literature. In a replacement password, each character in clear text is replaced by a different character to form a redaction. The variant used by Caesar is password conversion in three locations. Each letter is forwarded with three locations so that the letter A is replaced by the letter D, the letter B is E, and so on. The letter X is replaced by A.

In the early 1970s, IBM customers requested for encryption, and IBM set up an "encryption group" led by Horst-Feistel. A password called Lucifer" was created by the encryption group. In 1973, the National Standards Administration, now renamed to National Institute of Standards and Technology (NIST) proposed group passwords, which became national standards. IBM realized that lots of commercial products are purchased with no encryption support at all. Lucifer also known as DES (Data Encryption Standard) was adopted to provide that support.

DES was hacked in 1997 in a desperate search attack. The DES encryption key size was very small. As computing power grows, all the different combinations of the brute force are calculated to obtain possible clear text information. In the 1980s, there was only one option which was DES. But times have changed. Today, we have more choices, stronger algorithms, faster, and better design. Now, the question is how to choose.

In 1997, NIST again proposed a new group password scheme and 50 proposals came. In 2000, Rijndael was adopted and renamed to AES (Advanced Encryption Standard) [1-2]. In order to provide security in the database, this paper investigates the applications of cryptography in the context of databases and then offers suggestions that can provide security and privacy. The organization of the paper is as follows: section II presents cryptography definition and types, section III presents some of the existing cryptographic algorithm that provide security to the database, section IV is the analysis of cryptography, section V shows the method of analysis, section VI offers some suggestion to enhance security and privacy in the database and section VII is the conclusion.

II. The Definition of Cryptography

Encryption is converting data to make it unrecognizable and useless to unapproved persons; the most secure technique is to use mathematical algorithms and variable values called "keys". The keys you choose, in general, are random strings that are entered through encryption and are integrated into the data transformation. To decrypt the data, you must enter the exact same key.

Cryptography begins with plaintext which is not encrypted. The plaintext is converted into ciphertext with encryption as shown in fig. 1. This ciphertext can be converted back to usable plaintext using decryption. The encryption and decryption is based upon the type of cryptography scheme being employed and some form of key. This process is sometimes written as follows:

$$C = E_k(P) \qquad (1)$$

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

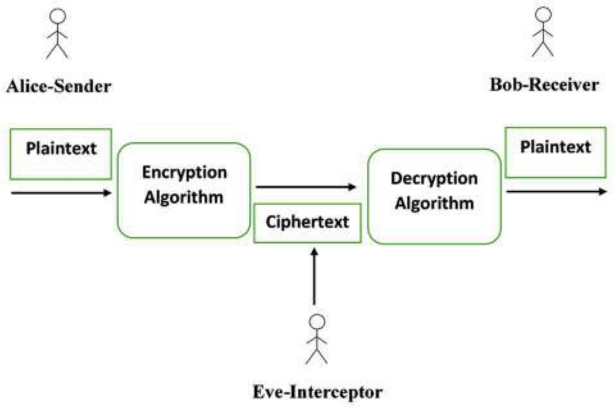


Fig. 1. Cryptography process

$$P = D_k(C) \quad (2)$$

P = plaintext, raw data, data to be encrypted
C=ciphertext, some camouflage or transformed output of clear text
E=encryption process, the process of turning clear text into a red tape in some way.
D=decryption process, the process of restoring redaction into clear text
k=the key, specialized tools used in encryption or decryption

Cryptography resembles similar mathematical algorithms for encrypting and decrypting messages, while the scientific analysis of cryptography and the destruction of encryption schemes. Cryptography is a secret term that refers to a wide range of research writing, including encryption and cryptographic analysis [3].

Three types of cryptography algorithms are discussed here:

A. Symmetric Cryptography

When making clear or secret edict transformations of information, the password system using the same key is solved and decrypted. Figure 2 shows the symmetric cryptography where both Alice and Bob share the same key.

Security in this scheme depends on the secret key and the encryption algorithm security.

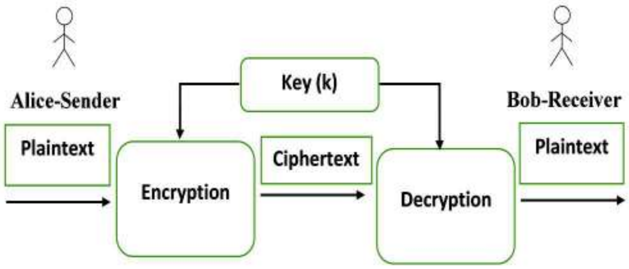


Fig. 2. Symmetric Cryptography

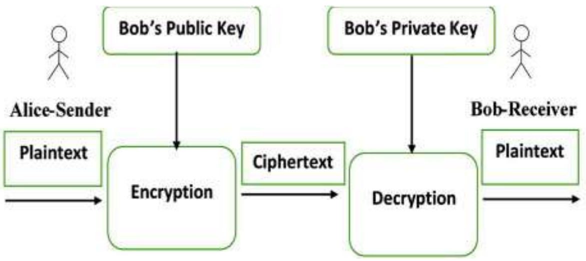


Fig. 3. Asymmetric encryption

Advantages: Algorithm open, fast, high level of secrecy, small footprint.

Disadvantages: Key distribution and management is complex.

Purpose: Encryption with large amount of information.

Representing algorithms: DES algorithm, 3DES algorithm, IDEA algorithm, AES algorithm.

Problem: If the receiver forges a message and falsely is sending it, the sender cannot excuse that it cannot resolve the confirmation of the message and cannot achieve digital signature.

B. Public-Key Cryptography

Encryption and decryption keys are not the same password system when transforming information with clear or secret text. In the asymmetric password system, every user uses one key for encryption and another key for decryption. The key used for encryption is called a public key and the key used for decryption is called private keys which is a secret key for each user. Figure 3 shows the asymmetric encryption where Alice encrypts the plaintext using Bob public key and Bob decrypt the ciphertext using his own private key [4].

Advantages: The sender and the receiver do not require to set up a secure channel to exchange their key, the key space is generally small which reduces the challenges in the key management procedure.

Disadvantages: Slow implementation, not suitable for heavy communication load situation

Purpose: Encrypting critical, core confidential data.

Representing algorithms: RSA algorithm, ElGamal algorithm, elliptic curve encryption algorithm.

Problem: Because the public key is open to the public, if a person uses his own public key encrypted data to send to us, we cannot determine who sent it. If we use the private key to encrypt the data, anyone who knows our public key can decrypt our data.

C. Cryptographic Hashing

The cryptographic hashing is a hash function that is suitable for encryption. Hashing is mainly used to provide integrity property which ensures that the information is correct and no unauthorized person or malicious software has altered the data. This is a mathematical algorithm of any size, a string of map data fixed size and one-way functions, that is, a function is actually not feasible conversion. Hashing algorithms are useful to protect data from unauthorized modification.

III. The Encryption Algorithms

Day by day people's awareness of the importance of information security is growing with the development of information technology and digital society. In 1997, the National Bureau of Standards announced the implementation of the "American Data Encryption Standard (DES)". Civil forces began to fully intervene in the research and application of cryptography, and the use of encryption algorithms such as DES, RSA, SHA, etc. increased. With the increasing demand for encryption strength, AES, ECC, and so on have recently emerged [4-7].

The following section shows the comparison of the two main types of encryption algorithm used in symmetric systems - DES and AES and two encryption algorithms used in asymmetric encryption algorithms- RSA and DSA.

A. *Data Encryption Standard (DES)*

DES encryption algorithm is a grouping password, with 64 bits for grouping data encryption and key length is 56 bits. Both encryption-decryption uses a similar algorithm. The DES encryption algorithm is the secret of the key, while the public algorithm, including the encryption and decryption algorithm. In this way, only those who have the same key as the sender can interpret the redaction data encrypted by the DES encryption algorithm. Therefore, the DES deciphering encryption algorithm is the encoding of the search key. For a 56-bit-long key, if searched by the method of exhaustion, the number of operations is 2 of 56 squares [8-9].

B. *Advanced Encryption Standard (AES)*

In the late 20th century, it is the rapid development of computers, component manufacturing process progress makes the computer processing power much more strong, DES will not provide enough security. On January 2, 1997, the National Institute of Standards and Technology: NIST announced its desire to recruit advanced encryption standards: AES), to replace DES. The U.S. federal government adopted this block encryption algorithm as a standard as well as has been analyzed by multiple parties and widely used worldwide.

AES encryption is an advanced encryption standard algorithm in cryptography, the encryption algorithm adopts the symmetric packet password system, the key length is supported by at least 128, 192, 256, the group length is 128 bits, and the algorithm should be easy to implement a variety of hardware and software.

The cryptographic algorithm requirements are reversible so that the decryption algorithm can correctly recover the cleartext. Take AES, in the case of key fixing, clear text and redaction are one-to-one correspondence throughout the input space.

Therefore, the various parts of the algorithm are also reversible, and then the sequence of operation of each part is designed to be reversible, the redaction can be correctly decrypted. DES and AES algorithm comparison summary highlighted in Table 1.

TABLE I. Comparison of DES and AES Algorithms

Type	*Key Length (bits)*	*Security*	*Computing Speed*	*Resource Consumption*
DES	56	Low	Medium	Medium
AES	128 192 256	High	High	Low

C. *Rivest Shamir Adleman (RSA)*

RSA encryption algorithm is the most influential public key cryptography algorithm and is generally regarded as one of the best public key schemes. It was proposed by Ron Rivest, Adi Shamir and Leonard Adleman in 1977. All three were working at the Massachusetts Institute of Technology at the time. RSA is made up of the initial letters of their three surnames. RSA has been recommended by ISO as the first algorithm that can be used simultaneously for both encryption and number signature and is resistant to all known password attacks so far and RSA encryption algorithm used as a public key data encryption standard. RSA uses a very simple numerical fact which is to multiply two large prime numbers, but very difficult to decompose its product type where the product can be exposed as an encryption key.

D. *Digital Signature Algorithm (DSA)*

DSA is based on integer finite domain discrete pairs, an important feature of DSA is that two prime numbers are exposed, so that when using someone else's p and q, even if you do not know the private key, you can confirm whether they are random, or do it. This is not possible with the RSA algorithm. Compared to RSA, DSA is used only for signatures, while RSA can be used for signature and encryption. Table II shows the comparison between RSA and DSA.

TABLE II. Comparison of RSA and DSA Algorithms

Type	*Maturity*	*Security*	*Computing Speed*	*Resource Consumption*
RSA	High	High	Low	High
DSA	High	High	High	Only for Digital Signature

IV. The Analysis of Cryptography

According to the attacker's knowledge of the clear text, redacted and other information, password analysis are four types: ciphertext-only attack, known-plaintext attack, chosen-plaintext attack, and chosen-ciphertext attack.

A. *Ciphertext-only Attack*

The attacker had no supporting information in his hands other than the intercepted message. Secret attacks are one of the most common types of password analysis and the most difficult.

This way can be used to attack both the symmetric password system and the asymmetric password system.

B. Known-Plaintext Attack

The attacker uses some ciphertext and also knows the relationship between the part of plaintext and ciphertext. For example, if you are in a conversation that follows a communication protocol, because the protocol uses a fixed keyword, such as "login" and "password", analysis can determine the redaction corresponding to that keyword. If the transmission is a legal document, unit notice, and other types of documents, because most of the documents have a fixed format and some agreed text, in the interception of more documents, you can infer some text, phrases corresponding to the secret.

C. Chosen-Plaintext Attack

The attacker knows the encryption algorithm and can select the clear text and get the redaction corresponding to the corresponding clear text. This is a more common type of Mima analysis. For example, an attacker who intercepts valuable redaction and acquires an encrypted use device and enters any clear text into the device to obtain the corresponding reduction is based on the attacker's attempt to crack the valuable redaction. Select clear text attacks are often used to crack information content encrypted with a public key password system.

D. Chosen-Ciphertext Attack

The attacker knows the encryption algorithm and can select a redaction and get the corresponding clear text. Using this method of selecting a redaction attack, the attacker's target is usually the key used by the encryption process. Digital signatures based on a public key password system are vulnerable to this type of attack.

V. The Method of Analysis

From the analysis of the password, three methods can be used in the process of password analysis, namely, the method of poor attack, the statistical analysis method, and the mathematical analysis method.

A. Poor Attack Method

The idea of cracking the poor attack method is to try all the possibilities to find out the clear text or key. The poor-lifting attack method can be divided into two categories, the poor-lifting key, and the poor-lifting explicit text. The poor key refers to the attacker, in turn, using various possible decryption keys to intercept the secret text, to try to translate, if a decryption key can produce meaningful clear text, then the corresponding key is the correct decryption key. The poor text means that an attacker encrypts all possible clear text while keeping the encryption key unchanged, and if the result of a piece of clear text encryption is consistent with the intercepted message, the corresponding clear text is the message sent by the sender [10-11].

In order to combat the attack, the modern cryptographic system is often designed by expanding the key space or improving the complexity of encryption and decryption algorithms. When the key space is expanded,

- The method of raising the key in the process of cracking need to try more decryption key,
- Improve the complexity of the encryption, decryption algorithm,
- Will enable the attacker whether the use of a poor key or poor method of the password system to crack,
- Each crack attempt requires a higher computational overhead for a perfect modern cryptographic system, and
- The cost of using poor attack methods to crack is likely to exceed the value of a secret edit.

B. Statistical Analysis Method

Statistical analysis is a method to crack by analyzing the statistical laws of clear and text. Some classical cryptographic system encrypted information, secret letters and letter combination of statistical laws and clear text is exactly the same, such a cryptographic system is easy to be cracked by statistical analysis. The statistical analysis method first needs to obtain the statistical law of the dense text, on the basis of which, the statistical law of the dense text is compared with the known clear statistical law, and the correspondence of the clear and dense text is extracted, and then the secret text is cracked.

In order to combat statistical analysis attacks, the password system should be designed to avoid the consistency of the clear text in the statistical law, so that the attacker cannot analyze the statistical law of the paper to infer the clear text content.

C. Mathematical Analysis Method

Most modern cryptographic systems take mathematical problems as the theoretical basis. Mathematical analysis refers to the method by which an attacker solves an unknown amount such as deciphering a key by mathematically using some known quantities, such as the correspondence of some clear and texts, against the mathematical basis and cryptography characteristics of the cryptography. Mathematical analysis is an important way to crack a password system based on mathematical difficulties.

VI. Database And Cryptography

In this part, this paper researches on the database about cryptography. Some knowledge related to cryptography is studied and the combination of this knowledge and database is used [5, 12-13].

The database is a storehouse of data that is prepared, deposited, and managed according to the data structure, and data is the most central property in information systems. A database, like the human brain, is at the heart of all information systems. Once the brain is damaged, it is bound to affect the body function of the whole person. Similarly, if data is lost, destroyed, or leaked in the database, it is bound to cause incalculable damage to the enterprise. Therefore, the encryption protection of data in the database has become an important part of database security. There are three main dangers to the database:

Loss of availability - Loss of availability means that legitimate users cannot use database objects.

Loss of integrity - Integrity loss occurs when a database accidentally or maliciously performs unacceptable operations.

This can happen when you create, insert, update, or delete data. The outcome is corrupted data which results in incorrect decisions.

Loss of confidentiality – It is due to illegal or unintended exposure to confidential information. Loss of confidentiality can lead to unlawful practices, security coercions, and harm to public confidentiality.

Access control – It is a security mechanism of the database management system (DBMS) to prevent unapproved access. After the login process is cleared only through a valid password-protected user account, the user can access the database.

So we will need to take steps to control the risks. The first step is traffic control. It is a distributed system containing a large amount of traffic from one site to another, as well as within a site. The second step is traffic control prevents data from being transmitted as unapproved agents.

The flow policy outlines the channels where information can flow and describes the security classes for data and transactions.

The last step is data encryption refers to encoded sensitive data transmitted on a public channel. Since data is in an incomprehensible format an unauthorized agent can't understand even if he gets access to the data.

How is the database is encrypted? In most cases, the database will use transparent data encryption. For example, the Oracle database used a lot for the encryption as shown in Fig. 4. Database transparent encryption refers to the encryption and decryption of data in the library, the access to the database program is completely unaware. In particular, the application system, which does not need to make any modifications and compilation, can be directly applied to the encryption library.

Two types of data types are unstructured data and structured data. Examples of unstructured are documents and pictures, and the example of structured data is data in a database. Data in both types are important and require encryption protection. Structured data is usually hosted centrally and contains valuable business-sensitive information, so it is especially important to protect it with encryption.

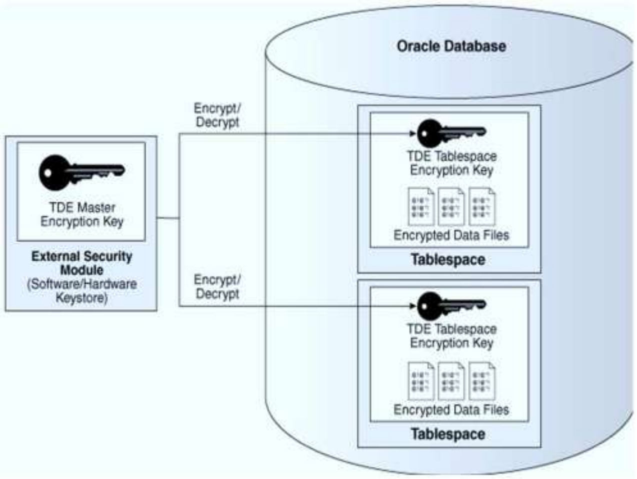

Oracle Database
Encrypt/ Decrypt
TDE Tablespace Encryption Key
Encrypted Data Files
Tablespace
TDE Master Encryption Key
External Security Module (Software/Hardware Keystore)
Encrypt/ Decrypt
TDE Tablespace Encryption Key
Encrypted Data Files
Tablespace

Fig. 4. Cryptography in Oracle Database [14]

Even though the process of decrypting will damage the efficiency of the use of the database the encryption of the database is still necessary to protect the necessary measures to avoid the harsh reality of frequent leakage of sensitive data. Encryption of the Database will increase security significantly. Encryption prevents data from data leakage and malicious destruction and stores data in a confidential manner and presents direct access to the data.

The majority of the encryption solutions application code requires calls to cry forward functions. This is costly because it often requires a deeper knowledge of the application and expertise to program and maintain software. In general, most enterprises spend minimal time on this or don’t have relevant expertise to modify existing applications and invoke encryption routines. Oracle transparent data encryption uses nested encryption capabilities deep into the Oracle database to solve the encryption problem.

Application logic executed through SQL does not need to be changed and still works. It can be further explained that the application doesn’t need to worry about encryption and decryption. The programmer will write the usual code to insert/update or delete the date in the application table. Oracle database will encrypt data as it writes to the disk and similarly decrypts reading data from disk to maintain the functionality of the application. This is important because current applications typically expect unencrypted application data. Displaying encrypted data can at least confuse application users and even break existing applications [15-16].

VII. Conclusion

Nowadays 5G is taking over its predecessor, the cloud era is coming, research suggests that the future of IT architecture, more is cloud-based design. If more data is stored on the cloud, then if cloud vendors steal and analyze user data, it will seriously violate our privacy. Malicious DBAs and developers, as also mentioned in the threat model, tend to have higher database permissions, even if they don't have permissions, and if they have a way to read the cache, the data will also leak. So how do we protect our data?

Therefore, the current performance of homomorphic encryption cannot meet the normal needs, if commercial to database level, this survey suggests that the need for further study by cryptographers. Encryption is only a small part of database security, more content needs to be the joint efforts of everyone, but also hope to see more people engaged in the field of database security in the future.

References

[1] K. Thakur, M. Qiu, K. Gai, and M. L. Ali, 2015, November. An investigation on cyber security threats and security models. In 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing (pp. 307-311). IEEE.

[2] C. Paar, and J. Pelzl, , 2009. Understanding cryptography: a textbook for students and practitioners. Springer Science & Business Media.

[3] G. C. Kessler. "An overview of cryptography." *the Handbook on Local Area Networks, Auerbach* (1998).

[4] V. Gorbach, M. L. Ali, and K. Thakur, 2020, September. A Review of Data Privacy Techniques for Wireless Body Area Networks in

Telemedicine. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.

[5] U. Maurer. "The role of cryptography in database security." In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pp. 5-10. 2004.

[6] M. A. Saleh, N.M. Tahir, E. Hisham, and H. Hashim, 2015, April. An analysis and comparison for popular video encryption algorithms. In 2015 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 90-94). IEEE.

[7] Kadhem, Hasan, Toshiyuki Amagasa, and Hiroyuki Kitagawa. "A novel framework for database security based on mixed cryptography." In *2009 Fourth International Conference on Internet and Web Applications and Services*, pp. 163-170. IEEE, 2009.

[8] L. Thakur, S, Kopecky, M. Nuseir., M. L. Ali, and M. Qiu, 2016, June. An analysis of information security event managers. In 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud) (pp. 210-215). IEEE..

[9] Ciampa, Mark. "Guide to Network Security." *URL https://books. google. com/books/about/Security+ _Guide_to_ Network_Security_Fund. html*.

[10] K. Thakur, M. L. Ali, N. Jiang, and M. Qiu. "Impact of cyber-attacks on critical infrastructure." In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 183-186. IEEE, 2016.

[11] K. Thakur, M. L. Ali, K. Gai, and M. Qiu. "Information security policy for e-commerce in Saudi Arabia." In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 187-190. IEEE, 2016.

[12] D. Deshmukh., A. Pasha, and D. Qureshi, 2013. Transparent Data Encryption--Solution for Security of Database Contents. arXiv preprint arXiv:1303.0418.

[13] K. Thakur, M. L. Ali, S. Kopecky, A. Kamruzzaman, and L Tao. "Connectivity, Traffic Flow and Applied Statistics in Cyber Security." In *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 295-300. IEEE, 2016.

[14] D. Adams, 2014. Oracle Database Online Documentation 12c Release 1 (12.1). Application Development.. Available at https://docs.oracle.com/database/121/ASOAG/introduction-to-transparent-data-encryption.htm#ASOAG10117. Last accessed-Feb 4, 2021

[15] G. J. Simmons., 1979. Symmetric and asymmetric encryption. ACM Computing Surveys (CSUR), 11(4), pp.305-330.

[16] L. Li, K. Thakur, and M. L. Ali, 2020, September. Potential Development on Cyberattack and Prospect Analysis for Cybersecurity. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.

Approach for the Wound Area Measurement with Mobile Devices

Filipe Ferreira
R&D Unit in Digital Services, Applications and Content
Polytechnic Institute of Castelo Branco
Castelo Branco, Portugal
filipemiguel2801@gmail.com

Ivan Miguel Pires
Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
Computer Science Department
Polytechnic Institute of Viseu
Viseu, Portugal
UICISA:E Research Centre
School of Health
Polytechnic Institute of Viseu
Viseu, Portugal
impires@it.ubi.pt

Vasco Ponciano
R&D Unit in Digital Services, Applications and Content
Polytechnic Institute of Castelo Branco
Castelo Branco, Portugal
Altranportugal
Lisbon, Portugal
vasco.ponciano@ipcbcampus.p

Mónica Costa
&D Unit in Digital Services, Applications and Content
Polytechnic Institute of Castelo Branco
Castelo Branco, Portugal
monicac@ipcb.pt

Nuno M. Garcia
Instituto de Telecomunicações
Universidade da Beira Interior
Covilhã, Portugal
ngarcia@di.ubi.pt

Abstract — **There are various causes related to the appearance of wounds in people that are cause several health problems. Earlier detection and easy monitoring is essential for the correct treatment of this problem. The pandemic time that we live in is causing the people not to go to the hospital to treat the various wounds. This paper performs the revision of the methods for measuring the wound size, but the wounds' color can detect other parameters. This study also proposed a technique for further implementing the detection of wounds with the camera available in off-the-shelf mobile devices. As future work, the method will be implemented to correct wound size in an Android application.**

Keywords **– Wound area; Mobile application; segmentation; image processing techniques.**

I. INTRODUCTION

The measurement of wound area is essential to reduce the appearance of other health problems in the future [1], [2], such as infected wounds and chronic wounds. The constant monitoring of wound areas is specialty important in chronic wounds [3], [4]. It is an injury to the body related to violence, accident, or surgery, typically involving laceration or breaking of a membrane, usually damaging to underlying tissues [5].

Mobile devices are equipments that embeds several sensors, including imaging sensors that are commonly called digital cameras [6], [7]. Currently, the mobile devices' cameras have high quality, allowing the reliable capture of images for further processing to identify the wound size [8], [9]. Other sensors are available in the mobile device to enable a more accurate method development for this type of measurement, determining the distance between the device and the surface to calculate the wound's correct size [10], [11].

This paper intends to propose an architecture of a method for the local measurement of the wound area with a mobile device with an Android operating system. However, this paper starts with reviewing the literature's implemented methods to find the best way to be implemented.

As results of this paper, it is verified that the most implemented techniques are the capture of the image, the conversion of image to grayscale, the performance of threshold to enhance the contrast of the image with Otsu's thresholding algorithm, the implementation of segmentation with threshold, the finding of contours of the wound, and the measurement of the wound area as the number of pixels.

This paragraph finalizes the introduction of the paper. Next, Section II presents the related work found in the literature. The methodology used to identify the best methods was introduced in Section III. Section IV presents the proposed method and the results of the analysis. We are finalizing this paper with the discussion and conclusions in Section V.

II. RELATED WORK

Various studies use mobile devices to measure the wound area with a camera embedded on mobile devices. The authors of [12] performing multiple stages for the measurement of the wound area, including the selection of saturation plan, the inversion and filtering of saturation values, the performance of segmentation with threshold.

In [13], the authors performed the detection of a wound. Initially, the authors inserted calibration marks to identify the localization of the region of the injury. After that, the authors used a method to convert the image to grayscale, and it is transformed from RGB to YCbCr color space. Next, the Principal Component Analysis (PCA) was performed on the Cb and Cr color channels. Thus, it is possible to find the fly's subspace with the maximal contrast between the wheel and the surrounding erythema. The authors performed the dimensionality reduction by finding an orthogonal projection of the high-dimensional data into a lower-dimensional subspace. The implementation finished with the median filtering, the threshold to enhance the contrast of the image with Otsu's thresholding algorithm, the erosion operation on the binarized image, and the suppression of the structures connected to the image border dilation operation.

The authors of [14] converted the image to HSV color space (hue, saturation, value), and they performed the segmentation. Next, they extracted the saturation space from

978-1-6654-4067-7/21/$31.00©2021

color space and implemented the threshold to enhance the image's contrast with Otsu's thresholding algorithm and the dilation operation. After that, the implemented method finds the contours from the binary image using the Suzuki85 algorithm. The number of black pixels inside the segmented image was calculated to design a plot healing curve.

In [15], the authors started with the color correction and color calibration on original 2D images to measure the wound area. Next, 3D reconstruction and segmentation were performed to compute the wound perimeter. Finally, the authors design the contour in a 2D plane, calculated the area as the number of pixels between the outlines, and translated the measured pixels to real-world measurement units.

The authors of [16] calculated the approximate size of a wound, inserting and adjusting a rectangle box in the image.

In [17], the images of wounds are segmented in the ulcer region with image decomposition using the parametric equations that define the toroidal geometry. After that, the resulted image is decomposed in different contrast levels. Next, the threshold technique was performed to enhance the contrast of the image with Otsu’s thresholding algorithm, and the contours of the wound were detected. In continuation, the resulted image was transformed from RGB to grayscale, creating an appearance model using linear combinations of discrete Gaussians (LCDG) and minimizing the noise with a Generalized Gauss-Markov Random Field (GGMRF) image model for the calculation of the area of the wounds.

The authors of [18] applied different techniques, including white balance, anti-glare, Enhance Local Contrast (CLAHE) algorithm, the level set algorithm to find the boundary of the wound, and the snakes model algorithm to define an energy function of the image to detect the size of an injury.

In [19], the authors performed several actions, including the detection of small bright rectangular regions in each image using thresholding and convolution, the combination into one patch of the overlapped areas, the detection of candidate patch sequences based on intensity change, and the segmentation to analyze the wound images.

The authors of [20] started with the extraction of RGB channels from images, extracting the blue channel. Next, the histogram equalization was performed. Before the application of segmentation techniques, the picture was converted from RGB to grayscale. Next, the average low pass filter is applied, and the threshold is calculated. In continuation, the degree intensity image is converted to a binary image, and Otsu’s thresholding algorithm is used to highlight the contrast of the picture. After that, the Asymmetry, Border irregularity, Color, and Diameter were extracted, and the image was denoised for the extraction of Region of Interest (ROI) of the wound. Finally, the Support Vector Machine (SVM) was implemented for the classification of the wounds.

In [21], the authors extracted an image from videos, and they computed the absolute scale of the wound with optics equations. Finally, the methods traced the contour with wound boundary detection for the calculation of the wound area.

The authors of [22] started applying the Mask-Recurrent Convolutional Neural Network (RCNN) model to segment the wound. In continuation, the 3D mesh is rasterized, generating a top view image and the matrix of face indices. The Expectation-Maximization (EM) approach was implemented, and a projective transform matrix of the image was calculated. Next, the image is segmented, and the RANSAC algorithm was used to calculate the best fitting hyperplane. Finally, the authors calculated the faces and the wound's boundary, computing its depth, area, volume, and axes.

In [23], the implemented method starts with applying a mask to select the relevant part. Next, matching common points on the input images was found, estimating each image's camera positions. In continuation, the camera calibration parameters were refined, building the point cloud model, a polygonal mesh, and the texture. Finally, the 3D reconstruction is performed for the measurement of the wound area.

The authors of [24] cropped the center of the wound and removed the unnecessary artifacts. Next, the image was resized and re-sampled to extract the saturation plane of the HSV color model. In continuation, the authors calculated the contour with the contrast between infected and normal skin, and the authors smoothed the image with a Gaussian filter. After that, the authors applied the snakes model algorithm to define an image's energy function, and they transformed it into a grayscale image. The grayscale image was segmented, building the contour of the wound for further measurement of wound size.

In [25], the authors performed segmentation, feature extraction and segmentation, sequentially.

The authors of [26] started with the scaling of the image and photogrammetric 3D reconstruction. Next, the position and orientation of each image in the 3D space were estimated. With these features, the authors performed the feature matching process. In continuation, the authors evaluated the surface of the object by a dense point cloud. After applying histogram equalization, the Otsu’s thresholding algorithm was used to enhance the image's contrast. The induration was cropped for the identification of the margin of the rough. Finally, the elliptical approximation of induration margins was calculated, and the wound area was measured.

In [27], the authors started with the segmentation of RGB images into foreground and background. After that, the Minimum Bounding Rectangle (MBR) of the region of interest is cropped. Next, the image is binarized and converted to grayscale. After that, the ISODATA algorithm was used to find a threshold for the image, and the Line Segment Detector (LSD) approach was implemented to find the ticks of the measurement tool. After grouping the ticks by angle, only the segments with more than 5 degrees of difference to the group with more elements. Finally, the distance between ticks in pixels serves to measure the wound area converted to centimeters.

The authors of [28] started with the normalization of the images. After that, they applied the SEED algorithm for superpixel segmentation to identify the superpixels that are skin. The skin area was reconstructed with superpixels, and the wound area was detected.

In [29], the OpenCV library was used to apply a coin detection algorithm. After that, the image is segmented, and a rectangle was created around the wound. Next, the grabCut algorithm was used to measure the wound area. Still, the intensity values of colors with OpenCV histogram were calculated. Finally, the coin detection algorithm was implemented to improve the wound area measurement.

The authors of [30] performed the 3D reconstruction of the body's wound part, and they mapped the 3D model to the 2D plane. After applying the image segmentation techniques, the scale conversion algorithm was implemented to measure the wound area.

III. METHODS

Various methodologies can be implemented for the identification of the wound. This paper proposes a method to be implemented in an Android application to identify the wound area.

The Android devices have different types of cameras with other qualities, and they can be used for the accurate recognition of the wound area.

This paper intends to identify the best method for the identification of the wound area. It is essential to identify the stages that are the most implemented in the various studies.

Based on the studies available about this subject in IEEE Xplore, ScienceDirect, Google Scholar, and PubMed Central by the following keywords "wound", "measurement", "size", "image processing", and "mobile device", we selected the studies with the best fit of the purpose for the identification of the wound area. After that, the stages of the various methods were presented in section II.

As only 19 studies were selected for the research and the statistical validity, we chose the methods implemented in more than 10% of the studies selected from the literature.

The various stages implemented in the different studies were statistically analyzed to identify the wound area measurement's best techniques.

IV. EXPECTED RESULTS

Table I presents the results related to the available literature methods in more than two studies (10%), showing that only eight techniques are included.

TABLE I. RANKING OF THE METHODS IMPLEMENTED FOR THE WOUND AREA MEASUREMENT

Method Implemented	Number of Occurrences
Perform segmentation with threshold	14
Measure the wound area as the number of pixels	12
Convert image to grayscale	5
Perform threshold to enhance the contrast of the image with Otsu's thresholding algorithm	5
Find contours of the wound	4
Perform 3D reconstruction	4
Perform the histogram equalization	3
Crop the center of the wound	3

In general, the majority of the studies implemented the segmentation of the images with threshold. It is always related to the identification of the wound area as the number of pixels. However, the implementation of these methods was revealed more efficiently with images in grayscale. Also, Otsu's thresholding algorithm is the most implemented method for the threshold. Only four studies identified the contours, but the ways that identified the shapes represent the methods with a smaller number of strategies implemented.

In general, we can say that identifying the wound area can be performed with the sequence of methods presented in Figure 1. The accuracy of the proposed technique will be analyzed and reported in the future.

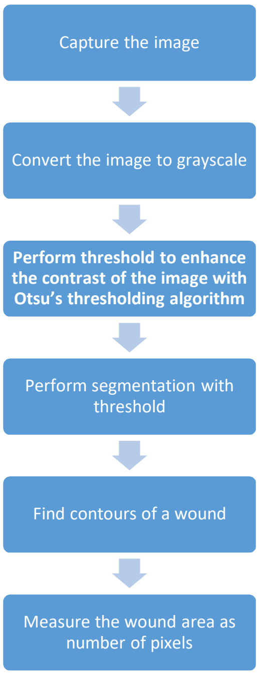

Capture the image
Convert the image to grayscale
Perform threshold to enhance the contrast of the image with Otsu's thresholding algorithm
Perform segmentation with threshold
Find contours of a wound
Measure the wound area as number of pixels

Fig. 1. Proposed sequence of methods for wound area measurement.

V. DISCUSSION AND CONCLUSIONS

This paper analyzed the sequence of methods implemented for the wound to measure images captured from mobile devices. A smaller number of studies for the wound area measurement with mobile devices can be improved with the various experiments that will be performed.

In general, it is expected that the best results will be achieved with the implementation of the following sequence of methods: capture the image, convert image to grayscale, perform threshold to enhance the contrast of the image with Otsu's thresholding algorithm, perform segmentation with threshold, find contours of the wound, and measure the wound area as the number of pixels.

As final work, it will be necessary to implement a method for converting pixels and centimeters considering the distance between the camera and the surface captured. It can be performed with the proximity sensors available in the mobile device or other methods that will be studied.

ACKNOWLEDGMENT

This work is funded by FCT/MEC through national funds and co-funded by FEDER – PT2020 partnership agreement under the project **UIDB/50008/2020** (*Este trabalho é financiado pela FCT/MEC através de fundos nacionais e cofinanciado pelo FEDER, no âmbito do Acordo de Parceria PT2020 no âmbito do projeto UIDB/50008/2020*). This work is also funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project **UIDB/00742/2020**. This article is based upon work from COST Action IC1303–AAPELE–Architectures, Algorithms and Protocols for Enhanced Living Environments and COST Action CA16226–SHELD-ON–Indoor living space improvement: Smart Habitat for the Elderly, supported by COST (European Cooperation in Science and

Technology). More information in www.cost.eu. Furthermore, we would like to thank the Politécnico de Viseu for their support.

REFERENCES

[1] G. S. Schultz *et al.*, "Wound bed preparation: a systematic approach to wound management," *Wound Repair Regen*, vol. 11, no. s1, pp. S1–S28, Mar. 2003, doi: 10.1046/j.1524-475X.11.s2.1.x.

[2] S. Monstrey, H. Hoeksema, J. Verbelen, A. Pirayesh, and P. Blondeel, "Assessment of burn depth and burn wound healing potential," *Burns*, vol. 34, no. 6, pp. 761–769, Sep. 2008, doi: 10.1016/j.burns.2008.01.009.

[3] M. E. Porter and E. O. Teisberg, "How physicians can change the future of health care," *Jama*, vol. 297, no. 10, pp. 1103–1111, 2007.

[4] M. A. Fonder, G. S. Lazarus, D. A. Cowan, B. Aronson-Cook, A. R. Kohli, and A. J. Mamelak, "Treating the chronic wound: A practical approach to the care of nonhealing wounds and wound care dressings," *Journal of the American Academy of Dermatology*, vol. 58, no. 2, pp. 185–206, 2008.

[5] P. K. Stefanopoulos, D. E. Pinialidis, G. F. Hadjigeorgiou, and K. N. Filippakis, "Wound ballistics 101: the mechanisms of soft tissue wounding by bullets," *European journal of trauma and emergency surgery*, vol. 43, no. 5, pp. 579–586, 2017.

[6] J. A. Burke *et al.*, "Participatory sensing," 2006.

[7] C. W. Chen, "Internet of Video Things: Next-Generation IoT With Visual Sensors," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6676–6685, 2020.

[8] D. N. Breslauer, R. N. Maamari, N. A. Switz, W. A. Lam, and D. A. Fletcher, "Mobile phone based clinical microscopy for global health applications," *PloS one*, vol. 4, no. 7, p. e6320, 2009.

[9] L. Wang, P. C. Pedersen, D. M. Strong, B. Tulu, E. Agu, and R. Ignotz, "Smartphone-based wound assessment system for patients with diabetes," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 477–488, 2014.

[10] V. Williamson, "Mobile Wound Assessment and 3d Modeling from a Single Image," The University of Wisconsin-Milwaukee, 2020.

[11] I. M. Pires and N. M. Garcia, "Wound Area Assessment using Mobile Application.," in *BIODEVICES*, 2015, pp. 271–282.

[12] M. Casal-Guisande, A. Comesaña-Campos, J. Cerqueiro-Pequeño, and J.-B. Bouza-Rodríguez, "Design and Development of a Methodology Based on Expert Systems, Applied to the Treatment of Pressure Ulcers," *Diagnostics*, vol. 10, no. 9, p. 614, Aug. 2020, doi: 10.3390/diagnostics10090614.

[13] O. Bulan, "Improved wheal detection from skin prick test images," San Francisco, California, USA, Mar. 2014, p. 90240J, doi: 10.1117/12.2038442.

[14] A. Gupta, "Real time wound segmentation/management using image processing on handheld devices," *JCM*, vol. 17, no. 2, pp. 321–329, Apr. 2017, doi: 10.3233/JCM-170706.

[15] E. Sirazitdinova and T. M. Deserno, "System design for 3D wound imaging using low-cost mobile devices," Orlando, Florida, United States, Mar. 2017, p. 1013810, doi: 10.1117/12.2254389.

[16] M. Tang, K. Gary, O. Guler, and P. Cheng, "A Lightweight App Distribution Strategy to Generate Interest in Complex Commercial Apps: Case Study of an Automated Wound Measurement System," presented at the Hawaii International Conference on System Sciences, 2017, doi: 10.24251/HICSS.2017.418.

[17] B. Garcia-Zapirain, A. Shalaby, A. El-Baz, and A. Elmaghraby, "Automated framework for accurate segmentation of pressure ulcer images," *Computers in Biology and Medicine*, vol. 90, pp. 137–145, Nov. 2017, doi: 10.1016/j.compbiomed.2017.09.015.

[18] C.-H. Huang, S.-D. Jhan, C.-H. Lin, and W.-M. Liu, "Automatic Size Measurement and Boundary Tracing of Wound on a Mobile Device," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, Taichung, May 2018, pp. 1–2, doi: 10.1109/ICCE-China.2018.8448729.

[19] T. Kanade *et al.*, "Cell image analysis: Algorithms, system and applications," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, Kona, HI, USA, Jan. 2011, pp. 374–381, doi: 10.1109/WACV.2011.5711528.

[20] S. T Y, S. D, and G. P. M N, "Early Detection of Melanoma using Color and Shape Geometry Feature," *JBEMi*, vol. 2, no. 4, Aug. 2015, doi: 10.14738/jbemi.24.1315.

[21] A. Yee, M. Patel, E. Wu, S. Yi, G. Marti, and J. Harmon, "iDr: An Intelligent Digital Ruler App for Remote Wound Assessment," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Washington, DC, USA, Jun. 2016, pp. 380–381, doi: 10.1109/CHASE.2016.78.

[22] S. Zahia, B. Garcia-Zapirain, and A. Elmaghraby, "Integrating 3D Model Representation for an Accurate Non-Invasive Assessment of Pressure Injuries with Deep Learning," *Sensors*, vol. 20, no. 10, p. 2933, May 2020, doi: 10.3390/s20102933.

[23] R. Dendere, T. Mutsvangwa, R. Goliath, M. X. Rangaka, I. Abubakar, and T. S. Douglas, "Measurement of Skin Induration Size Using Smartphone Images and Photogrammetric Reconstruction: Pilot Study," *JMIR Biomed Eng*, vol. 2, no. 1, p. e3, Dec. 2017, doi: 10.2196/biomedeng.8333.

[24] N. D. J. Hettiarachchi, R. B. H. Mahindaratne, G. D. C. Mendis, H. T. Nanayakkara, and N. D. Nanayakkara, "Mobile based wound measurement," in *2013 IEEE Point-of-Care Healthcare Technologies (PHT)*, Bangalore, India, Jan. 2013, pp. 298–301, doi: 10.1109/PHT.2013.6461344.

[25] N.-M. Cheung, V. Pomponiu, D. Toan, and H. Nejati, "Mobile image analysis for medical applications," *SPIE Newsroom*, Jul. 2015, doi: 10.1117/2.1201506.005997.

[26] S. Naraghi, T. Mutsvangwa, R. Goliath, M. X. Rangaka, and T. S. Douglas, "Mobile phone-based evaluation of latent tuberculosis infection: Proof of concept for an integrated image capture and analysis system," *Computers in Biology and Medicine*, vol. 98, pp. 76–84, Jul. 2018, doi: 10.1016/j.compbiomed.2018.05.009.

[27] M. T. Cazzolato *et al.*, "Semi-Automatic Ulcer Segmentation and Wound Area Measurement Supporting Telemedicine," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA, Jul. 2020, pp. 356–361, doi: 10.1109/CBMS49503.2020.00073.

[28] Y.-W. Chen, J.-T. Hsu, C.-C. Hung, J.-M. Wu, F. Lai, and S.-Y. Kuo, "Surgical Wounds Assessment System for Self-Care," *IEEE Trans. Syst. Man Cybern, Syst.*, pp. 1–16, 2019, doi: 10.1109/TSMC.2018.2856405.

[29] W. Wu, kenneth Y. W. Yong, M. A. J. Federico, and S. K.-E. Gan, "The APD Skin Monitoring App for wound monitoring: Image processing, area plot, and colour histogram," *spamd*, May 2019, doi: 10.30943/2019/28052019.

[30] C. Liu, X. Fan, Z. Guo, Z. Mo, E. I.-C. Chang, and Y. Xu, "Wound area measurement with 3D transformation and smartphone images," *BMC Bioinformatics*, vol. 20, no. 1, p. 724, Dec. 2019, doi: 10.1186/s12859-019-3308-1.

On the Security of Cyber-Physical Systems Against Stochastic Cyber-Attacks Models

Qasem Abu Al-Haija
Department of Data Science and Artificial Intelligence, Faculty of Information Technology
University of Petra (UoP), Amman 11196, Jordan)
qasem.abualhaija@uop.edu.jo

***Abstract*—Cyber Physical Systems (CPS) are widely deployed and employed in many recent real applications such as automobiles with sensing technology for crashes to protect passengers, automated homes with various smart appliances and control units, and medical instruments with sensing capability of glucose levels in blood to keep track of normal body function. In spite of their significance, CPS infrastructures are vulnerable to cyberattacks due to the limitations in the computing, processing, memory, power, and transmission capabilities for their endpoint/edge appliances. In this paper, we consider a short systematic investigation for the models and techniques of cyberattacks and threats rate against Cyber Physical Systems with multiple subsystems and redundant elements such as, network of computing devices or storage modules. The cyberattacks are assumed to be externally launched against the Cyber Physical System during a prescribed operational time unit following stochastic distribution models such as Poisson probability distribution, negative-binomial probability distribution and other that have been extensively employed in the literature and proved their efficiency in modeling system attacks and threats.**

***Index Terms*—Cyber Physical Systems, System Security, Stochastic Process, Cyberattacks, Poisson distribution, Negative Binomial(NB) distribution.**

I. Introduction

Cyberattacks are meant to deliberate an exploitation of system resources and/or components in order to compromise necessary data and lead to utilize the system in illegal usage (i.e., cyber-crimes). Examples of illegal uses of the systems includes information and identity theft, buffering overflow, damaging system's utilities and others. Usually, the cyberattack follows a vulnerability to realize a threat which is an event that results in a security violation. Practically, cyberattacks are launched through a series of actions to compromise the security services (such as availability, integrity, and confidentiality) of the cyber physical systems such as military system, banking systems, telecommunication systems, transportation systems and smart power grids. Therefore, the cyberattacks have become potential threat that affect the cyber physical systems (CPSs) of our society with wide range of consequences [1]. To support this argument, Figure 1 shows the total cost of cyber-crimes in seven countries for three consecutive years 2013-2015 according to the cost of cyber-crime study published by Ponemon institute [2], where the cost expressed in millions of dollars (US$) and comprised around 252 different companies.

Due to the rapid advances of cyberattacks and threats techniques such as, advanced persistent threats, file-less malwares, zero-days exploits [3], cyberattack data analytic has emerged as an important area of cybersecurity research field. Thus, utilizing the cyberattack data needs to characterize the statistical attributes exhibited by the data through an analytical modeling mechanisms that enable us to investigate the cybersecurity implications. Indeed, several techniques were investigated to model the incoming cyberattacks rate against cyber-systems where most of them have been developed using probabilistic (i.e., stochastic) approaches such as Model based Markov chains [4], [5], Model based Hawkes Process [6], Model based Poisson [7] and Model based negative binomial distributions models [8] as well as other probabilistic processes.

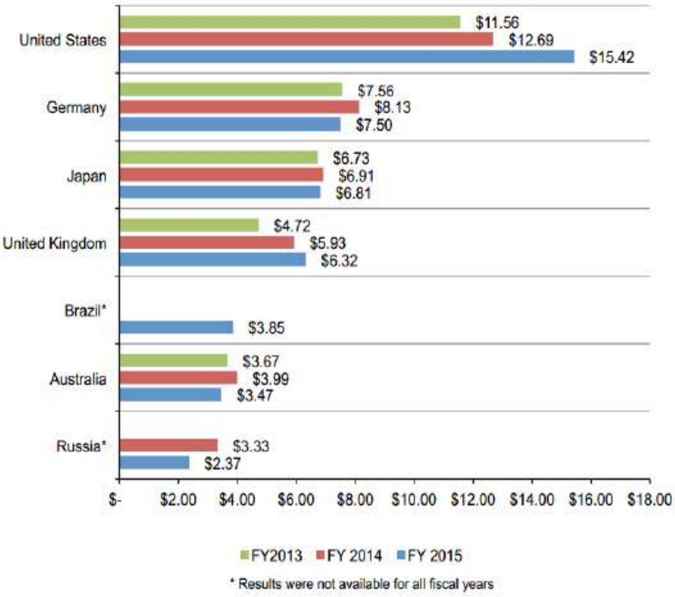


Fig. 1. Total cost of cyber-crimes in seven countries

Indeed, cyber-crimes and cyber-attacks have become a fast-growing area of crime, since criminals can destructively exploit the speed and anonymity of the Internet world to commit a wide range of criminal activities that cause serious harm and real threats to victims worldwide. Recently, the prediction study developed in [9] showed that cyber-crimes are going to increase in a linear trend and thus more security efforts should be emphasized in face of such crimes and

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

cyber threats.
In this paper, we propose a short review of models and techniques for *cyber-attacks and threats rate* against any cyber-physical systems such as, network of computing or storage modules. The cyberattacks are assumed to be externally launched against the Cyber Physical System during a prescribed operational time unit following stochastic distribution models such as Poisson probability distribution, negative-binomial probability distribution and other that have been extensively employed in the literature and proved their efficiency in modeling system attacks and threats [8].

The rest of this paper is organized as follows. Section II provides a brief mathematical background for probabilistic system models. Section III describes the state-of-art stochastic techniques to model the cyber-attacks in the used in the area of attacks-modeling. Finally, Section VI concludes the paper.

II. Related Research

CPS communication networks are usually used to communicate essential information of stringent surroundings and services in real time system connectivity of either synchronous or asynchronous modes [10]. Nevertheless, CPS infrastructures are vulnerable to cyberattacks and instructions occasioned by the restrictions of their computation, processing and communication abilities [11]. Typically, Cyber physical system's intrusion detection system (CPS-IDS) are established to distinguish various external cyberattacks and intrusions formulas lunched against the system components or applications with random rates of attacks.
In the field of cyber physical system design [12], the redundancy mechanisms to develop solutions for fault tolerant structures have been studied extensively. For instance, Triple-Modular-Redundancy (TMR) [13] technique was proposed earlier to generate true outputs on condition that no less than two out of three segments (modules) generate true outputs. This in turn is taking place by employing three replicas of a every system module and using a voter to process the results of the system outputs. For more about the fault-tolerant system techniques and structures, the reader is directed to navigate through [14] which provides a respectable chronological review for such systems.
While considering the design of secure cyber physical systems or cyberattack-tolerant CPSs, only few researches have been deliberated on applying redundancy mechanisms to prolong the system survivability against the different cyberattacks. Recalling the Transparent Runtime Randomization (TRR) proposed in [15], where multiple randomized redundant memory places are allocated for executing program's and software packages' trying to prevent attackers of being able to locate and exploit the decisive address locations of a susceptible software and programs. Another noticeable work on system security have been investigated in [16] to evaluating the survivability specifications of a cyber-system using a model-driven techniques. Moreover, several probabilistic methodologies to model the cybersecurity requirements and factors for several CPSs has been deemed and investigated such as the work presented in [17] employing the Markov Chain-based scheme [18].

Recently, machine learning-based techniques have been widely employed to develop cybersecurity solutions to defend and protect the communication networks of IoT and CPS against several cyber-attacks and intrusions. Examples of cyberattacks detection based machine learning techniques employed to address CPS security concern: Shallow Neural Network (SNN) [19], Convolutional Neural Network (CNN) [20], K-Nearest Neighbors (KNN) [21], Decision Tree Method (DTM) [22], Support Vector Machine (SVM) [23], and others. For instance, Q. Abu Al-Haija et. al. [24] presented a novel deep-learning-based detection and classification system for cyber-attacks in IoT communication networks employing convolutional neural networks. They evaluated their model, using NSL-KDD dataset scoring accuracy results of almost 99.3% and 98.2% for the binary-class classifier (two categories) and the multiclass classifier (five categories), respectively. Besides, they validated their system using a 5-fold cross-validation, confusion matrix parameters, precision, recall, F1-score, and false alarm rate. As a result, they showed that their system outperformed many recent machine-learning-based IDCS systems in the same area of study.
In this paper, we propose a short review of models and techniques for *cyber-attacks and threats rate* against any cyber-system of interest (such as, network of computing or storage modules) that are launched during the time unit of interest (such as, hours, days, moths) using *probabilistic distribution* modeling such as *Poisson probability distribution and negative-binomial probability distribution* that proved their efficiency in modeling system attacks and threats [8].

III. System Security Analysis

Cyber-Physical Systems (CPSs) are composed of integrated set of interacting subsystems, such as computation, processing, networking, physical processes, and other, to achieve a collective specified task and accomplish a common goal/application. The interconnection of system elements is often architected in series configuration, parallel configuration, or $k-out-of-n$ configurations. The security of the system can be described as a measurement of probability that the system will perform its objective(s) (i.e. survive) at a certain satisfactory level within a given period of time. The system security is considered one of the major factors to characterize the cyber physical system abilities in achieving the design objectives and deliver the agreed-on services.

A. Security of Series Systems

Series systems are formed by cascading system elements in a sequential interconnection fashion as illustrated in Figure 2. The major characteristic of series systems is that they can function properly only when all their connected elements function properly, i.e., fully inter-dependent systems. This is similar to layered hierarchical systems that communicate

TABLE I
SYSTEM MODELING NOTATIONS

Notation	Description
E_i	The Element i of the system model
P_i	The Probability that element i is attacked
n_i	The number of components used for sub-system i.
k_i	The number of operating elements used for sub-system i.
X	The random variable for the number of attacked elements.
$f(k; n; p)$	The binomial distribution function of three parameters.
S	The number of sub-systems.
$n_i - k_i$	The number of redundant elements in subsystem i.
$p(x)$	Probability of subsystem attack following Poisson process.
λ_{ij}	The attack rate of the j^{th} element of sub-system i.
E_{ij}	The j^{th} element of sub-system i.

information from one level to the next. As a result, the security of a series system is entirely depending on the security of its interconnected elements.

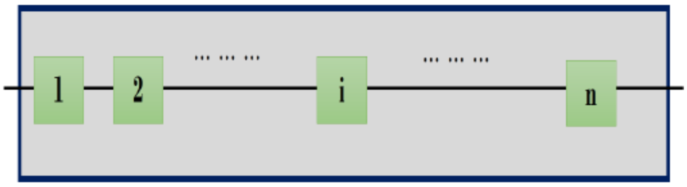


Fig. 2. System Model Using Series Interconnection of several subsystems

To measure the system security of series models, lets consider that we have a system with n elements connected in series (E$_1$, E$_2$, E$_3$, ... E$_n$) with P$_i$ is the probability that element E$_i$ is attacked, then:

The probability that element E$_1$ survive = (1-P$_1$)
The probability that element E$_2$ survive = (1-P$_2$)
The probability that element E$_i$ survive = (1-P$_i$)

Since the elements of the system are attacked independently, then the system security can be computed as:

$$(1 - P_1)(1 - P_2)...(1 - P_{n-1})(1 - P_n)$$

$$SystemSecurity = \prod_{i=1}^{n}(1 - P_i) \tag{1}$$

B. Security of Parallel Systems

Parallel systems are formed by connecting system elements in parallel interconnection fashion as illustrated in Figure 3. The major characteristic of parallel systems is that the system is attacked only if all of its components are compromised, that is, it can still function properly if some of the connected elements function properly. This similar to networked system that communicate information from node to node. Thus, the security of a parallel system depends on the comprising (due to attack) of all system elements.

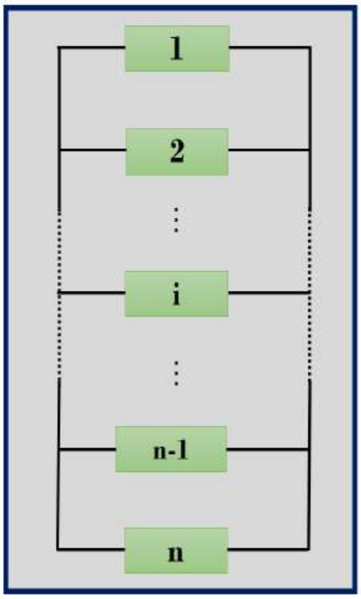


Fig. 3. System Model Using Parallel Interconnection of several subsystems

To measure the system security of parallel models, lets consider that we have a system with n elements connected in parallel (E$_1$, E$_2$, E$_3$, ... E$_n$) with P$_i$ is the probability that element E$_i$ is attacked, then:

The probability that element E$_1$ is attacked = (P$_1$)
The probability that element E$_2$ is attacked = (P$_2$)
The probability that element E$_i$ is attacked = (P$_i$)

Since the elements of the system are attacked independently , then the system security can be computed as the probability of at least one element is surviving, as follows:

$$1 - [(P_1)(P_2)...(P_{n-1})(P_n)]$$

$$SystemSecurity = 1 - \prod_{i=1}^{n}(P_i) \tag{2}$$

C. Security of k-out-of-n Systems

Even though series and parallel models can be used to represent the connectivity structure for many elements of the cyber systems, however, real systems may include a combination of both representations configured in $k - out - of - n$ systems where k is the minimum number of operating elements in the cyber system, i.e. active/operational elements. Thus, such system is compromised if k or more out of n components are compromised due to cyberattacks. In other words, the system is attacked when $(n - k + 1)$ of its elements are attacked/compromised. Indeed this is a generalized model representation that can be used to generate the different schemes/scenarios by setting the values of k and

n as follows:

Generalization - *For any cyber physical system* with k active/operating elements and includes n as total number of elements for both operating and sandy-redundant elements, the system can be developed in one of teh following interconnection architectures:

$$\begin{cases} Series - system & \text{if} \quad k = n \\ Parallel - system & \text{if} \quad k = n \\ k - out - of - n - system & \text{if} \quad k < n \end{cases}$$

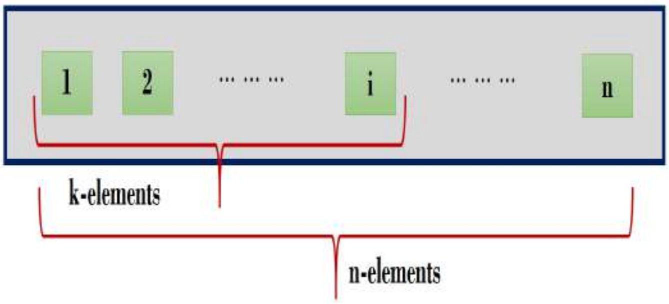


Fig. 4. System Model Using $k - out - of - n$ Interconnection of several subsystems

To measure the system security of $k - out - of - n$ models, lets consider a cyber system with n elements (E$_1$, E$_2$, E$_3$, ... E$_n$) where the elements are attacked independently with probability P$_i$, i.e., the probability that element E$_i$ is attacked/compromised. For simplicity, lets assume that all elements have the same probability of attack, that is: P$_1$ = P$_2$ = ... P$_i$ = p. In this way, the the probability of attack follows the $probabilisticbinomialdistribution$ with parameters k, n and p.
Let X be the random variable representing the number of attacked elements following the binomial distribution with parameters $n \in N$ and $p \in [0, 1]$, then, we write $X \sim B(n, p)$. The probability of getting exactly k successes in n trials is given by the probability mass function:

$$f(k, n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (3)$$

for $k = 0, 1, 2, \ldots, n$, where $\binom{n}{k}$ is the binomial coefficient is defined by the next expression:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Now, for the system to survive, i.e. to be secure, the number of attacked elements should not exceed $k - 1$ of the system elements, and thus the system security can be calculated as follows:

$$SystemSecurity = P[X < k] = \sum_{i=0}^{k-1} f(i, n, p) \qquad (4)$$

D. *Security of Parallel-Series Systems*

Generally, large system are composed of several subsystems that are interconnected to accomplish their outcome based missions with maximum probability of success. The use of series systems in connecting the elements of subsystem is highly undesired since series system is risky and highly vulnerable for cyberattacks sine the compromising of only one element of the system will cause the system to shut down. To increase system security, the system can be designed with several sub-systems connected in series, but have an internal parallel structure for the interconnected elements (i.e. $k = n$), as illustrated by the scheme given in Figure 4.
To measure system security, consider a system composed of (S) sub-systems each with n$_i$ elements connected in parallel where P$_{ij}$ is the probability of attack of the j^{th} element of sub-system (i). Since the elements of system are attacked independently, the security of the system can be calculated as follows:

System Security = Security (subsystem$_1$) x Security (subsystem$_2$) x ... x Security (subsystem$_S$)

System Security = $(1 - \prod_{j=1}^{n_1} P_{1j})$ x $(1 - \prod_{j=1}^{n_2} P_{2j})$ x ... x $(1 - \prod_{j=1}^{n_i} P_{ij})$ x ... x $(1 - \prod_{j=1}^{n_s} P_{Sj})$

$$SystemSecurity = \prod_{i=1}^{S}(1 - \prod_{j=1}^{n_i} P_{ij}) \qquad (5)$$

E. *Security of Complex Systems*

Practically, many real system are configured by connecting several subsystems in series, but having an internal $k - out - of - n$ structure(i.e. $k < n$).
In this case, for the system to be secure, the number of attacked elements in each subsystem (i) should not exceed k$_i$-1 of the subsystem elements, therefore, the cyberattacks rate over the system elements should be modeled a probabilistic distribution model. Indeed, there two common approaches to model the cyberattacks rate over the system element; namely, Binomial probability distribution and Poisson probability distribution.

In the case of Binomial probability distribution: All elements in every subsystem are attacked with equally likely probability (p), then:

System Security = Security (subsystem$_1$) x Security (subsystem$_2$) x ... x Security (subsystem$_S$)

System Security = $(\sum_{j=0}^{k_1-1} f(j, n_1, p_{1j}))$ x $(\sum_{j=1}^{k_2-1} f(j, n_2, p_{2j}))$ x ... x$(\sum_{j=1}^{k_{sj}-1} f(j, n_s, p_s))$

$$SystemSecurity = \prod_{i=1}^{S}\sum_{j=1}^{k_i-1} f(j, n_i, p_{ij}) \qquad (6)$$

In the case of Poisson probability distribution: The elements in every subsystem are attacked with probability (p) according to Poisson process distribution as follows:

$$p(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^{x}}{x!} & \text{if} \quad x = 0, 1, 2, ..., L \\ 0 & \text{if} \quad otherwise \end{cases} \tag{7}$$

System Security = Security (subsystem$_1$) x Security (subsystem$_2$) x ... x Security (subsystem$_S$)

System Security = $(1 - \sum_{j=1}^{k_1-1} p_{1j})$ x $(1 - \sum_{j=1}^{k_1-1} p_{1j})$ x ... x$(1 - \sum_{j=1}^{k_1-1} p_{sj})$

System Security = $\prod_{i=1}^{S}(1 - \sum_{j=1}^{k_i-1} p_{ij})$

$$SystemSecurity = \prod_{i=1}^{S}(1 - \sum_{j=1}^{k_i-1} \frac{e^{-\alpha}\alpha^{j}}{j!}) \tag{8}$$

IV. Stochastic Models of Cyber-attacks

In the area of *attacks modeling* for designing security solutions for cyber-system, *stochastic techniques* [25], [26] have been well studied. An example of using *stochastic* technique to model cybre attacks, where Poisson variables [27] where used to characterize the attack processes in cloud computing networks. Such a system take into consideration the inter-dependencies amongst vulnerabilities/attacks in a given attack path. Similarly, Poisson Process have been widely and employed in [7], [28], [29] to model several types of system attacks and vulnerabilities, for instance, the process of the virus-infection [7] that infect a network of computers for a period of time due to individual infection.

In the same context, work in [8] modeled the cyber-attacks using negative binomial (NB) probability distribution, where, the proposed a generalized linear models (GLMs) using to predict the number of successful cyber-intrusions into computer network. Moreover, several other stochastic approaches were used to model the several attacks on the system such as the log-normal distribution [30] and Pareto distribution [30], Circular Statistics [31], Hawkes distribution [6] and Beta distribution for cyber-security risk analysis [28].

In more specific situations such as the modeling extreme cyber-attack, both works in [32], [33] proposed a model based extreme value theory (EVT) to model and investigate the extreme cyber-attack rates phenomenon where [32] utilized the value-at-risk as a natural measure of intense cyber-attacks. Such a model can describe and predict the extreme cyber-attack rates at a very satisfactory accuracy, whereas [33] showed that EVT can offer long-term predictions compared with some other models such as graybox time series theory (TST) models with higher rates of accuracy. Moreover, [34] proposed vine copula approach to model model the cyber-security risks with high-dimensional dependency. Furthermore, contributors of [35] proposed a new concept of stochastic cyber-attack process using Long-Range Dependence (LRD).

As a case study, they applied the model on the low interaction honeypot data-set and thus they confirmed that LDR model is feasible to predict cyber-attacks in order to provide an early-warning for defenders to adjust their system security. Finally, Table I summarizes the techniques used to model cyber-attacks for different applications using probabilistic approaches.

TABLE II
Stochastic Models for Cyber-Attacks and Threats Rates

Research	Modeling Technique	Attack Type	Application
[5] [7] [26] [28]	Poisson-Process Model	Cyber-threats, Viruses, Attacks, Vulnerabilities	Virtual network, Computer network, Cloud computing, Network Porosity
[8]	Negative-Binomial Model	Virus Infection	Organization's Computer Networks
[4] [27]	Markov-Chains Model	Cyber-Threats (Bot, Low rate attacks-DoS)	Multiport Scan (KISA) & Computer network
[29]	Log-Normal Distribution Model	Invasive Software (viruses, worms, Trojan horses)	large-scale computer network
[29]	Pareto-Distribution Model	Cyber-Intrusions	large-scale computer network
[30]	Circular-Statistics Model	Malware infections	Live Network Monitoring by Spamhaus
[5]	Beta-Distribution Model	Cyber-Risk	Cybersecurity Risk analysis
[6]	Hawkes-Process Model	Cyber Attacks	Information systems
[31] [32]	Extreme-Value Theory (EVT) Model	Extreme-Cyber Attacks	Network telescope & honeypot data-sets
[33]	Vine-Copula-GARCH Model	Mmultivariate Cybersecurity Risks	Network of virtual nodes
[34]	Long-Range Dependence (LDR)	Cyber-attacks	Honeypot data-set

Recently, more appreciated solutions appeared with the emergence of deep-learning techniques, such as the recent intelligent self-reliant cyberattacks detection and classification system for IoT communication using deep convolutional neural network [36]. The authors of this up-to-date work, has no longer assuming any stochastic approach to predict the incoming cyberattacks, instead, they developed a new intelligent system that can detect the slightly mutated cyberattacks using deep convolutional neural network (CNN) leveraging the power of CUDA based Nvidia-Quad GPUs for parallel computation and processing. Indeed, they obtained a very attractive and superior results by recording a very-high accuracy results for the classification of cyberattacks.

V. Conclusions

This paper presented an investigation study for CPS security against external cyberattacks for different redundant architecture scenarios including series, parallel, series-parallel, and k-out-of-n systems. Besides,the paper presented a methodical study for the stochastic approaches employed in the modeling of cyberattacks rates over the cyber physical Systems with multiple subsystems and redundant elements. Hence, this paper provides an important insight for researchers who works on developing security system solutions for cyber physical systems.

References

[1] A. Nourian and S. Madnick, *A Systems Theoretic Approach to the Security Threats in Cyber Physical Systems Applied to Stuxnet*. IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 1, pp. 2-13, 1 Jan.-Feb. 2018. doi: 10.1109/TDSC.2015.2509994.

[2] L. Ponemon, *2015 Cost of Cyber Crime Study: Global*. Ponemon Institute, Measuring Trust in Privacy and Security, Retrieved Online: https://www.ponemon.org/blog.

[3] C.H. Chi, R. Wong, K.Y. Lam, *Guest Editorial: Special Issue on Behavior Data Analytics for Cybersecurity* Journal of Internet Technology, Taiwan Academic Network Publishing, vol. 19, no. 5 , pp. 1543-1544, Sep. 2018.

[4] D.H. Kim,et. al., *Cyber threat trend analysis model using HMM*. Third International Symposium on Information Assurance and Security, IEEE, Manchester, UK, 2007, pp. 177–182

[5] N. Chaturvedi, H. Mohant, *A mathematical model for randomly-occurring low-rate denial of service attack*. International Journal of Computer & Communication Technology, Vol. 2, No.5, pp. 13–17,2011.

[6] J. M. Chen, *Hawkes Process, Risk Modeling & Applications*. Insurance Market Conferences, Property and Casualty, Oct-2015.

[7] A.K. Rauta, et. al., *A Probabilistic Approach Using Poisson Process for Detecting the Existence of Unknown Computer Virus in Real Time*. The International Journal Of Engineering And Science (IJES), Vol.4, No. 6, PP.47-51, 2015.

[8] N.O. Leslie, et. al., *Statistical models for the number of successful cyber intrusions*. Journal of Defense Modeling and Simulation (JDMS), Special Issue on Applications, Methodology, Technology, pp.1–16, 2017.

[9] Q. A. Al-Haija and L. Tawalbeh, *Autoregressive Modeling and Prediction of Annual Worldwide Cybercrimes for Cloud Environments*. 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2019, pp. 47-51, doi: 10.1109/IACS.2019.8809125.

[10] Q. A. Al-Haija, C. D. McCurry and S. Zein-Sabatto, "A Real Time Node Connectivity Algorithm for Synchronous Cyber Physical and IoT Network Systems," 2020 SoutheastCon, Raleigh, NC, USA, 2020, pp. 1-8, doi: 10.1109/SoutheastCon44009.2020.9249730.

[11] Q. A. Al-Haija, I. Marouf, M. M Asad, K. Al Nasr, "Implementing a lightweight Schmidt-Samoa cryptosystem (SSC) for sensory communications", International Journal on Smart Sensing & Intelligent Systems (IJSSIS), Exeley, vol.12, No.1, 2019. doi: 10.21307/ijssis-2019-006.

[12] W. Kuo, V. R. Prasad, F. A. Tillman and C. L. Hwang, "Optimal Reliability Design: Fundamentals and Applications", Cambridge University Press, 2001.

[13] J. V. Neumann, "Probabilistic Logics and the synthesis of reliable organisms from unreliable components", Automata Studies Ann. of Math. Studies, no. 34, pp. 43-98, 1956.

[14] J. H. Lala and R. E. Harper, "Architectural principles for safety-critical real-time applications", Proc. of the IEEE, vol. 82, no. 1, pp. 25-40, 1994.

[15] J. Xu, Z. Kalbarczyk and R. K. Iyer, "Transparent Runtime Randomization for Security", Proc. 22nd Intl Symp. Reliable Distributed Systems (SRDS 03) IEEE CS Press, pp. 260269, 2003.

[16] S. Bernardi, L. Dranca and J. Merseguer, "A model-driven approach to survivability requirement assessment for critical systems", Institution of Mechanical Engineers Part O: Journal of Risk and Reliability Sage Journals, vol. 230, no. 5, pp. 485-501, 2016.

[17] X. Chang et al., "Survivability Model for Security and Dependability Analysis of a Vulnerable Critical System", IEEE 27 th International Conference on Computer Communication and Networks (ICCCN) , 2018.

[18] C. M. Grinstead and J. L. Snell, Introduction to Probability, American Mathematical Society, 1997

[19] C. C. Aggarwal, "Machine Learning with Shallow Neural Networks", Neural Networks and Deep Learning. Springer, Cham, doi.org/10.1007/978-3-319-94463-0_2. (2019).

[20] Fei-Fei. CS231n: Convolutional Neural Networks for Visual Recognition. Computer Science, Stanford University. Available online: http://cs231n.stanford.edu. (2019).

[21] J. S. Meneses, Z.R. Chavez, J.G. Rodriguez, "Compressed kNN: K-Nearest Neighbors with Data Compression" Entropy 21, no. 3: 234. https://doi.org/10.3390/e21030234. (2019).

[22] Y.Y. Song, Y. Lu, "Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry" vol. 27(2):130-5. PMID: 26120265; PMCID: PMC4466856. (2015)

[23] A. Ghose, "Support Vector Machine (SVM) Tutorial: Learning SVMs from examples". Medium: towards data science. (2017).

[24] Abu Al-Haija, Q.; Zein-Sabatto, S. An Efficient Deep-Learning-Based Detection and Classification System for Cyber-Attacks in IoT Communication Networks. Electronics 2020, 9, 2152. https://doi.org/10.3390/electronics9122152.

[25] E. Jonsson, T. Olovsson, *A quantitative model of the security intrusion process based on attacker behavior*. Software Engineering, IEEE Transactions on , Vol. 23, No.4, pp.235-245, 1997.

[26] D.J. Leversage, E. James, *Estimating a System's Mean Time-to-Compromise*. IEEE Transactions on Security & Privacy, Vol. 6, No.1, pp.52-60, 2008.

[27] A. Zimba, Z. Wang, H. Chen, *Bayesian-poisson based modeling of cyber attacks in cloud computing networks*. IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, China, 2017.

[28] M. Xu, L.Hua, *Cybersecurity Insurance: Modeling and Pricing*. Society of Actuaries, 2017.

[29] J.F. Riordan, et. al., *A Model of Network Porosity*. MIT Lincoln Laboratory, Technical Report(1217), 2016.

[30] H. Holm, *A large-scale study of the time required to compromise a computer system*. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, Vol.11, No. 1, pp. 2-15, 2014.

[31] L. Pan, A. Tomlinson, A.A. Koloydenko, *Time Pattern Analysis of Malware by Circular Statistics*. ACM/IEEE Symposium on Architectures for Networking and Communications,2017.

[32] C. Peng, et. al., *Modeling and predicting extreme cyber attack rates via marked point processes*. Journal of Applied Statistics, Taylor & Francis Group, 2016.

[33] Z. Zhan. M.Xu, S. Xu, *Predicting Cyber Attack Rates with Extreme Values*. IEEE Transactions on Information Forensics and Security, Vol. 10, No.8, pp.1666 - 1677, 2015.

[34] C. Peng, et. al., *Modeling multivariate cybersecurity risks*. Journal of Applied Statistics, Taylor & Francis Group, 2018.

[35] Z. Zhan. M.Xu, S. Xu, *Characterizing Honeypot-Captured Cyber Attacks: Statistical Framework and Case Study*. IEEE Transactions on Information Forensics and Security, Vol. 8, No.11, pp.1775 - 1789, 2013.

[36] Q. Abu Al-Haija, C.D. McCurry, S. Zein-Sabatto, *Intelligent Self-Reliant Cyber-Attacks Detection and Classification System for IoT Communication Using Deep Convolutional Neural Network*. 12th International Network Conference 2020 (INC2020), Springer Lecture Notes in Networks and Systems, Sep. 2020.

Stock Price Prediction Using a Multivariate Multistep LSTM: A Sentiment and Public Engagement Analysis Model

Bipin Aasi, Syeda Aniqa Imtiaz, Hamzah Arif Qadeer, Magdalean Singarajah, Rasha Kashef
Electrical, Computer, and Biomedical Engineering Department
Ryerson University
Toronto, Canada
{bipin.aasi, syeda.aniqa.imtiaz, hamzah.qadeer, msingarajah, rkashef}@ryerson.ca

***Abstract*—The impact of many factors on stock price has made the prediction of the stock market a problematic and highly complicated task to achieve. IoT analytics has enabled predictive analysis concerning the stock market, with internet search trends, reactions to current events, Twitter data, and historical stock returns as input data. Although inconsistencies remain as to which data sources are deemed most adequate, data preprocessing techniques have successfully overcome data integrity issues and unstructured data formats in specific applications. Additionally, advancements in computational power and machine learning technologies have led to the ability to handle tremendous amounts of information, accompanied by the growth of interest in this specific domain. In this paper, a Multivariate Multistep Output Long-Short-Term-Memory (MMLSTM) model is proposed to provide a one-week prediction on the stock close value for the technology company, "Apple Inc." with the stock name "AAPL". A large variety of data sources enabled by IoT platforms have been employed to model the impact of public sentiment and engagement on the closing price of this particular stock by looking at Google Search Trends, e-News headlines, and Tweets involving AAPL and its products. The proposed MMLSTM has improved the Mean Square Error (MSE) of up to 65% compared to ARIMA and Random Forest models. In addition, the proposed MMLSTM has outperformed most of the LSTM models introduced in the literature.**

***Index Terms*—forecasting, stock market, Tweets, Google, sentiment, LSTM, IoT Analytics**

I. Introduction

In-depth market analysis is continually refined to forecast stock behaviour. Multiple factors are involved in predicting the stock market trends, such as physical factors, physiological, rational, and irrational behaviour. Combining these factors ensures the volatile nature of stock prices; however, it becomes very difficult to predict it accurately. The knowledge of how a stock will perform provides insight to investors looking to invest. It can also aid the marketing teams of the stock's respective company. Public reaction to stock information and the sentiment of the information on the stock itself are two ways to decipher stock trajectories, which have yet to be thoroughly studied in combination with predictive models. Various financial institutions and traders have created their prediction models for competitive advantages. However, it has been hard to achieve higher-than-average results on the Returns on Investment (ROI). Nevertheless, the appeal of predicting the stock market as an accurate forecasting process can potentially lead to profits of millions of dollars. The forecasting problems are typically classified into one of three categories:

- Short term forecasting (seconds, minutes, days, weeks, or months)
- Medium-term forecasting (1-2 years)
- Long term forecasting (<2 years)

Numerous research papers have attempted to solve the stock forecasting problem over the years [1] - [5]. The researchers in [1] use various machine learning algorithms and social media news to perform stock forecasting. The work outlined in [2] compares multiple machine learning algorithms' effectiveness in the stock market forecasting of the SP 500 Index. Additionally, current market research tends to leverage knowledge obtained from local sources, and so a more diversified approach is needed to account for a global network of influencers and investors [3]. The inclusivity of public opinion that is not limited to English web text mining was a key feature that this research paper aimed to implement by looking at Google searches done worldwide and including Twitter sentiment analysis of Tweets from all languages [4][6].

In this paper, we propose a stock price prediction model, explicitly focusing on the technology-based company, Apple Inc. The actual stock values are forecasted instead of just the positive and negative fluctuations of the stock close. Various versions of LSTM models were initially explored before a specific model was selected. A Multistep Output Long-Short-Term-Memory (MMLSTM) is proposed in this paper to provide efficient stock forecasting. The stock market volatility [4] is strongly influenced by social and news media, which are extensively reviewed in this paper. The proposed model is applied to Google Trends, news headlines, and Twitter datasets. For the latter two datasets, VADER is used to detect the headlines and Tweets' sentiment, respectively. A daily average of the news sentiments and the Tweet sentiments are used as two separate features to the model. The VADER's ability to detect sentiment on colloquial terms is essential to the work discussed in this paper, as inclusivity is prioritized. Through this method, a more accurate sentiment of Tweets written using everyday

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

terminology will be incorporated as opposed to only professional terminology.

Experimental results show that the inclusion of both features pertaining to sentiment data and certain Google Trends aided in optimizing the model with the lowest error. The one-week forecast using MMLSTM achieved an average MAPE of 0.05932. The key aspects of this model are the diversity of the data sources and demographics. The structure of the paper is presented as follows: Section II and III provide a background and literature review of related work in stock forecasting using a variety of different textual and numeric datasets. Section IV provides a framework of the proposed model. Section V provides the experimental analysis. A discussion on the results is presented in section VI. Finally, the conclusion and future directions are presented in section VII.

II. LITERATURE REVIEW

In this section, a literature review on the usage of sentiment analysis for stock forecasting is discussed, along with an analysis of the potential of Google Trends' predictive capabilities.

A. Sentiment Analysis for Stock Forecasting

Over the years, various methods have been explored for predicting the price of a stock, but researchers have recently started using sentiment analysis for stock forecasting. In specific, these researchers have performed sentiment analysis on data obtained from two sources: news and social media. For example, in [7], sentiment classification of Tweets, which were labeled as either negative, positive, or neutral, has been conducted to train and tune a classifier. The classifier is based on SVM employing a wrapper approach, which constructs two linear-kernels trained to distinguish between positive and negative-or-neutral and vice versa. The Tweet is labeled as neutral when the two classifiers disagree, which occurs rarely. To process the data, they followed a typical bag-of-words computation procedure by applying tokenization, lemmatization using LemmaGen, and n-gram construction, which included unigrams and bigrams, the feature set, and the TF-IDF weighting scheme. The computed sentiment values were then used to model the relation between Twitter sentiment and DJIA.

The authors in [8] also used bag-of-words to perform sentiment analysis on Twitter data for stock forecasting. Still, they pre-processed the data to remove Tweets that had been retweeted to reduce redundancy. Tweets that were not in the English language were also discarded. Although bag-of-words is widely used to perform sentiment analysis, it fails to capture the sentiment by ignoring the sentences' structure. Thus, the authors in [1] have proposed an approach that uses the Sentiment Treebank to obtain sentiment from Tweets and news headlines. Their approach incorporates the sentence structure using deep learning to create a model for every sentence in the textual dataset. Compared to the bag-of-words approach, the suggested approach can use sentiment label phrases with ≈ 10% accuracy [1].

Similarly, the researchers in [9] used bag-of-words, noun phrases, and named entities to create a representation of financial news headlines to predict stock prices. After performing comparative analysis, they concluded that using noun phrases performed better than bag-of-words to predict the stock price.

A significant drawback of using these approaches is that they fail to capture the sentiment conveyed through informal text such as slang and acronyms. However, an increase in the use of sentimental analysis for various applications using text from social media has led to a number of open-source software that can be integrated into any learning-based model with ease. For example, in [10], the researchers have proposed a simple yet effective tool that combines lexical features with a rule-based model to generate a numerical sentiment value that categorizes text with varying sentiment intensities instead of labeling it as 1 if positive, 0 if neutral, and -1 if negative. The lexical features used in VADER are a combination of acronyms and slang words used in micro-blogging, such as Tweets to convey a sentiment with commonly used lexicons such as LIWC [10].

B. Google Trend Analytics

Another aspect that is being considered in the forecasting of the stock market is data from Google Trends, an application released by Google in May 2006, which shows how frequently a given search term is entered into Google's search engine. Google Trends was the precursor to Google Insights for Search, which was released in 2008, which provides insights into search terms used in the Google search engine. However, in 2012, the applications were merged in Google Trends. Google Trends does this by analyzing the popularity of various search queries across many regions and languages. It then graphs this data to compare it against the search volume of different queries over a period of time. This allows the user to see and compare the relative search volume of searches of one or more terms. The research work in [6] proposes a method to evaluate internet search data's predictive capabilities via Google Trends. They offered to treat Google Trends as a useful proxy for investors' attention regarding a particular company's stock belonging to the SP 500 Index. They found that the overall directional movement of the SP 500 Index correlated directly with the search volume series; however, they did note that often it "depends on which specific term is being searched for, and by extension the sentiment of the term itself" [6]. Overall, they hypothesized that while Google Trends can be considered a valid indicator of investor interest, it is conditional on the specific search term's inherent sentiment. A trading strategy was proposed through the generated Google Trends forecasts, which resulted in a 40% outperformance of a traditional buy-and-hold strategy. This adds validity to the reasoning behind including Google Search Trends as a feature for the proposed model discussed in this paper.

III. STOCK FORECASTING MODELS

Various algorithms have been successfully applied for stock forecasting; each of them is suited for forecasting for multiple reasons and different periods. In this section, some of these techniques are explored.

Autoregressive Integrated Moving Average (ARIMA): The univariate ARIMA model in [11] was used to predict stock price fluctuations. This type of model is heavily based on the moving average of the dataset. It provides an overall idea of the trends in a data set through the average of a subset of numbers. This technique is extremely useful for forecasting long-term trends for any period, e.g., calculating a 3-year, 5-year, and 20-year moving average. Stock market analysis often provides a 50 or 200-days' moving average to help forecasting where the stock market could be heading. For stock forecasting, a predicted closing price is decided for each day based on the average of a set of previous days observed values. Each prediction uses the latest set of values, hence the term "moving average". While this method has its advantages for showing the current market trend, it fails to account for future changes that may impact the stock price, making it overly unsuitable for long-term forecasting [11].

Linear Regression: A linear regression model application in stock forecasting is discussed in [12]. Linear regression is one of the basic machine learning algorithms. This model determines the relationship between the independent variables and the dependent variable. Linear regression is a simple technique; however, one of the most significant issues with utilizing linear regression algorithms is that it is prone to outliers, making it easily influenced by one-time jumps in the stock data. An assumption that the data is linearly related also must be validated for linear regression to be effective.

Artificial Neural Network (ANN): Stock market forecasting requires analyzing huge amounts of nonlinear data. Hence, an outbreak of advanced intelligent techniques is being implemented in this field to identify the hidden and complex patterns within the stock market. An ANN model achieves this through a self-learning process by adjusting approximators to find the input-output relationship of large, complex datasets, which enables accurate stock price predictions [13], [14], [15].

Random Forest (RF): RF utilizes an ensemble technique for stock forecasting and, as a result, is capable of performing both regression and classification tasks. It operates by constructing multiple decision trees during the training time, which provides an output in the form of a mean regression of each decision tree. An RF model is implemented in [16], and this model is capable of making steady predictions when many trees are used.

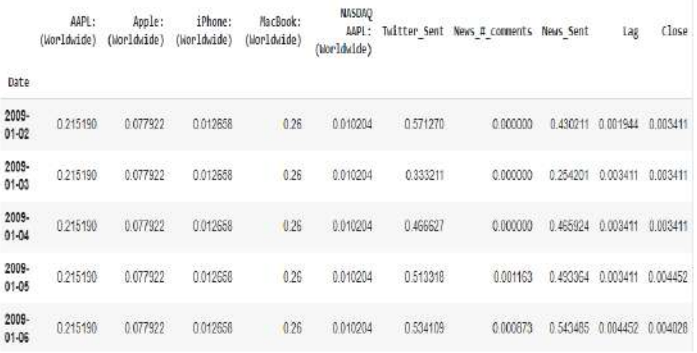
SVM: SVMs have also been used in more recent studies to forecast stock prices [2], [16], [17], as SVM is one of the best-known binary classifier models. The SVM model is an alternative to the ANN model. The difference between the two types of models is that while ANNs aim to minimize classification errors within the training data, SVMs can reduce classification errors within training data to achieve less error across test data. SVMs typically display higher accuracy in high-volatility stock markets.

IV. The Proposed MMLSTM model

An LSTM model is a special kind of Recurrent Neural Network (RNN), consisting of three gates, with the main difference being that the hidden layers in the LSTM model are fitted with LSTM cells to control the input flow. The three gates within the LSTM architecture are:

- Input Gate: Adds information to the cell state
- Forget Gate: Removes information that is deemed irrelevant to the model
- Output Gate: Selects the information to be shown as output

The LSTM models can provide a more efficient analysis of time-dependent variables. Additionally, this type of model can hold on to past information. Finally, this model can identify long-term dependencies for future prediction, making the model quite suitable for stock market analysis. The gates of an LSTM control how much input and memory are taken. An LSTM is chosen over other RNNs as it can tackle the vanishing gradient problem and the exploding gradient problem seen with other RNNs caused by an abundance of backpropagation. Research work in [13], [14], [18] show promising results of implementing an LSTM model for stock price prediction. There exist various LSTM models, such as univariate, multivariate, multi-step time series forecasting, etc. In this paper, a Multiple Input Multi-Step Output LSTM (MMLSTM) model is proposed. This particular model is used in multivariate time series forecasting problems. The output series is separate but depends on the input time series and requires multiple time steps to reach the output series. An LSTM model allows for multi-step prediction, which is useful for predicting multiple days in advance. This is beneficial, especially in a field such as stock forecasting, in which knowing the current and future state of the stock market is the key to profitability.



	AAPL: (Worldwide)	Apple: (Worldwide)	iPhone: (Worldwide)	MacBook: (Worldwide)	NASDAQ AAPL: (Worldwide)	Twitter_Sent	News_#_comments	News_Sent	Lag	Close
Date										
2009-01-02	0.215190	0.077922	0.012658	0.26	0.010204	0.571270	0.000000	0.430211	0.001944	0.003411
2009-01-03	0.215190	0.077922	0.012658	0.26	0.010204	0.333211	0.000000	0.254201	0.003411	0.003411
2009-01-04	0.215190	0.077922	0.012658	0.26	0.010204	0.466627	0.000000	0.465924	0.003411	0.003411
2009-01-05	0.215190	0.077922	0.012658	0.26	0.010204	0.513318	0.001163	0.493364	0.003411	0.004452
2009-01-06	0.215190	0.077922	0.012658	0.26	0.010204	0.534109	0.000673	0.543485	0.004452	0.004028

Fig.1 Dataset Excerpt

A. Data and Model Structure

Nine features were used to predict one target variable, the stock close value for Apple Inc. An example of the dataset is shown in Fig. 1. The first five features pertained to the level of interest in certain topics were obtained from Google Trends. The next feature depicts the weighted, aggregated sentiment from Tweets about Apple stock. The aggregated daily number of comments on SeekingAlpha news headlines and their respective sentiment values were the next two features. And finally, the lagged actual stock closing values were considered as a feature and the target variable. Each of the features was normalized between zero and one so that they were all on one common scale. The nine features mentioned above are listed as F0 through F8, respectively. The proposed MMLSTM model takes in 90 rows of feature data, representing 90 past days of data, to forecast one week of the stock close price. Therefore, at time t, features from t-89 to t are used to forecast the future stock value for time t to t+7.

B. Finalized Hyperparameters

Numerous combinations of MMLSTM units and epochs were tested to minimize the loss function measured by the MSE value, as discussed in section IV of this paper. Fig. 2 shows the loss plot for the final chosen configuration, which was 1 LSTM layer, with 64 LSTM units, followed by a dropout layer with a fraction of 0.4, and one dense layer, trained with 50 epochs and the “Adam” optimizer. It was found that adding LSTM stacks caused the model to be under-fit. Fig. 3 shows the loss plot of a model fit with a stacked LSTM, whereas Fig. 4 and 5 show significantly fewer under-fitting signs when a single LSTM layer is used. In these two figures, the training loss is lower than the validation loss, and the validation loss converges better. A flowchart of the proposed model is illustrated in Fig. 6. The final model parameters after using Bayesian Optimization were 43 LSTM units and 56 epochs of training.

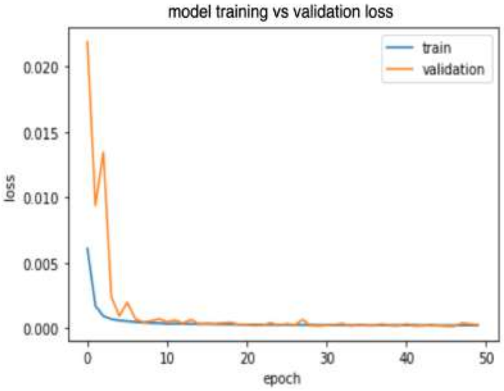


Fig. 2. Single Layer LSTM 64 units (Selected features).

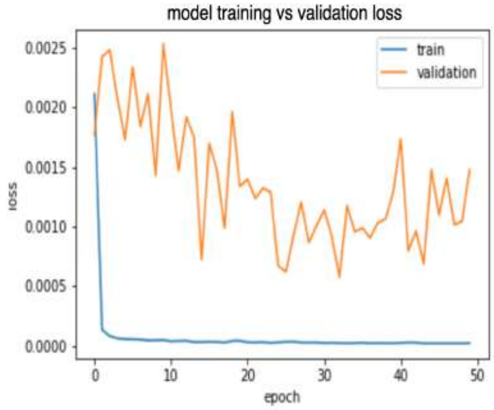


Fig. 3. Stacked 2-LSTM Layers, 64 units each, missing a dropout layer.

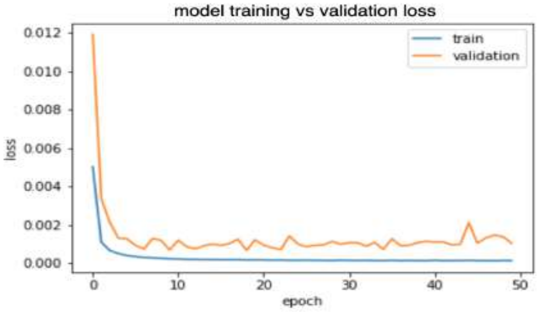


Fig. 4. Single Layer LSTM 64 units (all features).

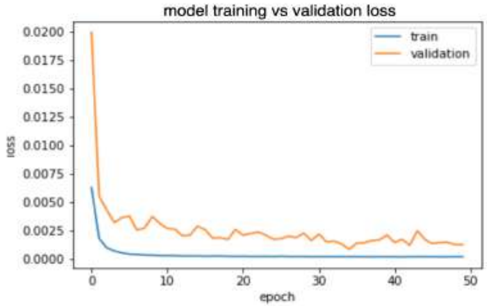


Fig. 5. Single Layer LSTM 128 units.

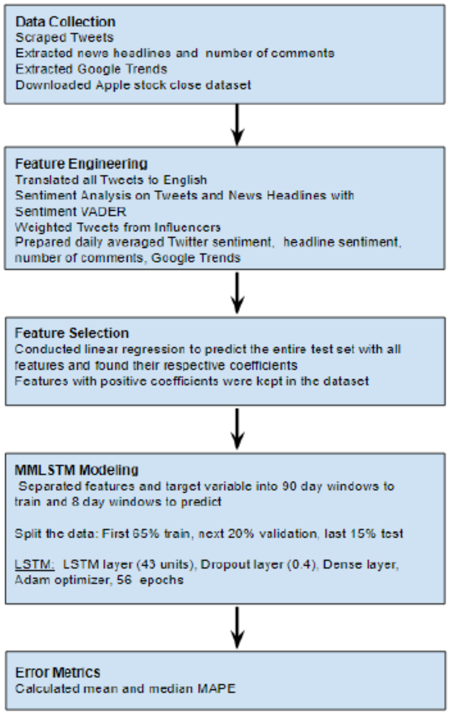

Data Collection
Scraped Tweets
Extracted news headlines and number of comments
Extracted Google Trends
Downloaded Apple stock close dataset
Feature Engineering
Translated all Tweets to English
Sentiment Analysis on Tweets and News Headlines with Sentiment VADER
Weighted Tweets from Influencers
Prepared daily averaged Twitter sentiment, headline sentiment, number of comments, Google Trends
Feature Selection
Conducted linear regression to predict the entire test set with all features and found their respective coefficients
Features with positive coefficients were kept in the dataset
MMLSTM Modeling
Separated features and target variable into 90 day windows to train and 8 day windows to predict
Split the data: First 65% train, next 20% validation, last 15% test
LSTM: LSTM layer (43 units), Dropout layer (0.4), Dense layer, Adam optimizer, 56 epochs
Error Metrics
Calculated mean and median MAPE

Fig. 6. Flowchart of the proposed model

V. Experimental Analysis and Design

Under-fitting is described as when a model does not comprehensively learn patterns from the training data. When the loss is plotted, this can be identified visually as the point before convergence between the training loss and validation loss. This portion is generally where the training and validation data are similar, alluding to a high bias, albeit a low variance. Overfitting occurs when a network is over-trained, which can be identified as the curve behaviour past the convergence point. The train and test error move in opposite directions, with the test error continually increasing. Telltale signs of over-fitting are the properties of low bias and high variance. The high variance source is noise in the training data, which might not be part of the test data [19]. After following linear regression coefficient ranking, as previously mentioned, the train versus validation loss plot illustrates a lower loss, faster convergence, and lack of over and under-fitting. There is a slight spike at the beginning of the validation plot, reflecting an unrepresentative validation dataset, which questions the model's ability to generalize [20].

A. Data Processing

Twitter data: The workflow of processing Tweets from Twitter can be seen in Fig. 7. The Tweets have been scraped from Twitter using the "SNSCRAPE" library for a date range between 2009-01-01 and 2020-11-11. A total of four different keywords have been used to generate .csv files containing the date, text from the tweet, and the respective Twitter user id. In the first stage of data processing, the duplicated Tweets were removed using the "drop duplicates ()" function from the pandas library. This was necessary in some cases where Tweets contained more than one of the keywords used to scrape, and as a result, the Tweets were scraped more than once. Tweets containing less than six characters had to be discarded from the dataset to detect the Tweets' language. The number of discarded Tweets were verified during every run of this workflow, and in all cases, only a small percentage of Tweets (approximately 5 %) were discarded.

To further justify the discarding, a Tweet that contained less than six characters was also assumed to be more likely to be a neutral Tweet. The Twitter user id of each Tweet was compared with a list of stock influencer user ids that were manually curated using online forums [21], [22], [23]. Tweets were assigned a weight, which was multiplied by the compound sentiment of that respective Tweet. A weight of 2 was applied to the Tweets obtained from stock influencers, and a weight of 1 was applied to all other Tweets. This was done to mimic the increased impact that influencer Tweets were assumed to have on public opinion. The average daily sentiment was then calculated from the weighted compound sentiments.

News Headline data: The news headlines in the proposed framework have been processed, as shown in Fig 8. The text for each headline, along with the data and the number of comments, have been curated in a .csv file, with all data from SeekingAlpha. The headlines have been processed to remove duplicate headlines, and a sentiment value was added using VADER. Similar to the Tweet sentiment aggregation, the sentiment for the news headlines was aggregated daily.

Google Trends data: The Google trends for the queries listed in the workflow in Fig. 9 were obtained monthly. The same overall level of interest for the month was a valuable indicator for all days of that month.

B. Feature Selection Methods

Two different feature selection methods were implemented to decide on the inclusion of the obtained features.

Correlation analysis: To conduct feature selection through correlation analysis, the following heat map seen in Fig. 10 was obtained. This method deemed that Twitter Sentiment and News Sentiment were not valuable features due to their low correlation. When the finalized model was run, omitting these two features, the error obtained was approximately 10.4%.

Linear regression coefficient ranking: The second feature selection method attempted is Linear Regression (LR) coefficient ranking. In this method, all features were used in a linear regression algorithm to predict the target variable. The resulting coefficients for the model and its corresponding features were stored. A threshold for the coefficient value is then used to determine which features should be omitted from the dataset.

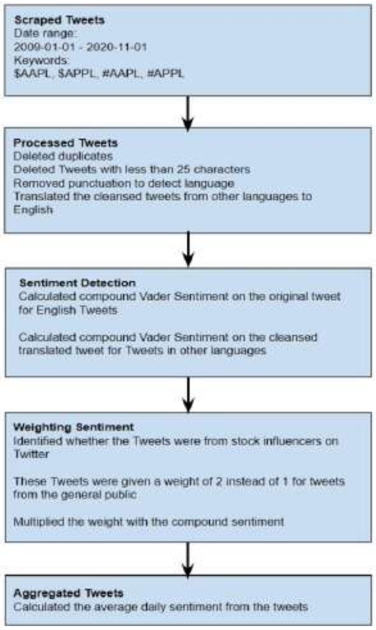

Scraped Tweets
Date range:
2009-01-01 - 2020-11-01
Keywords:
$AAPL, $APPL, #AAPL, #APPL
Processed Tweets
Deleted duplicates
Deleted Tweets with less than 25 characters
Removed punctuation to detect language
Translated the cleansed tweets from other languages to English
Sentiment Detection
Calculated compound Vader Sentiment on the original tweet for English Tweets
Calculated compound Vader Sentiment on the cleansed translated tweet for Tweets in other languages
Weighting Sentiment
Identified whether the Tweets were from stock influencers on Twitter
These Tweets were given a weight of 2 instead of 1 for tweets from the general public
Multiplied the weight with the compound sentiment
Aggregated Tweets
Calculated the average daily sentiment from the tweets

Fig. 7. Twitter Extraction, Pre-processing, Daily Sentiment Aggregation.

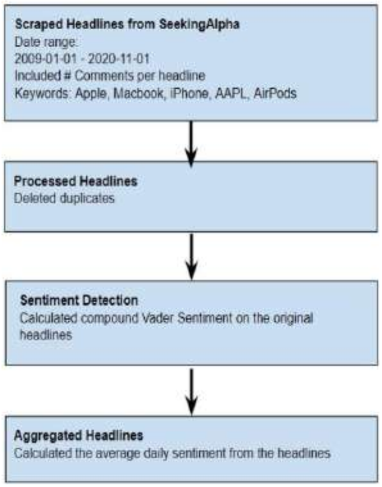

Scraped Headlines from SeekingAlpha
Date range:
2009-01-01 - 2020-11-01
Included # Comments per headline
Keywords: Apple, Macbook, iPhone, AAPL, AirPods
Processed Headlines
Deleted duplicates
Sentiment Detection
Calculated compound Vader Sentiment on the original headlines
Aggregated Headlines
Calculated the average daily sentiment from the headlines

Fig. 8. News Headline Extraction, Pre-processing, Daily Sentiment Aggregation.

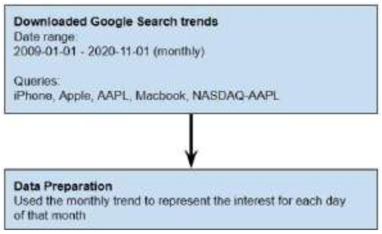

Downloaded Google Search trends
Date range:
2009-01-01 - 2020-11-01 (monthly)
Queries:
iPhone, Apple, AAPL, Macbook, NASDAQ-AAPL
Data Preparation
Used the monthly trend to represent the interest for each day of that month

Fig. 9. Google Search Trends Extraction and Data Preparation.

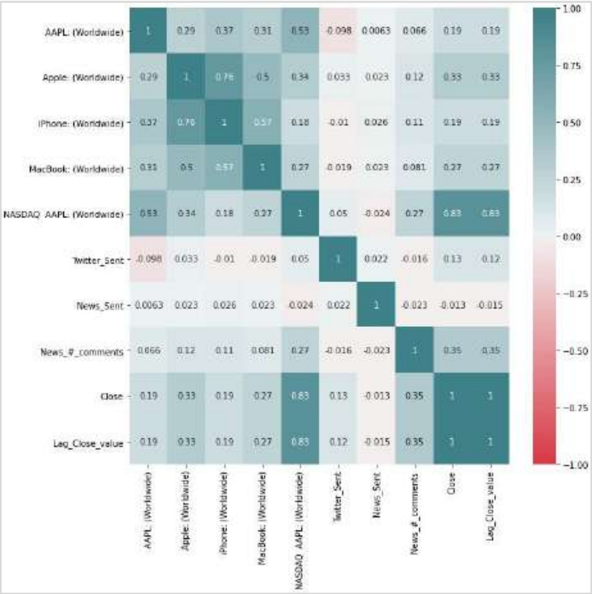


Fig. 10. Heatmap of Correlation between Features and Target Variable.

Using this method and keeping all features with positive coefficients, only five features remained in the dataset. These features include the Google Search Trends for iPhone and NASDAQ-AAPL, Twitter and News Sentiment, as well as the previous day stock close price. When the final model was run using only these features, the error amount was approximately 5.9%. When comparing the two feature selection methods with the error obtained when using all features, it was found that using LR coefficient ranking had the best result. For this reason, the subset of features corresponding to the LR coefficient ranking method was chosen for the final model.

C. Data Sources and Types

Four different datasets have been used in the proposed model. The first dataset consists of the stock price, while the second dataset consists of Google Search Trends data. The third and fourth datasets consist of sentiment values derived from news/financial headlines and Tweets from Twitter, respectively.

Stock price dataset: The $AAPL stock price has been obtained from Yahoo! Finance for the time period between 2009-01-01 to 2020-11-11. In particular, the daily closing price has been extracted as it is commonly used for stock price prediction by researchers, for example, in [1] and [8]. Stock prices are not updated on the weekends, and so the preceding weekday values are kept the same throughout the weekends.

Google Trends dataset: The number of people searching for terms related to the AAPL stock was retrieved from Google Trends from January 2009 to November 2020. The dataset consists of numerical values representing the number of times a term has been searched on Google monthly. The terms used to obtain the search trend include AAPL, Apple, iPhone, MacBook, and NASDAQ-AAPL.

News headlines dataset: The news headlines were searched using the stock symbol for Apple and collected from SeekingAlpha over a time period of 11 years. SeekingAlpha was used because of the strong correlation established between the price of a stock and articles published on this platform by the researchers in [24] and [25]. This final dataset consists of the title of the news article, the date, and the number of comments on each article. A sentiment value was also added to the dataset for each headline.

Twitter dataset: Tweets from Twitter have also been obtained for the same time period of 11 years. The Tweets were searched using three different keywords, which were aggregated into one dataset. Specifically, the hashtag and the ticker symbol (e.g., $AAPL) were used as one of the keywords since it has been found effective to obtain Tweets relevant to predicting stock price [1]. Other keywords used to search for Tweets include #AAPL, $APPL, and #APPL. This final dataset consisted of the user's id, text in each Tweet, date, and sentimental value.

D. Data Analysis and Discussion

The ARIMA one-day prediction and Random Forest with 1000 trees models were created to conduct a comparative analysis. The ARIMA model (results included in Table 1) was altered such that the true values were appended to the model history to match the MMLSTM prediction period. The ARIMA model achieved a MAPE of 35.452%. The Random Forest model had a notably higher error than the MMLSTM as it predicts the entire test set at once, as opposed to steps-like as the MMLSTM. The proposed MMLSTM achieved an improvement of more than 65% as compared to the Random Forest model.

The mean absolute percentage error is regarded as one of the most widely used metrics for prediction/forecast accuracy. This is because it is both scale-independent and is easy to interpret. However, the MAPE has some significant disadvantages, such as the fact that it produces infinite or undefined values for zero and/or close-to-zero values, which leads to percentages and ratios that are not intuitive. Prediction models rely on minimizing errors through numerical optimization, and the MAPE throws off these optimizers. Thus, there is a need to explore a new metric to be used for stock price prediction. Another drawback is that the MAPE "puts a heavier penalty on forecasts that exceed the actual than those that are less than the actual", as mentioned in [26]. Considering an example where the observation is 150 and the forecast is 50 results in a relative error of 50/150=0.33, and the result would be 50/100=0.50 in the opposite situation. Researchers in [27] propose a new measure for stock price prediction, which they called the mean arctangent absolute percentage error (MAAPE). This MAAPE measure was derived from looking at the MAPE but from a different angle. The critical difference between MAAPE and MAPE is that the MAAPE inherently preserves the MAPE properties while overcoming the problem with MAPE caused by the division by zero. The MAAPE accomplishes this by using bounded influences for outliers by considering the ratios as an angle rather than a slope. Listed below are the formulas of the MAPE and MAAPE for comparison:

$$MAPE_t = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (1)$$

Aₜ denotes, the actual value at data point *t*, and *Ft* represents the predicted/forecasted value at data point *t*, with *N* indicating the number of data points.

$$MAAPE_t = \frac{1}{n}\sum_{t=1}^{n}(AAPE_t) \quad (2)$$

$$AAPE_t = arctan\left|\frac{A_t - F_t}{A_t}\right| \quad (3)$$

TABLE I: COMPARATIVE ANALYSIS

	MMLSTM	ARIMA	Random Forest
Mean MAPE %	6.328	34.117	18.343
Mean MAAPE %	6.311	32.115	17.483

TABLE II: A COMPARISON OF THE MODELS WITH DIFFERENT FEATURES

	Mean MAPE %	Mean MAAPE %
All Features	13.547	13.363
Without F5 and F7	10.414	10.359
Without F0, F1, F3, F6	6.328	6.311

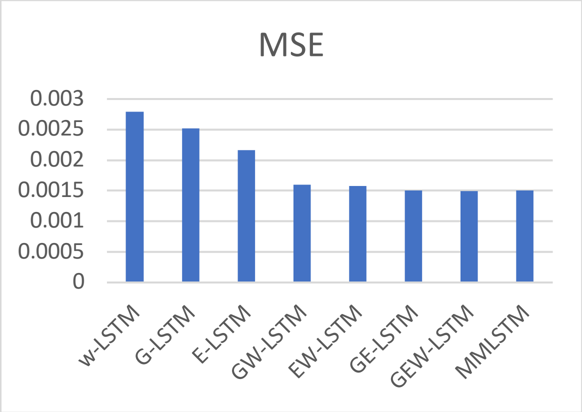
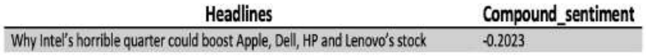


Fig. 10. A Comparative Analysis with the LSTM Models in [28]

As shown in Eq.3, if *Aₜ* is set to zero or even close to zero, the MAPE would result in an infinite or undefined value, which leads to an overall increase in outliers. Table II compares the two different error metrics for the proposed model and the different feature selection techniques. The inclusion of features F5 and F7: Twitter Sentiment, and News Headline Sentiment, respectively, resulted in the lowest percentage error values for both metrics. This confirms that public sentiment has an impact on the price of Apple stock. In [28], various LSTM models were created and trained with windows of 22 days of data to predict for the next 1-day volatility of the KOSPI 200 index, using explanatory variables as input data. Their LSTM architecture consisted of three LSTM layers with 0.3, 0.8, and 0.8 drop-out values, respectively, and two fully connected layers. The LSTM layers consisted of 10, 4, and 2 units, and the model was trained with 150 epochs. Their validation data consisted of 20% of the training data. It is trained on the previous 90 days to forecast the next seven days instead of just a one-day forecast. Given the additional days of forecast, the proposed MMLSTM has achieved similar performance to the GEW-LSTM and GE-LSTM models introduced by Kim and Won [28], and it outperforms all other models, including the W-LSTM, G-LSTM, E-LSTM, GW-LSTM, and the EW-LSTM. One of the main advantages of the proposed MMLSTM is that it has a much simpler architecture than what is described in [28] with comparable prediction performance.

VI. Discussions

In this paper, forecasting Apple Stock using an MMLSTM has effectively predicted the stock pieces for the next 7 days. It was found that the Google translate library used in our experiments limits the number of characters analyzed per day. An alternative to this would be to find libraries with sentiment detection capabilities in multiple languages such that translation error is minimized. This would also aid in combating translation errors, as the sentiment in phrases of one language may not always be the same as when it is translated. This would happen if the headlines or Tweets contained idioms that are generally language-specific and do not make sense when taken out of linguistic context. The model's generalizability can also be improved by implementing a time series cross-validation or merely increasing the dataset's size and diversity. This research work could be further expanded by incorporating the prediction of multiple stocks. Additionally, using the volatility of other companies' stock could potentially complement the forecast of Apple stock. An example shown in Fig. 12 is when the news headline's sentiment [30] is labeled negative; however, the headline favors Apple.



Headlines	Compound_sentiment
Why Intel's horrible quarter could boost Apple, Dell, HP and Lenovo's stock	-0.2023

Fig. 12. Inconsistent News Sentiment label for Apple Stock.

VII. CONCLUSION AND FUTURE DIRECTIONS

Deep learning forecasting algorithms are gaining importance in deciphering relevant features in obtaining the closest forecast to reduce the uncertainty from the complex and dynamic market information [29]. In this paper, forecasting Apple Stock using a proposed MMLSTM model has been shown to predict the stock pieces for the next week effectively. The key aspects of this model are the diversity of the data sources and demographics. A variety of data sources enabled by IoT platforms such as social media networks have been employed, particularly Google

Search Trends, e-News headlines, and Tweets involving AAPL and its products. The sentiment detection methods used could also be expanded to aggregate multiple methods, such as a Google API or another sentiment detection library, such as TextBlob.

Additionally, other machine learning models, such as SVMs, renowned for stock prediction, could be tested using the same features as in the proposed LSTM model. Sensitivity analysis using different weights for the sentiment of Apple experts' Tweets could also be performed to achieve higher forecast accuracy. The proposed MMLSTM model, along with the selected features, could be extended to forecast cryptocurrency [31][32] in future work. The trend of Bitcoin price, one particular type of cryptocurrency, is similar to that of Apple stock, this paper's subject. Both Apple stock and Bitcoin price saw exponential growth, albeit this growth in Bitcoin occurred later, around 2018. The size of specific cryptocurrency price datasets may not be as vast as the Apple stock dataset used in this paper, which dated back to 2009 due to the more recent development of cryptocurrency. The scope of the MMLSTM models can be extended to numerous use cases, including forecasting other stocks and cryptocurrencies [29][33].

References

[1] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and Social media, news," J. Ambient Intell. Humaniz. Comput., no. 0123456789, 2020, doi: 10.1007/s12652-020-01839-w.

[2] I. Kumar, K. Dogra, C. Utreja, and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Apr. 2018, pp. 1003–1007, doi: 10.1109/ICICCT.2018.8473214.

[3] H. Shu and J. Chang, "Spillovers of volatility index: evidence from U.S., European, and Asian stock markets", Applied Economics, vol. 51, no. 19, pp. 2070-2083, 2018. Available: 10.1080/00036846.2018.1540846.

[4] O. Kraaijeveld and J. De Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," Journal of International Financial Markets, Institutions Money, vol. 65, pp. 101188, 2020.

[5] N. Bakar and S. Rosbi, "Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction", International Journal of Advanced Engineering Research and Science, vol. 4, no. 11, pp. 130137, 2017. Available: 10.22161/ijaers.4.11.20.

[6] M. Huang, R. Rojas and P. Convery, "Forecasting stock market movements using Google Trend searches", Empirical Economics, vol. 59, no. 6, pp. 2821-2839, 2019.

[7] G. Ranco et al, "The Effects of Twitter Sentiment on Stock Price Returns," PloS One, vol. 10, (9), pp. e0138441-e0138441, 2015.

[8] M. Skuza and A. Romanowski, "Sentiment analysis of Twitter data within big data distributed environment for stock prediction," Proc. 2015 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2015, vol. 5, pp. 1349–1354, 2015, doi: 10.15439/2015F230.

[9] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," ACM Trans. Inf. Syst., vol. 27, no. 2, 2009, doi: 10.1145/1462198.1462204.

[10] C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).”," Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014, 2014, [Online]. Available: http://sentic.net/.

[11] Zhang, G. Peter. (2003) "Time series forecasting using a hybrid ARIMA and neural network mode." Neurocomputing 50 : 159-175.

[12] Seber, George AF and Lee, Alan J. (2012) "Linear regression analysis." John Wiley Sons 329.

[13] Li, Lei, Yabin Wu, Yihang Ou, Qi Li, Yanquan Zhou, and Daoxin Chen. (2017) "Research on machine learning algorithms and feature extraction for time series." IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC): 1-5.

[14] Oyeyemi, Elijah O., Lee-Anne McKinnell, and Allon WV Poole. (2007) "Neural network-based prediction techniques for global modeling of M (3000) F2 ionospheric parameter." Advances in Space Research 39 (5) : 643-650.

[15] Hamzac¸ebi, Cos¸kun, Diyar Akay, and Fevzi Kutay. (2009) "Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting." Expert Systems with Applications 36 (2) : 3839-3844.

[16] Liaw, Andy, and Matthew Wiener. (2002) "Classification and regression by Random Forest." R news 2 (3) : 18-22.

[17] Madge, S. and Bhatt, S., 2015. Predicting Stock Price Direction Using Support Vector Machines. [online] Cs.princeton.edu.

[18] Selvin, Sreelekshmy, R. Vinayakumar, E. A. Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. (2017) "Stock price prediction using LSTM, RNN and CNN-sliding window mode." International Conference on Advances in Computing, Communications and Informatics (ICACCI): 1643-1647.

[19] W. Khan, S. Chung, M. Awan and X. Wen, "Machine learning facilitated business intelligence (Part II)", Industrial Management Data Systems, vol. 120, no. 1, pp. 128-163, 2019.

[20] J. Brownlee, "How to use Learning Curves to Diagnose Machine Learning Model Performance", Machine Learning Mastery, 2019.

[21] M. O'Mahony, "9 Best Investing Twitter Accounts: Think Like An Investor", 2020, https://blog.mywallst.com/best-investing-twitteraccounts/.

[22] T.Parker, "10 Twitter Feeds Investors Should Follow", Investopedia, April 2020, https://www.investopedia.com/financial-edge/0712/10twitter-feeds-investors-should-follow.aspx.

[23] O. Whitehouse, "10 Trading Twitter Accounts To Follow In 2020", Traderlife, 2020 https://traderlife.co.uk/features/lunch-break-reads/10trading-twitter-accounts-to-follow-in-2020/.

[24] H. Chen, P. De, Y. J. Hu, and B.-H. Hwang, "Customers as Advisors: The Role of Social Media in Financial Markets," SSRN Electron. J., 2012, doi: 10.2139/ssrn.2024086.

[25] G. Wang et al., "Crowds on wall street: Extracting value from collaborative investing platforms," CSCW 2015 - Proc. 2015 ACM Int. Conf. Comput. Coop. Work Soc. Comput., pp. 17–30, 2015, doi: 10.1145/2675133.2675144.

[26] J. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons", Long Range Planning, vol. 26, no. 1, p. 150, 1993.

[27] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts", International Journal of Forecasting, vol. 32, no. 3, pp. 669-679, 2016.

[28] H. Kim, C. Won, Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models, Expert Systems with Applications, Volume 103, 2018, Pages 25-37, ISSN 09574174.

[29] T. Jerez and W. Kristjanpoller, "Effects of the validation set on stock returns forecasting," Expert Systems with Applications, vol. 150, pp. 113271, 2020.

[30] Pano, T.; Kashef, R. A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. Big Data Cogn. Comput. 2020, 4, 33.

[31] Ibrahim, A., Kashef, R., Li, M., Valencia, E., & Huang, E. (2020). Bitcoin Network Mechanics: Forecasting the BTC Closing Price Using Vector Auto-Regression Models Based on Endogenous and Exogenous Feature Variables. Journal of Risk and Financial Management, 13(9), 189.

[32] Tan, Xue, and Rasha Kashef. "Predicting the closing price of cryptocurrencies: a comparative study." Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems. 2019.

[33] Ahmed Ibrahim, Rasha Kashef, Liam Corrigan," Predicting market movement direction for bitcoin: A comparison of time series modeling methods, Computers & Electrical Engineering, Volume 89, 2021, 106905.

Galois Field Arithmetic Operations using Xilinx FPGAs in Cryptography

Hari Krishna Balupala
Design Engineer
Xilinx India Pvt. Limited
Hyderabad, India
krishnab@xilinx.com

Kumar Rahul
Senior Design Manager
Xilinx India Pvt. Limited
Hyderabad, India
kumarr@xilinx.com

Santosh Yachareni
Distinguished Engineer
Xilinx India Pvt. Limited
Hyderabad, India
santoshy@xilinx.com

***Abstract*—Cryptography algorithms are standards for any security-based industry. Internationally widely accepted and used cryptography algorithms like AES, DES rely heavily on finite field arithmetic which needs to be performed efficiently, to meet execution speed and design constraints. This paper aims to provide a concise perspective on designing efficient architectures in finite field arithmetic. In this paper, we propose Galois field arithmetic using irreducible polynomial to generate the S-box for AES using 128, 192, and 256-bit Keys. Cryptographic algorithms are more prone to side-channel attacks, so we implemented this algorithm instead of using a lookup table-based approach. The proposed Galois Field implementation of arithmetic operations are unique which can be extended to any primitive polynomial of any word size GF(2^n). A novel scheme is proposed for AES S-box, Inverse S-box, and validated using a Xilinx Virtex-7 FPGA.**

Keywords*—*Advanced Encryption Standard (**AES**), Data Encryption Standard (**DES**), Extended Euclidean Algorithm (**EEA**), Affine-transformation, Greatest Common Divisor (**GCD**), Galois field (**GF**), Substitution Box (**S-box**), Inverse Substitution, Field Programmable Gate Array (**FPGA**)

I. Introduction

Cryptography allows the organization to protect data and keep confidence in the electronic world. The increase in technology especially in the field of communication, information transmitted electronically resulted in an increased reliance on Cryptography and authentication. Cryptography means the transformation of information into a format that is unreadable for an unauthorized user. There are several algorithms designed to protect data like DES, Triple DES, RSA, Blow fish, Two fish, and AES. These are unbreakable encryption algorithms of the future. Some algorithms can be hacked using various crypto analyses like Side Channel attacks, Differential Power Analysis, etc.

Any field consisting finite number of elements is defined as finite field. F={elements} on which if we perform any binary operations like addition, subtraction, multiplication, division on elements result is said to be field. For any prime number 'p' there exists a finite field of p elements. GF(P) is represented as a field of p elements. GF(P^m) represents a field of P^m elements. Finite fields are also called Galois fields. The number of elements in a field is defined as an order of the field. Arithmetic operations are explained in GF(2^8) order. Inverse does not exist for infinite field, so it is mapped to zero. We represent the field in the form of the equation.

AES algorithm is a symmetric encryption algorithm that uses a single key for encryption and decryption. Encryption and decryption steps are described in figure Fig. 1. AES uses the keys of various lengths like 128,192,256 bits. 128-bit data plain text input is converted in a 4x4 byte matrix, followed by substitution and arithmetic operations in Galois field GF(2^8) like addition, multiplication. All these are done in a systematic way called Round. The round includes a series of logical and arithmetic operations in GF(2^8).

II. Advanced Encryption Standard

A. Implementation of AES

Encryption algorithms use the concept of encryption and decryption [6] i.e.

Plain text + Cipher key →Cipher text

Cipher text + Cipher key→Plaintext

Complete algorithm lies in creating these cipher text by adding cipher key to plain text. The main important operations of AES are S-box, inverse S-box, mix-columns, and inverse mix-column [5].

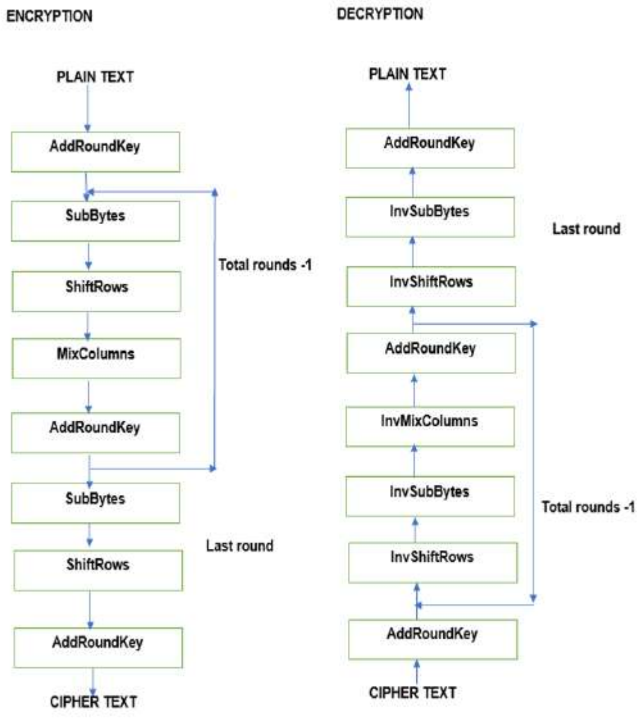


Fig. 1. AES algorithm flow chart

B. S-Box operations

AES algorithm is implemented in rounds, consists of the following computations for encryption.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Substitute from S-box: Shifting of bytes in a matrix. **Shift rows of the matrix:** Replacement of bytes from the boxes. **Mix columns:** Multiplying with constant. **Add round key:** Bitwise exclusive or operation.

Similarly, for decryption we need to perform steps in reverse order.

Substitute from inverse S-box: Shifting of bytes in a matrix. **Inverse Shift rows of the matrix:** Replacement of bytes from the boxes. **Inverse mix columns:** Multiplying with inverse constant. **Add round key:** Bitwise exclusive or operation.

The number of rounds depends on the key listed below.

128 bits key – 10 rounds

192 bits key – 12 rounds

256 bits key – 14 rounds

Depends on the key data undergoes several rounds like sub bytes, shift rows, mix columns and add round key for encryption and decryption. A separate algorithm is used for key expansion in each round not mentioned here.

C. *Substitution from S-box*

S-box is implemented from series of steps where the inverse is calculated using the Extended Euclidean Algorithm. EEA needs bits multiplication, addition, and division in the Galois field which is discussed in section III.

The S-box is generated by using multiplicative inverse of a given number in GF(2^8) = GF(2) [k]/ P(k), where P(k) is a primitive polynomial ($k^8 + k^4 + k^3 + k + 1$).

III. Implementation of S-box and Inverse S-box

We get 8 bits of data as an input for sub byte transformation operation, we need to follow the sequence of steps for encrypting data.

A. *Steps in the creation of S-box*

- Convert the given binary number into field GF(2^8).
- Check field equation is not zero for finding inverse.
- Perform inverse GF(2) [k] / ($k^8 + k^4 + k^3 + k + 1$) for finding inverse of an equation.
- Perform affine transformation and convert field equation to number again.

B. *Steps in the creation of Inverse S-box*

- Convert the given binary number into field GF(2^8).
- Perform inverse affine transformation.
- Check field equation is not zero for finding inverse.
- Perform inverse GF(2) [k] / ($k^8 + k^4 + k^3 + k + 1$) for finding inverse of an equation and convert to number again.

C. *Inverse using Extended Euclidean Algorithm*

Extended Euclidean Algorithm is an extension to Euclidean Algorithm. This algorithm is used in calculating coefficients of Bezouts Identity.

If a, b are coprime in the equation ax+by=gcd(a,b) then x = inverse of a modulo(b), y=inverse of b modulo(a). This equation can be extended to polynomial in any Galois field. In this paper we used the concept of EEA for finding inverse in S-box and Inverse S-box steps.

D. *Algorithm for Galois field GF(2^8) multiplication*

Multiplication is defined as a multiplication of two polynomials modulo and irreducible reducing polynomial in the finite field [3].

Let A(k), B(k) and C(k) belongs to Galois field GF(2^n) and P(k) be the irreducible polynomial generating GF(2^n) multiplication in GF(2^n) is defined as polynomial multiplication modulo the irreducible polynomial P(k), namely.

C(k) = A(k) multiply B(k) mod P(k). In this S-box implementation mod P(k) is ignored because A(k) multiply B(k) is always less than P(k).

The algorithm for Verilog implementation is implemented in the below figure Fig. 2 and variables are mentioned below.

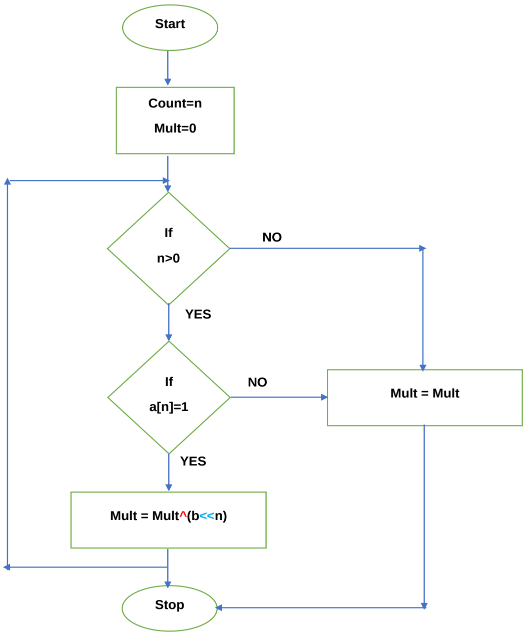


Fig. 2. Galois Field Multiplication

- a is a multiplier, b is the multiplicand
- Count = “n” which is the bit length of multiplicands
- “^” is xor operation and “<<” is bitwise left shift
- Final product will be stored in Mult

An illustration is mentioned with help of an example

A=10101, B=10101, n= 5

Example: 10101 X 10101

10101
00000x
10101xx
00000xxx
10101xxxx

Mult 100010001

E. Algorithm for Galois field GF(2^8) division

The division algorithm in the Galois field is similar to binary division. The algorithm for Verilog implementation is implemented in the below figure Fig. 3, Fig. 4, and some variables are used mentioned below.

Example:

Divisor = 001010011

Dividend = 100011011

Normal division algorithm deals with these steps, these cannot be implemented directly we need to undergo some steps for finding a quotient.

K^6+K^4+K+1) $K^8+K^4+K^3+K+1$ (K^2+1

$\underline{K^8+K^6+K^3+K^2}$

$K^6+K^4+K^2+K+1$

$\underline{K^6+K^4+K+1}$

K^2

For finding quotient we need to find the count value for shifting of the divisor and adding an extra bit at MSB.

Initial →

Count =1, symbol '||' denotes binary or operation.

Divisor → 0001010011 checking the logic (dividend || divisor) > (divisor<<1'b1)

The dividend →0100011011 If yes, we need to shift the divisor and increase the count value.

Final →

Count =3

Divisor after shifting → 0101001100

Dividend → 0100011011

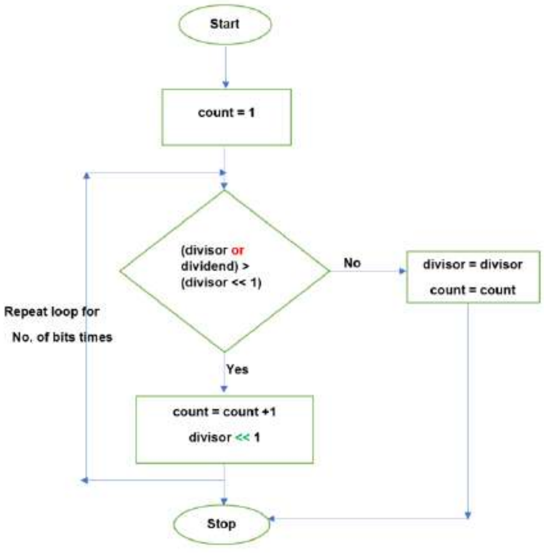


Fig. 3. Count value

Now we need to do actual division

Quotient → 000000000 count =3

Divisor → 101001100

Dividend → 100011011

until the count value >0 we need to do steps

Step1: quot→0000000000

Checking the logic ((dividend ^ divisor) < dividend)

yes, quot → 0000000000 || 0000000001

quot → 0000000001

dividend →dividend ^ divisor

divisor → divisor >> 1'b1

count = count -1 →2

Step2:

quot → 0000000001

Checking the logic ((dividend ^ divisor) < dividend)

No, quot →0000000010

divisor → divisor >> 1'b1

count = count -1 →1

Step3: quot→0000000100

Checking the logic ((dividend ^ divisor) < dividend)

yes, quot → 0000000100 || 0000000001

quot → 0000000101

dividend →dividend ^ divisor

divisor → divisor >> 1'b1

count = count -1 →0

End of steps Final quotient →0000000101 →K^2+1 (shown in the above example)

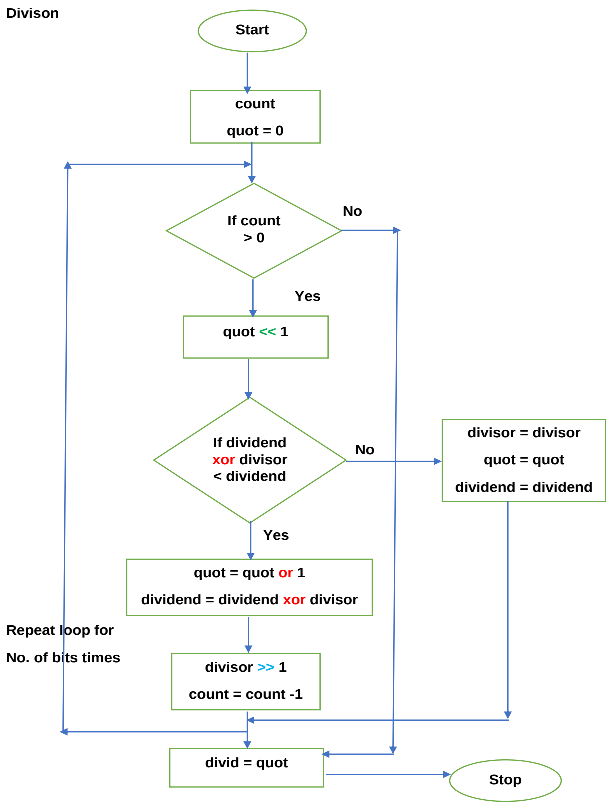


Fig. 4. Galois Field Divison

F. Algorithm for creating inverse GF(2^8)

Finding a multiplicative inverse needs operations addition, multiplication, and division. Implementation is defined in both binary and decimal.

The inverse of 197 mod 3000 using Extended Euclidean algorithm

Step1: Find GCD of 3000 and 197

3000 = 15(197) + 45
197 = 4(45) + 17
45 = 2(17) + 11
17 = 1(11) + 6
11 = 1(6) + 5
6 = 1(5) + 1 (Inverse exits for 197)

Step2: Writing in reverse order

1 = 6 - 1(5)
= 6 - 11(-6) =2(6)-1(11)
= 2(17) - 3(11)
= 8(17) - 3(45)
= 8(197) - 35(45)
= 533(197) - 35(3000)

Applying mod → 1= 533(197) mod 3000 - 35(3000) mod 3000 →1= 533(197) mod 3000
533 is the inverse of given 197 mod 3000

Similarly, we can use the same algorithm for the Galois field to find the inverse

Find the inverse of 01010011 mod 100011011

$K^6+K^4+K+1)$ $K^8+K^4+K^3+K+1$ (K^2+1
$\underline{K^8+K^6+K^3+K^2}$
$K^6+K^4+K^2+K+1$
$K^6\underline{+K^4+K+1}$
K^2)K^6+K^4+K+1 (K^3+K^2
$\underline{K^6+K^4}$
$K+1)$ K^2 (K
$\underline{K^2+1}$
1

(Inverse exists of a given number)
By following similar steps, we find inverse value.

A=$k^8+k^4+k^3+k+1$ B= k^6+k^4+k+1

k^2=A - B(k^2+1); k^2+1=0 -1(k^2+1)

k+1 = B – k^2(k^4+k^2); k^6+k^2+1= 1- (k^4+k^2)(k^2+1)

1 = k^2–(k+1) (k+1); k^7+k^6+k^3+k= (k^2+1)–(k+1) (k^6+k^2+1)

The inverse of 01010011 mod 100011011 is 11001010

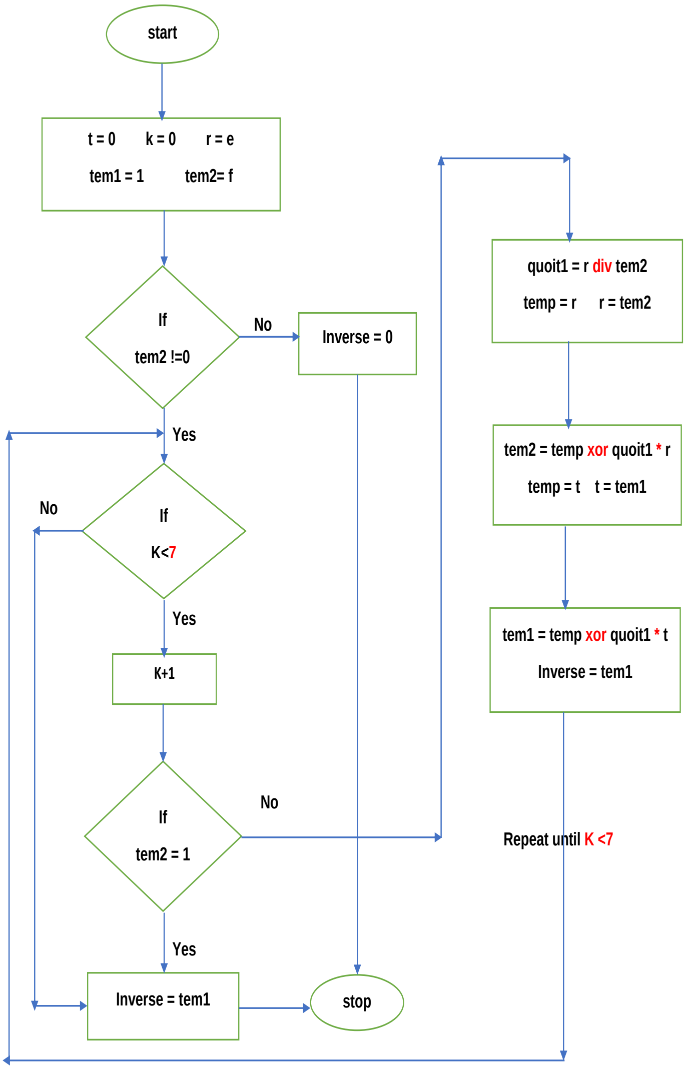


Fig. 5. Galois field Multiplicative Inverse

The above-mentioned algorithm is implemented in a sequence of steps using primitive polynomial e i.e. k^8+ k^4+ k^3+ k+ 1.

- F is a polynomial for which inverse must be found.
- If f = 0 inverse is 0.
- Some variables are used k, t, tem1, r, tem2.
- Division "div" and multiplication "*" algorithms are used.
- Until the value of K reaches less than 7 we keep on performing the steps for reaching the remainder 1.
- The final inverse of the polynomial will be tem1 in the following series of steps, temp and quoit1 variables are used for calculation purposes.

G. *Affine transform and S-box*

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \text{Multiply} \begin{bmatrix} k_7 \\ k_6 \\ k_5 \\ k_4 \\ k_3 \\ k_2 \\ k_1 \\ k_0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Fig. 6. Affine transform

Where [k_7 k_6 k_5 k_4 k_3 k_2 k_1 k_0] is the multiplicative inverse as a vector. S-box values are listed below.

S-box Values in Hexadecimal			
Inputs	Outputs	Inputs	Outputs
00	63	1A	A2
11	82	1B	AF
22	93	1C	9C
33	C3	1D	A4
44	1B	1E	72
55	FC	1F	C0
66	33	BB	EA
77	F5	CC	4B
88	C4	DD	C1
99	EE	EE	28
AA	AC	FF	16

Fig. 7. Substitution Box values

H. *Inverse Affine transform and Inverse S-box*

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \text{Multiply} \begin{bmatrix} k_0 \\ k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Fig. 8. Inverse Affine transform

Where [k_7 k_6 k_5 k_4 k_3 k_2 k_1 k_0] is the multiplicative inverse as a vector. Inverse S-box values are listed in Fig. 9.

Inverse S-box Values in Hexadecimal			
Inputs	Outputs	Inputs	Outputs
00	52	1A	43
11	C3	1B	44
22	94	1C	C4
33	66	1D	DE
44	86	1E	E9
55	ED	1F	CB
66	D3	BB	FE
77	02	CC	27
88	97	DD	C9
99	F9	EE	99
AA	62	FF	7D

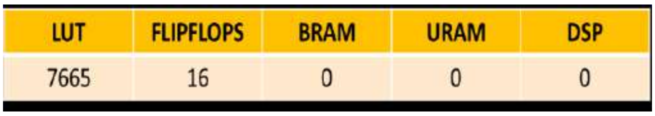
Fig. 9. Inverse Substitution Box values

IV. Synthesis and Simulation Results

We successfully implemented sequential logic of S-box in SystemVerilog using these Galois field arithmetic. We verified design functionality using Xilinx Vivado, Questasim, and VCS compiler.

Synthesis and Implementation results are mentioned in Fig. 10. These are synthesized in Xilinx Virtex-7 FPGA.

Design requires 7665 Look Up Tables, 16 Flipflops

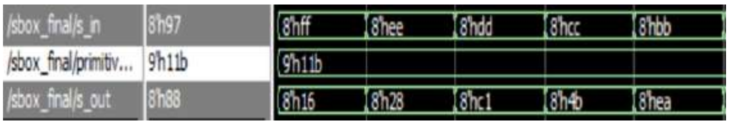


LUT	FLIPFLOPS	BRAM	URAM	DSP
7665	16	0	0	0

Fig. 10. Synthesis Results of S-box

A. *RTL simulation of S-box*

In Fig. 11 s_in indicates 8-bit input data, primitive polynomial as $k^8+ k^4+ k^3+ k+ 1$ and s_out indicates 8bit output data.

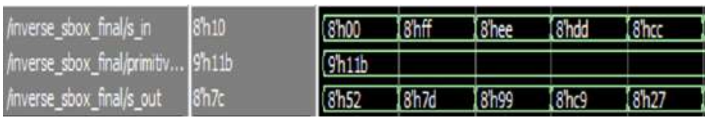


```
/sbox_final/s_in         8'h97    8'hff   8'hee   8'hdd   8'hcc   8'hbb
/sbox_final/primitiv...  9'h11b   9'h11b
/sbox_final/s_out        8'h88    8'h16   8'h28   8'hc1   8'h4b   8'hea
```

Fig. 11. Simulation S-box values

B. *RTL simulation of Inverse S-box*

In Fig. 12 s_in indicates 8-bit input data, primitive polynomial as $k^8+ k^4+ k^3+ k+ 1$ and s_out indicates 8bit output data.



```
/inverse_sbox_final/s_in         8'h10    8'h00   8'hff   8'hee   8'hdd   8'hcc
/inverse_sbox_final/primitiv...  9'h11b   9'h11b
/inverse_sbox_final/s_out        8'h7c    8'h52   8'h7d   8'h99   8'hc9   8'h27
```

Fig. 12. Simulation Inverse S-box values

V. Conclusion

Every GF(2^n) has many valid irreducible polynomials. The proposed Galois field arithmetic method is unique and can be used with any valid irreducible polynomial in GF(2^n) to generate the S-box and inverse S-box for AES for 128, 192, and 256-bits Key. We implemented this algorithm instead of using a lookup table-based approach to overcome side-channel attacks. These algorithms can be synthesized in all FPGAs and implemented as ASIC. The proposed algorithms have been used and validated to generate S-box and inverse S-box on a Xilinx Vertix-7 FPGA. This design can be verified in all FPGA's. Future work concerns deeper analysis in improving power, performance, and area [PPA] of these Galois field arithmetic operations in various cryptographic applications.

References

[1] A. Pradeep, V. Mohanty, A. M. Subramaniam and C. Rebeiro, "Revisiting AES SBox Composite Field Implementations for FPGAs," in IEEE Embedded Systems Letters, vol. 11, no. 3, pp. 85-88, Sept. 2019, doi: 10.1109/LES.2019.2899113.

[2] C. Yu and M. Ciesielski, "Formal Analysis of Galois Field Arithmetic Circuits-Parallel Verification and Reverse Engineering," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 38, no. 2, pp. 354-365, Feb. 2019, doi: 10.1109/TCAD.2018.2808457.

[3] D. Q. Huy, N. M. Duc, L. D. Khai and V. D. Lung, "Hardware Implementation of AES with S-Box Using Composite-Field for WLAN Systems," 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), Danang, Vietnam, 2019, pp. 1-6, doi: 10.1109/RIVF.2019.8713711.

[4] S. V. GADED and A. Deshpande, "Composite Field Arithematic Based S-Box For AES Algorithm," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 1209-1213, doi: 10.1109/ICECA.2019.8821862.

[5] L. Sarti, L. Baldanzi, B. Carnevale and L. Fanucci, "An automated S-box optimization based on composite field arithmetic," 2017 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), Giardini Naxos, 2017, pp. 85-88, doi: 10.1109/PRIME.2017.7974113.

[6] M. R. Rao and R. Sharma "FPGA implementation of combined s-box and invs-box of aes" Signal Processing and Integrated Networks (SPIN) 2017 4th International Conference on pp. 566-571 2017

[7] S. Oukili, S. Bri and A. V. S. Kumar, "High speed efficient FPGA implementation of pipelined AES S-Box," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 901-905, doi: 10.1109/CIST.2016.7805015.

[8] R. Ueno, N. Homma, Y. Sugawara and T. Aoki, "Formal Design of Galois-Field Arithmetic Circuits Based on Polynomial Ring Representation," 2015 IEEE International Symposium on Multiple-Valued Logic, Waterloo, ON, Canada, 2015, pp. 48-53, doi: 10.1109/ISMVL.2015.16.

[9] Pu Baoxing and Chen Jiye, "Implementation of arithmetic operation of finite field GF(2n)," Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), Shenyang, China, 2013, pp. 1710-1714, doi: 10.1109/MEC.2013.6885331.

[10] Mehran Mozaffari-Kermani "A Low Power high performance concurrent fault detection approach for the composite field S-Box and Inverse S-Box" IEEE Transactions on computers vol. 60 no. 9 september 2011.

[11] X. Zhang and K. K. Parhi "High-speed vlsi architectures for the aes algorithm" IEEE transactions on very large scale integration (VLSI) systems vol. 12 no. 9 pp. 957-967 2004

[12] W. McLoone and J. V. McCanny "Rijndael fpga implementation utilizing look-up tables" in Signal Processing Systems 2001 IEEE Workshop on IEEE pp. 349-360 2001.

[13] N. F. Pub "197: Advanced encryption standard (aes)" Federal Information Processing Standards Publication vol. 197 no. 441 pp. 0311 2001.

[14] J. Daemen and V. Rijmen, AES Proposal: Rijndael, AES Algorithm Submission, September 3, 1999

[15] E. Mastrovito, "VLSI architectures for computation in Galois fields," Ph.D. dissertation, Dept. Elect. Eng., Linkoping University, Link ¨ oping, ¨ Sweden, 1991.

[16] D. R. Stinson, "On bit-serial multiplication and dual bases in GF(2m)," IEEE Trans. Inform. Theory, vol. 37, pp. 1733–1936, Nov. 1991.

[17] Lidl, Rudolf; Niederreiter, Harald (1983), Finite Fields, Addison-Wesley, ISBN 0-201-13519-1 (reissued in 1984 by Cambridge University Press ISBN 0-521-30240-4).

[18] J. Ax, "Zeros of polynomials over finite fields," Amer. J. of Math., vol. 86, pp. 255–261, 1964.

Multiuser Data Dissemination in OFDMA System Based on Deep Q-Network

Yuan Xing, Haowen Pan, Bin Xu, Tianchi Zhao, Cristiano Tapparello, Yuchen Qian

***Abstract*—In this paper, a multiuser data dissemination problem is analyzed in an orthogonal frequency division multiple access(OFDMA) downlink system. By dynamically allocating subchannels and power to the mobile users, the system aims to minimize the time consumption in order to successfully deliver data to multiple mobile users under the restriction of total energy consumption. Both the objective and the constraint of the optimization are related to the real-time resource allocation strategies. Moreover, neither the statistics nor the full channel state information is known to the base station. In order to solve the global optimization problem with partial channel information, a Deep Q-Network algorithm is adopted. The numerical results show that compared with the other algorithms, Deep Q-Network can learn the optimal resource allocation strategies and achieve very good system performance.**

***Index Terms*—OFDMA, Energy efficiency, Data dissemination, Global optimization, Deep Q-Network.**

I. Introduction

As a promising multi access technique, orthogonal frequency division multiple access(OFDMA) is applied to many broadband wireless communication systems. Adapting OFDMA on multiple users wireless transmission can sufficiently exploit multiple users diversity in order to enhance the overall system performance. In [1], the optimization is formulated as maximizing a weighted sum information rates while satisfying both power and additional receiver-specific rate requirements in multiple parallel Gaussian broadcast channels.

In order to reduce the energy consumption in communication system, energy-efficient communications attracts much attentions. It is an effective metric to evaluate efficiency of the energy consumption [2]. The energy efficiency problem in OFDMA wireless communication systems has recently been discussed [3]–[5]. In [3], the authors study the energy-efficient resource allocation in OFDMA cellular networks. The energy efficiency is maximized under certain quality-of-service (QoS) requirements. In [5], the authors maximize the energy efficiency of the worst-case link under the information rate, total transmit power and available subcarrier constraints. In most of these works, the OFDMA resource allocation optimization problems are solved with statistical or complete knowledge about the environment. Very few papers formulate the optimization as global optimization problems. In [6], the optimization aims at maximizing the total energy efficiency of T time instants. Adapting to the dynamics of the user arrivals and channel state information, a real-time transmission policy can be found out.

In this paper, we consider a base station disseminate data to multiple mobile users. The power allocation and subchannel assignment are designed to meet the system requirements. The optimization problem is formulated as minimizing the time consumption in order to successfully deliver the required data to each user under the total energy consumption constraint. For the optimization, both the objective and the constraint are affected by the resource allocation strategy in each specific time slot. If the statistics of the channel is known to the base station, the problem is formed as a NP-hard problem, like traveling salesman problem [7]. However, the full channel information is unknown to the base station, which makes this problem even harder to be solved. We propose to apply the Deep Q-Network algorithm to solve the problem. Deep Q-Network has been widely adopted to solve complicated communication problems [6], [8]–[11]. In our model, Deep Q-Network is adopted in the optimization framework to make the dynamic resource allocation decision. The simulation results demonstrate that Deep Q-Network outperforms the other existing algorithms in solving the optimization problem.

II. System Model

In a downlink broadcast system: the base station serves K mobile users. There are N subchannels. $\mathcal{N} = \{1, 2, ..., N\}$. By properly adjusting the power allocation p_k and the subchannel assignment $\mathcal{I}_k$ to each user, the base station aims to disseminate the information payload B_k to each user in the shortest time. The system model is shown in Fig. 1.

The broadband spectrum is divided into several subchannels. Each subchannel has identical bandwidth W. Any subchannel can be allocated to any user k. The subchannels assigned to user k form the set $\mathcal{I}_k$. $\mathcal{I}_k$ is the subset of $\mathcal{N}$.

$$\mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_K \in \mathcal{N} \tag{1}$$

OFDM is utilized to convert the frequency selective wireless channels into multiple parallel flat channels over different subcarriers. We assume time is slotted and channel fading is approximately the same within one subchannel and independent across different subchannels in each time slot [11]. We

Y. Xing is with the Department of Engineering and Technology, University of Wisconsin-Stout, Menomonie, WI, 54751, USA. Email: xingy@uwstout.edu.

H. Pan and B. Xu are with Shanghai Legit Network Technology Co. Email: hp2414@columbia.edu, binxu1989@gmail.com.

T. Zhao is with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, 85721, USA. Email: tzhao7@email.arizona.edu.

C. Tapparello is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, 14627, USA. Email: cristiano.tapparello@rochester.edu.

Y. Qian is with the Department of Electrical and Computer Engineering, Baylor University, Waco, TX, 76706, USA. Email: Yuchen_Qian@baylor.edu.

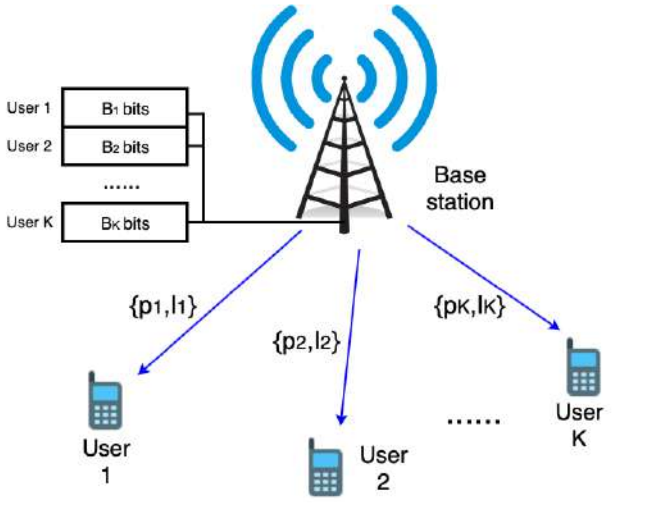


Fig. 1. Transmission from base station to multiple mobile users over parallel frequency channels.

define the channel gain between the transmitter and the kth user on the nth sub-channel at time t as:

$$h_{kn}(t) = \alpha_k g_{kn}(t) \tag{2}$$

where $g_{kn}(t)$ denotes the frequency dependent small-scale fading power component and assumed to be exponentially distributed, due to the Rayleigh fading channel feature.

$$\alpha_k = l_k(t)^{-\beta} \tag{3}$$

which denotes the path loss between the base station and user k. β is the path-loss exponent.

We define a subchannel selection threshold $\hat{p}_{sub}$ for each subchannel. For each individual user k, all N subchannels are classified into two categorical sets at time t: good subchannel set $\mathcal{N}_k^G(t)$ and bad subchannel set $\mathcal{N}_k^B(t)$:

$$\mathcal{N}_k^G(t) = \{\arg_{n\in\mathcal{N}} |h_{kn}(t)|^2 > \hat{p}_{sub}\} \tag{4}$$

$$\mathcal{N}_k^B(t) = \{\arg_{n\in\mathcal{N}} |h_{kn}(t)|^2 \leq \hat{p}_{sub}\} \tag{5}$$

In the real-time broadband communication system, acquiring the entire channel state information(CSI) is at extremely high cost. Hence, in this paper, we suppose only $\mathcal{N}_k^G(t)$ is available at the base station, the full CSI is unknown. The unique good subchannel set is denoted as

$$\mathcal{N}_k^u(t) = \{n | n \in \mathcal{N}_k^G(t), n \notin \mathcal{N}_j^G(t), j \in \mathcal{K}/k\} \tag{6}$$

where all the subchannels are only good for user k. We define the shared good subchannels as those good subchannels are qualified for multiple users. We evenly distribute shared good channels to all users. After distribution, the remaining extra subchannels are randomly assigned. From the shared good channels, the subchannels assigned to user k is included in set $\mathcal{N}_k^s(t)$. The ultimate assigned subchannel set for user k is defined as

$$\mathcal{N}_k^{tot}(t) = \mathcal{N}_k^u(t) \cup \mathcal{N}_k^s(t) \tag{7}$$

The power can only be allocated to the subchannels in $\mathcal{N}_k^{tot}(t)$ for user k. The total power allocated for user k at time t is denoted as $p_k(t)$, which is equally allocated to each subchannel in $\mathcal{N}_k^{tot}(t)$, the power allocated to each particular subchannel is $\frac{p_k(t)}{\mathcal{N}_k^{tot}(t)}$. In each time slot, b bits loaded information is transmitted on each subchannel. At time t, the Bit Error Rate(BER) of transmission from base station to user k on subchannel n in the case of M-ary QAM is given by:

$$BER_{k,n}(t) = 0.2e^{-1.6\frac{p_k(t)}{\mathcal{N}_k^{tot}(t)(2^b-1)}\frac{h_{kn}(t)}{\sigma_n^2(t)}} \tag{8}$$

where

$$\sigma^2(t) = N_o(t)W \tag{9}$$

denotes the variance of additive white Gaussian noise at time t. $N_0(t)$ is the noise power spectrum density.

We define the data successfully delivered to user k at time t as $R_k(t)$.

$$R_k(t) = \sum_{n\in\mathcal{N}_k^{tot}(t)} r_k^n(t) \tag{10}$$

$r_k^n(t) = b$ when all b bits data are received with no error on subchannel n. Otherwise, $r_k^n(t) = 0$ and a retransmission has to be issued.

III. Optimization Problem Formulation

The base station holds B_k bits information for each user. By appropriately allocating power $p_k(t)$ in real-time, the optimization problem is to minimize the time consumption T to successfully deliver B_k bits data to each user under the total energy consumption constraint $\hat{E}$. the optimization is formulated as:

$$\mathcal{P}_1: \begin{array}{ll} \underset{\{p_k(t)\}}{\text{minimize}} & T \\ \text{subject to} & \sum_{t=1}^{T} R_k(t) \geq B_k, \quad \forall k \in \mathcal{K} \\ & \sum_{t=1}^{T}\sum_{k=1}^{K} p_k(t)d_c \leq \hat{E} \end{array} \tag{11}$$

where d_c denotes the channel coherence time.

$\mathcal{P}_1$ is a complicated long term optimization problem. Data delivery can be seen as the shortest path problem. The energy consumption constraint affect the transmit decision at each time slot. It's unable to be solved by Dynamic Programming since very limited channel information is known to the base station. All these obstacles makes the Deep Q-Network(DQN) a fitting tool to solve $\mathcal{P}_1$.

In DRL, the system state s is consist of: the number of subchannels assigned to users $|\mathcal{N}_k^{tot}(t)|$, the accumulated throughput

$$R_k^{acc}(t) = \sum_{u=1}^{t} R_k(u) \tag{12}$$

the accumulated energy consumption

$$E^{acc}(t) = \sum_{u=1}^{t}\sum_{k=1}^{K} p_k(u)d_c \tag{13}$$

the accumulated sum throughput

$$R^{acc}(t) = \sum_{k=1}^{K} R_k^{acc}(t) \tag{14}$$

The system state s is denoted as:

$$\begin{aligned} s_t = [&|\mathcal{N}_1^{tot}(t)|, ..., |\mathcal{N}_K^{tot}(t)|, R_1^{acc}(t), ..., R_K^{acc}(t), \\ &E^{acc}(t), R^{acc}(t)] \end{aligned} \tag{15}$$

where $\mathbf{s}_t \in \mathbf{R}^{1\times(2K+2)}$. The set of all states is denoted by $\mathcal{S}$.

We define the power allocated $p_k(t)$ to each user k in a particular time slot t as the action $\mathbf{a}_t$:

$$\mathbf{a}_t = [p_1(t), p_2(t), ..., p_k(t)] \in \mathbf{R}^{1\times K} \tag{16}$$

where $p_k(t) \in [0, P]$, $\sum_{k=1}^{K} p_k(t) \leq P$. The whole action set is denoted as $\mathcal{A}$.

The evolution of our system can be described by a Markov process. The first system state starts at $t = 0$:

$$\mathbf{s}_0 = [|\mathcal{N}_1^{tot}(0)|, ..., |\mathcal{N}_K^{tot}(0)|, 0, ..., 0] \tag{17}$$

and the final state $\mathbf{s}_T$ in which $t = T$, i.e.,

$$\begin{aligned} s_T = [&|\mathcal{N}_1^{tot}(T)|, ..., |\mathcal{N}_K^{tot}(T)|, B_1, ..., B_k, \\ &E^{acc}(T), KB] \end{aligned} \tag{18}$$

At time slot t the system is in a generic state $\mathbf{s}$, the transmitter selects an action $\mathbf{a}_t \in \mathcal{A}$, and the system moves to a new state $\mathbf{s}'$. $w(\mathbf{s}, \mathbf{a}_t, \mathbf{s}^{'})$ denotes the reward when the current state is $\mathbf{s} \in \mathcal{S}$, action $\mathbf{a}_t \in \mathcal{A}$ is selected and the system moves to state $\mathbf{s}^{'} \in \mathcal{S}$.

Since we aim at minimizing data delivery time consumption under total energy constraint, both the optimization target and the constraints have to be related to the reward [11]. In order to design a proper reward function, optimization $\mathcal{P}_1$ is reformed as:

$$\mathcal{P}_2: \quad \begin{aligned} &\underset{\{p_k(t)\}}{\text{minimize}} && T \\ &\text{subject to} && \textstyle\sum_{t=1}^{T} R_k(t) \geq B_k, \quad \forall k \in \mathcal{K} \\ & && \frac{KB}{\sum_{t=1}^{T}\sum_{k=1}^{K} p_k(t)d_c} \geq \hat{\Gamma} \end{aligned} \tag{19}$$

where

$$\hat{\Gamma} = \frac{\sum_{k=1}^{K} B_k}{\hat{E}} \tag{20}$$

which denotes the energy efficiency threshold. The energy efficiency at time t is denoted as:

$$\Gamma(t) = \frac{\sum_{k=1}^{K} R_k^{acc}(t)}{\sum_{u=1}^{t}\sum_{k=1}^{K} p_k(u)d_c} \tag{21}$$

In each time slot, $\Gamma(t)$ is compared with $\hat{\Gamma}$, which effectively tracks the energy consumption condition. We define $\sigma(t)$ as the energy constraint satisfactory indicator.

The reward function is defined as:

$$w(\mathbf{s}, \mathbf{a}_t, \mathbf{s}^{'}) = \min(R_1^{acc}(t), ..., R_K^{acc}(t))\sigma(t) \tag{22}$$

$$\sigma(t) = \begin{cases} 1 & \Gamma(t) \geq \hat{\Gamma} \\ 0 & \Gamma(t) < \hat{\Gamma} \end{cases} \tag{23}$$

The optimization problem $\mathcal{P}_2 = (\mathcal{S}, \mathcal{A}, p, w)$ can then be seen as a Markov Decision Process from state $\mathbf{s}_0$ to state $\mathbf{s}_T$ on the Markov chain with states $\mathcal{S}$ and probabilities $\{p_{\mathbf{s},\mathbf{s}'}(\mathbf{a})\}$, actions $\mathbf{a} \in \mathcal{A}$, and rewards $w(\mathbf{s}, \mathbf{a}, \mathbf{s}^{'})$. Without knowledge about $\{p_{\mathbf{s},\mathbf{s}'}(\mathbf{a})\}$, our objective is to find, for each possible state $\mathbf{s} \in \mathcal{S}$, an optimal action $\mathbf{a}^*(\mathbf{s})$ so that the optimization goal is achieved. A generic policy can be written as

$$\pi = \{\mathbf{a}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\} \tag{24}$$

IV. Solving Optimization with Deep Q Network

Procedure 1 : Deep Q Network algorithm training process

```
1. Randomly generate the weight parameter $\theta$ for the $eval\_net$. The $target\_net$ clones the weight parameters $\theta' = \theta$. $u = 1$. $t = 1$. $D = d = 1$.
2. The base station acquires $\mathcal{N}_k^G$ from all the users in order to calculate $\mathcal{N}_k^{tot}$. $\mathbf{s} = \mathbf{s}_0$.
3. Randomly generates a probability $p \in [0, 1]$.
   IF $D > D_{ini}$ and $p \geq \epsilon_{ch}$:
      Play the action $\mathbf{a}$ as: $\mathbf{a} = \max_{\mathbf{a}\in\mathcal{A}} Q(\mathbf{s}, \mathbf{a})$
   ELSE:
      Randomly play the action from action set $\mathcal{A}$.
4. $R_k^{acc}(t)$ and $E^{acc}(t)$ are renewed and feedbacked to the base station. At the end of each time slot, the channel updated, base station acquires $\mathcal{N}_k^G$ from all the users and calculate $\mathcal{N}_k^{tot}$. The system state is updated to $s'$.
5. $ep(d) = \{s, \mathbf{a}, w(s, \mathbf{a}, s'), s'\}$. $d = d + 1$. If $D$ reaches the maximum of experience pool, $D$ remain constant, $d = 1$, otherwise, $D = d$. $s = s'$. $t = t + 1$..
6. After experience pool accumulates enough data, from $D$ experiences, randomly select $D_s$ experiences to train the neural network $eval\_net$. Back-propagation method is applied to minimize the loss function Loss($\theta$). Clone the weight parameters from $eval\_net$ to $target\_net$ after several time intervals.
7. IF $s' = s_T$:
      $s = s_0$. $t = 1$. $u = u+1$. If $u = U$, algorithm terminates, otherwise, go back to step 2.
   ELSE:
      go to step 3.
```

In this section, Deep Q-Network is applied to solve the proposed optimization. The main idea of DQN is to train a neural network to find the cost function(Q function) of a particular system state and action combination. When the system is in state $\mathbf{s}$, and action $\mathbf{a}$ is selected, the Q function is denoted as $Q(\mathbf{s}, \mathbf{a}, \theta)$. θ denotes the parameters of the Q network. The purpose of training the neural network is to make:

$$Q(\mathbf{s}, \mathbf{a}, \theta) \approx Q^*(\mathbf{s}, \mathbf{a}) \tag{25}$$

According to the DQN algorithm [12], two neural networks are used to solve the problem: the evaluation network and the target network, which are denoted as $eval\_net$ and $target\_net$, respectively. Both the $eval\_net$ and the $target\_net$ are set up with several hidden layers. The input of the $eval\_net$ and the $target\_net$ are denoted as s and s', which describe the current system state s and the next system state $\mathbf{s}'$, respectively. The output of $eval\_net$ and $target\_net$ are denoted as $Q_e(\mathbf{s}, \mathbf{a}, \theta)$ and $Q_t(\mathbf{s}, \mathbf{a}, \theta')$, respectively. The evaluation network is continuously trained to update the value of θ, however, the target network only copy the weight

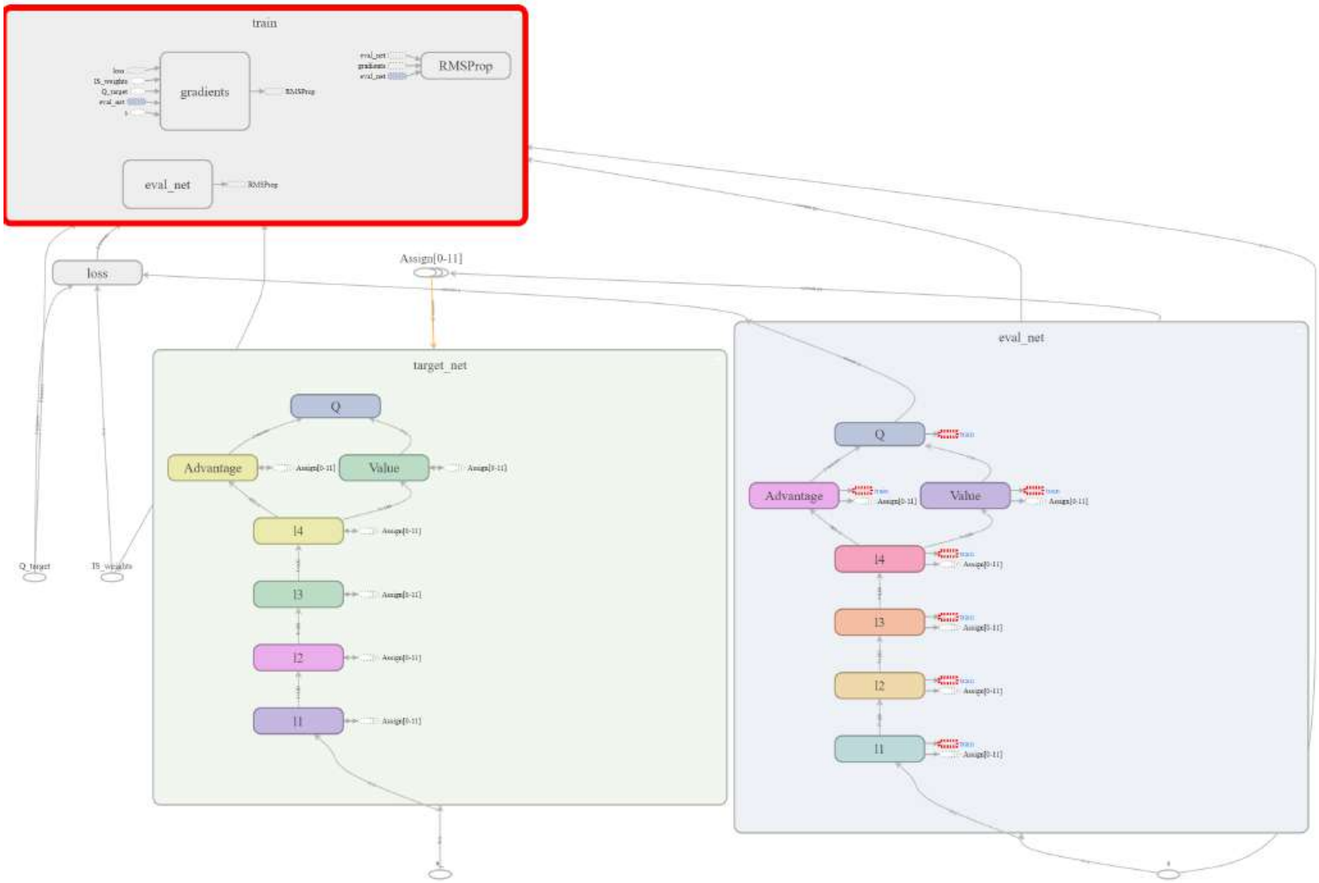


Fig. 2. The framework of Deep Q-Network.

parameters from the evaluation network intermittently (i.e., $\theta' = \theta$). In each neural network learning epoch, the loss function is defined as:

$$\text{Loss}(\theta) = E\left[(y - Q_e(\mathbf{s}, \mathbf{a}, \theta))^2\right]. \tag{26}$$

y represents the real Q value, and is is calculated as:

$$y = w(\mathbf{s}, \mathbf{a}, \mathbf{s}') + \gamma \max_{\mathbf{a}' \in \mathcal{A}} Q_t(\mathbf{s}', \mathbf{a}', \theta') \tag{27}$$

where γ is the reward discount. As the loss function updates, the values are back-propagated to the neural network to update the weight of the $eval\_net$.

In order to better train the neural network, we apply the experience reply method to remove the correlation between different training data. Each experience consists of the current system state s, the action $\mathbf{a}$, the next system state $\mathbf{s}'$, and the corresponding reward $w(\mathbf{s}, \mathbf{a}, \mathbf{s}')$. The experience is denoted by the set

$$ep = \{\mathbf{s}, \mathbf{a}, w(\mathbf{s}, \mathbf{a}, \mathbf{s}'), \mathbf{s}'\} \tag{28}$$

The algorithm records D experiences and randomly select $D_\mathbf{s}$ (with $D_\mathbf{s} < D$) experiences from D for training. $D_{ini} = 10000$. After the training is finished, $target\_net$ clones all the weight parameters from the $eval\_net$ (i.e., $\theta^{'} = \theta$).

The algorithm used for the DQN training process is presented in Algorithm 1. In the algorithm, we define in each training iteration, we generate D usable experiences ep, and select D_s of all for training the $eval\_net$. In total, we suppose there are U training iterations. We consider that for both the $eval\_net$ and the $target\_net$, there are N_l layers in the neural network. The structure of DQN is shown in Fig. 2.

V. Simulation Results

We conduct the experiment with the National Instruments/Ettus Research Universal Software Radio Peripheral(USRP) N210 with the CBX daughterboard in order to measured the indoor channel. The number of the mobile user is $K = 3$. The total number of available subchannels is $N = 16$. Each subchannel has same bandwidth $W = 10^3$Hz. The users are randomly distributed around the base station and remain stationary throughout the whole T data transmission time slot. The channel gains from the base station to all users are approximately in range $[-80, -60]$dB. The path loss exponential is $\beta = 2$. The noise power spectrum density is $N_0 = 170$ dBm/Hz. $b = 9$ bits are transmitted by each symbol on each subchannel. The channel coherence time is $d_c = 1$ms. The subchannel selection threshold is $\hat{p}_s = 10^{-7}$. The required delivered data for all users are same: $B_1 = ... = B_K = B = 450$ bits. $p_k(t)$ is selected from $[0.3, 0.2, 0.15, 0.1, 0.075, 0.05, 0.0375, 0.025, 0.0125, 0.00625]$

TABLE I
DQN SIMULATION PARAMETERS

Deep Q Network	Value
Number of hidden layers (N_L)	4
Number of nodes of each hidden layer	100
Learning rate (ϵ)	0.00005
Mini-batch size	10
Training starting step	10000
Experience pool	60000
Initial exploration rate(ϵ_c)	1
Final exploration rate(ϵ_c)	0.1
Exploration interval	0.001
$target\_net$ weight replacement interval	500
Training episodes	60000

mW. $\sum_{n=1}^{K} p_k(t) \leq P$. $P = 0.3$mW. The energy consumption constraint $\hat{E}$ are regulated as 3.9×10^{-4} J and 4.5×10^{-4} J.

The simulation parameters used for DQN are presented in Table I. The exploration rate ϵ_c decreases from 1 to 0.1 with 0.001 interval. The software environment for simulation is TensorFlow 0.12.1 with Python 3.6 in Jupyter Notebook 5.6.0.

In Fig. 3, the time consumption T and moving average of energy constraint satisfactory $\bar{\eta}$ are observed throughout the learning process. An training episode lasts T time slots, an episode is over if B bits data are successfully received at each user. If at the end of the eth episode, $\sigma(T) = 1$, we define $\eta(e) = 1$, otherwise, $\eta(e) = 0$.

$\bar{\eta}$ is the moving average of $\eta(e)$ in $e = 100$ episodes.

$$\bar{\eta} = \frac{\eta(e-99) + ... + \eta(e)}{100} \tag{29}$$

Higher $\bar{\eta}$ denotes the higher probability of satisfactory of energy consumption constraint. We observe that as the training goes on, both T and $\bar{\eta}$ converges. With lower energy consumption constraint $\hat{E}$, more time is consumed to finish data transmission.

We apply $N_t = 1000$ test data to test the performance of DQN trained with differnt reward discount γ. Of N_t test data, N_s data satisfy the energy consumption constraint $\hat{E}$. ζ_E is defined as:

$$\zeta_E = \frac{N_s}{N_t} \tag{30}$$

$\bar{T}$ is the average time consumption of N_t test data. If $\gamma = 0$, the myopic solution is learnt. The strategy only maximize the immediate reward at each particular system state without considering system dynamics. If $\gamma = 1$, each system state are equally considered. Both $\gamma = 0$ and $\gamma = 1$ don't achieve good learning effect. Fig. 4 shows that when $\gamma = 0.75$, the base station consumes least time to successfully deliver required data to all users, while has the highest probability of energy constraint satisfactory. Hence, the optimal reward discount is $\gamma = 0.75$.

In fig. 5 and fig. 6, the performance of Deep Q-Network is compared with the state of the art. For Maximum power transmission method, we equally allocate 0.1 mW power to each user and continuously transmit with that strategy until the required data are delivered to all the users. For Minimum power transmission method, each user is allocated with 0.0125 mW power. Random transmission need to randomly select an

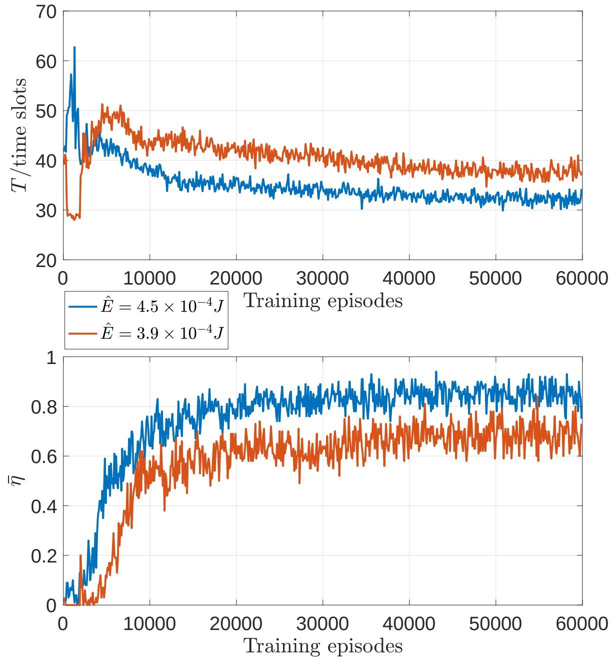


Fig. 3. The convergence on time consumption T and $\bar{\eta}$(moving average of η) in Deep Q-Network training process. $\gamma = 0.75$.

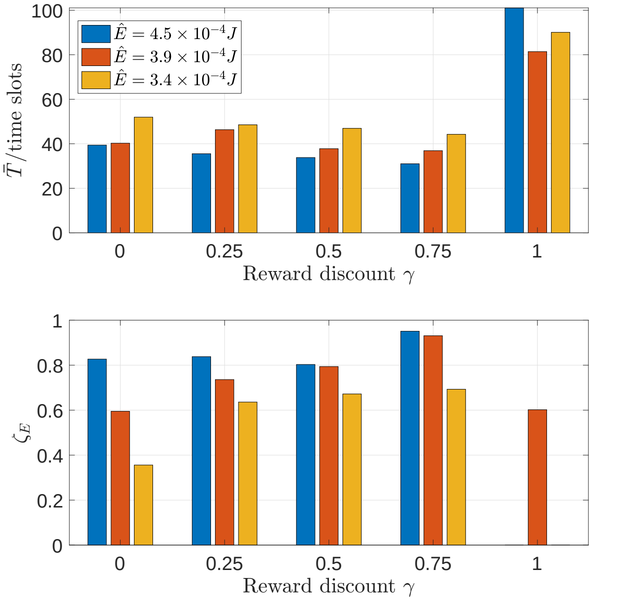


Fig. 4. The performance comparison of average time consumption $\bar{T}$ and energy consumption constraint satisfactory probability ζ_E on reward discount γ.

action in each time slot. We allow the energy consumption constraints to vary. We can observe that Deep Q-Network outperforms all the other four algorithms. Maximum and Minimum power transmission can only guarantee a good performance on either time consumption and constraint satisfactory. The myopic solution doesn't consider the long term reward, hence achieves bad performance on both time consumption and energy consumption. Compared with DQN, the random transmission can achieve similar energy consumption performance, however consumes more time to finish data dissemination. Deep Q-Network is proved to learn the optimal solution.

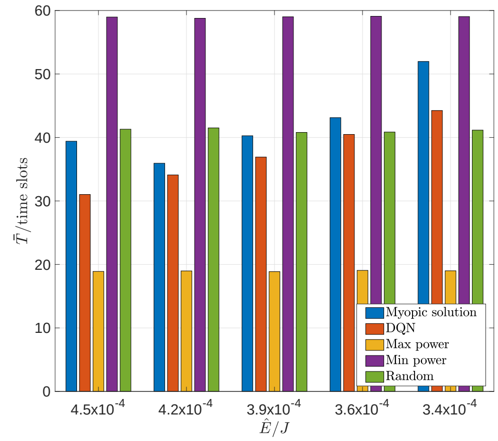


Fig. 5. The comparison of time consumption between Deep Q-Network and other algorithms on energy consumption constraint $\hat{E}$.

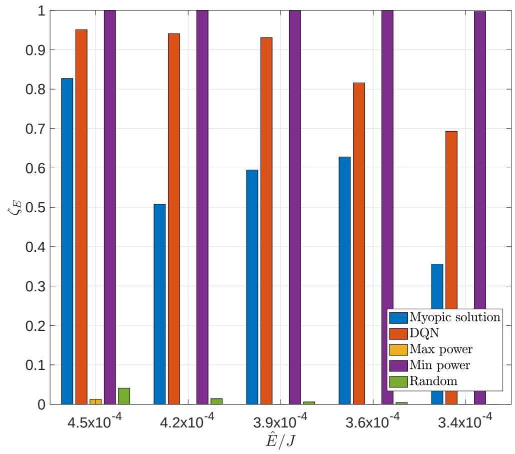


Fig. 6. The comparison of energy consumption between Deep Q-Network and other algorithms on energy consumption constraint $\hat{E}$.

VI. Conclusion

The multiple users data dissemination problem is explored in a OFDMA downlink system. The system aims to minimize the data transmission time under the total energy consumption constraint. In order to solve this global optimization in real-time, a Deep Q-Network method is proposed to determine the optimal transmission strategy in a real-time. Without the prior knowledge about the channel information, the Deep Q-Network can learn the variation of the wireless channel and dynamically provide the resource allocation strategy for the base station. Compared with the state of the art, the proposed Deep Q-Network achieves the best system performance on both time and energy consumption.

References

[1] G. Wunder and T. Michel, "Optimal resource allocation for parallel gaussian broadcast channels: minimum rate constraints and sum power minimization," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4817–4822, 2007.

[2] C. C. Zarakovitis, Q. Ni, and J. Spiliotis, "Energy-efficient green wireless communication systems with imperfect csi and data outage," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3108–3126, 2016.

[3] C. Xiong, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient resource allocation in ofdma networks," *IEEE Transactions on Communications*, vol. 60, no. 12, pp. 3767–3778, 2012.

[4] L. Dong, "Spectral-and energy-efficient transmission over frequency-orthogonal channels," in *2016 IEEE Online Conference on Green Communications (OnlineGreenComm)*. IEEE, 2016, pp. 13–20.

[5] Y. Li, M. Sheng, C. W. Tan, Y. Zhang, Y. Sun, X. Wang, Y. Shi, and J. Li, "Energy-efficient subcarrier assignment and power allocation in ofdma systems with max-min fairness guarantees," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3183–3195, 2015.

[6] H. Li, T. Lv, and X. Zhang, "Deep deterministic policy gradient based dynamic power control for self-powered ultra-dense networks," in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–6.

[7] J. K. Lenstra and A. R. Kan, "Complexity of vehicle routing and scheduling problems," *Networks*, vol. 11, no. 2, pp. 221–227, 1981.

[8] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y.-C. Liang, "Intelligent resource scheduling for 5g radio access network slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691–7703, 2019.

[9] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 2137–2145.

[10] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud rans," in *Proc. of IEEE ICC*. IEEE, 2017, pp. 1–6.

[11] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.

[12] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning." in *AAAI*, vol. 2. Phoenix, AZ, 2016, p. 5.

Numerical and Simulation Verification for Optimal Server Allocation in Edge Computing

Dawei Li*, Chigozie Asikaburu*, Jiacheng Shang*, and Ning Wang†
*Department of Computer Science, Montclair State University, Montclair, NJ, USA
†Department of Computer Science, Rowan University, Glassboro, NJ, USA

dawei.li@montclair.edu, asikaburug1@montclair.edu, shangj@montclair.edu, and wangn@rowan.edu

***Abstract*—In this paper, we consider the server allocation problem in edge computing. We consider a system model where there are a number of areas or locations, each of which has an associated Base Station (BS), where we can deploy an edge cloud with multiple servers. Each edge cloud will process application requests received at the corresponding BS from users in the corresponding area. The system manager/operator has a budget to deploy a given number of servers to the BSs. Our goal is to come up with a server allocation plan, i.e., how many servers to deploy at each of the BSs, such that the overall average turnaround time of application requests generated by all the users is minimized. In order to achieve the optimal solution for the problem, we resort to queueing theory and model each edge cloud as an M/M/c queue. Analysis on the problem motivates a Largest Weighted Reduction Time First (LWRTF) algorithm to assign servers to edge clouds. Numerical comparisons among various algorithms verify that *Algorithm LWRTF* has near-optimal performances in terms of minimizing the average turnaround time. Simulation results using the CloudSim Plus simulation tool also verify that *Algorithm LWRTF* achieves better performances compared to other reasonably designed heuristic algorithms.**

***Index Terms*—Edge computing, edge cloud, queueing theory, numerical method, simulation approach.**

I. Introduction

Facilitated by emerging communication, networking, and information technologies, such as 5G [1], Internet of Things (IoT) [2], [3], Artificial Intelligence (AI) [4], etc, our current era is witnessing a proliferation of services and applications that are provided at the edge of the Internet. These services and applications include augmented reality [5], virtual reality [6], cognitive assistance [7], gesture recognition [8], mobile gaming [9], and online social networks [10], [11]. All these applications can be computation-intensive and/or communication-intensive, increasing the requirements for various resources from mobile devices. However, mobile devices (i.e., mobile phones, tablets, IoT devices, etc.) are still limited in terms of storage size, computation capacity, communication bandwidth, and battery life. Cloud computing has been applied to execute the applications offloaded from the mobile devices. However, various issues arise due to long latency and limited bandwidth from mobile devices to remote clouds.

Recently, the *edge computing* paradigm [12], [13], which extends cloud computing to the edge of the Internet, has been proposed to support applications originating from the edge of the Internet. By running applications on relatively small computing systems deployed at the Internet edge, edge computing allows users to exploit computing power outside of the mobile devices without incurring long access delays to remote clouds. Several other terms have been used, such as fog computing, cloudlets, edge-centric computing, mobile edge computing, etc [5], [14]–[17]. In our discussion, we will consistently refer to them as *edge computing*. The small computing systems deployed at the Internet edge will be called *edge clouds*. In our discussion, the mobile devices are only treated as clients and will not help other mobile devices process tasks.

In edge computing, a fundamental problem is the server allocation problem, which has been considered by various existing works [18]–[21]. In most existing works, the system model for an edge cloud is largely simplified. In [21], the numbers of servers in all the edge clouds are fixed and equal to each other. In [19] and [18], the numbers of servers in the edge clouds are not the same; however, they are still fixed; the problem is to derive the edge cloud placement given some capacity constraints to minimize the edge cloud access delay, while the queueing time and execution time of application requests are not considered. In this paper, we assume that the number of servers at an edge cloud is flexible, and we want to arrive at the optimal server allocation plan so that the average turnaround time for all application requests is minimized.

Our main contributions in this paper are as follows. We consider the server allocation problem using queueing theory; each edge cloud is modeled as an M/M/c queueing system. We design a Largest Weighted Reduction Time First (LWRTF) algorithm to allocate servers to edge clouds. Numerical comparisons among various heuristic algorithms verify that *Algorithm LWRTF* has near-optimal performances in terms of minimizing the average turnaround time. We demonstrate the accuracy and validity of a simulation based approach (using the CloudSim Plus simulation tool [22]) by comparing the numerical results from queueing theory and the simulation results. Simulation results using the CloudSim Plus simulation tool also verify that *Algorithm LWRTF* achieves the best performance compared to other reasonably designed heuristics.

The rest of the paper is organized as follows. Section II presents the system model. In Section III, we present some preliminary analyses on the server allocation problem. In Section IV, we propse the Largest Weighted Reduction Time

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

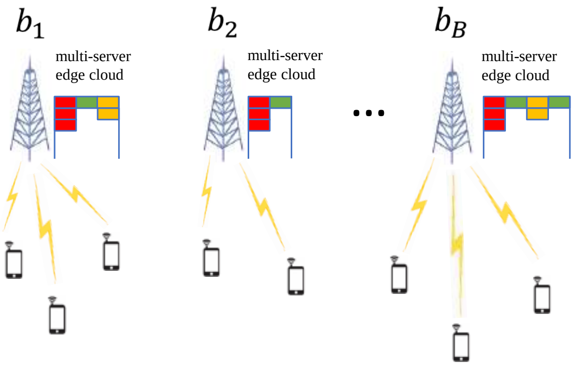


Fig. 1. System model with multi-server edge clouds.

TABLE I
IMPORTANT NOTATIONS USED IN THE PAPER

Notations	Meaning
N	total number of servers
B	total number of BSs
b_i	the ith BS
t_e	application request's average execution time on a server
μ	request departure rate on a server; $\mu = 1/t_e$
T	time period under discussion
m_i	the number of requests generated in b_i during time period T
λ_i	request arrival rate at b_i; $\lambda_i = m_i/T$
$\boldsymbol{\lambda}$	the request arrival rate vector for the B BSs
n_i^{min}	the minimum number of servers that must be deployed at b_i
n_i	the number of servers deployed at the ith edge cloud
$\boldsymbol{n}$	the server allocation vector
$t_{s,i}$	the average turnaround time for application requests at b_i
t	the average turnaroudn time for all application requests

First (LWRTF) algorithm in detail. In Section V, we introduce several reasonably designed heurisitics for the same problem. Comparisions among the LWRTF algorithm and other heuristic algorithms using both numerical methods and high-fidelity simulations are presented in Section VI. We conclude our paper and give future work directions in Section VII.

II. SYSTEM MODEL

We denote the set of all BSs in the target region as $\mathcal{B} = \{b_1, b_2, \cdots, b_B\}$, where B is the cardinality of the set, i.e., the total number of BSs. We can deploy an edge cloud at each BS to serve the area/location covered by that BS. For notational convenience, we will also use b_i to refer to the edge cloud deployed at the BS, the location of the BS, and the user area covered by the BS. The system model with multi-server edge clouds is shown in Fig. 1.

We assume that all the servers in the system have the same execution speed, and that all the application requests/tasks have an average execution time of t_e following the exponential distribution. Equivalently, when the requests are executed on an edge server, the depature rate of the requests can be calcualted as $\mu = 1/t_e$.

During a time period of T, m_i application requests are generated in user area b_i and will be offloaded to the edge cloud for processing. Assume that the application request arrivals follow a Poisson distribution. We can calculate the application request arrival rate at b_i as $\lambda_i = m_i/T$. The *turnaround time* for a given application request consists of the queueing time and the execution time of the request at the corresponding edge cloud.

The service provider has a budget to deploy N servers across all the edge clouds. Let n_i be the number of servers to be deployed at b_i, $\forall i = 1, 2, \cdots, B$. Since deploying more servers will definitely help to reduce the average turnaround time, we will deploy all the N servers, i.e., $\sum_{i=1}^{B} n_i = N$. The service provider wants to derive a server allocation plan in order to minimize the average turnaround time for all application requests generated within the target region. Table I lists the important notations used in the paper; some of them will be introduced later.

III. PRELIMINARIES

Given a valid N, we aim to derive an optimal server allocation plan to minimize the average turnaround time of all application requests. A naive approach is allocate the servers such that the numbers of servers of the edge clouds are proportional to their request arrival rates. However, we find that this approach can result in very poor performances in terms of minimizing the average turnaround time for all the application requests. Besides, we find that other reasonably designed heuristics may be unstable and result in poor performances in many situations. This movitates us to delve deeper into the problem. We resort to queuing theory in order to derive the optimal solution for the problem.

The edge cloud at a BS is modeled as an M/M/c queue, i.e., the Erlang C model [23]. Let $\rho_i = \frac{\lambda_i}{n_i \mu}$. The probability that an arriving application request is forced to join a queue at b_i (i.e., all n_i servers are currently occupied) is given by:

$$\begin{aligned} C(n_i, \lambda_i/\mu) &= \frac{\frac{(n_i\rho_i)^{n_i}}{n_i!}\frac{1}{1-\rho_i}}{\sum_{k=0}^{n_i-1}\frac{(n_i\rho_i)^k}{k!} + \frac{(n_i\rho_i)^{n_i}}{n_i!}\frac{1}{1-\rho_i}} \\ &= \frac{1}{1+(1-\rho_i)\frac{n_i!}{(n_i\rho_i)^{n_i}}\sum_{k=0}^{n_i-1}\frac{(n_i\rho_i)^k}{k!}}, \end{aligned} \tag{1}$$

which is the Erlang C formula. The average turnaround time (consisting of the queueing time and the execution time) of requests that are processed at b_i is

$$t_{s,i} = \frac{C(n_i, \lambda_i/\mu)}{n_i\mu - \lambda_i} + \frac{1}{\mu}. \tag{2}$$

Thus, the average turnaround time of all application requests generated within the entire region is:

$$t = \frac{\sum_{i=1}^{B}(\lambda_i t_{s,i})}{\sum_{i=1}^{B}\lambda_i}. \tag{3}$$

Besides, for the queueing system to be stable at b_i, we must have $n_i\mu - \lambda_i > 0$. Let

$$n_i^{min} = \lfloor \lambda_i/\mu \rfloor + 1, \forall i = 1, 2, \cdots, B. \tag{4}$$

Here, n_i^{min} is the minimal number of servers that must be allocated to edge cloud b_i.

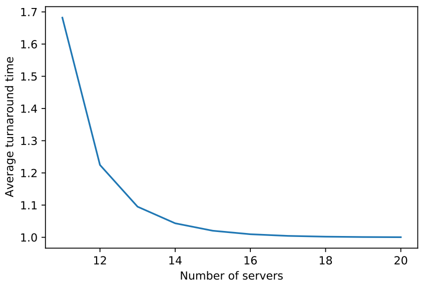


Fig. 2. Average turnaround time with respect to the number of servers.

The optimization problem can be formulated as follows:

$$\begin{aligned} \underset{n_i, \forall i=1,\cdots,B}{\text{minimize}} \quad & t \\ \text{subject to} \quad & n_i \geq n_i^{min}, \forall i = 1, \ldots, B. \\ & \sum_{i=1}^{B} n_i = N. \\ & \text{Equations (1), (2), and (3).} \end{aligned}$$

The complexity of the above optimization problem lies in the following aspects. First, it is an Integer Programming problem with an exponentially large solution search space; these kinds of problems are generally NP-Hard. Second, the closed-form expression of the objective function is quite complex, resulting from the complexity of the Erlang C formula; the fact that $C(n_i, \lambda_i/\mu)$ involves factorial and exponentiation operations on the solution variables (n_i's) makes various relaxation-based convex optimization techniques inapplicable here.

It is intuitive and easy to verify that when the application request arrival rates at the BSs are equal to each other, the server allocations at the BSs should be balanced. In other words, the number of servers at one location will not differ from that at another location by more than 1. The balanced server allocation for the case with equal request arrival rates has been verified in our previous works [24]–[26].

Next, we consider the general case where the request arrival rates at the BSs are not necessarily equal to each other. For each queue at each edge cloud to be stable, the number of servers deployed at b_i should satisfy the following constraint,

$$n_i \geq n_i^{min} = \lfloor \lambda_i/\mu \rfloor + 1, \forall i = 1, 2, \cdots, B. \qquad (5)$$

We must have $N \geq \sum_{i=1}^{B} n_i^{min}$. Any valid server allocation plan should allocate at least n_i^{min} servers at b_i. Thus, the server allocation problem becomes how to allocate the remaining $N - \sum_{i=1}^{B} n_i^{min}$ servers to the B edge clouds.

IV. Algorithm LWRTF

We design a Largest Weighted Reduction Time First (LWRTF) strategy to assign the remaining servers to the edge clouds one by one. Our design is motivated by how the average turnaround time changes with respect to the number of servers in a multi-server queueing system, as shown in Fig. 2. In Fig. 2, we set $\mu = 1$ and $\lambda = 10$. When we have 10 servers, the system is unstable, i.e., the average turnaround time is infinite. For the queueing system to be stable, we should have at least 11 servers. When the number of servers is small, adding one server will result in a large reduction in the average turnaround time. When the number of servers is large, the average turnaround time will be already close to 1, and adding more servers will provide minimal reduction in the average turnaround time. This indicates that when we consider allocating servers to the edge clouds, we should take into consideration the amount of reduction in the average turnaround time. Intuitively, we should allocate our next server to the edge cloud that can achieve a large reduction in the average turnaround time. For the M/M/c queueing system at b_i, adding one server to the system will reduce the average turnaround time. The reduction time can be calculated as follows:

$$\Delta t_{s,i} = \frac{C(n_i, \lambda_i/\mu)}{n_i\mu - \lambda_i} - \frac{C(n_i + 1, \lambda_i/\mu)}{(n_i + 1)\mu - \lambda_i}. \qquad (6)$$

Algorithm LWRTF

Require: N: the total number of servers; B: the number of BSs; t_e: application request's average execution time; m_i: the number of applicaiton requests at b_i, $\forall \lambda_i, i = 1, \cdots, B$, during time period T.

Ensure: **n** = $(n_1, \cdots, n_B)$: the server allocation vector.

```
 1: $\mu = 1/t_e$;
 2: $\lambda_i = m_i/T$;
 3: for $i = 1, \cdots, B$ do
 4:     $n_i^{min} = \lfloor \lambda_i/\mu \rfloor + 1$;
 5:     $n_i = n_i^{min}$;
 6:     Calculate $\Delta t_{s,i}$ according to Equation (6);
 7: if $N < \sum_{i=1}^{B} n_i^{min}$ then
 8:     return null;
 9: for $j = 1, \cdots, N - \sum_{i=1}^{B} n_i^{min}$ do
10:     $idx = \underset{i}{\text{argmax}}\{\lambda_i \Delta t_{s,i} | i = 1, \cdots, B\}$;
11:     $n_{idx} = n_{idx} + 1$;
12:     Update $\Delta t_{s,idx}$ according to Equation (6) with updated
        $n_{idx}$;
13: return n;
```

Besides, different edge clouds have different request arrival rates, λ_i's. A larger λ_i will have a more significant influence on the overall average turnaround time given the same amount of reduction in the average turnaround time. Thus, our strategy is to assign the server to the edge cloud with the largest $\lambda_i \Delta t_{s,i}$ value among all the edge clouds.

The detailed procedures are described in **Algorithm** LWRTF. The algorithm will first assign n_i^{min} servers to edge cloud b_i, and calculate $\Delta t_{s,i}$ in a for loop (lines 3 to 6). The algorithm returns null if it finds the input invalid (line 7). Line 10 gets the index of the edge cloud with the largest weighted reduction time. Line 11 assigns the next server to the corresponding edge cloud. Line 12 updates the reduction time after the server has been assigned to the edge cloud.

```
Algorithm Proportional
Require: $N$: the total number of servers; $B$: the number of BSs; $t_e$: application request's average execution time; $m_i$: the number of applicaiton requests at $b_i$, $\forall \lambda_i, i = 1, \cdots, B$, during time period $T$.
Ensure: n = $(n_1, \cdots, n_B)$: the server allocation vector.
1: $\mu = 1/t_e$;
2: $\lambda_i = m_i/T$;
3: for $i = 1, \cdots, B$ do
4:     $n_i^{min} = \lfloor \lambda_i/\mu \rfloor + 1$;
5:     $n_i = n_i^{min}$;
6: if $N < \sum_{i=1}^{B} n_i^{min}$ then
7:     return null;
8: for $j = 1, \cdots, N - \sum_{i=1}^{B} n_i^{min}$ do
9:     $idx = \underset{i}{\text{argmax}}\{\lambda_i N/(\sum_{i=1}^{B} \lambda_i) - n_i | i = 1, \cdots, B\}$;
10:    $n_{idx} = n_{idx} + 1$;
11: return n;
```

After all the remaining servers have been assigned, the server allocation vector, $\boldsymbol{n}$ will be returned.

It is worth noting that when the application request arrival rates, λ_i's are equal to each other, *Algorithm LWRTF* will also generate a balanced server allocation, which is the optimal solution for this special case. This demonstrates that *Algorithm LWRTF* is applicable to both special cases and general cases.

V. Comparing Heuristics

In this section, we explore other reasonable heuristic algorithms that also try to minimize the average turnaround time for all applicaiton requests. All the heuristics will first allocate n_i^{min} servers to the ith edge cloud. After that, they use different methods to assign the remaining $N - \sum_i^B n_i^{min}$ servers to the selected edge clouds one by one.

Algorithm Proportional: This algorithm tries to achieve a server allocation where the number of servers of edge clouds are proportional to their request arrival rates. That is, n_i should be close to $\lambda_i N/(\sum_{i=1}^{B} \lambda_i)$. The algorithm will assign the next server to the edge cloud with the largest $\lambda_i N/(\sum_{i=1}^{B} \lambda_i) - n_i$ value among all edge clouds. The details of the algorithms are presented in **Algorithm** Proportional.

Algorithm Largest Server Shortage Severeness First (LSSSF): The server shortage severeness of an edge cloud is defined as $1/(n_i\mu - \lambda_i)$. This algorithm assigns the next server to the edge cloud with the largest $1/(n_i\mu - \lambda_i)$ value among all edge clouds. The intuition is to assign the next server to the edge cloud that has the severest shortage of servers. Algorithm LSSSF only needs to replace line 9 in **Algorithm** Proportional with "$idx = \underset{i}{\text{argmax}}\{1/(n_i\mu - \lambda_i) | i = 1, \cdots, B\}$;".

Algorithm Largest Weighted Server Shortage Severeness First (LWSSSF): This method adds the weight λ_i to the server shortage severeness. It assigns the next server to the edge cloud with the largest $\lambda_i/(n_i\mu - \lambda_i)$ value among all edge clouds. Algorithm LWSSSF only needs to replace line 9 in **Algorithm** Proportional with "$idx = \underset{i}{\text{argmax}}\{\lambda_i/(n_i\mu - \lambda_i) | i = 1, \cdots, B\}$;".

Algorithm Largest Reduction Time First (LRTF): This method is the simpler case of *Algorithm LWRTF*. It assigns the next server to the edge cloud with the largest $\Delta t_{s,i}$ value among all edge clouds. Algorithm LRTF only needs to repace line 10 in **Algorithm** LWRTF with "$idx = \underset{i}{\text{argmax}}\{\Delta t_{s,i} | i = 1, \cdots, B\}$;".

VI. Evaluations

A. Numerical Comparisons Among Different Server Allocation Algorithms

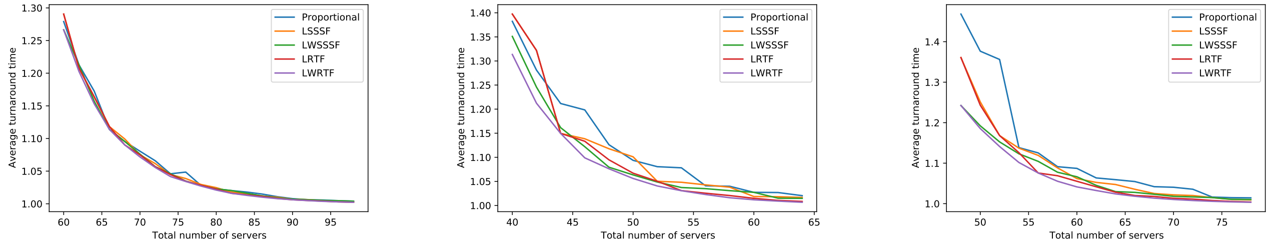
In this subsection, we use numerical results from queueing theory to verify that *Algorithm LWRTF* acheives the best performances in terms of minimizing the average turnaround time compared to other reasonably designed heuristics.

In the first numerical comparison setting, we let $\mu = 1$ and $B = 8$. We randomly generate floating point λ_i values within the range of $[1, 10)$. In our second comparison setting, we let $\mu = 0.1$ and $B = 8$. We randomly generate floating point λ_i values within the range of $[0.1, 1)$. In both settings, we generate the total numbers of servers in the following way. For a given request arrival rate vector $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_B)$, we vary the number of servers, N, within the range of $[\lceil 1.2 \sum_{i=1}^{B} n_i^{min} \rceil, 2 \sum_{i=1}^{B} n_i^{min}]$, with a step size of 2. We start with $\lceil 1.2 \sum_{i=1}^{B} n_i^{min} \rceil$, instead of $\sum_{i=1}^{B} n_i^{min}$, because doing so can avoid the cases where all the algorithms generate the same server allocation with very large average turnaround time, making their comparisons in other configurations hard to be observed.

Fig. 3 shows the results for the first comparison setting. Fig. 3(a), Fig. 3(b), and Fig. 3(c) show the comparisons with request arrival rate vectors of $\boldsymbol{\lambda}_a^1$, $\boldsymbol{\lambda}_b^1$, and $\boldsymbol{\lambda}_c^1$, respectively, where $\boldsymbol{\lambda}_a^1 = (7.82, 7.21, 8.05, 3.96, 3.46, 4.70, 4.98, 6.54)$, $\boldsymbol{\lambda}_b^1 = (3.90, 2.32, 2.06, 7.82, 1.35, 1.51, 3.13, 6.84)$, and $\boldsymbol{\lambda}_c^1 = (4.15, 2.05, 1.25, 9.54, 2.78, 8.91, 1.81, 5.44)$. The variances of $\boldsymbol{\lambda}_a^1$, $\boldsymbol{\lambda}_b^1$, and $\boldsymbol{\lambda}_c^1$ are 2.80, 5.27, and 9.08, respectively. $\boldsymbol{\lambda}_a^1$, $\boldsymbol{\lambda}_b^1$, and $\boldsymbol{\lambda}_c^1$ represent the cases with a relatively low variance, medium variance, and large variance, respectively.

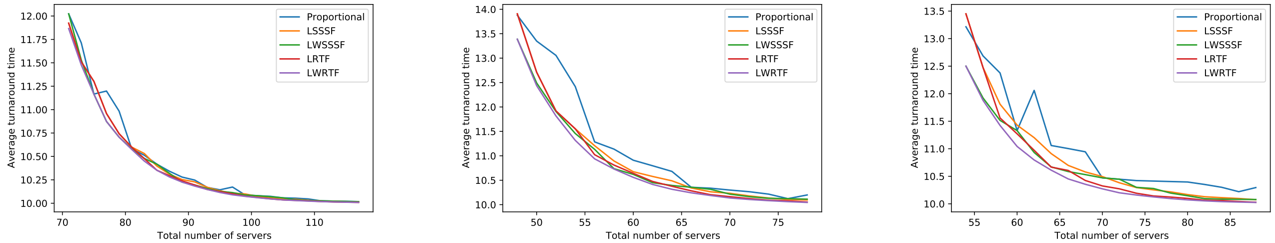
Fig. 4 shows the results for the second comparison setting. Fig. 4(a), Fig. 4(b), and Fig. 4(c) show the comparisons with request arrival rate vectors of $\boldsymbol{\lambda}_a^2$, $\boldsymbol{\lambda}_b^2$, and $\boldsymbol{\lambda}_c^2$, respectively, where $\boldsymbol{\lambda}_a^2 = (0.53, 0.83, 0.61, 0.96, 0.79, 0.32, 0.55, 0.89)$, $\boldsymbol{\lambda}_b^2 = (0.28, 0.90, 0.70, 0.14, 0.17, 0.46, 0.39, 0.68)$, and $\boldsymbol{\lambda}_c^2 = (0.22, 0.28, 0.17, 0.65, 0.96, 0.12, 0.98, 0.79)$. The variances of $\boldsymbol{\lambda}_a^2$, $\boldsymbol{\lambda}_b^2$, and $\boldsymbol{\lambda}_c^2$ are 0.0413, 0.0654, and 0.1156, respectively. $\boldsymbol{\lambda}_a^2$, $\boldsymbol{\lambda}_b^2$, and $\boldsymbol{\lambda}_c^2$ represent the cases with a relatively low variance, medium variance, and large variance, respectively.

All the numbers reported here have been trimmed without affecting meaningful comparisons. According to the results, generally, when the relative variance of $\boldsymbol{\lambda}$ is small, all the algorithms have close performances. The performance differences among the algorithms become clearer when the variance of $\boldsymbol{\lambda}$ increases. In all evaluated configurations, not all of which have been included in the paper, *Algorithm LWRTF* results in the



(a) $\boldsymbol{\lambda}_a^1$ (b) $\boldsymbol{\lambda}_b^1$ (c) $\boldsymbol{\lambda}_c^1$

Fig. 3. Average turnaround time comparisons among the five heuristic algorithms (first comparison setting).



(a) $\boldsymbol{\lambda}_a^2$ (b) $\boldsymbol{\lambda}_b^2$ (c) $\boldsymbol{\lambda}_c^2$

Fig. 4. Average turnaround time comparisons among the five heuristic algorithms (second comparison setting).

most stable performances and always achieves the minimum average turnaround time.

B. Validation for the Simulation Based Approach

In this subsection, we validate the accuracy of the CloudSim Plus [22] simulation platform. CloudSim Plus, an independent fork of CloudSim [27], is a discrete-event simulation tool designed for modeling and simulating cloud computing infrastructures and services. It defines many classes/interfaces for entities in the cloud computing environments, including *Datacenter*, *Host*, *Vm*, and *Cloudlet* (the interface/class that corresponds to an application request). It also defines the interfaces/classes for common concepts/techniques used in cloud computing, such as *CloudInformationServices* (serving as a registry of all resources in a cloud), DatacenterBroker (acting as a cloud customer that accepts application requests and submits them to the cloud), *VmAllocationPolicy* (used by the data center to allocate hosts for VMs), *VmSchedulingPolicy* (defining how a host schedules the VMs assigned to it), *CloudletScheduler* (defining how a VM schedules the cloudlets, i.e., application requests, assigned to the VM), etc. For most Java interfaces, the simulation tool provides some basic implemtations, and makes implementations of customized classes quite convenient. The simulation tool can also be easily tweaked and modified to provide high-fidelity simulations for edge computing.

To validate the accuracy of the CloudSim Plus simulation tool, we first generate random sets of application requests. The application requests are expected to have an average execution time of 10 seconds. Equivalently, the average departure rate of the requests when executed on a server is expected to be $\mu = 0.1$ (tasks/second). In other words, every second, it is expected that a server can finish 0.1 application requests. Notice that the actual average execution time and average departure rate of the reqeusts may slightly differ from the expected values, since the requests are randomly generated. We consider a time period of $T = 1000$ seconds. We conduct four groups of compharsions between the results from numerical methods and the simulations. We vary the numbers of requests as 350, 450, 550, and 650 in these four groups, respectively. Correspondingly, the request arrival rates for the four groups are $\lambda_1 = 0.35$, $\lambda_2 = 0.45$, $\lambda_3 = 0.55$, and $\lambda_4 = 0.65$, respectively. In each comparison group, we vary the number of servers that execute the requests. We consider 5 numbers of servers starting from $\lceil \lambda_i / \mu \rceil + 1$. Having smaller numbers of servers, for example, $\lceil \lambda_i / \mu \rceil$, will tend to make the queueing system unstable and result in very high turnaround time, which is not practal in the real world. For a specific scenario with a given set of random requests and the given number of servers, we first calculate the average turnaround time using the numerical method, i.e., using Equation (2), and then derive the average turnaround time through CloudSim Plus simulations.

The comparisons between numerical and simulation results are presented in Table II. We can tell from the table that, in most comparison scenarios (except for a few rare scenarios when the number of servers is too small), the numerical results and the simulation results are very close. We briefly touch upon some potential reasons for the differences between the numerical and simulation results. First, queueing theory

TABLE II
COMPARISONS BETWEEN NUMERICAL AND SIMULATION RESULTS

$\lambda_1 = 0.35$	number of servers	5	6	7	8	9
	average turnaround time (numerical)	12.29	10.58	10.11	9.97	9.93
	average turnaround time (simulation)	11.03	10.30	10.06	9.98	9.94
	$\lvert$simulation - numerical$\rvert$/numerical	10.25%	2.65%	0.49%	0.10%	0.10%
$\lambda_2 = 0.45$	number of servers	6	7	8	9	10
	average turnaround time (numerical)	13.08	11.00	10.40	10.20	10.12
	average turnaround time (simulation)	13.08	10.94	10.37	10.18	10.11
	$\lvert$simulation - numerical$\rvert$/numerical	0.00%	0.55%	0.29%	0.20%	0.10%
$\lambda_3 = 0.55$	number of servers	7	8	9	10	11
	average turnaround time (numerical)	12.80	10.88	10.28	10.06	9.98
	average turnaround time (simulation)	12.87	10.84	10.29	10.07	9.97
	$\lvert$simulation - numerical$\rvert$/numerical	0.55%	0.38%	0.10%	0.10%	0.10%
$\lambda_4 = 0.65$	number of servers	8	9	10	11	12
	average turnaround time (numerical)	13.52	11.26	10.54	10.26	10.15
	average turnaround time (simulation)	12.64	11.09	10.52	10.27	10.16
	$\lvert$simulation - numerical$\rvert$/numerical	6.51%	1.51%	0.19%	0.10%	0.10%

characterizes a system with an infinite time length, assuming that there will be an infinite number of requests coming and leaving all the time, which is different from the CloudSim Plus simulation, where we only consider one time period with duration $T = 1000$ seconds. We can imagine that when we simulate multiple time periods consecutively, the interferences between different time periods will come into play and bring the simulation results closer to the numerical results. Second, queueing theory assumes that $t_e = 1/\mu$ strictly follows the exponential distribution and λ_i's strictly follow the Poison distribution, while a specific set of requests, though randomly generated, will not follow these distributions perfectly. Despite these obvious reasons that may result in differences between the numerical results and simulations results, they are actually very close, as we can tell from the table. In most of the scenarios, the difference between the numerical and simulation results differ by less then 1%. This validates the reliability and accuracy of the simulation platform, and justifies using CloudSim Plus simulations to evaluate various server allocation algorithms.

C. Simulation Comparisons Among Different Server Allocation Algorithms

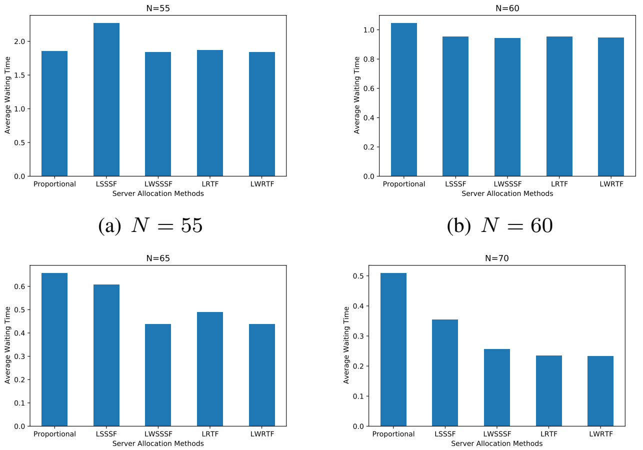
In this subsection, we use CloudSim Plus simulations to compare different server allocation algorithms. We design the simulation settings as follows. The number of BSs is $B = 8$. The average execution time of all requests is $t_e = 10$ seconds. Equivalently, $\mu = 0.1$. For CloudSim Plus simulations, we consider a time period of $T = 1000$ seconds. During this time period, the vector of the numbers of requests generated within the 8 user areas is: $(150, 250, 400, 500, 550, 600, 900, 1000)$. We have sorted the numbers of requests in the ascending order, since it will not affect the simulation results in any way. We can then calculate the request arrival rate vector as $\boldsymbol{\lambda} = (0.15, 0.25, 0.4, 0.5, 0.55, 0.6, 0.9, 1.0)$. We have $\sum_{i=1}^{8} \lambda_i/\mu = 43.5$. Thus, the absolute minimum number of servers needed is 44. In the real world, if the number of available servers is small, i.e., too close to 44, the edge cloud will tend to become overloaded and unstable, according to the results from queueing theory. We consider four scenarios where the numbers of servers are $N = 55$, 60, 65, and 70, respectively.

With the values for μ, $\boldsymbol{\lambda}$, and N, we can derive the server allocation plans using the 5 heuristic algorithms. Then, we use each of the server allocation plans to run the CloudSim Plus simulation and derive the simulated average turnaround time. For the given request set, the average execution time is the same for different heurisitics. To clearly see the differences among the different algorithms, we compare the average waiting time (which is equal to the average turnaround time minus the average execution time) of different algorithms.

The comparison results for the 4 scenarios are presented in Fig. 5. When the number of servers is comparatively small (Fig. 5(a) and Fig. 5(b)), the server allocations are not that flexible because we need to allocate an enough number of servers to each cloud to make the edge clouds stable. In these cases, most of the heuristics can arrive at the best server allocation and they achieve similar average turnaround time; still, some algorithms (for example, *Algorithm LSSSF* in Fig. 5(a) and *Algorithm Proportional* in Fig. 5(b)) may not be able to achieve the minimum average turnaround time. When the number of servers is large, all algorithms have more room to explore for a better server allocation. However, all the other heuristic algorithms are not stable, and may arrive at very large average waiting times when a much smaller average waiting time is possible. Nevertheless, *Algorithm LWRTF* always achieves the optimal average waiting time, and thus, the optimal average turnaround time. The results are consistent with that of Section VI-A. We can conclude that *Algorithm LWRTF* is superior to other reasonably designed heuristic algorithms and achieves near-optimal performances in minimizing average turnaround time for all the application requests.

VII. CONCLUSION AND FUTURE WORK

In this paper, we consider the server allocation problem in edge computing. We design the Largest Weighted Reduction Time First (LWRTF) algorithm to generate the server allocation plan. Extensive evaluations by both the numerical method and the simulation approach verify that it achieves



(a) $N = 55$

(b) $N = 60$

(c) $N = 65$

(d) $N = 70$

Fig. 5. Average waiting comparison among the five heuristic algorithms using CloudSim Plus simulations.

the minimum average turnaround time compared to other reasonably designed heuristics.

We plan to conduct future work in the following directions. First, the applicaiton request model can be extended. For example, each application request may require multiple processing units and each server can have multiple processing units. Second, the communication delays between the mobile users and the edge clouds can be taken into consideration. Third, more practical system models for the edge clouds should be explored; these include queueing network, as compared to separate queueing systems used in our current study, and time-shared queueing model, where the processing power of the servers can be time-shared for all arriving requests.

References

[1] Q. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[2] X. Sun and N. Ansari, "Mobile edge computing empowers internet of things," *ArXiv e-prints*, vol. arXiv:1709.00462, September 2017.

[3] Y. Liu, M. Peng, G. Shou, Y. Chen, and S. Chen, "Toward edge intelligence: Multiaccess edge computing for 5g and internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6722–6747, 2020.

[4] Q. Liang, P. Shenoy, and D. Irwin, "Ai on the edge: Characterizing ai-based iot applications using specialized edge architectures," in *2020 IEEE International Symposium on Workload Characterization (IISWC)*, 2020, pp. 145–156.

[5] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, October - December 2009.

[6] X. Yang, Z. Chen, K. Li, Y. Sun, and H. Zheng, "Optimal task scheduling in communication-constrained mobile edge computing systems for wireless virtual reality," in *Proceedings of the 23rd Asia-Pacific Conference on Communications (APCC)*, December 2017, pp. 1–6.

[7] Z. Chen, L. Jiang, W. Hu, K. Ha, B. Amos, P. Pillai, A. Hauptmann, and M. Satyanarayanan, "Early implementation experience with wearable cognitive assistance applications," in *Proceedings of the Workshop on Wearable Systems and Applications (WearSys)*, May 2015, pp. 33–38.

[8] M. Shahzad, A. X. Liu, and A. Samuel, "Secure unlocking of mobile touch screen devices by simple gestures: You can see it but you can not do it," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking (MobiCom)*, September - October 2013, pp. 39–50.

[9] K. Lee, D. Chu, E. Cuervo, J. Kopf, Y. Degtyarev, S. Grizan, A. Wolman, and J. Flinn, "Outatime: Using speculation to enable low-latency continuous interaction for mobile cloud gaming," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, May 2015, pp. 151–165.

[10] W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, "Trust evaluation in online social networks using generalized network flow," *IEEE Transactions on Computers*, vol. 65, no. 3, pp. 952–963, March 2016.

[11] W. Jiang and J. Wu, "Active opinion-formation in online social networks," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, May 2017.

[12] Z. Pang, L. Sun, Z. Wang, E. Tian, and S. Yang, "A survey of cloudlet based mobile computing," in *Proceedings of International Conference on Cloud Computing and Big Data (CCBD)*, November 2015, pp. 268–275.

[13] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1628–1656, Third quarter 2017.

[14] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC Workshop on Mobile Cloud Computing*, August 2012, pp. 13–16.

[15] S. Yi, C. Li, and Q. Li, "A survey of fog computing: Concepts, applications and issues," in *Proceedings of the Workshop on Mobile Big Data (Mobidata)*, June 2015, pp. 37–42.

[16] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere, "Edge-centric computing: Vision and challenges," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 5, pp. 37–42, October 2015.

[17] Y. Cui, J. Song, K. Ren, M. Li, Z. Li, Q. Ren, and Y. Zhang, "Software defined cooperative offloading for mobile cloudlets," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1746–1760, June 2017.

[18] Z. Xu, W. Liang, W. Xu, M. Jia, and S. Guo, "Capacitated cloudlet placements in wireless metropolitan area networks," in *Proceedings of the 40th IEEE Conference on Local Computer Networks (LCN)*, October 2015, pp. 570–578.

[19] ——, "Efficient algorithms for capacitated cloudlet placements," *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, vol. 27, no. 10, pp. 2866–2880, October 2016.

[20] M. Jia, W. Liang, Z. Xu, and M. Huang, "Cloudlet load balancing in wireless metropolitan area networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, April 2016, pp. 1–9.

[21] M. Jia, J. Cao, and W. Liang, "Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks," *IEEE Transactions on Cloud Computing*, vol. 5, no. 4, pp. 725–737, October 2017.

[22] M. C. S. Filho, R. L. Oliveira, C. C. Monteiro, P. R. M. Inácio, and M. M. Freire, "Cloudsim plus: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," in *IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, May 2017, pp. 400–406.

[23] L. Kleinrock, *Queueing Systems. Volume 1: Theory*. Wiley-Interscience, January 1975.

[24] D. Li, B. Dong, E. Wang, and M. Zhu, "A study on flat and hierarchical system deployment for edge computing," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 2019, pp. 0163–0169.

[25] E. Wang, D. Li, B. Dong, H. Zhou, and M. Zhu, "Flat and hierarchical system deployment for edge computing systems," *Future Generation Computer Systems*, vol. 105, pp. 308–317, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X19304042

[26] D. Li, C. Asikaburu, B. Dong, H. Zhou, and S. Azizi, "Towards optimal system deployment for edge computing: A preliminary study," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, 2020, pp. 1–6.

[27] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, vol. 41, no. 1, pp. 23–50, January 2011.

A C
C

Ramprakash Pavithrakannan, Nikitta Baker Fenn, Sriram Raman, Varadharajan Kalyanaraman, Vignesh Kumar Murugananthan, Jeevanandham Janarthanan

A
C
Toronto, Canada
c0793484@mylambton.ca, c0790287@mylambton.ca, c0787336@mylambton.ca, c0793756@mylambton.ca, c0793760@mylambton.ca, c0787513@mylambton.ca

— In real world datasets missing al es are so common. ost achine Learning algorithms won't wor with missing al es, and so they sho ld be handled before training the model. It is a common practice to imp te the missing al es with central tendencies ean, edian, ode , b t choosing a partic lar one among them is not an easy choice to ma e. This paper analy es the impact of sing each central tendency for different distrib tions of data. S ewness and the presence of o tliers are considered for selecting the data for analysis. Certain pres mptions ha e been made before the examination, and performance metrics s ch as acc racy, UC OC, precision, recall, and score are analy ed to pro e dispro e the ass mptions.

—

. Introd ction

Real-world data may have more missing values compared to the processed data available in machine learning repositories. In Machine Learning, handling missing values is vital, and it has an excellent chance of impacting the model's performance. Hence, analyzing the performance for different values becomes extremely necessary.

In this paper, we have analyzed the performance of classification models on imputing with different central tendencies. The analysis is made on two categories: skewness of the data and presence of outliers. Considering skewness, data that is highly skewed with zero and right-skewed are analyzed. Considering outliers, data with outliers on both sides of the boxplot's whiskers and few outliers are analyzed. Certain assumptions have been made for these categories, and they are checked for truthness.

. andling missing al es

There are several reasons like a failure in recording the data and data corruption, due to which the data gets lost. Missing data handling is crucial to keep up the representativeness of the sample. If left untreated, the missing values can have a severe impact on the analysis. Moreover, the datasets should be preprocessed to fit the machine learning algorithms.

There are three main approaches to dealing with missing values: Imputation, Omission, and Analysis.

2.1 Imputation is filling the missing value with some value that makes sense, mostly a central tendency. Features with continuous numeric data can be imputed with the mean, mode, or median and categorical data can be imputed with the mode or median of the data available.

Pros:

- Imputation comes in handy when the dataset is small and implementation is straightforward.
- Data is not lost when compared to the other handling methods.

Cons:

- Imputation can cause a leakage of data
- It can add bias and variance due to the approximations.

2.2 Omission is the process of removing the observations with the missing value from the dataset. It is one of the most used methods to treat null values. Two techniques of omission are – Listwise deletion and Pairwise deletion. If there is at least one value missing in the listwise deletion technique, that record is dropped, and the model is run on record with a complete set of values. Whereas in pairwise deletion, the record is not deleted completely, and instead, the omission is based on the values included in the analysis.

Pros:

- Omission of missing data builds an accurate and robust model.

Cons:

- Data and Information loss
- Cannot use this method when the missing value percentage is high.

2.3 Analysis is the process of using predictive modelling techniques to predict the missing value for the records. Analysis can help in building a more accurate model.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Pros:

- Creating a predictive model is not required for each attribute with a missing value.

Cons:

- The analysis method is time-consuming.

Imputation is a crucial method to analyze further because it’s not time-consuming and easy to implement. In this paper, we will examine the impact of imputing the missing values with each central tendency.

. ss mptions

Considering the fact that the range of the data and outliers will have an impact on the mean and median, certain assumptions have been made. They are:

ero s ewed data – Missing values in data that is highly skewed with the presence of zero must be imputed with zero (median and mode). For example, a dataset containing school students’ information will have zero as the income column’s value in most records. But a group of competent students working on sponsored projects may get a stipend, and it changes the mean of the record. If any record is missing, we can assume that the student is not earning and impute with zero.

ight s ewed data – Missing values in skewed data should not be imputed with mean. If the data’s tail has a large difference with the peak, the data’s mean will be significantly impacted. Hence, it would not be a better choice to impute the missing values.

ith o tliers – Same as skewed data, data’s mean is not immune to outliers’ presence. To reduce the impact of an outlier, we can remove it before calculating the mean. But in data with a sizable number of records outside the whiskers of a box plot, we must account for the outliers to make a decision. In that case, imputing the missing values with mean is not a better choice.

itho t few o tliers – Data with most of the values between the boxplot’s whiskers will have an almost similar value for the central tendencies. Hence, every central tendency will be giving similar results.

. Experiment set p

Python project with minimal human intervention is used to make the analysis. It reduces manual errors, which may impact the results of the study. The project was designed to take dataset and input file as input, and perform analysis and return the results in a compressed file. The processes performed were also automatically documented for analyzing the results later.

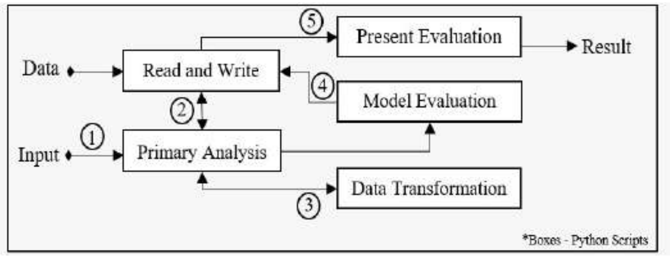


Fig. 1. Experiment setup

The primary analysis script will be reading the user(our) inputs to perform the analysis. The input file will contain particulars such as dataset file name and information about the dataset required to train the machine learning model. Then the primary analysis script will call read/write script to read the dataset. The pandas dataframe created from the data will then be sent to the data transformation script to perform analysis and imputation. This script will perform on-hot encoding, label encoding, scaling, and imputation based on the input file. Three dataframes, each with mean, median, and mode imputed, will be created at this stage to study the performance.

Three dataframes will be passed to the model evaluation script, where various machine learning models will be trained, and their metrics are captured. All three dataframes are trained with RandomForest, LogisticRegression, K nearest neighbors, and State Vector Machine algorithms. Training and evaluating the data with these different algorithms, each using distinct methods for prediction, will help us understand which central tendency has better performance on which algorithm. Evaluation metrics such as Accuracy, AUC-ROC score, precision, recall, and F1 score are captured for every trained model. Analyzing all of these parameters will give us a clear idea about the impact of each central tendency.

Dataframes are split into train and test set using the Stratified shuffle split, and the same is used to perform the cross-fold validation of 10. Stratified shuffle split will ensure that both train and test set contains the same amount of distribution of the imputed feature, and it ensures that the same amount of imputed records are found in both train and test set. The same is used to generate ten different train and test sets to recreate the cross-validation function[3]. Using ten-fold cross-validation will ensure that captured results are not particular to one set of records. It will help us get more data on the performance for the analysis.

Evaluation metrics will then be passed to the presenting evaluation script through the read/write script. This script will document the analysis results and then plot the results using the matplotlib library for comparison. The plotted graphs, steps performed, and input files are then packaged and compressed for later analysis.

This experimental setup is designed to impute given the feature and perform analysis on the classification models. With a little tweaking, this setup can be used to perform analysis on regression models too. Besides, it can also be modified to train data on various algorithms and return the best algorithm and the trained model, like GridsearchCV and RandomsearchCV.

. ataset sed

A dataset containing features will all the features mentioned above is hard to find. So, we decided to use two different datasets to analyze all required distributions. They are Census income data and Heart disease data.

Census income data contains all required distributions but the data without outliers. This data had some missing values on features that we are not analyzing. It being a large dataset, dropping those records didn’t lead to much data loss. If we could follow the same for every missing value in reality, this analysis would not have been needed. Unfortunately, we could come across a dataset where we can't afford to lose even a single record. So it becomes necessary to proceed with this research. To analyze the data without outlier, we have used heart disease data, which has a feature with only one value away from the whisker of the boxplot.

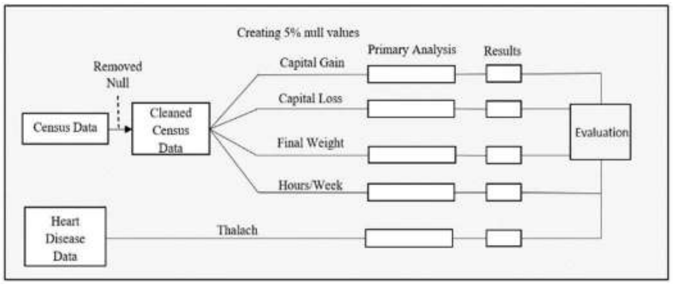


Fig. 2. Analysis performed

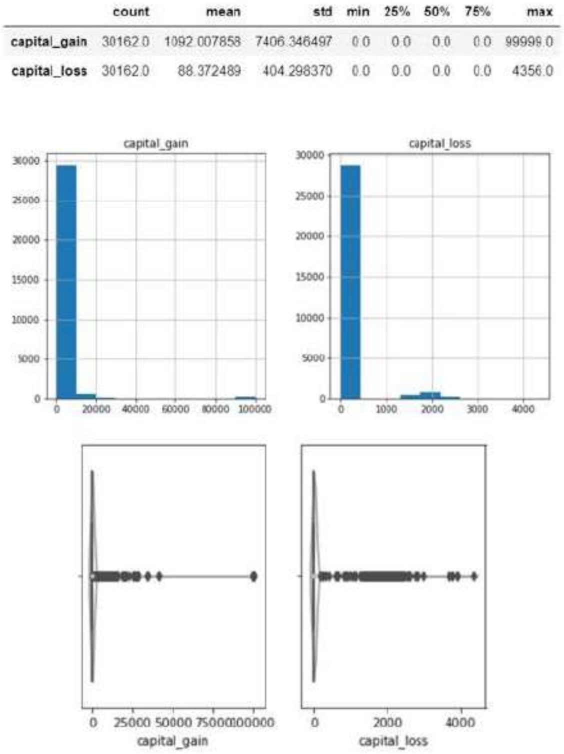
Since these datasets don't contain missing values on the required features, we have randomly removed the available values to create missing values. 5% of the data is removed to make a good impact when imputation is analyzed. When analyzing each feature, we ensured that other features are not touched to ensure results are independent of other factors.

• nalysis

Data identified has undergone an initial analysis to identify nominal and ordinal categorical features and numerical features. This information was included in the input text file for the primary analysis script to understand and analyze. The packaged results are then evaluated to get the below results.

ero s ewed data

In the census income dataset, capital gain and capital loss features have more than 75% of the values as zero. It can be seen at histogram and violin plot that most of the values of these fields are skewed near zero. For this kind of dataset, mode and median will be the same(zero), and we can see that the mean is far away from zero. In this kind of dataset, we can't use IQR to detect the outliers because both the 25th and 75th quartile values will be zero, so the IQR will also be zero. So we can't consider the results of this data for outliers, and we can consider it only for skewness.



	count	mean	std	min	25%	50%	75%	max
capital_gain	30162.0	1092.007858	7406.346497	0.0	0.0	0.0	0.0	99999.0
capital_loss	30162.0	88.372489	404.298370	0.0	0.0	0.0	0.0	4356.0

Fig. 3. Zero skewed distribution

According to the mean and standard deviation of the cross-validation (Table 1), the analysis results show only a minor difference in different central tendencies' performance. Imputing mean gives slightly less accuracy compared to others. But other metrics such as AUC-ROC, precision, recall, and F1 are better for imputing with mean. This dataset being unbalanced, we have to rely upon the latter metrics to select a better performing model. Contradictory to our assumption, the mean gives comparatively better results for zero skewed distribution. The results of KNN are deviating from the results of other models, which should be further analyzed to get a clear picture.

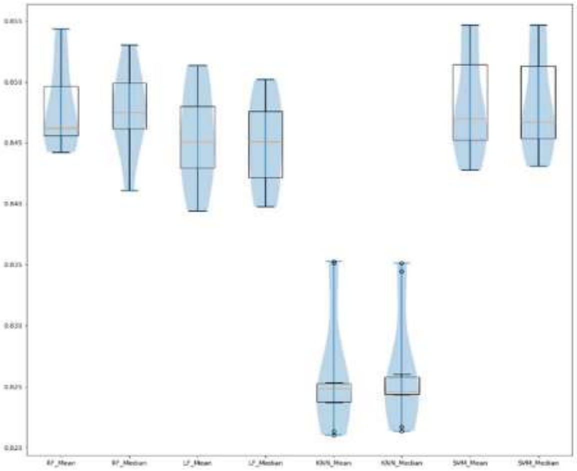


Fig. 4 a). Zero-skewed accuracy(capital_gain)

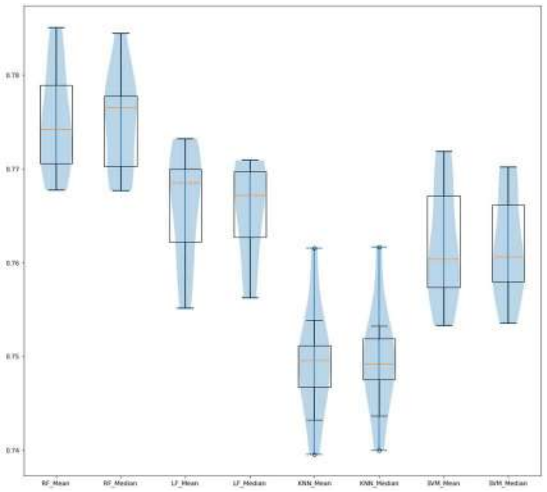
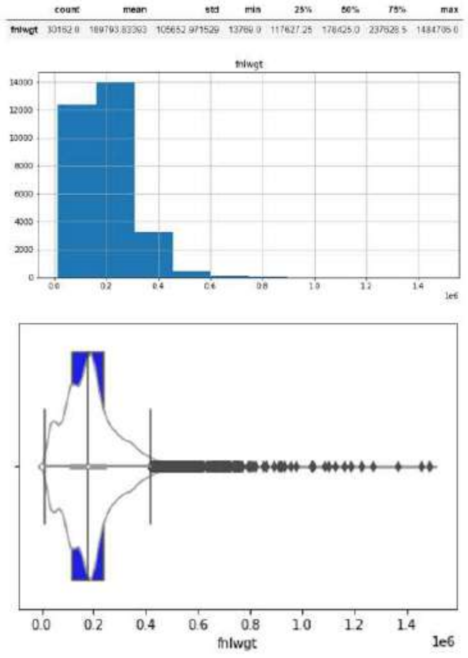


Fig. 4 b). Zero-skewed AUC(capital_gain)

But analyzing all the results of cross-validation gives a different picture. Some better-performing datasets show better results on imputing with mean than the median. In most cases, imputing with mean gives better results, i.e., looking at the boxplot's percentile lines shows that the mean's median is higher.

ight s ewed data

In the census income dataset, the fnlwgt feature is right-skewed, with most of its value peaking on the left side. It can be seen at histogram and violin plot that there are many values outside the whiskers. So we can consider the results of this data for both outliers and skewness.



	count	mean	std	min	25%	50%	75%	max
fnlwgt	30162.0	189793.83393	105652.971529	13769.0	117627.25	178425.0	237628.5	1484705.0

Fig. 5. Right skewed distribution

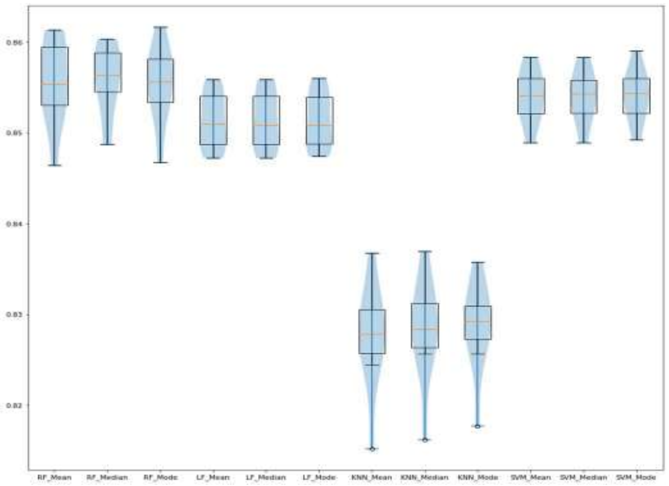


Fig. 6 a). Right skewed accuracy

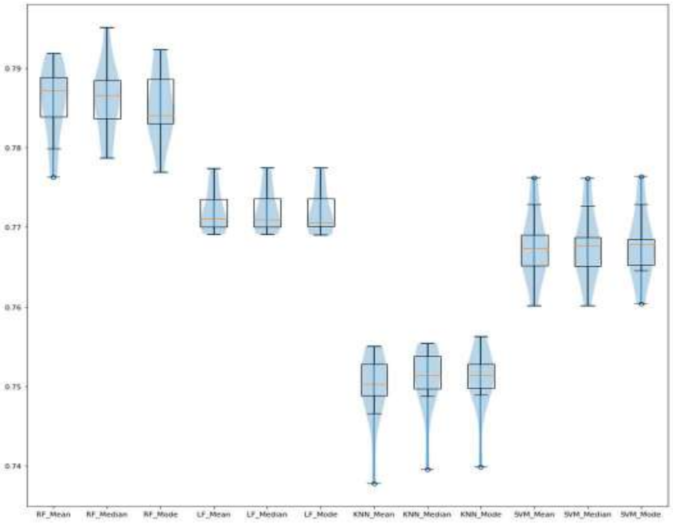


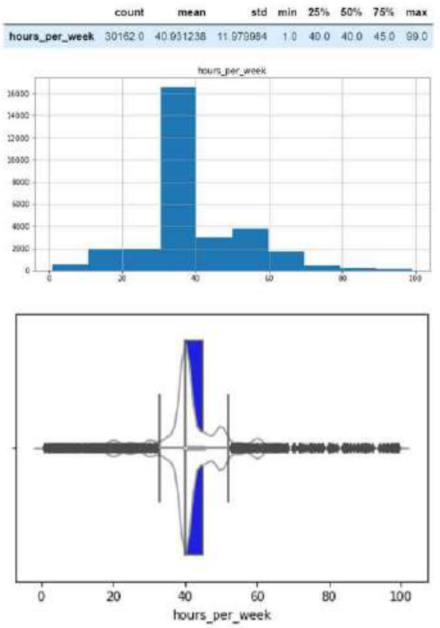
Fig. 6 b). Right skewed AUC

Considering both the mean of the results (Table 2) and the distribution of the results for each cross-validation (Fig. 6), there is no clear evidence that any central tendencies have precedence over the other. The impact of skewness should be analyzed further to get a clear understanding.

ith O tliers

In the census income dataset, hours per week have many outliers. From the boxplot, we can see that outliers are present at both sides of the whiskers, with a peak around the median. For this feature, the mode and median will be the same.

Except for the Random Forest, all other algorithms provide more or less the same results for mean and median. It's due to the fact that there is no significant difference between both's values. Having the same mode and median, with skewness of both sides, made mean to be close to mode. Comparing both, mode gives a better performance in this feature, which proves our assumption.



	count	mean	std	min	25%	50%	75%	max
hours_per_week	30162.0	40.931238	11.979984	1.0	40.0	40.0	45.0	99.0

Fig. 7. With outlier - distribution

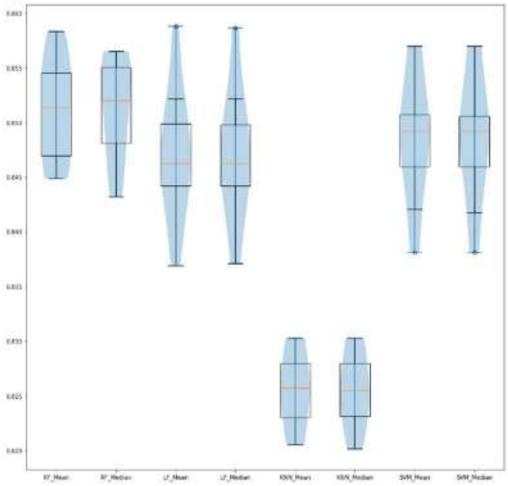


Fig. 8 a). With outlier accuracy

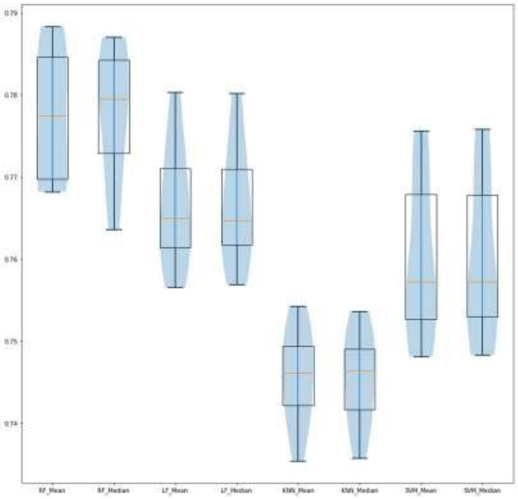
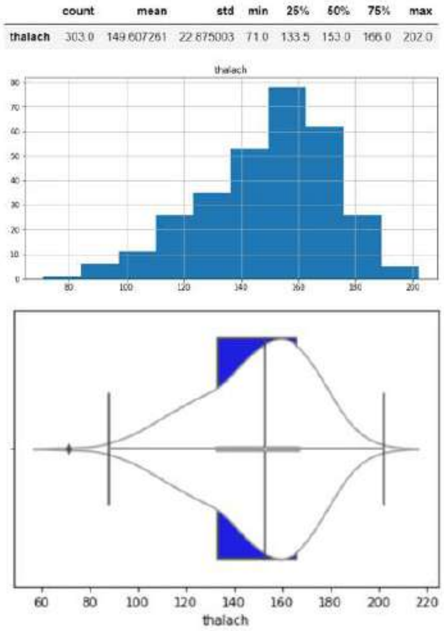


Fig. 8 b). With outlier AUC

itho t few o tliers

In the heart disease dataset, thalach has only one outlier. Since it’s not so distant from the whisker, it can be included in the analysis. From the histogram and boxplot, we can see that most of the values are peaked around the median.



	count	mean	std	min	25%	50%	75%	max
thalach	303.0	149.607261	22.875003	71.0	133.5	153.0	166.0	202.0

Fig. 9. Without outlier - distribution

In most cases, mode gives a better performance in this feature. If we closely look at Table 4, the mode offers better accuracy and AUC performance, and the same was reflected in the boxplot. In most algorithms, mode gave good results on the validation dataset that produced underperforming models. The percentile lines of the boxplots also support our claim that mode provides better performance. Contradictory to our assumption, not all central tendencies have near values, so they impacted the performance.

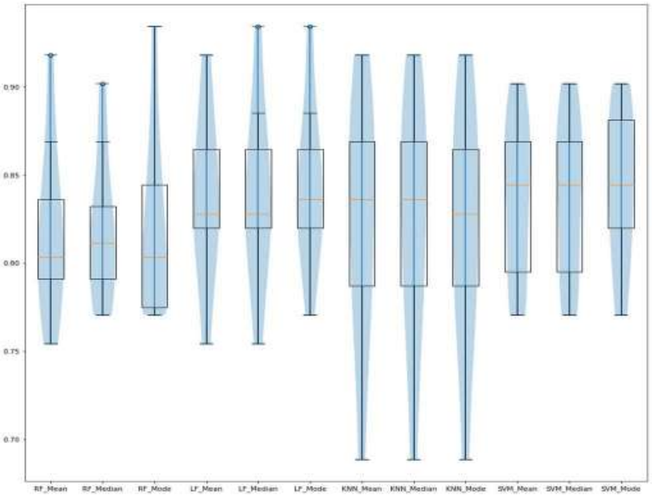


Fig. 10 a). Without outlier accuracy

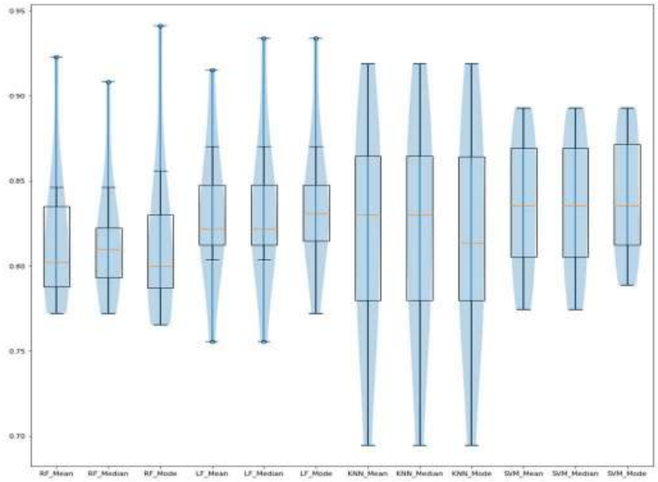


Fig. 10 b). Without outlier AUC

. t re wor s

We may have to consider the importance of a feature in the prediction to get an exact picture of the impact of imputation. Hence, we will perform the same analysis on the features with high importance for modelling.

Results are varying for different algorithms, and so we should analyze the impact of imputation on algorithms. The importance of features must be taken into account to get precise results.

Algorithm	Imputation	ACC_Mean	ACC_STD	AUC_Mean	AUC_STD	Prec_Mean	Prec_STD	Recall_Mean	Recall_STD	F1_Mean	F1_STD
KNeighbors	Mean	0.826029	0.004789	0.749289	0.005651	0.869154	0.003877	0.903614	0.003752	0.886047	0.003483
KNeighbors	Median	0.826161	0.00452	0.749531	0.005495	0.869288	0.003773	0.903639	0.003548	0.886128	0.003277
LogisticRegression	Mean	0.845388	0.003888	0.76601	0.005723	0.875002	0.005061	0.925723	0.003774	0.899634	0.002736
LogisticRegression	Median	0.844857	0.003496	0.765433	0.005002	0.874759	0.004815	0.925234	0.003533	0.899275	0.002529
RandomForest	Mean	0.847943	0.003324	0.775032	0.005331	0.880737	0.004164	0.92168	0.002029	0.900738	0.002412
RandomForest	Median	0.847528	0.003271	0.774995	0.00515	0.880868	0.003774	0.920857	0.002856	0.900412	0.002446
StateVector	Mean	0.848225	0.004066	0.761823	0.005876	0.871153	0.005051	0.935643	0.002988	0.902236	0.002888
StateVector	Median	0.848142	0.003866	0.761674	0.00538	0.871076	0.004773	0.935621	0.003172	0.902185	0.00278

I

Algorithm	Imputation	ACC _Mean	ACC _STD	AUC_Mean	AUC_STD	Prec _Mean	Prec_STD	Recall_Mean	Recall_STD	F1_Mean	F1_STD
KNeighbors	Mean	0.827538	0.005323	0.749617	0.004599	0.871245	0.004133	0.904267	0.005802	0.887439	0.003989
KNeighbors	Median	0.828334	0.00519	0.750866	0.004352	0.871883	0.004064	0.904623	0.00593	0.88794	0.003901
KNeighbors	Mode	0.828666	0.00462	0.750616	0.004103	0.871611	0.004092	0.905529	0.005596	0.888233	0.003514
LogisticRegression	Mean	0.851261	0.002878	0.772066	0.002619	0.879734	0.002979	0.929224	0.003864	0.903794	0.002264
LogisticRegression	Median	0.851277	0.002912	0.772099	0.002655	0.879752	0.002982	0.929224	0.00386	0.903804	0.002281
LogisticRegression	Mode	0.851277	0.002876	0.772033	0.002693	0.879704	0.003025	0.92929	0.003767	0.903811	0.002254
RandomForest	Mean	0.855574	0.004416	0.785735	0.004519	0.888152	0.003579	0.924317	0.004417	0.905868	0.003284
RandomForest	Median	0.855939	0.003623	0.786098	0.004477	0.888307	0.003338	0.924673	0.003614	0.90612	0.002706
RandomForest	Mode	0.855358	0.003951	0.785171	0.004313	0.887811	0.003473	0.924456	0.003974	0.905757	0.002914
StateVector	Mean	0.854048	0.002759	0.767686	0.00425	0.875812	0.003587	0.939064	0.003348	0.906327	0.002062
StateVector	Median	0.854015	0.002685	0.767663	0.004201	0.875807	0.003528	0.93902	0.003307	0.906304	0.002013
StateVector	Mode	0.854147	0.002694	0.767706	0.004229	0.875797	0.003511	0.939241	0.003296	0.906402	0.002001

Algorithm	Imputation	ACC _Mean	ACC _STD	AUC_Mean	AUC_STD	Prec_Mean	Prec_STD	Recall_Mean	Recall_STD	F1_Mean	F1_STD
KNeighbors	Mean	0.825498	0.003331	0.745612	0.005578	0.868682	0.003273	0.90448	0.002974	0.886215	0.002376
KNeighbors	Median	0.825448	0.003285	0.745402	0.005415	0.868549	0.003286	0.904589	0.003049	0.886198	0.00238
LogisticRegression	Mean	0.847047	0.005758	0.766395	0.007155	0.876656	0.00449	0.926819	0.004359	0.901036	0.003995
LogisticRegression	Median	0.847047	0.005673	0.76644	0.006994	0.876687	0.00443	0.926774	0.004323	0.901031	0.003949
RandomForest	Mean	0.851128	0.004457	0.777406	0.007324	0.883195	0.005161	0.924094	0.00238	0.903173	0.002932
RandomForest	Median	0.851327	0.004425	0.777812	0.007425	0.883423	0.005339	0.924095	0.003109	0.903289	0.002922
StateVector	Mean	0.848656	0.005565	0.759581	0.00923	0.871477	0.005287	0.936733	0.003774	0.902918	0.003662
StateVector	Median	0.848573	0.005574	0.759592	0.009248	0.87151	0.005281	0.936556	0.003756	0.902854	0.003664

Algorithm	Imputation	ACC _Mean	ACC _STD	AUC_Mean	AUC_STD	Prec_Mean	Prec_STD	Recall_Mean	Recall_STD	F1_Mean	F1_STD
KNeighbors	Mean	0.822951	0.065901	0.819501	0.064079	0.830553	0.060696	0.861142	0.072488	0.84353	0.055243
KNeighbors	Median	0.822951	0.065901	0.819501	0.064079	0.830553	0.060696	0.861142	0.072488	0.84353	0.055243
KNeighbors	Mode	0.819672	0.065163	0.816189	0.064061	0.828077	0.061399	0.858365	0.069722	0.840908	0.053996
LogisticRegression	Mean	0.836066	0.045194	0.830861	0.040512	0.824994	0.055417	0.894316	0.06515	0.856254	0.044756
LogisticRegression	Median	0.837705	0.048326	0.832713	0.04454	0.827683	0.060136	0.894316	0.06515	0.857618	0.047056
LogisticRegression	Mode	0.842623	0.043497	0.837354	0.040825	0.829407	0.058538	0.903597	0.055681	0.862979	0.04116
RandomForest	Mean	0.818033	0.046048	0.81685	0.04227	0.827836	0.065187	0.854784	0.069337	0.837726	0.042527
RandomForest	Median	0.818033	0.039786	0.816376	0.036668	0.827872	0.065738	0.853836	0.061262	0.837527	0.037765
RandomForest	Mode	0.818033	0.050501	0.816667	0.049135	0.828001	0.073777	0.854239	0.0641	0.83784	0.046766
StateVector	Mean	0.837705	0.043031	0.835085	0.037624	0.831108	0.052948	0.89455	0.074184	0.858226	0.038383
StateVector	Median	0.837705	0.043031	0.835085	0.037624	0.831108	0.052948	0.89455	0.074184	0.858226	0.038383
StateVector	Mode	0.842623	0.041602	0.840512	0.034764	0.837647	0.049863	0.89455	0.074184	0.861859	0.037768

eferences

[1] M.N. Norazian Ramli, Yahaya, A.S., Ramli, N.A., Yusof, N.F.F.M., and Abdullah, M.M.A., “ *V* *A* ,” Advances in Environmental Biology, Vol. 7(12), 2013, pp. 3861-3869.

[2] R. Pavithrakannan, “ *A* ”, GitHub repository, 2021. https://github.com/rampk/imputation-analysis

[3] A. Géron, “End-to-End Machine Learning Project” in w, 2nd ed. Sebastopol, CA, USA: O’Reilly, 2019, pp. 51-55.

[4] UCI Machine learning repository. Census income data set. Available: https://archive.ics.uci.edu/ml/datasets/Census+Income

[5] UCI Machine learning repository. Heart Disease Data Set. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease

[6] Akande, O., Li, F., & Reiter, J. (2017). *A* . The American Statistician, 71, 162– 170.

[7] Ghorbani S, Desmarais MC (2017) Appl Artif Intell 31(1):1–22.

[8] Gondara, L. and Wang, K. *A* . In PAKDD (3), volume 10939 of Lecture Notes in Computer Science, pp. 260–272. Springer, 2018.

[9] Hippel PT . Sociolog Methods Res. 2018; 1(1): 1- 20.

[10] Josse, J., Prost, N., Scornet, E., and Varoquaux, G. arXiv preprint arXiv:1902.06931, 2019.

[11] Huang J, Keung JW, Sarro F, Li Y-F, Yu YT, Chan WK, Sun H (2017) *C* . J Syst Softw 132:226–252

[12] L. A. Hunt, ” Data Science: Innovative Developments in Data Analysis and Clustering, pp. 3–14, 2017, Springer.

[13] Kahale LA, Diab B, Brignardello-Petersen R, Agarwal A, Mustafa RA, Kwong J, et al. J Clin Epidemiol. 2018; 99:14–23.

[14] Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, ´I. . CoRR, abs/1807.03653, 2018.

[15] Newgard, C.D.; Lewis, R.J. *A* JAMA 2015, 314,940–941.

[16] Spineli LM, Yepes-Nuñez JJ, Schünemann HJ. *A* BMC Med Res Methodol. 2018; 18:115.

[17] J. A. C. Sterne, I. R. White, J. B. Carlin et al., ” BMJ, vol. 338, no. jun29 1, p. b2393, 2009.

[18] Tang Y. *C* Stat Med. 2018; 37(9): 1467- 1481.

[19] van der Heijden GJ, Donders AR, Stijnen T, et al. J Clin Epidemiol 2006;59:1102-9.

Face Mask Recognition with Realistic Fabric Face Mask Data Set: A Combination Using Surface Curvature and GLCM

Regina Lionnie
Department of Electrical Engineering
Universitas Indonesia
Depok, Indonesia
regina.lionnie@ui.ac.id

Catur Apriono
Department of Electrical Engineering
Universitas Indonesia
Depok, Indonesia
catur.apriono@ui.ac.id

Dadang Gunawan
Department of Electrical Engineering
Universitas Indonesia
Depok, Indonesia
guna@eng.ui.ac.id

***Abstract*—Wearing a mask is a requirement in the Covid-19 pandemic for the general public. While it is one of the several must-do actions to prevent forward spread in the Covid-19 infections, at the same time, the effect of wearing a mask in naïve face recognition systems have shown lower system performance in several cases and conditions. Simultaneously, only a handful of research studies have focused on a non-medical face mask with realistic images data set. This research proposed a new data set of realistic fabric face mask data set to be evaluated using surface curvature and gray level co-occurrence matrix (GLCM). The classification applied support vector machine (SVM). One hundred seventy-six images in the data set were analyzed with various properties, resulting in several experiments. The experiments' parameters were color properties, approaches in surface curvature, i.e., Gaussian, mean and principal curvature, angle and distance in GLCM, GLCM properties, i.e., contrast, homogeneity, correlation and energy, also kernel functions in SVM. The best accuracy result, 87.5%, was derived from the combinations of these parameters. This research also improved the running time of the recognition process while maintaining the system's performance.**

***Keywords*—Covid-19, face mask recognition, gray level co-occurrence matrix, support vector machine, surface curvature**

I. Introduction and Motivation

Covid-19 is a contagious disease caused by a new strain of the coronavirus. The indications can include fever, cough, shortness of breath. In difficult situations, it can induce pneumonia and breathing complications. Close contact with respiratory droplets of an infected individual spreads the virus. Touching surfaces contaminated with the virus and touching their eyes, nose and mouth can also infect a person [1]. The Covid-19 outbreak was declared as Public Health Emergency of International Concern (PHIC) by WHO on Jan 30, 2020. This statement is WHO's most crucial alarm to unite all countries to take notice and take effort immediately [2]. To stop the spreading of the Covid-19, wearing a mask is one of several requirements to protect a healthy society and prevent forward spreads. Even in cases where visitors who do not live together come to a family's house, they should wear masks if the physical distance cannot be maintained or when the home's ventilation is in poor condition [3]. Furthermore, wearing a mask is needed for the general public, regardless of whether they have gotten two doses of Covid-19 vaccines [4].

As the public needs to wear a mask in their daily lives, naïve face recognition systems have shown lower system performance in several cases and conditions [5,6]. This reduction of performance might occur due to the nose and mouth on the face area being covered. Although the upper half of the face, i.e., eyes and eyebrows, has a more considerable influence on the recognition performances, the lower part of the face, i.e., nose and mouth, if covered, still weaken the recognition performance [6,7]. Wearing a mask is one of the examples of occlusion problems in the research of face recognition. Other occlusion problems are wearing glasses, attributes such as hats and hair accessories, having a mustache, situations where objects covering the face's area, pose variations and exceptional cases like low-resolution problems and blurred faces [8]. In this research, wearing a mask was treated as the focused problem for occlusions, and the data set of face mask images was created for masked face and barefaced images. For the part of the data set of masked face images, the recognition system's input included whole masked face images as they were not divided into occluded or un-occluded areas on the face. This principle also applied for part of the data set of barefaced images.

During Covid-19 pandemic global event, the number of research focusing on face mask challenges has been increased. These researches have focused on detecting whether individuals wear a mask in public to help authorities monitor and maintain safety during the Covid-19 pandemic [9-12]. Moreover, the number of research on identifying individuals while wearing a face mask has also improved significantly [5,13-15]. The purpose of identifying individuals while wearing a mask is to help the public use face recognition systems correctly but safely without taking off the mask. In [5], the performance evaluation was compared using two academic face recognition systems, i.e., ArcFace and SPhereFace and one commercial system, MegaMatcher 11.2 SDK Neurotechnology. Multi-Task Cascaded Convolutional Neural Network (MTCNN) and support vector machine (SVM) was studied in [13]. Improvement using cosine distance combined with transfer learning was analyzed in [14]. In [15], the authors studied de-occlusion distillation for knowledge transfer evaluated with three synthetic and realistic face mask data sets.

Contradictory with the rocketing numbers of research focusing on face detection and recognition during Covid-19, there are still very few data sets specifically designed using a real face mask that implements real-life situations. Sometimes, the data set was created with an image editing approach [16], and sometimes they are parts of the larger variations of the occlusion data set [17]. This research proposed creating a face mask data set containing respondents

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

wearing a real face mask with two variations of pattern and color of fabric or non-medical mask. The first reason behind choosing fabric or non-medical mask for the data set is the lack of fabric or non-medical face mask data set. The second is because the fabric face mask is recommended by WHO to be worn by the general public under the age of 60 and people who do not have underlying health conditions [3]. WHO recommends the medical mask to be worn by medical workers, people who have symptoms of Covid-19 and those who take care of them. WHO also recommends that medical masks be worn by older people (age 60 years and over) and people with underlying health conditions when the physical distancing of at least one meter cannot be achieved [18]. Furthermore, in the early days of the Covid-19 pandemic, the medical mask's availability was prioritized for the health workers. As a result, the government recommends using fabric masks for the general public [19].

In addition to constructing a realistic fabric face mask data set, this research also studied a combination of surface curvature and gray level co-occurrence matrix (GLCM). Both work as an extractor of features. The motivation behind this proposed combination of methods was due to the pattern and color of fabric masks that suitable for using methods for texture analysis, such as GLCM, that operated as a statistical approach. Moreover, the surface curvature further enhances the difference of arrangements in color and intensity of spatial relations in images. The effect of approaches using surface curvature, i.e., Gaussian curvature, mean curvature and principal curvature, were studied along with the angle, distance and properties of GLCM. The property of images, such as color, also affects the recognition performance. This research further evaluated the color properties by separating the color components. The classification method was completed by support vector machine (SVM). The SVM kernel function's effect was also assessed to achieve better recognition performance.

This paper's writing is settled as follows: Section I displays the research background and the authors' motivations. Section II explains the theory behind surface curvature and GLCM. Section III describes the creation of the data set and the design of the recognition system. Section IV discusses the results, and Section V concludes.

II. Surface Curvatures and Gray Level Co-occurrence matrix (GLCM)

A. Surface Curvatures

A surface in $\mathbb{R}^3$ is a set of points $S \subset \mathbb{R}^3$ that is two dimensional where each point $\boldsymbol{p} \in S$ has a neighborhood which can be parametrized by two coordinates [20],

$$\boldsymbol{x}: \mathbb{R}^2 \supseteq U \to \mathbb{R}^3: (u, v) \mapsto \boldsymbol{x}(u, v) \tag{1}$$

The curvature of a surface $S \subset \mathbb{R}^3$ at $\boldsymbol{p} \in S$ measures the rate at which S leaves the tangent plane to S at $\boldsymbol{p}$. Several approaches calculating the surface curvature are Gaussian curvature, mean curvature and principal curvature. In a way, Gaussian and mean curvature can be obtained from principal curvature and vice versa using different approaches [21]. In this research, Gaussian and mean curvature were obtained first using the first and second fundamental form. The principal curvature was calculated from Gaussian and mean curvature.

If $\boldsymbol{x}: U \to \mathbb{R}^3$ then the Gaussian curvature (*Gc*) and mean curvature (*Mc*) are calculated in (2) and (3), respectively [21].

$$Gc = \frac{LN - M^2}{EG - F^2} \tag{2}$$

$$Mc = \frac{EN + GL - 2FM}{2(EG - F^2)} \tag{3}$$

E, *F* and *G* are the first fundamental form's coefficients and *L*, *M* and *N* are the second fundamental form's coefficients.

The principal curvature (*P*) are the roots (*Pmax* and *Pmin*) of the quadratic equation of (4)

$$x^2 - 2Mcx + Gc = 0 \tag{4}$$

Hence, by calculating the quadratic equation's roots, *Pmax* (5) and *Pmin* (6) are obtained [21].

$$Pmax = Mc + \sqrt{Mc^2 - Gc} \tag{5}$$

$$Pmin = Mc - \sqrt{Mc^2 - Gc} \tag{6}$$

B. Gray Level Co-occurrence Matrix (GLCM)

Gray level co-occurrence matrix (GLCM) is a method to characterize the distance and angle relationship between the pixel of interest and its neighbors. The distance and angle relationship are essential between two pixels because this repeated distribution forms texture in the spatial position [22].

Assuming *I(k,k)* is the neighborhood of the pixel of interest (p_c, q_c), the co-occurrence value is the distribution of co-occurrence values at a certain distance (*d*) and angle (θ) from (p_c, q_c). The co-occurrence matrix for *I(k,k)* called C_M is defined in (7) and (8) [23].

$$C_M = \sum_{n=1}^{k} \sum_{m=1}^{k} \begin{cases} 1 & if\ I(n,m) = k\ and\ I(n + d_x, m + d_y) = k \\ 0 & else \end{cases} \tag{7}$$

$$d_x = d.cos(\theta), d_y = d.sin(\theta) \tag{8}$$

Fig. 1 shows the spatial relationships of distance and angle of the pixel of interest (p_c, q_c) and its neighborhood. The *d* represents the distance between the pixel of interest and its neighbor while the angle is formatted in the square bracket. In this research, variations of distance and angle can be seen in Fig. 1. [0,*d*] expresses *d* distance and 0^0 angles, [-*d*,*d*] expresses *d* distance and 45^0 angles, [-*d*,0] expresses *d* distance and 90^0 angles, [-*d*, -*d*] expresses *d* distance and 135^0 angles and [*d*,0] expresses *d* distance and 270^0 angles. The distances experimented in this research was *d*=1 for 0^0,45^0,90^0 and 135^0 ([0 1; -1 1;-1 0;-1 -1]) and *d*=2 for 0^0 and 270^0 ([2 0;0 2]).

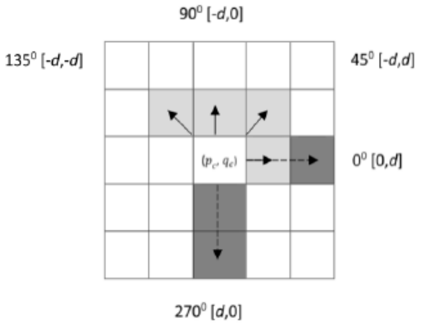


Fig. 1. Spatial relationships of distance and angle of the pixel of interest and its neighborhood; dashed arrow line is for [2 0;0 2] while the solid arrow line is for [0 1; -1 1;-1 0;-1 -1].

Properties in GLCM, i.e., contrast, homogeneity, correlation and energy used in this research for each C_M and angle (θ) as it can be seen in (9)-(12), respectively [24].

$$f_{contrast}^{\theta} = \sum_{i=1}^{L}\sum_{j=1}^{L}(i-j)^2 C_M \qquad (9)$$

$$f_{homogeinity}^{\theta} = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{C_M}{1+|i-j|} \qquad (10)$$

$$f_{correlation}^{\theta} = \sum_{i=1}^{L}\sum_{j=1}^{L} C_M\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i)^2(\sigma_j)^2}}\right] \qquad (11)$$

$$f_{energy}^{\theta} = \sum_{i=1}^{L}\sum_{j=1}^{L}(C_M)^2 \qquad (12)$$

where

$$\mu_x = \sum_{i=1}^{L}\sum_{j=1}^{L} xC_M\, , (\sigma_x)^2 = \sum_{i=1}^{L}\sum_{j=1}^{L} C_M(x-(\mu_x)^2) \quad (13)$$

III. Data Set and Design of Recognition System

Fig.2. shows the design of the recognition system. The training set consisted of RGB images of masked face and barefaced, while the testing set only consisted of RGB images of masked face. The images tested in this research were gathered from 8 respondents of adult male and female. Each respondent was taken 22 pictures when using a fabric face mask and when he/she was in a barefaced, with variations in positions and angle in tilting left, right, up and down. This research's total number of images was 176 images. 90% of total images were collected for the training set and the remaining 10% for the testing set. For masked face images, respondents were asked to wear two different colored and patterned fabric masks. While for barefaced images, respondents were asked to wear no attributes at all. The images were then resized for the final resolution of 200x150 pixels in RGB color space. The data set personally conducted in this research was then called Realistic Fabric Face Mask Data Set version 1.0 (RFFMDS v1.0). It is still a growing data set because more images for the fabric face mask and more respondents are still gathered. Fig. 3 is an example of images in the training and testing set from RFFMDS v1.0.

The next step of the design was to convert RGB images of both training and testing sets to one set of grayscale images and three sets of each RGB channel's separations. There were four sets to be evaluated by the system, gray set, red channel set, green channel set and blue channel set. The images then were altered into 2D arrays containing the surface value. This surface value was proportional to the value of the intensity of the appropriate image data. Fig. 4 is the example of alteration into 2D arrays with the surface value as the height. The *x* and *y* are the column and row of the image, while the *Z* is the height proportional to the image's intensity value where the maximum height is one and the minimum is zero, corresponding to the brightest and darkest pixel, respectively.

The images then were extracted to calculate four different surface curvature types, Gaussian curvature, mean curvature, max principal curvature and min principal curvature. As stated earlier in Section I, the surface curvature's motivation was to amplify the difference of arrangements in color and intensity of spatial relations in images. These extracted curvatures were arranged to improve the next step of the system's process, the GLCM. GLCM works best to analyze the texture of images hence chosen as the method in this research. The color and pattern of fabric face mask created a unique texture that can be examined with the GLCM. GLCM's properties in this research were contrast, homogeneity, correlation and energy. The distance and angle variations of GLCM were also assessed in this research.

The classification utilized support vector machine (SVM), and three different kernel functions of linear, radial basis function (rbf) and polynomial kernel (order 3) were engaged in the training phase. The system results were performance's calculation of accuracy and process's time. This research aimed to find the combinations of variations in color and gray input images with the extraction of curvature types with different properties, angles, and distances of GLCM and classified with different kernel functions of SVM that gave the best accuracy with the fastest time. All simulations in this research were conducted using Matlab with 16GM RAM and Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz, 2.90 GHz.

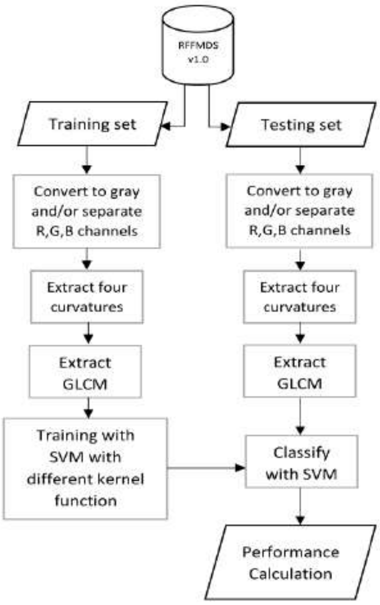
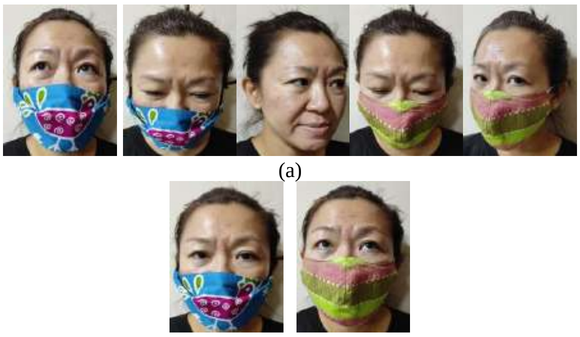


Fig. 2. The flowchart of the designed recognition system.


(a)

(b)

Fig. 3. Example of training set images (a) and testing set images (b) from one respondent from RFFMDS v1.0.

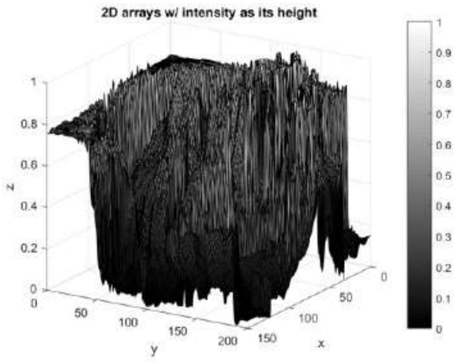


Fig. 4. Example of alteration into 2D arrays with the surface value as its height.

IV. Results and Discussion

A. Experiment 1

In the first experiment, comparing different combinations of methods for extracting features and classification was set side by side. The experiment was conducted on the same testing and training set of images from RFFMDS v1.0. In experiment 1, the experiment's highlight was to analyze the curvature extraction with the classification method and or the combination of curvature extraction with other feature extraction methods. There were two classification methods, i.e., distance calculation with Euclidean (*ED*) and support vector machine (*SVM*). Principal component analysis (*PCA*) to reduce dimension was also investigated in this experiment. Our proposed idea was to combine the extracted curvature features with gray level co-occurrence matrix (*GLCM*) to complement each other and improved recognition performance.

Table I shows the results of experiment 1. Curvature types, i.e., Gaussian curvature (*G*), mean curvature (*M*), max principal curvature (*Pmax*) and min principal curvature (*Pmin*), were evaluated for each combination of methods. The results were presented in a percentage format. From Table I, it can be observed that the simplest classification method of Euclidean distance did not give satisfactory results for each extracted curvature type. To further study the curvature effect in the recognition system, verified with another classification method such as *SVM* and the data analysis such as *PCA* were investigated. *PCA* mainly used as dimensional reduction method to eliminate the sorted eigenvalues that were less than a specific threshold. The idea behind was that the fabric face mask contained various patterns and colors, so maybe not all the information was useful to be put inside the recognition system hence the reductional dimension. To maximize the difference between the two categories, *SVM* was selected. This research operated in a one-vs-all type of manner in *SVM* classification.

Combined *PCA* with *ED*, recognition performance improved slightly, but some were not. Using *SVM* as the classification method also did not improve recognition performance. Until this attempt, no combination seems to improve recognition performance. The proposed idea in this research was to incorporate extracted curvature features with texture analysis method, *GLCM*, and then classified by *SVM*. The motivation behind the combination of surface curvature and *GLCM* was to amplify the variances in color and intensity of spatial relations in images using surface curvature before *GLCM*. The result improved significantly for *Pmax* (max principal curvature) feature, which resulted in 81.25% accuracy. This value was almost more than 2.5 times the previous best result (31.25%). While the max principal curvature feature increased, the mean curvature (*M*) was also improved but did not reach the satisfying rate. Contradictory, the Gaussian curvature (*G*) and min principal curvature (*Pmin*) produced a lower rate.

B. Experiment 2

Continuing from experiment 1, concentrating on the proposed idea that resulted in the best combination from experiment 1, i.e., extracted curvature with *GLCM* and classified by *SVM*, experiment 2 was conducted with the variation of *SVM* kernel functions. The kernel functions of *SVM* tested in this research were linear, rbf and polynomial order 3.

Table II shows the results from experiment 2. The previous experiment's result in Table I used rbf as the proposed idea's kernel function (81.25%). Table II shows that compared to the rbf kernel, using linear kernel function mostly lowered the rate and using polynomial function slightly lowered the recognition rate for the max principal curvature feature (75%) while other extracted curvatures gave varied results. It can be concluded from experiment 2 that rbf kernel offered the best result while the polynomial offered the second-best for max principal curvature feature. It appeared not surprisingly, the data in this research is not separable linear.

C. Experiment 3

Experiment 3 was conducted to study the color's effect on the research. The data evaluated in this research was the fabric face mask which had a unique pattern and color compared to the medical face mask. The color separation from RGB color space might affect recognition performance.

The training and testing set images were converted into gray images in the previous experiments (experiment 1 and 2). In this third experiment, the images were separated into red channel (*R*), green channel (*Gr*) and blue channel (*B*). Experiment 3 was conducted only for max principal curvature feature because it showed promising results from experiments 1 and 2.

Table III displays the results for these color variations. It shows that while in experiment 2, the *SVM* linear kernel function only offered the best result of 37.50%, in experiment 3, it achieved 68.75%. This effect also happened for the polynomial kernel function, it achieved 81.25%, while in experiment 2, it only achieved 75%. The rbf kernel in experiment 3 achieved the same performance as experiment 2. The red channel (*R*) and green channel (*Gr*) were preferred as they showed better results than gray image data.

D. Experiment 4

Experiment 4 was directed to witness the outcome of angle and distance variations for *GLCM*. In previous experiments (1-3), the angle and distance were set to be two pixels between pixel-of-interest and its neighbors while the angles' variation were 0^0 and 270^0 ([2 0;0 2]). In experiment 4, the distance was set to be one pixel between pixel-of-interest and its neighbors. The angles' variation were 0^0, 45^0, 90^0 and 135^0 ([0

1; -1 1;-1 0;-1 -1]). The effect of angle and distance was studied for each color separation and *SVM* kernel functions for the best feature, max principal curvature (*Pmax*).

Table IV displays the results for these variations. Varying the angle and distance in *GLCM* produced improved recognition performance for linear and rbf kernel while it lowered the polynomial recognition rate. It gave the highest accuracy result in the entire experiments (1-4) of 87.5% using red channel for rbf kernel. The recognition performance was improved by considering the relationship between the pixel of interest and its neighborhood choice.

E. Experiment 5

Experiment 5 was managed to calculate the time consumption for the recognition system process. The idea was to search for a combination of methods to reduce time consumption while keeping the same recognition performance (*Acc*) of 87.5%.

From Table V, it can be observed that variations were *Combo 1*, *Combo 2*, *T1* and *T2*. *Combo 1* consisted of max principal curvature (*Pmax*) feature with four properties of GLCM (contrast, homogeneity, correlation and energy) classified with rbf kernel of *SVM*. The distance was set to be 1 pixel between pixel-of-interest and its neighbors. The angles' variation were 0^0, 45^0, 90^0 and 135^0. Fundamentally, the best result from experiment 4. For *Combo 2*, the same principles were applied while only varied for GLCM's properties. *Combo 2* used only two properties, i.e., contrast and correlation. *T1* was the time spent testing only one image, while *T2* was time spent on training and testing all the training and testing set images.

Using only two properties of *GLCM* to reduce the time while keeping the same recognition performance (87.5%) was achieved in experiment 5. 0.0023 seconds was successfully shortened for *T1* and 0.116 seconds for *T2*.

TABLE I. ACCURACY RESULTS IN COMPARISON FOR DIFFERENT METHODS (%)

Methods	**Surface Curvature Types**			
	G	***M***	***Pmax***	***Pmin***
ED	12.50	6.25	31.25	31.25
PCA+ED	20	10	20	40
SVM	12.50	12.50	20	12.50
proposed (*GLCM+SVM*)	6	43.75	**81.25**	25

TABLE II. ACCURACY RESULTS IN COMPARISON FOR SVM KERNEL FUNCTIONS USING SURFACE CURVATURE AND GLCM (%)

SVM kernel functions	**Surface Curvature Types**			
	G	***M***	***Pmax***	***Pmin***
linear	18.75	31.25	37.50	18.75
rbf	6	43.75	**81.25**	25
polyno-mial	18.75	43.75	75	31.25

TABLE III. ACCURACY RESULTS IN COMPARISON FOR SVM KERNEL FUNCTIONS AND COLOR SPACE VARIATIONS FOR PMAX (%)

SVM kernel functions	**Max principal curvature (Pmax)**			
	gray	***R***	***Gr***	***B***
linear	37.50	56.25	68.75	50
rbf	**81.25**	68.75	68.75	68.75
polyno-mial	75	**81.25**	75	68.75

TABLE IV. ACCURACY RESULTS IN COMPARISON FOR SVM KERNEL FUNCTIONS AND COLOR SPACE VARIATIONS PMAX WITH DIFFERENT ANGLE AND DISTANCE (([0 1; -1 1;-1 0;-1 -1]) IN GLCM (%)

SVM kernel functions	**Max principal curvature (Pmax)**			
	gray	***R***	***Gr***	***B***
linear	56.25	56.25	50	56.25
rbf	81.25	**87.50**	81.25	75
polyno-mial	62.50	75	68.75	68.75

TABLE V. TIME CONSUMPTION RESULTS IN COMPARISON FOR RECOGNITION SYSTEM

Combi-nation Methods	**Max principal curvature (Pmax)**		
	Acc (%)	***T1 (s)***	***T2 (s)***
Combo 1	87.50	0.0929	4.6003
Combo 2	87.50	0.0906	4.4843

V. CONCLUSION

A face recognition system with realistic fabric face mask data set was built in this research. Combining surface curvature and GLCM as feature extractors and SVM for classification were applied to improve recognition performance. Several experiments were conducted, including the parameter for color properties, GLCM properties, GLCM angle and distance properties and SVM kernel functions properties. The best accuracy of 87.5% resulted from these various parameters in the feature extraction and classification methods. It can be concluded that combining surface curvature with GLCM improved recognition performance. Moreover, choosing only two properties of GLCM, i.e., contrast and correlation, shortened the running time for the recognition process while maintaining the system's performance.

Upcoming future research can try a deep learning algorithm if the total number of images in the data set increases significantly. The partial face recognition using essential areas in the face may also be evaluated.

REFERENCES

[1] L. Bender, "Key Messages and Actions for COVID-19 Prevention and Control in Schools," UNICEF, New York, March 2020, Accessed: Feb 13, 2021. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/key-messages-and-actions-for-covid-19-prevention-and-control-in-schools-march-2020.pdf

[2] WHO, "A year without precedent: WHO's COVID-19 response," Dec 23, 2020, Accessed: Feb 13, 2021. [Online]. Available:

https://www.who.int/news-room/spotlight/a-year-without-precedent-who-s-covid-19-response

[3] WHO, “Coronavirus disease (COVID-19): Masks,” Dec 1, 2020. Accessed: Feb 14, 2021. [Online]. Available: https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-masks

[4] CDC, “Frequently Asked Questions about COVID-19 Vaccination,” Jan 25 2021, Accessed: Feb 14, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/faq.html#

[5] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner and A. Kuijper, “The Effect of Wearing a Mask on Face Recognition Performance: an Exploratory Study,” 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2020, pp. 1-6.

[6] D. J. Carragher, P. J. B. Hancock, “Surgical face masks impair human face matching performance for familiar and unfamiliar faces,” Cognitive Research, vol. 5, article 59, 2020, doi: 10.1186/s41235-020-00258-x.

[7] P. Karczmarek, W. Pedrycz, A. Kiersztyn, P. Rutka, “A study in facial features saliency in face recognition: an analytic hierarchy process approach,” Soft Computing, vol. 21, 2017, pp. 7503–7517, doi: 10.1007/s00500-016-2305-9.

[8] L. Zhang, B. Verma, D. Tjondronegoro, V. Chandran, “Facial expression analysis under partial occlusion: a survey,” ACM Computing Surveys, vol. 51, no.2, article 25, 2018, doi: 10.1145/3158369.

[9] S. A. Sanjaya and S. Adi Rakhmawan, “Face Mask Detection Using MobileNetV2 in The Era of COVID-19 Pandemic,” 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 2020, pp. 1-5, doi: 10.1109/ICDABI51230.2020.9325631.

[10] A. Oumina, N. El Makhfi, M. Hamdi, “Control The COVID-19 Pandemic: Face Mask Detection Using Transfer Learning,” 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 2020, pp. 1-5, doi: 10.1109/ICECOCS50124.2020.9314511.

[11] M. R. Bhuiyan, S. A. Khushbu and M. S. Islam, “A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3,” 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225384.

[12] M. Loey, G. Manogaran, M. H. N. Taha, N. E. M. Khalifa, “A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic,” Measurement, vol. 167, 2021, doi: 10.1016/j.measurement.2020.108288.

[13] M. S. Ejaz and M. R. Islam, “Masked Face Recognition Using Convolutional Neural Network,” 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2019, pp. 1-6, doi: 10.1109/STI47673.2019.9068044.

[14] D. A. Maharani, C. Machbub, P. H. Rusmin and L. Yulianti, “Improving the Capability of Real-Time Face Masked Recognition using Cosine Distance,” 2020 6th International Conference on Interactive Digital Media (ICIDM), Bandung, Indonesia, 2020, pp. 1-6, doi: 10.1109/ICIDM51048.2020.9339677.

[15] C. Li, S. Ge, D. Zhang, J. Li, “Look Through Masks: Towards Masked Face Recognition with De-Occlusion Distillation,” Proceedings of the 28th ACM International Conference on Multimedia, Seattle, USA, 2020, pp. 3016-3024.

[16] A. Cabani, K. Hammoudi, H. Benhabiles, M. Melkemi, “MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19,” Smart Health, vol. 19, no. 100144, 2021, doi:10.1016/j.smhl.2020.100144

[17] S. Ge, J. Li, Q. Ye and Z. Luo, “Detecting Masked Faces in the Wild with LLE-CNNs,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 426-434, doi: 10.1109/CVPR.2017.53

[18] World Health Organization (WHO), Medical and fabric masks: who wears what when? (Jun 12, 2020), Accessed: Feb 14, 2021. [Online]. Available: https://www.youtube.com/watch?v=esM_ePHn0aw&feature=emb_logo

[19] A. Ikhsanudin, “Pemerintah Gencarkan Penggunaan Masker Kain, IDI: Akan Menghambat Droplet,” detikNews, Indonesia, Apr 06, 2020, Accessed: Feb 14, 2021. [Online]. Available: https://news.detik.com/berita/d-4966405/pemerintah-gencarkan-penggunaan-masker-kain-idi-akan-menghambat-droplet

[20] J. A. Bærentzen, J. Gravesen, F. Anton, H. Aanæs, “Differential Geometry,” in Guide to Computational Geometry Processing - Foundations, Algorithms, and Methods, London: Springer, 2012, ch. 3, sec. 3.1, pp. 45.

[21] A. Gray, “Shape and Curvature” in Modern Differential Geometry of Curves and Surfaces with Mathematica, 2nd ed., CRC Press, 1997, ch. 13, sec.13.4, pp. 410-414.

[22] X. Zhang and W. Wang, “Finger vein recognition method based on GLCM-HOG and SVM,” 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 2020, pp. 698-701, doi: 10.1109/ICISCAE51034.2020.9236798.

[23] L. S. Athanasiou, D. I. Fotiadis, L. K. Michalis, “Plaque Characterization Methods Using Intravascular Ultrasound Imaging,” in Atherosclerotic Plaque Characterization Methods Based on Coronary Imaging, Academic Press, 2017, ch, 4, pp. 71-94, doi: 10.1016/B978-0-12-804734-7.00004-X.

[24] M. Tuceryan, A. K. Jain, “Texture Analysis,” in The Handbook of Pattern Recognition and Computer Vision, 2nd ed., C. H. Chen, L. F. Pau, P.S. P Wang, Eds., World Scientific Publishing, 1998, ch. 2.1, pp. 207-248.

Phase Noise and Jitter Measurements in SEU-Hardened CMOS Phase Locked Loop Design

Naheem Olakunle Adesina
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
nadesi1@lsu.edu

Ashok Srivastava
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
eesriv@lsu.edu

Md Azmot Ullah Khan
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
mkhan42@lsu.edu

Jian Xu
Division of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803, USA.
jianxu1@lsu.edu

***Abstract*—Single event upset (SEU) is a significant problem in analog, digital, and mixed signal circuits. The extent of the attacks increases in radiation susceptible environments such as military and aerospace. Phase locked loop (PLL) is ubiquitous and usually employed as data recovery or clock signal in some electronic devices used in these environments. Single event transient causes ionizing particles to interact with the transistor and generate more leakage current that can result in malfunctioning of the transistor. A radiation hardened PLL is proposed whereby each block is designed to be SEU tolerant. Dual and triple redundancies are employed in the design of phase-frequency detector and frequency divider, respectively. The results show that the phase-locked loop operates from 3.5 to 4 GHz with the center frequency of 3.9 GHz. The phase noise of the voltage-controlled oscillator is estimated to be -109.5 dBc/Hz at 10 MHz offset frequency and the jitter is 128 ps at 3.9 GHz.**

Keyword*—*phase-locked loop (PLL), phase noise, Radiation effect, Single event upset (SEU), Tuning range, Voltage controlled oscillator (VCO)

I. INTRODUCTION

Short channel transistors, owing to their low noise margin and immunity, are more susceptible to single event upset (SEU) such as radiation and ionization effects. This causes energized particles to deposit charges in the transistor, which results in slight changes in its composition. The charge participates in the process of diffusion and generates electron-hole pair. The induced current increases and might alter the usual operation of the device. Studies have shown that SEU introduces errors into both combinational, sequential, and analog circuits. So, virtually every circuit in the field of VLSI design is prone to this effect. Ionization of particles or cosmic rays create oxide-traps and free carriers in silicon. The adverse effect is that there are formations of dangling bonds in $\mathrm{Si0_2}$ – bulk interface that affect the functionality of the transistor. Single-event transient is also referred to as unwanted asynchronous signals, which can have adverse effects on the performance of a circuit. It propagates through signal paths and causes undesirable responses in the circuit. SEU can change the state of digital circuits from 0 to 1 and vice versa, thereby, causing errors or indeterminate state. Radiation effects introduce particles such as proton, heavy ion, neutron etc. that can result in primary and secondary interactions. The charged particle (proton, electron, ion) participate in primary interactions and induces ionization while the neutron, because of its chargeless nature, bombards with heavy nuclei and generates secondary particles such as alpha, gamma, and beta. This can further produce additional current in transistor device by exciting the electrons in the bulk silicon. Initially SEE was not considered in older technologies, but the scaling down of transistor channel has pushed for its consideration. Short channel effect, such as drain induced barrier lowering (DIBL), is one of the major problems of CMOS technology scaling. The $\mathrm{I_{OFF}}$ current increases as a result of high leakage current, which eventually reduces the $\mathrm{I_{ON}}$ /$\mathrm{I_{OFF}}$ current ratio. In this case, it might be a little bit difficult to switch off the transistor and the device is no more suitable for digital applications. The effects of single event can produce additional leakage current that deteriorate the performance of the transistor. There are various studies that employed circuit-level simulation to discuss SEU and characterize its effects in a circuit. A comprehensive review of single event upsets is presented in [1] followed by improvements in single even transient (SET) modeling for future technologies. This approach incorporates energy disposition, peripheral charge sharing phenomenon, and its collection mechanism. Monte Carlo-based radiative energy disposition is used to simulate radiation effects with a purpose to predict its associated physical phenomena and examine the behavior of electronic circuits and devices under irradiation. It includes the upsets of muons and energetic electrons, which were not observed in experimental approach [2]. Song et al proposed an experimental approach to validate the simulation-based techniques. It is shown that the results are in good agreements, but the heavy ion experiments can resist different simulation conditions and is more accurate for SEU evaluation [3]. Phase locked (PLL) loop consists of both analog and digital building blocks that are easily prone to single event upset. This can unlock the PLL from its locking condition and alters its

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

features such as loop stability and dynamic response. In this work, we propose SEU-resistant phase locked loop structure, estimate the phase noise and jitter, and compare the results with related work in the literature.

The paper is organized as follows: Section II presents the proposed PLL architecture and examines the design of each component. Section III discusses the jitter and phase noise measurements and Section IV draws the conclusion.

II. PROPOSED SEU-HARDENED PHASE LOCKED LOOP

The proposed architecture, as shown in Fig. 1, is made of the usual blocks; voltage controlled oscillator, frequency divider, charge pump/loop filter, and phase frequency detector. However, each component is designed to resist single event transient by employing radiation-hardened logic gates, flip-flops, and triple modular redundancy (TMR).
Unlike conventional tri-state PFD, the phase frequency detector (PFD) we employed in this work is designed with two D flip flops, but without NAND gate. The reset path is also modified in order to reduce delay. Each D flip flop comprises of radiation hardened inverter and NAND gate, which are proposed in [4, 5].

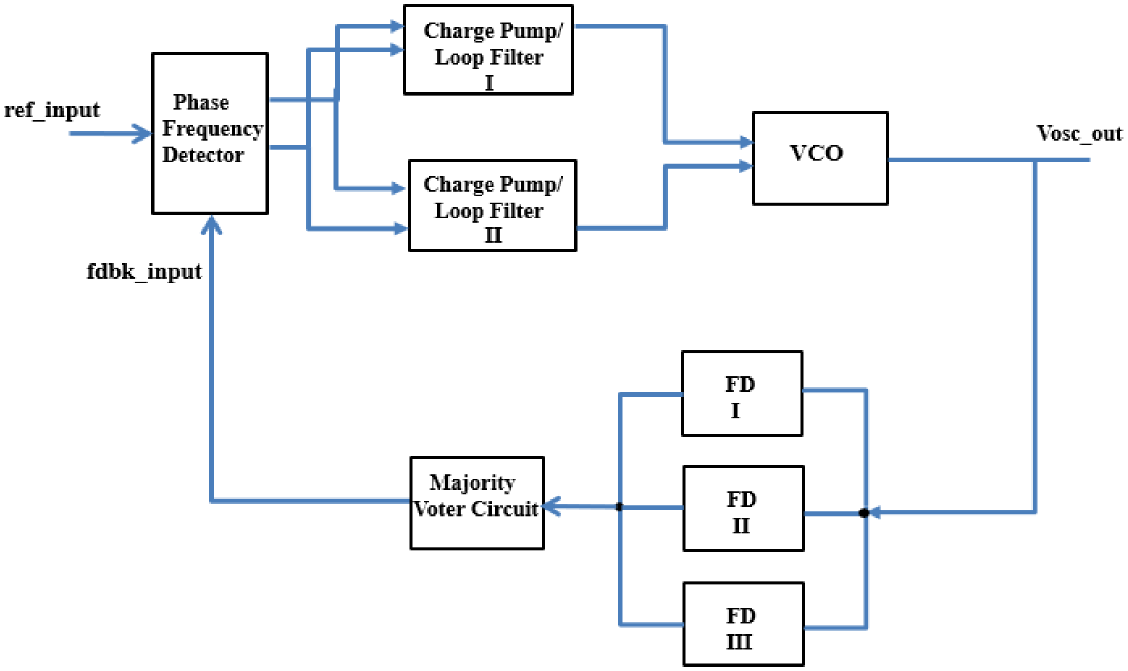
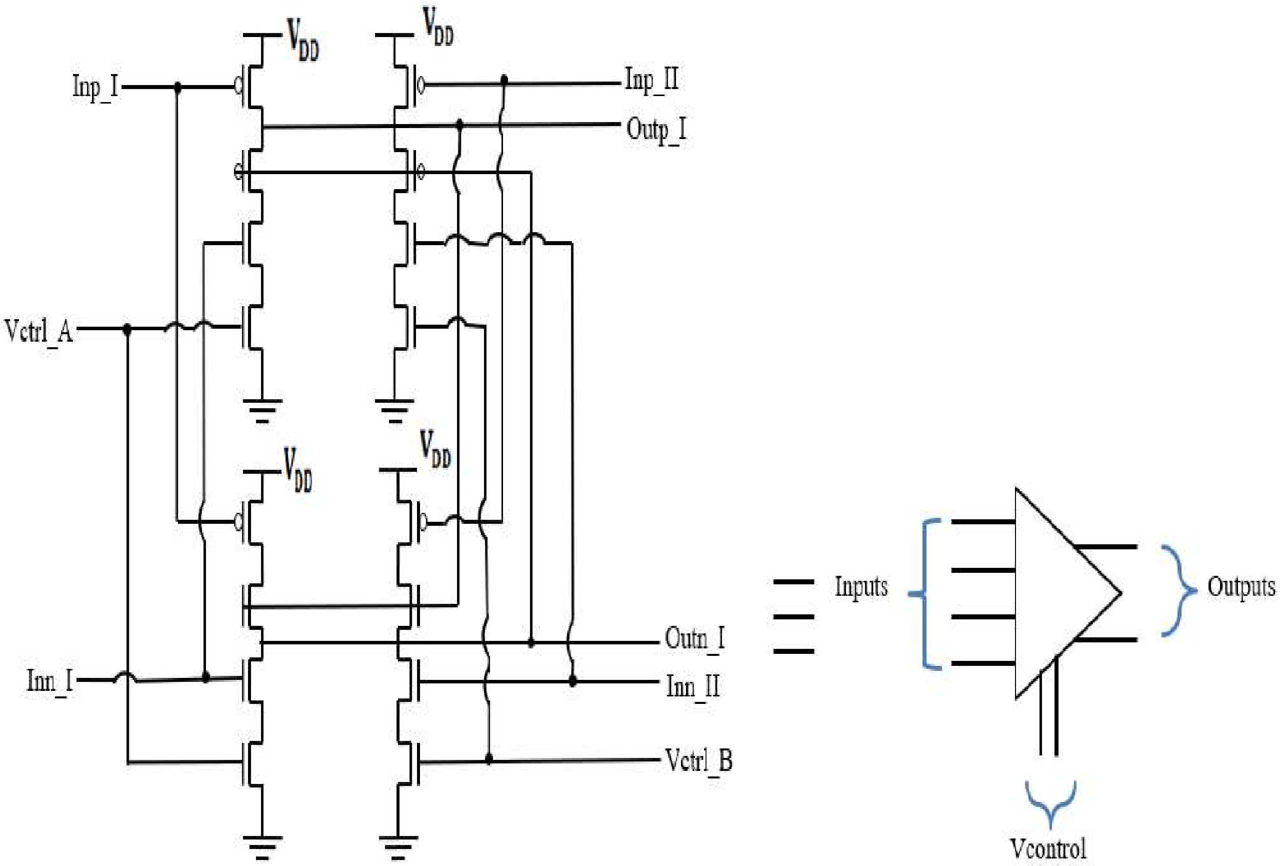


Fig. 1. Proposed single event tolerant PLL.

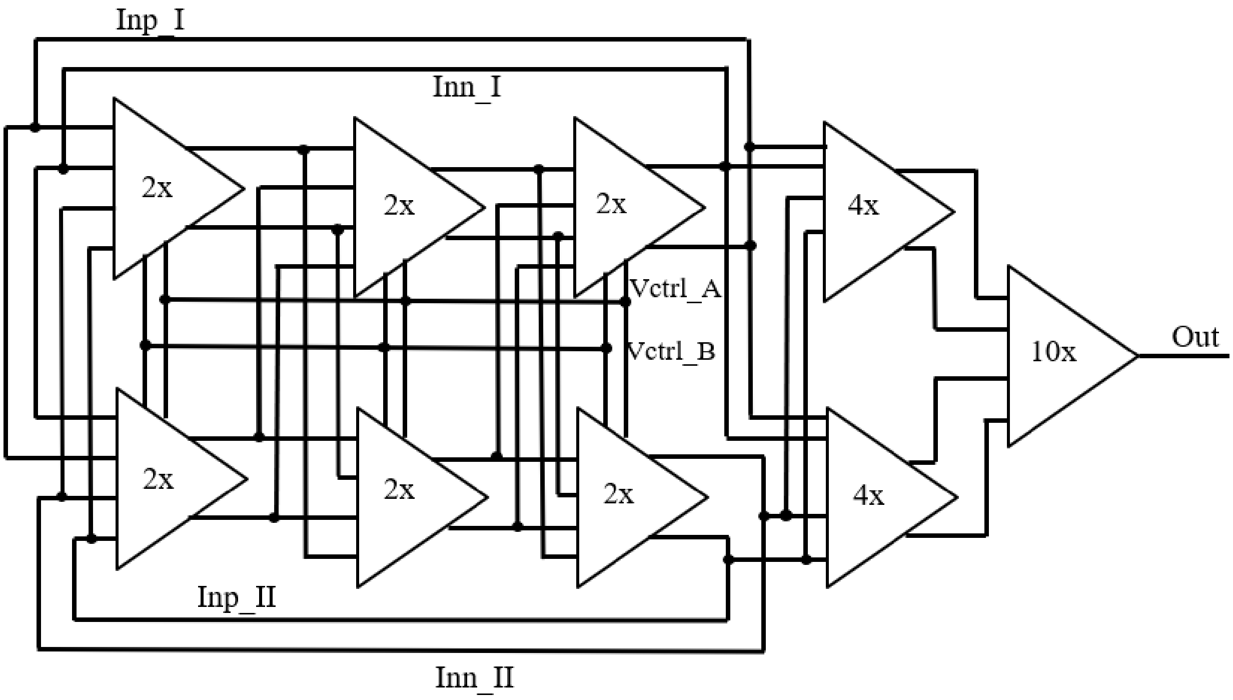
It is demonstrated experimentally that the inverter alleviates total ionizing dose (TID), has least variation in switching point, high energy efficiency, and occupies small silicon area. Similarly, the two input NAND gate is a standard cell radiation tolerant device, which consists of 4 inputs and 2 output.

The detailed information about its operation and why it is radiation resistant is already presented in [4]. The purpose of PFD circuit in phase locked loop is to sense the frequency and phase differences in the reference and feedback inputs. It generates outputs (UP and DOWN) that are applied to the input of charge pump. Since the CP/LF is an analog circuit, we duplicate the block so as to create alternative paths for the output from loop filter. By employing redundancy in the circuit design, one of the outputs is preserved even when the other is affected by radiation. The basic configuration of charge pump consists of biasing circuit, which supplies (draws) constant current to (from) the output when the UP (DOWN) signal is high. In this case, the two outputs of phase frequency detector are translated to charge pump current. The capacitors in the loop filter are discharged and charged through the current in order to generate the control voltage that serves as the input to voltage controlled oscillator [6]. For good noise rejection and optimum loop stability, the loop parameters are carefully chosen. The control voltage pulls up or pulls down the frequency of VCO depending on which of the two PFD's inputs is leading. The output continues to fluctuate until the PLL is locked. At this point both the reference input and feedback are in phase and the phase frequency detector generates no error signal. Therefore, the stable output frequency of VCO at lock condition is considered as its operating or center frequency. The voltage controlled oscillator is designed to be SEU tolerant by employing two different sets of three stage current starved VCO. Each oscillator has a modified version of the delay stage shown in Fig. 2(a). The inverter is also radiation hardened with a total of four inputs and two outputs. In addition, it has two other inputs (V_{ctrlA} and V_{ctrlB}) which controls the NMOS transistors in the inverter configuration. The ring oscillator has a similar cascaded structure like the conventional CSVCO, but only two inputs of the delay stage are controlled and driven by an inverter within the same ring, while the other inputs are controlled by an inverter from another ring [7]. For instance, if one of the inputs of ring oscillator is affected by single event transient, its output is minimally delayed because the alternate inverter of the ring inverter can still drive the output. As shown in Fig. 2(b), the outputs from both ring oscillators are driven by other delay stages until the drive strength is sufficient for a single inverter to generate the final output. The transistors are properly sized for the desired center frequency of 3.9 GHz and tuning range of 3.5 GHz to 4 GHz. This is discussed extensively in Session 3.

Frequency divider (FD) and VCO are highly vulnerable to single event upset because they operate at high frequency. If the divider is affected by SET, the feedback signal and charge pump are also affected, which leads to distortion in the phase of voltage controlled oscillator and increases the bit error rate (BER). Therefore, it is desirable to design a radiation hardened FD for phase locked loop, most especially for high radiation environment, such as spacecraft and military applications. The radiation tolerant flip flop is cascaded in three folds to obtain divide-by-eight frequency divider. Subsequently, the outputs of three different FD are applied to each of the inputs of voter circuit. By considering that one of the dividers is struck by radiation, the two remaining frequency dividers still produce the desired outputs which aids the voter circuit to elect the majority and output the correct value. Although the voter circuit is not modified for radiation effect, it can be made less radiation sensitive by adding dummy transistors or guard rings to its layout [8].



(a)



(b)

Fig. 2. (a) Radiation hardened delay (b) Current starved VCO.

III. PHASE NOISE AND JITTER MEASUREMENTS

A typical PLL is characterized by its phase noise, jitter, tuning range, and power consumption. The voltage controlled oscillator, being the heart of phase locked loop, contributes largely to the total phase noise in PLL [9]. The purity of spectral is also determined by VCO. The sources of noise from oscillator can be categorized into 3 major parts: thermal noise, flicker noise, and power supply noise. In this work, we have only considered device noise (flicker and thermal noise) because they contribute majorly to phase noise performance [10]. Both are random and act as baselines or minimum phase noise obtainable in VCO. We present phase noise in terms of deviation of the output of VCO from the reference input. It is expressed as:

$$L\{\Delta f\} \approx \frac{8kTV_{DD}}{3\eta P\, V_{char}} \frac{{f_0}^2}{\Delta f^2} \qquad (1)$$

where η , V_{char} , Δf are constant, device voltage characteristics, and offset frequency, respectively. f_0 is the centre frequency, P is power consumption and is given as:

$$P = N \times V_{DD} \times I_{DD} \qquad (2)$$

Jitter is also a critical issue in analog design because a clock with timing errors or an oscillator with high jitter in its waveform can drastically limit the speed of digital interface, alters the dynamic range of ADC, and increases BER. It is a measure of deviation, in time domain, of oscillator’s output from its ideal clock. Since the creation of redundancy in radiation-hardened VCO design employs more transistor, each transistor device contributes to circuit delay and increases the jitter. Depending on the operating frequency of oscillator and its delay, the rise and fall time of VCO fluctuates, which have adverse effects on the accumulated jitter. The equation for jitter is presented in (3);

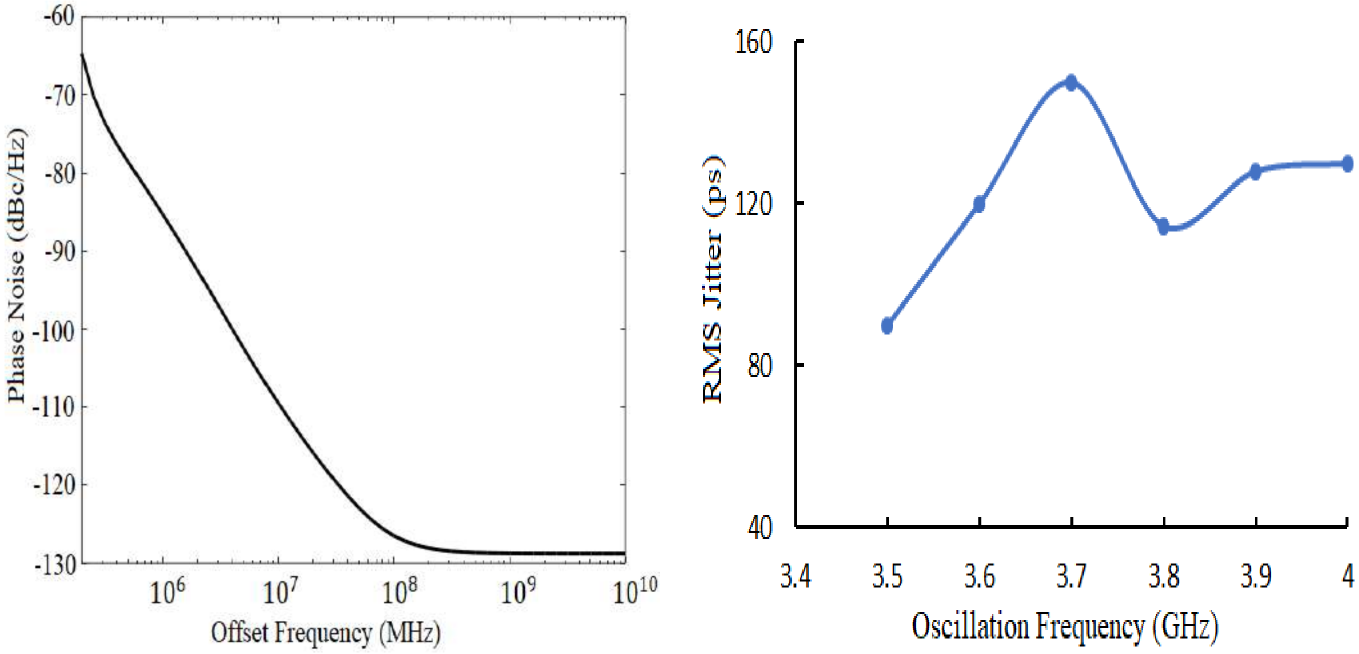
$$\sigma_{\Delta T} = \kappa\sqrt{\Delta T} \qquad (3)$$

and the proportionality constant, κ, is

$$\sqrt{\frac{8}{3\eta}}.\sqrt{\frac{kT\times V_{DD}}{P\times V_{Char}}} \qquad (4)$$

From Fig. 3, the phase noise -109.5 dBc/Hz measured at 10 MHz offset frequency and -125 dBc/Hz measured at 100 MHz offset, which is the corner frequency. It can be inferred that at low frequency, the phase noise is dominated with $^1/_f$ noise and the thermal noise is predominant at high frequency. Similarly, for the jitter, the minimum value is 90 ps at operating frequency of 3.5 GHz. The jitter increases linearly until it reaches maximum value at 3.7 GHz. The centre frequency of the radiation hardened VCO is 3.9 GHz and it produces a jitter of 128 ps. By using (5), we obtained the tuning range of the proposed PLL as 12.8%, which is lower when compared with non-radiation hardened structure [11, 12]. Fig. 4 is the layout of the delay stage of radiation-hardened voltage-controlled oscillator.

$$F_{TR} = \frac{f_{max} - f_{min}}{f_{center}} \tag{5}$$



(a) (b)

Fig. 3. (a) Phase noise (b) Jitter performance.

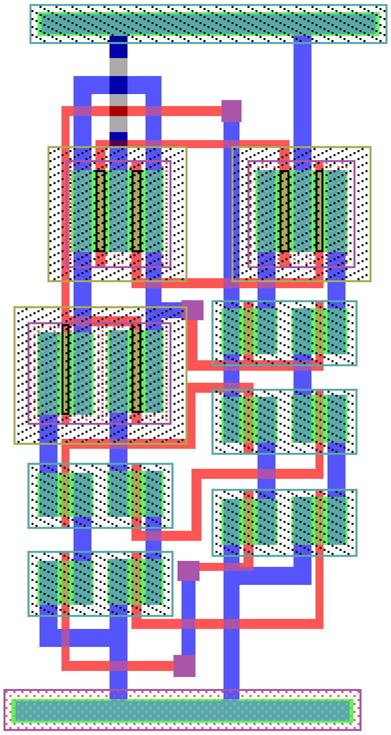


Fig. 4. Layout of SEU-hardened delay for current starved VCO.

IV. CONCLUSION

This work examines radiation effects on phase locked loop and propose a structure which is resistant to single event upsets. The phase frequency detector is designed with radiation hardened, D flip flop, NAND gate, and inverter. Two different CP/LPF are used in order to create alternative path for the control voltage to the voltage controlled oscillator. The proposed PLL structure operates from 3.5 to 4 GHz and the open loop phase noise is -109.5 dBc/Hz at 10 MHz offset. In addition, the results show that the maximum jitter is obtained at 3.7 GHz and reduces to 128 ps at centre frequency of 3.9 GHz.

ACKNOWLEDGEMENTS

The work is supported by the Louisiana State University (LSU) Economic Development Assistantships (002128), Louisiana Board of Regents Grant (LEQSF (2018-21)-RD-A-09), and the Office of Research & Economic Development (ORED) Grant (004475).

REFERENCES

1. L. Artola, M. Gaillardin, G. Hubert et al., IEEE Trans. Nucl. Sci., vol. 62, pp. 1528, 2015.
2. R. A. Reed, R. A. Weller, M. H. Mendenhall, D. M. Fleetwood, K. M. Warren, B. D. Sierawski, et al., "Physical Processes and Applications of the Monte Carlo Radiative Energy Deposition (MRED) Code," IEEE Trans. Nucl. Sci., vol. 62, no. 4, pp. 1441-1461, 2015.
3. R. Song, J. Shao, B. Liang, Y. Chi and J. Chen, "A Single-Event Upset Evaluation Approach Using Ion-Induced Sensitive Area," 2019 IEEE 13th International Conference on ASIC (ASICON), Chongqing, China, 2019.
4. R. Garg and S. Khatri, "A novel, highly SEU tolerant digital circuit design approach," 2008 IEEE International Conference on Computer Design, Lake Tahoe, CA, 2008.
5. S. Kim, J. Lee, I. Kwon, D. Jeon, " TID-Tolerant Inverter Designs for Radiation-Hardened Digital Systems," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 954, pp. 161473, 2020.
6. N. O. Adesina, and A. Srivastava, "Memristor-Based Loop Filter Design for Phase Locked Loop," J. Low Power Electron. Appl., vol. 9, no. 3, pp. 24, 2019.
7. R. Kumar, V. Karkala, R. Garg, T. Jindal, and S. P. Khatri, "A radiation tolerant phase locked loop design for digital electronics," Proc. IEEE Int. Conf. Comput. Design (ICCD*)*, pp. 505-510, 2009.
8. H. Yuan, Y. Guo, J. Chen, Y. Chi, X. Chen and B. Liang, "28nm Fault-Tolerant Hardening-by-Design Frequency Divider for Reducing Soft Errors in Clock and Data Recovery," IEEE Access, vol. 7, pp. 47955-47961, 2019.
9. N. O. Adesina and A. Srivastava, "A 250 MHz-to-1.6 GHz Phase Locked Loop Design in Hybrid FinFET-Memristor Technology," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, pp. 0901-0906, 2020.
10. L. Forbes, C. Zhang, B. Zhang and Y. Chandra, "Comparison of phase noise simulation techniques on a BJT LC oscillator," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 50, no. 6, pp. 716-719, 2003.
11. N. O. Adesina, and A. Srivastava, "Threshold Inverter Quantizer Based CMOS Phase Locked Loop with Improved VCO Performance," IEEE VLSI Circuits and Systems Lett., vol. 6, no. 3, pp. 1-13, 2020.
12. N. O. Adesina, A. Srivastava and M. A. U. Khan, "Evaluating the Performances of Memristor, FinFET, and Graphene TFET in VLSI Circuit Design," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, pp. 0591-0595, 2021.

Dual Tuned Switch for Dual Resonance 1H/13C MRI Coil

Gameel Saleh
Biomedical Engineering Department
College of Engineering
Imam Abdulrahman Bin Faisal University
P.O. Box 1982, Dammam, 31441, Saudi Arabia
gsmohammed@iau.edu.sa

Ashraf Abuelhaija
Department of Electrical Engineering
Faculty of Engineering and Technology
Applied Science Private University
Amman, Jordan
a_abualhijaa@asu.edu.jo

***Abstract*—This paper introduces two transmit/receive switch designs for 7 Tesla magnetic resonance spectroscopic imaging. Both designs based on microstripline-based couplers. The first is a dual-tuned 1H/13C switch with two concentric microstriplines on each side of the switch. A branch line technique from transmission line theory is applied to compact the switch to half of its initial dimension. The second proposed design is a dual tuned 1H/13C microstripline-based switch with one microstripline in each side of the switch. The second design benefits from the harmonics of an initial resonance signal to shift the first and second harmonics to the same Larmor frequencies of the 1H and 13C nuclear spins when exposed to 7 Tesla static magnetic field. The first design serves 1H and 13C RF Coils, working independently. Whereas the second design works with a dual resonance 1H/13C RF coil. The first and second designs achieved good matching less than -15 dB and -10 dB, respectively. They achieved low insertion loss less than -0.6 dB and -1.2 dB, respectively. The isolation between couplers in the first design is higher than 60dB. Furthermore, the isolation between the amplifier port and the receiver for both switches designs are higher than 60dB as well. The proposed switches are promising in reducing the number of T/R switches and the area of the place they need in the transmit front-end when multichannel RF coils are used.**

***Index Terms*—couplers, dual tuned switch, microstripline (MSL), magnetic resonance spectroscopic imaging, radiofrequency (RF) coil, pi-shaped technique.**

I. Introduction

13C MRI is a noninvasive molecular imaging method that investigates the metabolic differences in organic molecules. These changes in biochemical processes are a measure to the existence of different diseases [1]. Since 13C available in many organic molecules, its abnormal quantification (because of the metabolic changes) might indicates to many diseases. A hyperpolarized process can be used to enrich 13C concentration and improved its MR signal to detect diseases. These include: Prostate cancer [2], brain tumors [3], diabetes [4], variety of inflammatory conditions [5, 6, 7], cardiovascular disease [8], and cardiac metabolism [9].

RF coils in MRI are used to interrogate 1H and/or X-nuclei. To interrogate spins of any type of atomic nuclei, the coil should resonate at a frequency equals the speed of precession of the nuclear spins of the element under interest when exposed to a certain strength of the static magnetic field. A single or multichannel 1H RF coils that resonant at a single operating frequency are extensively introduced in [10, 11]. Similarly, commercial single or multichannel 13C RF coils are available [12]. Multichannel RF coils have been used to achieve better coils performance and hence enhance the quality of the scanned image. The performance of the coils improves by increasing the rotating magnetic field (B1) and the signal-to-noise-ration (SNR), and by decreasing the Electric field and hence the specific energy absorption rate (SAR) [13, 14, 15]. For 7 Tesla MRI, the 1H and 13C coils resonate at 298MHz and 75MHz, respectively. RF coils for 1H and 13C are designed independently and in dual resonance [16]. However, the available of commercially dual-tuned 1H/13C multichannel coils are essential to the SNR imaging of all body areas [1]. The challenge also is in the front end transmit RF circuit. From transmit/receive (T/R) switches perspective, the number of the required switches, and their size are also a challenge. The ability to design a dual tuned switch to handle the signal to/from a dual resonance coil at the same time is essential, accordingly.

Once RF coils are used to transmitting/receiving RF signals to/from the body, T/R switches must be used in the RF-transmitter front-end to handle the signal to/from the RF coils. For a 1H single resonance RF coil, a T/R switch is required to deliver the amplified signal power to the RF coil or to handle the received MR signal from the coil to the receiver. T/R switches of different topologies are introduced in the literature, some of which use PIN diodes [17], transistors [18, 19], and MEMS [20, 21]. A PIN diodes-based switch is designed to reduce the switching time to 1μs [17]. Transistors-based switches [18, 19] are proposed to minimize the operating current, power, and fields distortion but they obeyed less SNR values. A dual tuned MEMS-based 1H/19F T/R switch compared to PIN diodes-based topology is introduced in [20]. Results showed the advantage of MEMS topology in achieving higher isolation, and the advantage of PIN diode topology in achieving higher image SNR [21].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

A dual tuned 1H/31P switch for 3 Tesla MRI is designed with an insertion loss of 0.7 dB and 1.2 dB for 1H and 31P, respectively [22]. A T/R switch for preclinical 14.9 Tesla is presented in [23] with an insertion loss of 1 dB. A compact dual tuned 1H/23Na switch was introduced with a compact size to handle the signals into two RF coils of frequencies corresponding to the speed of rotation of 1H and 23Na atomic nuclei [24]. The advantage of the MSL-based coupler switch over the conventional couplers in heat dissipation and its capability in handling more power signal is explained in [25]. The first disclosed switch design in [25] is successfully used with 32 channels RF coils in [26].

In this paper, two different designs of T/R switches are proposed. In the first design, we introduce a dual tuned 1H/13C T/R switch to handle two signals at the same time to/from two different single resonance 1H and 13C RF coils. The proposed dual tuned switch based on microstripline coupler technology with two concentric MSLs on each side of the switch, each connected with a coil. A branch line technique from circuit theory is applied to compact the size of the proposed switch. In the second design, we introduce a novel dual tuned 1H/13C T/R switch with single MSL on each side of the switch that can deliver a dual resonance signal (corresponding to the 1H and 13C frequencies) to/from a dual resonance 1H/13C RF coil.

II. Dual-tuned Double-Coupler MSL-based 1H/13C T/R Switch

A. The Working Principle and Research Method

Although modern MRI scanners employ separate transmit and receive RF coils, some imaging of body parts such as head and knee imaging still used transmit/receive RF coil. This type of coil requires T/R switch to separate the transmitted and received signals. Once a dual-tuned 1H/X-nuclei T/R RF coil is used, a dual-tuned 1H/X-nuclei T/R switch is required for integration with this coil. To understand the working principle of dual tuned T/R switches, the block diagram of the first proposed design is drawn in Figure 1. This switch consists of pair of two concentric microstripline-based hybrid couplers. Each coupler in the left side has been joined with the corresponding coupler in the right side using metallic rods. The inner and outer MSLs couplers on each side of the switch are corresponding to 1H and 13C signals, respectively. During transmit, two couplers of MSLs on the left side are used. The 1H and 13C RF signals are directed from ports 1&5 (power amplifiers' ports) to ports 2&6 (RF coils' ports). This is accomplished by forward biasing all PIN diodes in the switch. During receive, all PIN diodes are reverse biased and the detected 1H and 13C signals are directed from ports 2&6 to the receiver ports 4&8. This switch has been designed using CST Microwave Studio as shown in Figure 2. This design relies on folded microstripline-based hybrid couplers to create a T/R switch with dimensions 240mm × 220mm. RO3010 Rogers substrate with height 1.27mm, ε_r=10.2 and tanδ=0.0022 has been used.

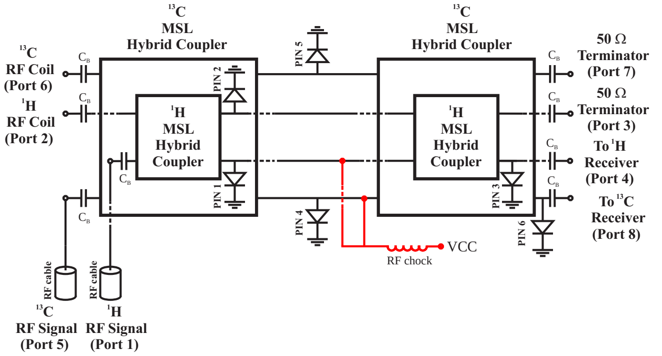


Fig. 1. Block diagram of a dual-tuned T/R switch

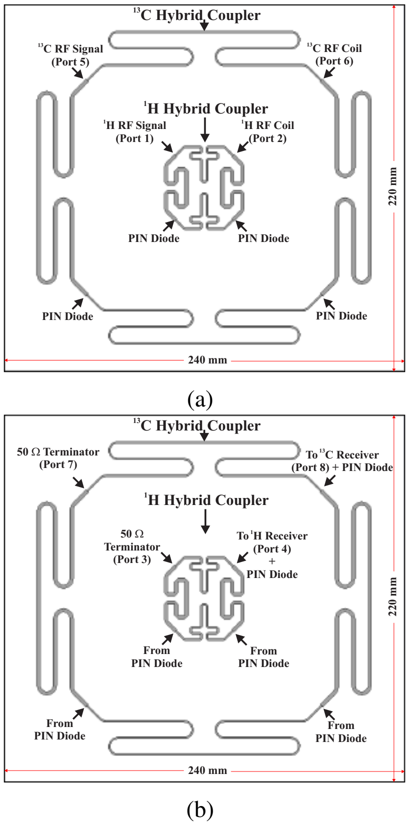
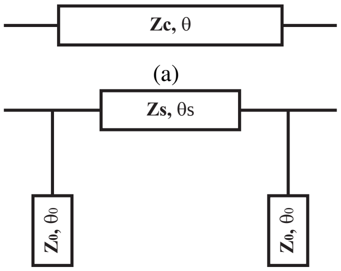


Fig. 2. The 1H/13C T/R switch using two concentric MSL couplers at (a) the top and, (b) the bottom face.

B. Compacting the Switch using the pi-Shaped Technique

Quadrature hybrid couplers are considered valuable passive devices in several modern communication systems. However, designing such couplers for applications operate at low frequencies increases the occupied area in the system. Therefore, several techniques have been proposed to miniaturize the overall size of the designed couplers [27, 28, 29, 30]. In this context, a pi-shaped technique [29] has been applied to reduce the size of the dual-tuned T/R switch in Figure 2. This technique makes use of the transmission line theory and



(a)

(b)

Fig. 3. (a) Conventional transmission line, (b) pi-shape equivalent transmission line.

TABLE I. 75 MHz coupler with θ_s=45

	Z_c=50 Ω, θ=90°		Z_c=35.355 Ω, θ=90°	
Z_s	70.71 Ω	w=0.45 mm	50 Ω	w=1.07 mm
θ_s	45°	l=189 mm	45°	l=184 mm
Z_0	70.71 Ω	w=0.45 mm	50 Ω	w=1.07 mm
θ_0	45°	l=189mm	45°	l=184 mm

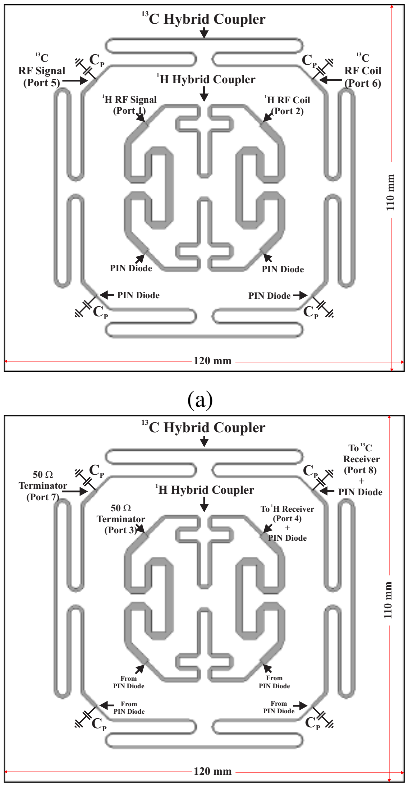
replaces each branch in the coupler by a pi-shaped equivalent circuit as shown in Figure 3. A pi-shaped equivalent circuit comprises one series transmission line associated with two open stubs at its terminals.

The pi-shaped equivalent transmission line parameters are calculated based on the following:

$$Z_s = \frac{Z_c\, sin\theta}{sin\theta_s} \tag{1}$$

$$\frac{tan\,\theta_o}{Z_o} = \frac{cos\theta_s\ - cos\theta}{Z_c\, sin\theta} \tag{2}$$

where Z_c is the characteristic impedance and θ is the electrical length of the conventional transmission line. In Equation 1, Z_s is the characteristic impedance of the series transmission line in the pi-shaped equivalent circuit with θ_s electrical length. The two open stubs have characteristic impedance Z_o and electrical length θ_o. The calculations of the pi-shaped equivalent transmission line parameters have been summarized in Table I. These calculations are based on predefined values of θ_s and θ_o which have been chosen to be 45°. Our analysis showed that using greater angles can more compact the structure . However more compacting leads to unrealistic microstriplines widths which might not be able to handle high power signals for this application. In addition, this table summarizes the corresponding width "w" and length "l" for each microstripline branch in the coupler. It is worth to mention that, these calculations have been done for the 13C coupler (the outer coupler). Figure 4 shows the compact dual-tuned T/R switch where C_p is a shunt capacitor replacing each adjacent open stubs in the pi-shaped equivalent transmission lines. This capacitor has a value of 72.5 pF. The dimensions of the T/R switch in Figure 4 show that the pi-shaped technique reduces the classical T/R switch design to the half.



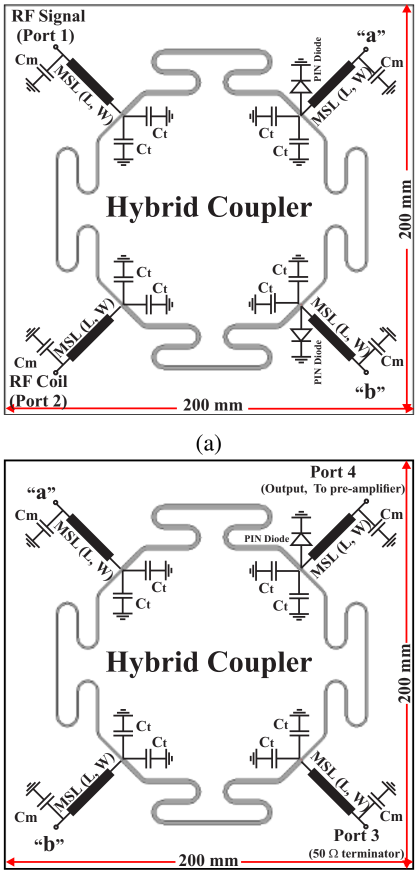
(a)

(b)

Fig. 4. The compact dual-tuned double-coupler MSL-based 1H/13C T/R switch (a) the top side, (b) the bottom side.

III. Dual-tuned Single-Coupler MSL-based 1H/13C T/R Switch

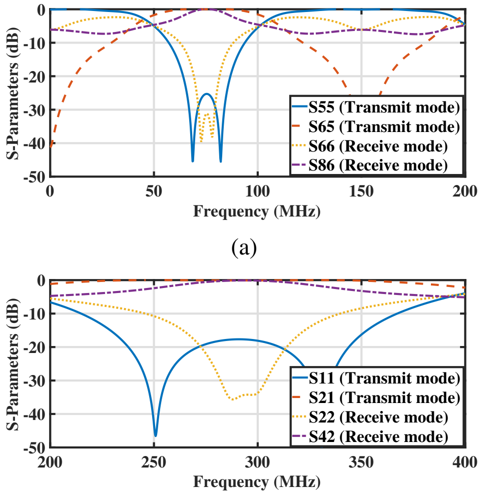
This switch is being designed to deliver a dual tuned signal, of 1H and 13C frequencies, to/from a dual resonance 1H/13C RF coils, simultaneously. It comprises a single microstripline-based hybrid coupler on each side of the switch rather than 2 couplers on each side compared to the previous switch, see the block diagram of the switch in Figure 5. During transmit, a signal of two resonance frequencies (corresponding to 1H and 13C at 7 Tesla) will bypass from the power amplifier(s) at port 1 to the dual resonance 1H/13C RF coil at port 2. During this mode, the PIN diodes are forward biased. During receive mode, the MR received signal from the 1H and 13C RF coils will be delivered from port 2 to the the receiver at port 4. During this mode, the PIN diodes are reverse biased. This switch relies on the first and second harmonics of the microstripline-based hybrid coupler. After an iterative simulations of the coupler, the obtained first and second harmonics can be shifted to two frequencies corresponding to 1H and 13C. The shifting process has been accomplished by shunt capacitors (C_t=61pF). The matching network has been accomplished by matching capacitor (C_m=13pF) and a MSL of specific length (l=30mm) and width (w=1.2mm). Figure 5 shows the dual-tuned single-coupler microstripline-based 1H/13C T/R Switch where the hybrid couplers have been



(a)

(b)

Fig. 5. Dual-tuned single-coupler MSL-based 1H/13C T/R switch. (a) The top side, (b) The bottom side.



(a)

(b)

Fig. 6. First design S-parameters (a) 13C at 75MHz (b) 1H at 298MHz.

designed initially to operate at 131 MHz .

The novelty of this paper is in (i) introducing new T/R switches based on microstriplines couples that proofed better performance in heat dissipation and handling high power compared to the classical couplers [25] (ii) compacting the size of the switch by 50% compared to the initial design, with the advantage in using the switch with half the number of the multichannel-single resonance RF coils (iii) designing a novel dual tuned switch that handle two frequencies signal to a dual resonance RF coil, at the same time and without tuning during operation. This proposed design has the advantage in using the switch with half the number of the multichannel-dual resonance RF coils.

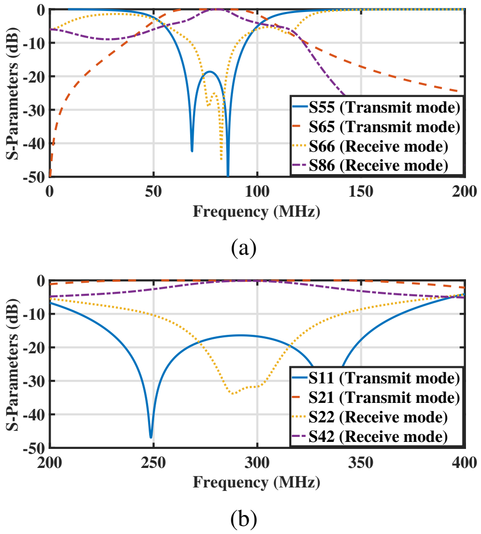
IV. The Simulation Results

A. *The Results of The Dual-Tuned Double-Coupler Switch*

The performance of the first dual-tuned 1H/13C T/R switch (without compacting) that is shown in Figure 2, has been verified using electromagnetic simulation for S-parameters as shown in Figure 6. During transmit mode, good matching for the 1H signal at port 1 (S11$\simeq$-18dB) and the 13C signal at port 5 (S55$\simeq$-25dB) has been achieved. Moreover, low insertion loss between ports 1&2 (S21$\simeq$-0.3dB) and ports 5&6 (S65<-0.2dB) has been achieved for the 1H and 13C amplifiers-coils delivered signals, respectively. During receive mode, good matching for the 1H RF coil signal at port 2 (S22$\simeq$-34dB) and for the 13C RF coil signal at port 6 (S66$\simeq$-30dB) has been achieved. In addition to that, low insertion loss between ports 2&4 (S42$\simeq$-0.3dB) and ports 6&8 (S86<-0.2dB) for the 1H and 13C coils-receivers signals have been achieved, respectively. The isolation between the inner and outer coupler is around 70 dB. In addition, the isolation between the amplifier port and the receiver during transmit mode is 62 dB and 70 dB for 1H and 13C couplers, respectively. This design has an insertion loss lower than that achieved by other switches, where 1.2dB was revealed for for the X-nuclei [22], and 1dB for the 1H [23].

B. *The Results of The Compact Dual-Tuned Double-Coupler Switch*

The S-parameters of the compacted switch is shown in Figure 7. In this design, the proposed dual tuned switch has been compacted in size to half, from 240mm × 220mm to 120mm × 110mm. The pi-shaped technique from transmission line theory is used. During transmit, good matching for the 1H signal at port 1 (S11$\simeq$-16dB) and the 13C signal at port 5 (S55$\simeq$-19dB) has been achieved. Moreover, low insertion loss between ports 1&2 (S21$\simeq$-0.3dB) and ports 5&6 (S65<-0.25dB) has been achieved, for the 1H and 13C amplifiers-coils delivered signals, respectively. During receive, good matching for the 1H RF coil signal at port 2 (S22$\simeq$-32dB) and for the 13C RF coil signal at port 6 (S66$\simeq$-25dB) has been achieved. Moreover, the insertion loss between ports 2&4 (S42$\simeq$-0.5dB) and ports 6&8 (S65<-0.6dB) has been achieved for the 1H and 13C coils-receivers delivered signals, respectively. The isolation between the inner and outer



(a)

(b)

Fig. 7. The compact first dual tuned switch during transmit and receive for (a) 13C at 75 MHz, and (b) 1H at 298MHz.

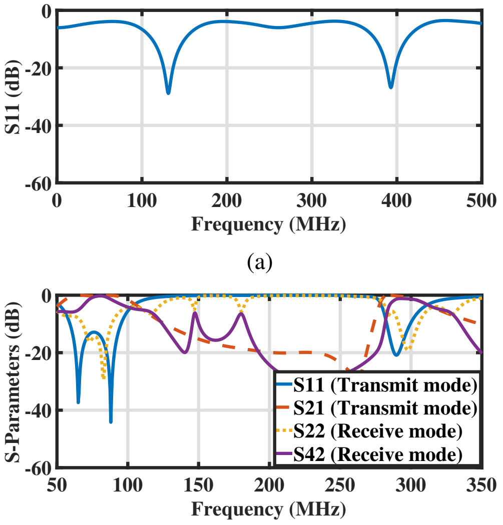
coupler is around 60 dB. In addition, the isolation between the amplifier port and the receiver during transmit mode is 62 dB and 67 dB for 1H and 13C couplers, respectively.

C. The results of the Dual-Tuned Single-Coupler Switch

In this design, a dual tuned 1H/13C switch is introduced to handle the signal to/from dual resonance 1H/13C RF coil. The switch is first designed to resonate at both 131MHz and 393MHz, as shown in Figure 8(a). The first (131MHz) and second (393 MHz) harmonics of the designed coupler are then shifted by the shunt capacitors to 75 MHz and 298 MHz, as shown in Figure 8(b) . During transmit, good matching at the power amplifier port 1 (S11≃-11.5 dB and -13dB) has been achieved for 1H and 13C frequencies, respectively. Low insertion loss between the power amplifiers and the dual-resonance coil ports 1&2 (S21≃-0.8 dB and -0.27 dB) has been achieved for 1H and 13C frequencies, respectively. During receive, good matching at port 2 (S22≃-19 dB and -15dB) has been achieved for the 1H and 13C frequencies respectively. Further, low insertion between the coil and receiver ports 2&4 (S42≃-1.2 dB and -0.8 dB) for 1H and 13C frequencies, respectively. In addition, the isolation between the amplifier port and the receiver during transmit mode is 62 dB at 1H and 13C frequencies.

V. Conclusion

In this paper, two designs of dual tuned 1H/13C T/R switches have been designed to serve 1H/13C RF coils working independently and in dual resonance at 7 Tesla MRI. Both switches based on microstriplines topology which is promising in heat dissipation and handling more power to/from the coils. The first switch has been designed with two concentric MSLs couplers on the top and bottom of the switch. This switch is used to deliver two signals of 1H and 13C frequencies to/from 1H and 13C RF coils, independently. Using one switch with two RF coils has the advantage of reducing the number of T/R switches to half the number of RF coils. A pi-shaped technique from transmission line theory is used to reduce the dimension of each dual tuned switch of the first type by 50%. This reduction in size is promising in using a reasonable area for the T/R switches when they are used near to multichannel RF coils. The second switch has a novel design with a single MSL coupler on the top and bottom of the switch. This switch is used to deliver a dual resonance signal (of 1H and 13C frequencies) to/from a dual resonance 1H/13C RF coils. Our proposed dual tuned T/R switch is a promising solution whenever multichannel dual resonance RF coils are required.



(a)

(b)

Fig. 8. The S-parameters for 1H/13C dual-tuned T/R switch.(a) Without the tuning capacitors, (b) with the tuning capacitors.

References

[1] Z. J. Wang, M. A. Ohliger, P. E. Larson, J. W. Gordon, R. A. Bok *et al.*, "Hyperpolarized 13C MRI: state of the art and future directions," *Radiology*, vol. 291, no. 2, pp. 273–284, 2019.

[2] S. J. Nelson, J. Kurhanewicz, D. B. Vigneron, P. E. Larson, A. L. Harzstark *et al.*, "Metabolic imaging of patients with prostate cancer using hyperpolarized [1-13C] pyruvate," *Science translational medicine*, vol. 5, no. 198, pp. 198ra108–198ra108, 2013.

[3] V. Z. Miloushev, K. L. Granlund, R. Boltyanskiy, S. K. Lyashchenko, L. M. DeAngelis *et al.*, "Metabolic imaging of the human brain with hyperpolarized 13C pyruvate demonstrates 13C lactate production in brain tumor patients," *Cancer research*, vol. 78, no. 14, pp. 3755–3760, 2018.

[4] L. M. Le Page, O. J. Rider, A. J. Lewis, V. Ball, K. Clarke *et al.*, "Increasing pyruvate dehydrogenase flux as a treatment for diabetic cardiomyopathy: a combined 13c hyperpolarized magnetic resonance and echocardiogra-

phy study," *Diabetes*, vol. 64, no. 8, pp. 2735–2743, 2015.

[5] K. Thind, M. D. Jensen, E. Hegarty, A. P. Chen, H. Lim *et al.*, "Mapping metabolic changes associated with early radiation induced lung injury post conformal radiotherapy using hyperpolarized 13c-pyruvate magnetic resonance spectroscopic imaging," *Radiotherapy and oncology*, vol. 110, no. 2, pp. 317–322, 2014.

[6] L. M. Le Page, C. Guglielmetti, C. F. Najac, B. Tiret, and M. M. Chaumeil, "Hyperpolarized 13C magnetic resonance spectroscopy detects toxin-induced neuroinflammation in mice," *NMR in Biomedicine*, vol. 32, no. 11, p. e4164, 2019.

[7] J. D. MacKenzie, Y.-F. Yen, D. Mayer, J. S. Tropp, R. E. Hurd, and D. M. Spielman, "Detection of inflammatory arthritis by using hyperpolarized 13c-pyruvate with MR imaging and spectroscopy," *Radiology*, vol. 259, no. 2, pp. 414–420, 2011.

[8] A. Z. Lau, J. J. Miller, M. D. Robson, and D. J. Tyler, "Simultaneous assessment of cardiac metabolism and perfusion using copolarized [1-13C] pyruvate and 13C-urea," *Magnetic resonance in medicine*, vol. 77, no. 1, pp. 151–158, 2017.

[9] C. H. Cunningham, J. Lau, A. Chen, B. Geraghty, W. Perks *et al.*, "Hyperpolarized 13C metabolic MRI of the human heart: Novelty and significance," *Circulation Research*, vol. 119, no. 11, pp. 1177–1182, 2016.

[10] B. Gruber, M. Froeling, T. Leiner, and D. W. Klomp, "Rf coils: A practical guide for nonphysicists," *Journal of magnetic resonance imaging*, vol. 48, no. 3, pp. 590–604, 2018.

[11] S. H. Rietsch, S. Brunheim, S. Orzada, M. N. Voelker, S. Maderwald *et al.*, "Development and evaluation of a 16-channel receive-only RF coil to improve 7T ultra-high field body MRI with focus on the spine," *Magnetic resonance in medicine*, vol. 82, no. 2, pp. 796–810, 2019.

[12] J. D. Sánchez-Heredia, R. B. Olin, M. A. McLean, C. Laustsen, A. E. Hansen *et al.*, "Multi-site benchmarking of clinical 13C RF coils at 3T," *Journal of Magnetic Resonance*, vol. 318, p. 106798, 2020.

[13] G. Saleh, K. Solbach, D. Erni, and A. Rennings, "Soft surface-EBG structure to improve the H/E field ratio of stripline coil for 7 Tesla MRI," in *21th Annual Meeting of the International Society for Magnetic Resonance in Medicine*, 2013.

[14] G. Saleh, F. Sibaii, N. Alashban, H. Alkhateeb, F. Hegazi, and M. Hegazi, "Effects of tissues and geometric shapes of phantoms on the specific energy absorption rate," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 28, no. 3, p. e21252, 2018.

[15] G. Saleh, K. Solbach, A. Rennings, and Z. Chen, "SAR reduction for dipole RF coil element at 7 Tesla by using dielectric overlay," in *2012 Loughborough Antennas & Propagation Conference (LAPC)*. IEEE, 2012, pp. 1–3.

[16] M. Oehmigen, M. E. Lindemann, M. Gratz, R. Neji, A. Hammers *et al.*, "A dual-tuned 13C/1H head coil for PET/MR hybrid neuroimaging: Development, attenuation correction, and first evaluation," *Medical physics*, vol. 45, no. 11, pp. 4877–4887, 2018.

[17] D. O. Brunner, L. Furrer, M. Weiger, W. Baumberger, T. Schmid *et al.*, "Symmetrically biased T/R switches for NMR and MRI with microsecond dead time," *Journal of Magnetic Resonance*, vol. 263, pp. 147–155, 2016.

[18] M. Twieg, M. Rooij, and M. Griswold, "Enhancement mode GaN on silicon (eGaN FETs) for coil detuning," in *Proc. Int. Soc. Magn. Reson. Med.*, vol. 22, 2014, p. 926.

[19] P. Grannell, M. Orchard, P. Mansfield, A. Garroway, and D. Stalker, "A FET analogue switch for pulsed NMR receivers," *Journal of Physics E: Scientific Instruments*, vol. 6, no. 12, p. 1202, 1973.

[20] H. Raki, K. T. V. Koon, I. Saniour, H. Souchay, S. A. Lambert *et al.*, "Serial and parallel active decoupling characterization using rf mems switches for receiver endoluminal coils at 1.5 T," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 511–10 520, 2020.

[21] A. Maunder, M. Rao, F. Robb, and J. M. Wild, "Comparison of MEMS switches and PIN diodes for switched dual tuned RF coils," *Magnetic resonance in medicine*, vol. 80, no. 4, pp. 1746–1753, 2018.

[22] B. Thapa, J. Kaggie, N. Sapkota, D. Frank, and E.-K. Jeong, "Design and development of a general-purpose transmit/receive (T/R) switch for 3T MRI, compatible for a linear, quadrature and double-tuned RF coil," *Concepts in Magnetic Resonance Part B: Magnetic Resonance Engineering*, vol. 46, no. 2, pp. 56–65, 2016.

[23] A. W. Magill, H. Lei, and R. Gruetter, "A high-power RF switch for arterial spin labelling with a separate tagging coil," in *Proceedings 20th Scientific Meeting of the International Society for Magnetic Resonance in Medicine*, no. CONF. International Society for Magnetic Resonance in Medicine, 2012, p. 2685.

[24] A. Abuelhaija and G. Saleh, "A pi-shaped compact dual tuned 1H/23Na Microstripline-based switch for 7Tesla MRI," *Antenna and Propagation (IRECAP)*, vol. 11, no. 1, 2021.

[25] A. Abuelhaija, G. Saleh, T. Baldawi, and S. Salama, "Symmetrical and unsymmetrical microstripline-based transmit/ receive switches for 7 Tesla magnetic resonance imaging," *submitted to be published in International Journal of Circuit Theory and Applications*, 2021, unpublished.

[26] S. Orzada, K. Solbach, M. Gratz, S. Brunheim, T. M. Fiedler, S. Johst, A. K. Bitz, S. Shooshtary, A. Abuelhaija, M. N. Voelker *et al.*, "A 32-channel parallel transmit system add-on for 7t mri," *Plos one*, vol. 14, no. 9, p. e0222452, 2019.

[27] S. Azizi, S. Rahim, and M. Sabran, "Realization of a compact branch line couple using semi-lumped element," in *2011 IEEE Symposium on Wireless Technology and Applications (ISWTA)*. IEEE, 2011, pp. 21–23.

[28] S. Gomha, E.-S. M. El-Rabaie, and A. A. T. Shalaby,

"Optimizing the performance of branch-line couplers using open ended stubs," in *2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE)*. IEEE, 2013, pp. 363–367.

[29] Y.-H. Chun and J.-S. Hong, "Compact wide-band branch-line hybrids," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 2, pp. 704–709, 2006.

[30] W.-L. Chang, T.-Y. Huang, T.-M. Shen, B.-C. Chen, and R.-B. Wu, "Design of compact branch-line coupler with coupled resonators," in *2007 Asia-Pacific Microwave Conference*. IEEE, 2007, pp. 1–4.

Assessment of Electromagnetic Energy Absorption Rate in Hibiscus Flower Model at 947.50 MHz and 1842.50 MHz

Nibedita Mukherjee
Electronics & Communication Engg. department
Budge Budge Institute of Technology
Budge Budge, India
mcnibedita@gmail.com

Ardhendu Kundu[#] & Bhaskar Gupta
Electronics & Telecommunication Engg. department
Jadavpur University
Kolkata, India
[#]ardhendukundu.1989@gmail.com,
gupta_bh@yahoo.com

Monojit Mitra
Electronics & Telecommunication Engg. department
IIEST Shibpur
Howrah, India
monojit_m1@yahoo.co.in

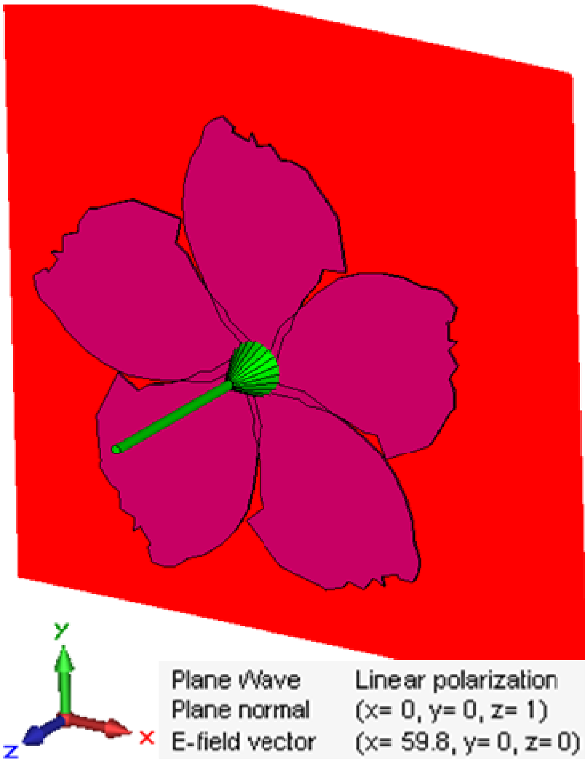
***Abstract*—Living biological tissues in plants, fruits and flowers possess considerably high permittivity and electrical conductivity over wide microwave frequency spectrum. Plants, being immobile, are continuously exposed to electromagnetic radiation emitted from cell tower antennas. As a consequence, plants are expected to absorb quite a large amount of incident electromagnetic energy over multiple frequency bands. Different global and national regulatory bodies have put limits on maximum permissible electromagnetic exposure to restrict absorption in humans and minimize related health risks. However, electromagnetic energy absorption in plants along with associated physiological and molecular effects has not been considered in these guidelines. Plants, in general, possess higher surface area while compared to humans or other living objects. In particular, flowers own reasonably high surface to volume ratio and consequently dissipate quite high amount of electromagnetic energy in a relatively less tissue mass. Hence, this paper aims at estimating specific absorption rate data for a hibiscus flower model at 947.50 MHz and 1842.50 MHz. Linearly polarized plane waves at those two frequencies impinge on the hibiscus flower model in separate simulations as per the International Commission on Non-Ionizing Radiation Protection guidelines. Maximum local point, averaged over 1g contiguous tissue and whole body averaged specific absorption rate data at those two frequencies are significantly different than the data reported earlier at 2450 MHz – indicating, the nature of dependence of specific absorption rate data on frequency of irradiation, incident field strength and dielectric properties of constituent flower tissue.**

***Keywords*—conductivity, hibiscus, permittivity, ICNIRP, electromagnetic radiation, specific absorption rate**

I. Introduction

With the increased use of different wireless communication systems, estimating electromagnetic energy absorption rate in human phantoms is now a standard practice for protecting human health [1-7]. Electromagnetic energy absorption rates in human phantoms are estimated and measured primarily due to near-field radiation from cell phones, wearable or implantable radio frequency devices. However, electromagnetic energy absorption rate estimations in different plant and fruit models have also been undertaken in recent time [8-13] – as the constituent tissues possess reasonably high dielectric properties [8-15] and moreover, plants are being continuously exposed to mobile tower radiations in far-field. The radiated power density in far-field or equivalent electric field strength, due to emission from antennas installed on mobile towers, is governed in accordance with the prescribed global or national electromagnetic regulatory guidelines [16-19]. Therefore, electromagnetic energy absorption rates estimation in different plant, fruit and flower models is absolutely necessary as per the global and national guidelines [8-13]. Electromagnetic energy absorption rate is quantified in terms of specific absorption rate (SAR) – i.e. the rate of electromagnetic energy absorption by a biological object when microwave radiation impinges on the same. SAR is in general averaged over point mass, 1g tissue mass, 10g tissue mass or the whole body mass as prescribed in protocol. Mathematical expression for point SAR is $\sigma|E|^2/2\rho$ – where, σ represents electrical conductivity of tissue, E represents peak value of internally developed electric field strength and ρ represents tissue density [13].

It should be noted that electromagnetic energy absorption rate estimations in typical flowers haven't been performed



Plane Wave   Linear polarization
Plane normal   (x= 0, y= 0, z= 1)
E-field vector   (x= 59.8, y= 0, z= 0)

Fig. 1. Typical hibiscus flower model exposed to plane wave with linear polarization at 947.50 MHz as per ICNIRP guidelines [12, 16].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Table I: Dielectric properties and tissue density of hibiscus flower [12]

Tissue	Density (Kg/m^3)	Frequency (MHz)	Permittivity	Loss Tangent
Hibiscus Flower	644.50	947.50	43.53	0.279
		1842.50	42.33	0.221

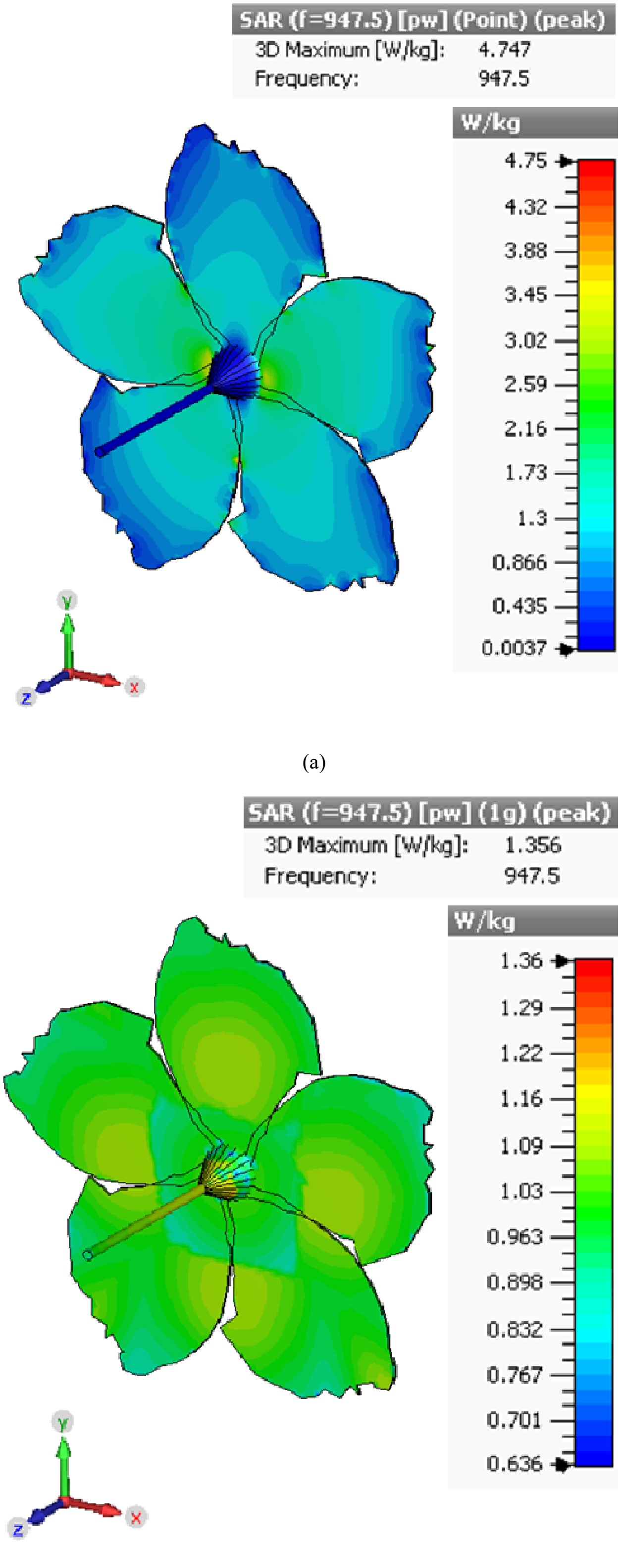
Table II: ICNIRP prescribed reference electromagnetic regulatory limits for public exposure zone [16]

Frequency (MHz)	Maximum permissible power density (W/m^2)	Equivalent peak E-field (V/m)
947.50	4.7375	59.77
1842.50	9.2125	83.34

earlier except one or two selective cases [12-13]. Broadband dielectric properties measurement of hibiscus flower tissue and subsequent SAR data assessment only at 2450 MHz was reported in a recent article [12] – but, SAR data estimations at other important telecommunication frequency bands such as 900 MHz and 1800 MHz downlink bands weren't reported for the designed hibiscus flower model. Hence, this article extends that work by estimating maximum local point SAR (MLP SAR), 1g averaged SAR (1g SAR) and whole body averaged SAR (WBA SAR) data at 947.50 MHz and 1842.50 MHz respectively. All SAR data have initially been planned to be simulated for a plane wave incidence with plane wave equivalent field strength specified as per electromagnetic regulatory guidelines prescribed by the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [16].

II. SAR Simulation Technique

The typical hibiscus flower model reported earlier in an article has been used here for SAR simulation in CST Microwave Studio 2016 [20] – moreover, frequency dependent dielectric properties at 947.50 MHz and 1842.50 MHz along with the measured tissue density data have been taken from the same published article [12]. The typical hibiscus flower model contains five petals along with a twig and possesses total mass of 2.30g – the same is illustrated in Fig. 1. The broadband dielectric properties (i.e. permittivity and loss tangent) of hibiscus flower tissue have been reported to be measured using the open ended coaxial probe technique; moreover, the tissue density characterization technique has also been outlined in the same article [12]. The dielectric properties at 947.50 MHz and 1842.50 MHz along with the measured tissue density are tabulated in Table I. Plane wave with linear polarization propagating along z-axis and electric field variation along x-axis impinges on the designed flower model at above mentioned frequencies in two different simulation environments as per the ICNIRP guidelines [16]. ICNIRP prescribed maximum permissible reference power densities in public zone are different at 947.50 MHz and 1842.50 MHz – detailed data have been tabulated in Table II [16]. Time domain solver in CST Microwave Studio 2016 has been employed to simulate SAR data [20]. Complex geometry, high permittivity and significant loss tangent of the designed flower model are the prime reasons to choose time domain solver for proper



(a)

(b)

Fig. 2. (a) Point SAR distribution on three dimensional surface of the typical hibiscus flower model due to plane wave (linearly polarized) irradiation at 947.50 MHz as per ICNIRP guidelines [16] (b) 1g SAR distribution on three dimensional surface of the typical hibiscus flower model due to plane wave (linearly polarized) irradiation at 947.50 MHz as per ICNIRP guidelines [16] meshing of the structure [13] – this solver is developed based on a computational method known as Finite Integration Technique [21-22]. Spatial distance of one wavelength in the tissue material has been segmented into 20 parts and the flower model has been discretized with hexahedral meshes of different sizes. Next, four perfectly-matched-layers (PMLs), each possessing 0.0001 reflection coefficient, have been used in absorbing boundary during SAR simulations. All open (add space) boundaries have been set very close to the structure so that the plane wave excitation can be placed as close as possible. Then, -40 dB inverse transformation accuracy has been set to obtain frequency domain characteristics from the time domain results (after steady state energy criterion is satisfied). MLP SAR, 1g SAR and WBA SAR data have been simulated at 947.50 MHz and 1842.50 MHz respectively using IEEE/IEC 62704-1 protocol with an average cell mass of 0.00034g.

III. Simulated SAR Results and Discussions

Simulated SAR data at 947.50 MHz and 1842.50 MHz are of significant values – moreover, SAR values at 1842.50 MHz increase by a large scale (3 to 4 fold) as illustrated in Table III. Simulated MLP SAR and 1g averaged SAR distributions on the three dimensional surface of the hibiscus flower model at 947.50 MHz (as per ICNIRP regulations) are illustrated in Figs. 2(a) and 2(b) respectively. To be specific, MLP SAR values are 4.75 W/Kg and 19.12 W/Kg respectively at 947.50 MHz and 1842.50 MHz – an increase of 4 fold. This increase in SAR value is due to multiple reasons. Firstly, wavelength gets shorten at higher frequency resulting in more number of electric field peaks within the hibiscus flower model [12-13]; Secondly, maximum permissible incident field strength significantly increases at 1842.50 MHz compared to 947.50 MHz – as a consequence, SAR value also increases [12-13]. In addition, dielectric properties of hibiscus flower also alter with frequency and this further contributes in altering SAR value with frequency [12]. However, it should be noted that earlier published SAR data at 2450 MHz for the same hibiscus flower model are of higher values while compared to the present reported SAR data at 947.50 MHz and 1842.50 MHz – and the reasons are obvious and have been discussed above [12].

Even at a particular frequency, SAR distribution on three dimensional surfaces of the hibiscus flower model possesses a wide spatial variation – as observed in Figs. 2(a) and 2(b) respectively. The hibiscus flower model contains regions with relatively higher SAR values near the junction of the petals and the twig; moreover, moderately high SAR values are also noted near the sharp edges of the petals – these observations resemble with the earlier reported SAR distribution at 2450 MHz [12]. It is so because strong charge accumulation takes place near sharp edges of any arbitrary shaped lossy dielectric biological object – consequently, higher electric field strength develops locally at regions with concentrated charge distribution resulting in increased SAR value near sharp edges of the flower structure [12-13]. Thus, increased point SAR values near sharp edges can have localized consequences and therefore spatial SAR averaging over 1g or larger mass can overlook this real scenario – therefore not recommended [12].

Table III: Simulated SAR results for the typical hibiscus flower model at 947.50 MHz and 1842.50 MHz

Frequency (MHz)	Peak E-field (V/m)	SAR averaging mass (g)	Max SAR (W/Kg)
947.50	59.77	Point	4.74682
		1	1.35591
		WBA	1.02422
1842.50	83.34	Point	19.1214
		1	4.82687
		WBA	3.39038

Simulated SAR results at 947.50 MHz and 1842.50 MHz shouldn't be further underestimated by averaging over 6 or 30 minutes of time duration – it is so as flowers in plants get continuous electromagnetic irradiation throughout their lifespan [12-13]. Here, reported SAR data have been noted for plane wave irradiation with linear polarization i.e. wave propagation along z-axis and electric field variation along x-axis. But, simulated SAR data can definitely alter if the direction of wave propagation or the polarization alters – as the designed flower structure is not symmetric along all axes and SAR data significantly depends on geometry of the biological object [11].

IV. Conclusions

There are a number of international and national electromagnetic regulatory organizations across the globe and ICNIRP is one of those – these electromagnetic guidelines are of wide contrast and prescribed reference power density limits at public zone differ a lot [13, 16-19]. Here, all simulated SAR data have been reported for ICNIRP public exposure guidelines – however, simulated SAR data would alter in case other electromagnetic regulatory guidelines are considered [13, 16-19]. Prescribed reference power density level as well as expression for point SAR changes proportionately with the square of electric field magnitude – the initial parameter is correlated to the square of incident electric field magnitude whereas SAR depends on square of developed internal electric field magnitude. SAR values, due to simultaneous exposure at multiple frequencies, add up in real scenarios and thus the concern is further raised [13]. Moreover, estimated SAR values would increase many folds in case occupational scenarios are taken into account [16].

Simulated SAR data can be treated as initial reference for practical SAR measurement in future – however, custom made phantom model, equivalent dielectric liquid, design of electric field probe along with its calibration are some of the challenges to be dealt with. To conclude, significant SAR values in hibiscus flower model insist to investigate biological effects of long duration as well as short span electromagnetic exposure on plants, fruits and flowers.

Acknowledgment

Authors acknowledge Rashtriya Uchchatar Shiksha Abhiyan (RUSA) 2.0 scheme, Govt. of India (internal project ref. no. R-11/269/19, Jadavpur University) for providing research grant.

References

[1] S. S. Stuchly, M. A. Stuchly, A. Kraszewski, and G. Hartsgrove, “Energy Deposition in a Model of Man: Frequency Effects,” IEEE Transactions on Biomedical Engineering, vol. 33, no. 7, pp. 702-711, 1986.

[2] K. Meier, V. Hombach, R. Kästle, R. Y. Tay, and N. Kuster, “The Dependence of Electromagnetic Energy Absorption upon Human-Head Modelling at 1800 MHz,” IEEE Transactions on Microwave Theory and Techniques, vol. 45, no. 11, pp. 2058-2062, 1997.

[3] A. Christ, A. Klingenböck, T. Samaras, C. Goiceanu, and N. Kuster, “The Dependence of Electromagnetic Far-Field Absorption on Body Tissue Composition in the Frequency Range from 300 MHz to 6 GHz,” IEEE Transactions on Microwave Theory and Techniques, vol. 54, no. 5, pp. 2188-2195, 2006.

[4] T. Iyama, T. Onishi, Y. Tarusawa, S. Uebayashi, and T. Nojima, “Novel Specific Absorption Rate (SAR) Measurement Method Using a Flat Solid Phantom,” IEEE Transactions on Electromagnetic Compatibility, vol. 50, no. 1, pp. 43-51, 2008.

[5] A. Y. Simba, T. Hikage, S. Watanabe, and T. Nojima, “Specific Absorption Rates of Anatomically Realistic Human Models Exposed to RF Electromagnetic Fields From Mobile Phones Used in Elevators,” IEEE Transactions on Microwave Theory and Techniques, vol. 57, no. 5, pp. 1250-1259, 2009.

[6] R. Asadi, H. Aliakbarian, G. Khayambashi, and P. Majdolashrafi, “A case-study on the effect of averaging duration on RF dosimetry of general public environments,” IEEE Electromagnetic Compatibility Magazine, vol. 9, no. 3, pp. 45-54, 2020.

[7] G. Tognola et al., “Numerical Assessment of RF Human Exposure in Smart Mobility Communications,” IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology, doi: 10.1109/JERM.2020.3009856.

[8] A. Kundu, and B. Gupta, “Comparative SAR Analysis of Some Indian Fruits as per the Revised RF Exposure Guideline,” IETE Journal of Research, vol. 60, no. 4, pp. 296-302, 2014.

[9] A. Kundu, “RF Energy Absorption in Plant Parts due to Cell Tower Radiation,” LAP Lambert Academic Publishing, Germany, 2015.

[10] A. Kundu, B. Gupta, and A. I. Mallick, “Specific Absorption Rate Evaluation in a Typical Multilayer Fruit: Coconut with Twig due to Electromagnetic Radiation as per Indian Standards,” Microwave Review (Mikrotalasana Revija), vol. 23, no. 2, pp. 24-32, 2017.

[11] A. Kundu, B. Gupta, and A. I. Mallick, “Dependence of electromagnetic energy distribution inside a typical multilayer fruit model on direction of arrival and polarization of incident field,” IEEE Radio and Antenna Days of the Indian Ocean 2019, Reunion Island, 2019.

[12] N. Mukherjee, A. Kundu, B. Gupta, and M. Mitra, “Specific Absorption Rate Estimation for a Typical Hibiscus Flower Model as per ICNIRP Electromagnetic Guidelines,” 2020 IEEE Calcutta Conference (CALCON), India, pp. 240-243, 2020.

[13] A. Kundu, B. Gupta, and A. I. Mallick, “Contrast in Specific Absorption Rate for a Typical Plant Model Due to Discrepancy Among Global and National Electromagnetic Standards,” Progress In Electromagnetics Research M, vol. 99, pp. 139-152, 2021.

[14] A. Kundu, and B. Gupta, “Broadband dielectric properties measurement of some vegetables and fruits using open ended coaxial probe technique,” 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC), Kolkata, India, 2014.

[15] A. Kundu, K. Patra, and B. Gupta, “Broadband Dielectric Properties Evaluation of Catharanthus Roseus Leaf, Flower and Stem Using Open Ended Coaxial Probe Technique,” Journal of Physical Science, vol. 18, pp. 62-69, 2014.

[16] ICNIRP, “Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz),” Health Physics, vol. 118, no. 5, pp. 483-524, 2020.

[17] R. F. Cleveland (Jr.), D. M. Sylvar, and J. L. Ulcek, “Evaluating Compliance with FCC Guidelines for human exposure to Radiofrequency Electromagnetic Fields,” FCC OET Bulletin 65, Edition 97-01, Washington D.C., 1997.

[18] IEEE, “IEEE standard for safety levels with respect to human exposure to electric, magnetic, and electromagnetic fields, 0 Hz to 300 GHz,” IEEE Std C95.1-2019 (Revision of IEEE Std C95.1-2005/ Incorporates IEEE Std C95.1-2019/Cor 1-2019), United States, pp. 1-312, 2019.

[19] DoT, “Mobile Communication Radio Wave Safety,” India, pp. 1-15, 2012.

[20] CST STUDIO SUITE 2016, https://www.3ds.com/products-services/simulia/products/cst-studio-suite/

[21] T. Weiland, “A discretization method for the solution of Maxwell’s equations for six-component fields,” Electronics and Communications AEU, vol. 31, no. 3, pp. 116-120, 1977.

[22] M. Clemens, and T. Weiland, “Discrete Electromagnetism with the Finite Integration Technique,” Progress in Electromag. Res., vol. 32, pp. 65-87, 2001.

Extended Three-Stage Recursive Least Squares Identification Algorithm for multiple-input single-output CARARMA Systems

Munya Ali Arwin
Department of Electrical and Computer Engineering
Libyan Academy
Tripoli, Libya
munya.arwini@academy.edu.ly

Nasar Aldian Ambark Shashoa
Department of Electrical and Computer Engineering
Libyan Academy
Tripoli, Libya
nasar-shashoa@ieee.org

***Abstract*—this paper derives extended three-stage recursive identification algorithm of MISO for (CARARMA) systems. Based on The decomposition technique, four subsystems are obtained and the parameters of each subsystem are identified. Some model validation methods are computed to measure the model value and Akaike's Final Prediction Error Criterion (FPE) is used to verify the selection of system order. The algorithm has a high computational efficiency because the covariance matrices dimensions become small in each subsystem. Finally, this algorithm effectiveness is demonstrated in simulation example**.

Keywords*— *parameter estimation, decomposition technique, model validation, covariance matrix, Final Prediction Error Criterion

I. Introduction

System identification has important effect on control theory, optimization, state filtering, modern control and other fields [1-4], deals with issue of designing mathematical models of dynamic systems from the true data for the input and the output of the system [5-7]. Parameter estimation is significant for system identification and system modeling [3, 8], and there are many Parameter estimation algorithms in the literature. For instance, LS algorithms, the auxiliary model based identification algorithms and gradient based algorithms [9, 10]. LS methods can be divided into two classes. First class can be used for offline identification and is called iterative identification methods; another can be used for online identification and is called recursive identification methods. RLS method considers important development which has made the LS algorithm one of the very important and largely used for real-time implementations. RLS method can be used only for Auto – Regressive model (ARX). for other models, such as, autoregressive moving average model(ARMAX) or controlled autoregressive autoregressive moving average (CARARMA) model, RLS method cannot be used. One such algorithms that deals with these types of models in system identification is multiple-stage algorithms [8] and the basic idea is based on the decomposition technique that can convert the main identification issue into small sub issues, which are simpler to solve[11]. For instance, Yao and Ding derived Two-stage least squares based iterative identification algorithm for controlled autoregressive moving average (CARMA) systems [12]. This work proposes extended three-stage recursive identification algorithm of MISO for (CARARMA) systems. The decomposition technique is used as the basic idea and thus the covariance matrices dimensions of each subsystem become small and computational efficiency of the proposed algorithm become high. This paper is structured as follows. Section 2 introduces multi-input single-output CARARMA system and its identification model is given. Extended three-stage recursive identification algorithm for MISO CARARMA systems is derived in Section 3. Section 4 presents model validation and order selection. In section 5, a simulation example is given to demonstrate the efficacy of this algorithm. Finally, the conclusions are given in section 6.

II. Problem Formulation

In this work, MISO system, linear, time-invariant, discrete-time system, described by (CARARMA) model [13] as shown in fig. 1 is considered and given as

$$A(z)y(k) = \sum_{j=1}^{2} B_j(Z)\, u_j(k) + \frac{D(z)}{C(z)} v(k) \qquad (1)$$

where

$u_j(k)$, $j = 1, 2$, $y(k)$ are the inputs, output of the system and $\nu(k)$ is the white noise with $m = 0$ and $\sigma^2 = 1$ [9]. $A(z), B_j(z), C(z)$and $D(z)$ are polynomials as [8]:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n_a} z^{-n_a} ,$$

$$B_j(z) = b_{j1} z^{-1} + b_{j2} z^{-2} + \cdots + b_{jn_j} z^{-n_j} ,$$

$$C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_{n_c} z^{-n_c} ,$$

$$D(z) = 1 + d_1 z^{-1} + d_2 z^{-2} + \cdots + d_{n_d} z^{-n_d} ,$$

First, the inner variable is defined as

$$w(k) = \frac{D(z)}{C(z)} v(k), \qquad (2)$$

Thus, equation (1) can be rewritten

$$A(z)y(k) = \sum_{j=1}^{2} B_j(z)\, u_j(k) + w(k). \qquad (3)$$

Or

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

$$y(k) = -\sum_{i=1}^{\mathrm{n_a}} a_i y(k-i) + \sum_{j=1}^{2}\sum_{i=1}^{n_{b_j}} b_{ij}\, u_j(k-i) + w(k) (4)$$

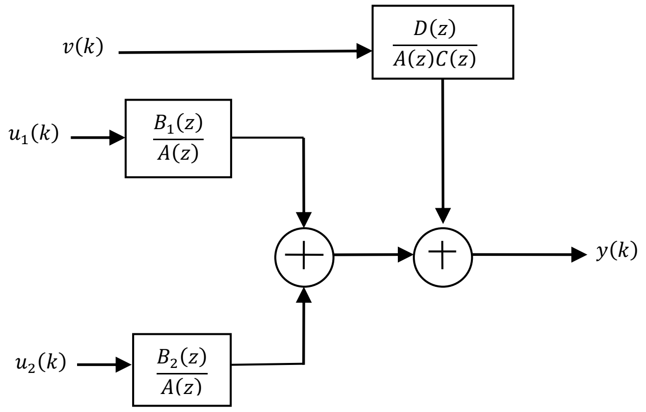


Fig. 1. The MISO (CARARMA) system.

It can be described in a linear regression form as

$$y(\mathrm{k}) = {\varphi_1}^T(k)\theta_1 + {\varphi_2}^T(k)\theta_2 + {\varphi_3}^T(k)\theta_3 + w(k) \quad (5)$$

Where

The information vectors are [8],

$$\varphi_1(\mathrm{k}) = [-\mathrm{y}(\mathrm{k}-1), -\mathrm{y}(\mathrm{k}-2), \dots, -\mathrm{y}(\mathrm{k}-\mathrm{n_a})]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_a}},$$

$$\varphi_2(\mathrm{k}) = [\mathrm{u_1}(\mathrm{k}-1), \mathrm{u_1}(\mathrm{k}-2), \dots, \mathrm{u_1}(\mathrm{k}-\mathrm{n_1})]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_{b_1}}},$$

$$\varphi_3(\mathrm{k}) = [\mathrm{u_2}(\mathrm{k}-1), \mathrm{u_2}(\mathrm{k}-2), \dots, \mathrm{u_2}(\mathrm{k}-\mathrm{n_2})]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_{b_2}}},$$

The parameter vectors are

$$\theta_1 := [\mathrm{a_1}, \mathrm{a_2}, \dots, \mathrm{a_{n_a}}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_a}},$$

$$\theta_2 := [\mathrm{b_{11}}, \mathrm{b_{12}}, \dots, \mathrm{b_{1n_1}}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_{b_1}}},$$

$$\theta_3 := [\mathrm{b_{21}}, \mathrm{b_{22}}, \dots, \mathrm{b_{2n_2}}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_{b_2}}},$$

In addition, equation (2) can be rewritten

$$w(k) = [1 - C(z)]w(k) + D(z)v(k)$$

$$= -c_1 w(k-1) - c_2 w(k-2) - \dots - c_{n_c} w(k - n_c) + d_1 v(k-1) + d_2 v(k-2) + \cdots + d_{n_d} v(k-n_d) + v(k) \quad (6)$$

And it can be written in a linear regression form as

$$w(k) = {\varphi_4}^T(k)\theta_4 + v(k) \quad (7)$$

Where

$$\varphi_4(\mathrm{k}) = [-\mathrm{w}(\mathrm{k}-1), -\mathrm{w}(\mathrm{k}-2), \dots, -\mathrm{w}(\mathrm{k}-\mathrm{n_c}), \mathrm{v}(\mathrm{k}-1), \mathrm{v}(\mathrm{k}-2), \dots, \mathrm{v}(\mathrm{k}-\mathrm{n_d})]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_c+n_d}},$$

$$\theta_4 := [\mathrm{c_1}, \mathrm{c_2}, \dots, \mathrm{c_{n_c}}, \mathrm{d_1}, \mathrm{d}, \dots, \mathrm{d_{n_d}}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{n_c+n_d}},$$

Finally, equation (7) is substituted into equation (5) [14], $y(k)$ can be written as [6]

$$y(k) = {\varphi_1}^T(k)\theta_1 + {\varphi_2}^T(k)\theta_2 + {\varphi_3}^T(k)\theta_3 + {\varphi_4}^T(k)\theta_4 + v(k) \quad (8)$$

$$= \varphi^T(k)\theta + v(k).$$

$$\varphi^T(k) = [{\varphi_1}^T(k)\ {\varphi_2}^T(k)\ {\varphi_3}^T(k)\ {\varphi_4}^T(k)] \in \mathbb{R}^{\mathrm{n}},$$

$$\theta = \begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{bmatrix} \in \mathbb{R}^{\mathrm{n}}, \quad \mathrm{n} = \mathrm{n_a} + \mathrm{n_{bj}} + \mathrm{n_c} + \mathrm{n_d}$$

Equation (8) is the identification model of MISO (CARARMA) system, it includes parameter vector θ that contains all parameters of the system to be identified [15].

III. Extended Three-Stage Recursive Identification Algorithm

Extended three-stage recursive identification algorithm is based on the decomposition technique as the basic idea and thus multiple-input single-output (CARARMA) System is decomposed into four subsystems.

Four intermediate variables are defined as,

$$y_1(k) = y(k) - {\varphi_2}^T(k)\theta_2 - {\varphi_3}^T(k)\theta_3 - {\varphi_4}^T(k)\theta_4, \quad (9)$$

$$y_2(k) = y(k) - {\varphi_1}^T(k)\theta_1 - {\varphi_3}^T(k)\theta_3 - {\varphi_4}^T(k)\theta_4, \quad (10)$$

$$y_3(k) = y(k) - {\varphi_1}^T(k)\theta_1 - {\varphi_2}^T(k)\theta_2 - {\varphi_4}^T(k)\theta_4, \quad (11)$$

$$y_4(k) = y(k) - {\varphi_1}^T(k)\theta_1 - {\varphi_2}^T(k)\theta_2 - {\varphi_3}^T(k)\theta_3, \quad (12)$$

From (9)–(12), equation (8) can be decomposed into four sub-identification models [8],

$$y_i(k) = \varphi_i^T(k)\theta_i + v(k), \quad i = 1,2,3,4. \quad (13)$$

These includes the parameters vectors $\theta_1, \theta_2, \theta_3$ and θ_4 [12].

Then, four criterion functions are defined as,

$$J_i(\theta_i) := \sum_{j=1}^{k} [y_i(j) - {\varphi_i}^T(j)\theta_i]^2, \quad i = 1,2,3,4.$$

Let the partial derivatives of $J_i(\theta_i), i = 1,2,3,4$ with respect to θ_i be zero

$$\left.\frac{\partial J_i(\theta_i)}{\partial \theta_i}\right|_{\theta_i = \hat{\theta}_i(k)} = -2\varphi_i(j)\sum_{j=1}^{k}[y_i(j) - {\varphi_i}^T(j)\hat{\theta}_i(k)]$$

$$= 0, \quad i = 1,2,3,4. \quad (14)$$

Let $\hat{\theta}(k) := \begin{bmatrix} \hat{\theta}_1(k) \\ \hat{\theta}_2(k) \\ \hat{\theta}_3(k) \\ \hat{\theta}_4(k) \end{bmatrix} \in \mathbb{R}^n$ be the estimate of $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \in \mathbb{R}^n$

Minimizing the criterion functions and as a result, RLS algorithm can be obtained for computing $\hat{\theta}_i(k)$:

$$\hat{\theta}_i(k) = \hat{\theta}_i(k-1) + L_i(k)\big[\big(y_i(k) - \varphi_i{}^T(k)\hat{\theta}_i(k-1)\big](15)$$

$$L_i(k) = P_i(k-1)\varphi_i(k)[1 + \varphi_i{}^T(k)P_i(k-1)\varphi_i(k)]^{-1},$$

$$P_i(k) = [I - L_i(k)\varphi_i{}^T(k)]P_i(k-1),$$

$$P_i(0) = p_0 I, \;\; i = 1,2,3,4$$

The equations from (9)–(12) are substituted into equation (15) with $= 1,2,3,4$, obtains

$$\hat{\theta}_1(k) = \hat{\theta}_1(k-1) + L_1(k)\big[y(k) - \varphi_2{}^T(k)\theta_2 - \varphi_3{}^T(k)\theta_3 - \varphi_4{}^T(k)\theta_4 - \varphi_1{}^T(k)\hat{\theta}_1(t-1)\big] \quad (16)$$

$$\hat{\theta}_2(k) = \hat{\theta}_2(k-1) + L_2(k)\big[y(k) - \varphi_1{}^T(k)\theta_1 - \varphi_3{}^T(k)\theta_3 - \varphi_4{}^T(k)\theta_4 - \varphi_2{}^T(k)\hat{\theta}_2(k-1)\big] \quad (17)$$

$$\hat{\theta}_3(k) = \hat{\theta}_3(k-1) + L_3(k)\big[y(k) - \varphi_1{}^T(k)\theta_1 - \varphi_2{}^T(k)\theta_2 - \varphi_4{}^T(k)\theta_4 - \varphi_3{}^T(k)\hat{\theta}_3(k-1)\big] \quad (18)$$

$$\hat{\theta}_4(k) = \hat{\theta}_4(k-1) + L_4(k)\big[y(k) - \varphi_1{}^T(k)\theta_1 - \varphi_2{}^T(k)\theta_2 - \varphi_3{}^T(k)\theta_3 - \varphi_4{}^T(k)\hat{\theta}_4(k-1)\big] \quad (19)$$

Equations (16) – (19) include the unknown parameter θ_i $i = 1,2,3,4$. The Replacement of the unknown θ_i in (16)–(19) with their estimates $\hat{\theta}_i(k-1)$ is the solution:

$$\hat{\theta}_1(k) = \hat{\theta}_1(k-1) + L_1(k)\;\big[y(k) - \varphi_2{}^T(k)\hat{\theta}_2(k-1) - \varphi_3{}^T(k)\hat{\theta}_3(k-1) - \varphi_4{}^T(k)\hat{\theta}_4(k-1) - \varphi_1{}^T(k)\hat{\theta}_1(k-1)\big]$$

$$= \hat{\theta}_1(k-1) + L_1(k)\big[y(k) - \varphi^T(k)\hat{\theta}(k-1)\big],$$

$$\hat{\theta}_2(k) = \hat{\theta}_2(k-1) + L_2(k)\big[y(k) - \varphi_1{}^T(k)\hat{\theta}_1(k-1) - \varphi_3{}^T(k)\hat{\theta}_3(k-1) - \varphi_4{}^T(k)\hat{\theta}_4(k-1) - \varphi_2{}^T(k)\hat{\theta}_2(k-1)\big]$$

$$= \hat{\theta}_2(k-1) + L_2(k)\big[y(k) - \varphi^T(k)\hat{\theta}(k-1)\big],$$

$$\hat{\theta}_3(k) = \hat{\theta}_3(k-1) + L_3(k)\big[y(k) - \varphi_1{}^T(k)\hat{\theta}_1(k-1) - \varphi_2{}^T(k)\hat{\theta}_2(k-1) - \varphi_4{}^T(k)\hat{\theta}_4(k-1) - \varphi_3{}^T(k)\hat{\theta}_3(k-1)\big],$$

$$= \hat{\theta}_3(k-1) + L_3(k)\big[y(k) - \varphi^T(k)\hat{\theta}(k-1)\big],$$

$$\hat{\theta}_4(k) = \hat{\theta}_4(k-1) + L_4(k)\big[y(k) - \varphi_1{}^T(k)\hat{\theta}_1(k-1) - \varphi_2{}^T(k)\hat{\theta}_2(k-1) - \varphi_3{}^T(k)\hat{\theta}_3(k-1) - \varphi_4{}^T(k)\hat{\theta}_4(k-1)\big],$$

$$= \hat{\theta}_4(k-1) + L_4(k)\big[y(k) - \varphi^T(k)\hat{\theta}(k-1)\big],$$

$\varphi_4(k)$ includes the unknown noise terms $w(k-i)$ and$v(k-i)$, and thus these estimated parameters cannot be generated based on the above algorithms in (16)–(19). Replacing $w(k-i)$ and $v(k-i)$ with their estimates $\widehat{w}(k-i)$ and$\hat{v}(\mathrm{k}-\mathrm{i})$ is the solution, and obtain

$$\hat{\varphi}_4(k) := [-\widehat{w}(k-1), -\widehat{w}(k-2), \ldots, -\widehat{w}(k-n_c), \hat{v}(k-1), \hat{v}(\mathrm{k}-2), \ldots, \hat{v}(\mathrm{k}-\mathrm{n_d})]^T \in \mathbb{R}^{n_c+n_d}.$$

$$\hat{\varphi}(k) := \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \hat{\varphi}_4 \end{bmatrix} \in \mathbb{R}^n.$$

Now, equation (5) can be written as,

$$w(k) = y(k) - \varphi_1{}^T(k)\theta_1 - \varphi_2{}^T(k)\theta_2 - \varphi_3{}^T(k)\theta_3,$$

And $\qquad v(k) = w(k) - \varphi_4{}^T(k)\theta_4,$

The estimated $w(k)$ and $v(k)$can be computed by

$$\widehat{w}(k) = y(k) - \varphi_1{}^T(k)\hat{\theta}_1 - \varphi_2{}^T(k)\hat{\theta}_2 - \varphi_3{}^T(k)\hat{\theta}_3$$

$$\hat{v}(k) = \widehat{w}(k) - \hat{\varphi}_4{}^T(k)\hat{\theta}_4$$

Thus, extended three-stage recursive parameter estimation algorithm for computing the estimated parameters $\theta_1, \theta_2, \theta_3$ and θ_4 of CARARMA model is obtained as

$$\hat{\theta}_1(k) = \hat{\theta}_1(k-1) + L_1(k)\big[y(k) - \hat{\varphi}^T(k)\hat{\theta}(k1)\big], \quad (20)$$

$$L_1(k) = P_1(k-1)\varphi_1(k)[1 + \varphi_1{}^T(k)P_1(k-1)\varphi_1(k)]^{-1}, \quad (21)$$

$$P_1(k) = [I - L_1(k)\varphi_1{}^T(k)]P_1(k-1), P_1(0) = p_0 I_{n_a}, (22)$$

$$\varphi_1(k) = [-y(k-1), -y(k-2), \ldots, -y(k-n_a)]^T \in \mathbb{R}^{n_a}, \quad (23)$$

$$\hat{\theta}_2(k) = \hat{\theta}_2(k-1) + L_2(k)\big[y(k) - \hat{\varphi}^T(k)\hat{\theta}(k-1)\big], \; (24)$$

$$L_2(k) = P_2(k-1)\varphi_2(k)[1 + \varphi_2{}^T(k)P_2(k-1)\varphi_2(k)]^{-1}, \quad (25)$$

$$P_2(k) = [I - L_2(k)\varphi_2{}^T(k)]P_2(k-1), P_2(0) = p_0 I_{n_1}, (26)$$

$$\varphi_2(k) = [u_1(k-1), u_1(k-2), \ldots, u_1(k-n_1)]^T \in \mathbb{R}^{n_1}, \quad (27)$$

$$\hat{\theta}_3(k) = \hat{\theta}_3(k-1) + L_3(k)\big[y(k) - \hat{\varphi}^{\mathrm{T}}(k)\hat{\theta}(k-1)\big], \; (28)$$

$$L_3(k) = P_3(k-1)\varphi_3(k)[1 + \varphi_3{}^T(k)P_3(k-1)\varphi_3(k)]^{-1}, \quad (29)$$

$$P_3(k) = [I - L_3(k)\varphi_3{}^T(k)]P_3(k-1), P_3(0) = p_0 I_{n_2}, (30)$$

$$\varphi_3(k) = [u_2(k-1), u_2(k-2), \ldots, u_2(k-n_2)]^T \in \mathbb{R}^{n_2}, \quad (31)$$

$$\hat{\theta}_4(k) = \hat{\theta}_4(k-1) + L_4(k)\big[y(k) - \hat{\varphi}^T(k)\hat{\theta}(k-1)\big], \; (32)$$

$$L_4(k) = P_4(k-1)\hat{\varphi}_4(k)\left[1 + \hat{\varphi}_4{}^T(k)P_4(k-1)\hat{\varphi}_4(k)\right]^{-1}, \quad (33)$$

$$P_4(k) = \left[I - L_4(k)\hat{\varphi}_4{}^T(k)\right]P_4(k-1), \mathrm{p}_4(0) = p_0 I_{n_c+n_d}, \quad (34)$$

$$\hat{w}(k) = y(k) - \varphi_1{}^T(k)\hat{\theta}_1 - \varphi_2{}^T(k)\hat{\theta}_2 - \varphi_3{}^T(k)\hat{\theta}_3. \quad (35)$$

$$\hat{v}(k) = \hat{w}(k) - \hat{\varphi}_4{}^T(k)\hat{\theta}_4. \quad (36)$$

$$\hat{\varphi}_4(k) = \begin{bmatrix} -\hat{w}(k-1), -\hat{w}(k-2), \ldots, -\hat{w}(k-n_c), \\ \hat{v}(k-1), \hat{v}(\mathrm{k}-2), \ldots, \hat{v}(\mathrm{k}-\mathrm{n_d}) \end{bmatrix}^T \quad (37)$$

$$\hat{\upvarphi}(\mathrm{k}) \coloneqq [\upvarphi_1^{\mathrm{T}}(k), \upvarphi_2^{\mathrm{T}}(k), \upvarphi_3^{\mathrm{T}}(k), \hat{\upvarphi}_4^{\mathrm{T}}(k)]^T \in \mathbb{R}^{\mathrm{n}}, \quad (38)$$

$$\hat{\theta}(k) := \left[\hat{\uptheta}_1^{\mathrm{T}}(k), \hat{\uptheta}_2^{\mathrm{T}}(k), \hat{\uptheta}_3^{\mathrm{T}}(k), \hat{\uptheta}_4^{\mathrm{T}}(k)\right]^T \in \mathbb{R}^n. \quad (39)$$

IV. MODEL VALIDATION AND ORDER SELECTION

A. *Model validation*

Model validation is a fundamental part of simulation model development, and generally implemented in parallel with the model design process [16]. There are many validation methods in the literatures and in this section, some criteria's will be studied [5]:

1. *Root-mean-square error*

RMSE is the predominant statistical methods [17] and it has widely been used in evaluating the accuracy of the model [18]**.** RMSE is computed as follows [19]:

$$RMS - Error = \sqrt{\frac{\sum_{i=1}^{N}\left((\hat{y}_i - y_i)\right)^2}{N}} \quad (40)$$

2. *The Cross-Correlation Test*

the estimated residual is used with the input sequence u(k), to develop a test and check the independence between the residual and the input. The cross-correlation matrix is given by

$$\hat{R}_{\epsilon u}(\tau) = \frac{1}{N-\tau}\sum_{k=\tau+1}^{N} \hat{\epsilon}\left(k, \hat{\theta}\right)u\left(k-\tau, \hat{\theta}\right)^T \quad (41)$$

if the cross-correlation function is approximately zero, then, the model is perfect [20]**.**

B. *Model order selection*

Akaike's Final Prediction Error (FPE) Criterion is used in this work for the model order selection and is defined as:

$$FPE = \left(\frac{1}{N}\sum_{k=1}^{N} e^2(k)\left(\frac{1+(n/N)}{1-(n/N)}\right)\right) \quad (42)$$

Where

N is the number of samples, n is the number of model parameters, V is the residuals variance and is computed as [21].

$$V = \left(\frac{1}{N}\sum_{k=1}^{N} e^2(k)\right) \quad (43)$$

V. SIMULATION RESULTS

An example is provided in this part to evaluate the capability of this algorithm. Consider the following MISO for (CARARMA) systems,

$$\mathrm{A(z)y(k)} = \sum_{\mathrm{j=1}}^{2} \mathrm{B_j(Z)\,u_j(k)} + \frac{\mathrm{D(z)}}{\mathrm{C(z)}}\mathrm{v(k)}$$

$$\mathrm{A(z)} = 1 + 0.65\mathrm{z}^{-1} + 0.85\mathrm{z}^{-2},$$

$$\mathrm{B_1(z)} = -0.15\mathrm{z}^{-1} - 0.25\mathrm{z}^{-2},$$

$$\mathrm{B_2(z)} = 0.18\mathrm{z}^{-1} - 0.28\mathrm{z}^{-2},$$

$$\mathrm{C(z)} = 1 + 0.7\mathrm{z}^{-1},$$

$$\mathrm{D(z)} = 1 + 0.4\mathrm{z}^{-1},$$

$$\uptheta = [\mathrm{a_1, a_2, b_{11}, b_{12}, b_{21}, b_{22}, c_1, d_1}]^{\mathrm{T}},$$

$$= [0.65, 0.85, -0.15, -0.25, 0.18, -0.28, 0.7, 0.4]^{\mathrm{T}},$$

The inputs $\{u_1(k)\}$, $\{u_2(k)\}$ are generated as a white sequence with $m = 0$ and $\sigma^2 = 1$, while $n(k)$ is generated as a Gaussian white noise with $m = 0$ and $\sigma^2 = 0.20^2$. First, RMS-Error and the cross-correlation test are used for evaluating the validation of the model. Fig. 2 shows root mean square errors versus all sequences (k). This figure demonstrates that, RMS-Error become smaller as the sequence increases, which indicate that the model accuracy is high, in addition, the correlation between the residual and one of the inputs sequences is showed in fig. 3. The figure shows that, the values are very low (the model is typical).

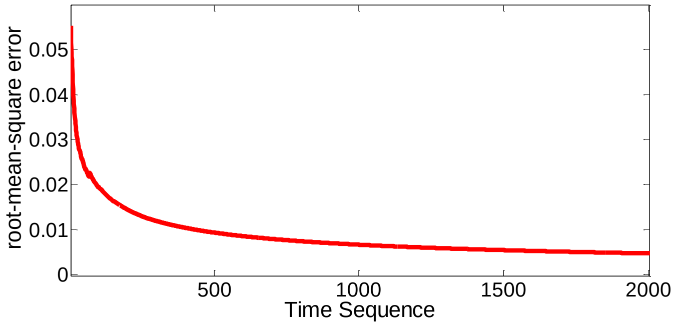


Fig. 2. Root mean square errors versus time sequences

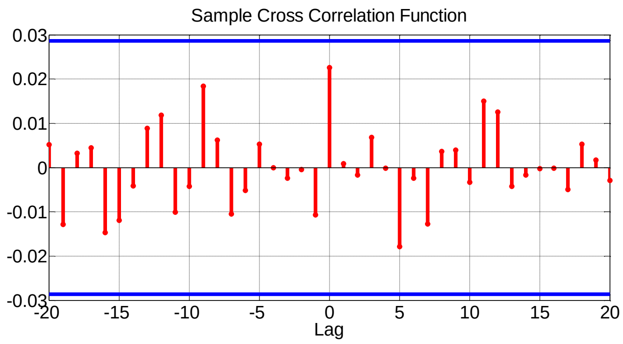


Fig. 3. Cross-correlation between the residual and one of the inputs sequences

Fig. 4 shows the true output and the estimated output of three-stage RLS algorithm and extended three-stage RLS algorithm. Window from n =1080 to n =1100 has been taken for more clarification as shown in Fig. 5,it shows that the estimated output of extended three stage RLS algorithm very close to the true output compared with the estimated output of three stage RLS algorithm, that means, the effectiveness of this algorithm is high. This conclusion has been confirmed by computing and plotting the residual of extended three-stage RLS algorithm and three-stage RLS algorithm as illustrated in Fig 6.

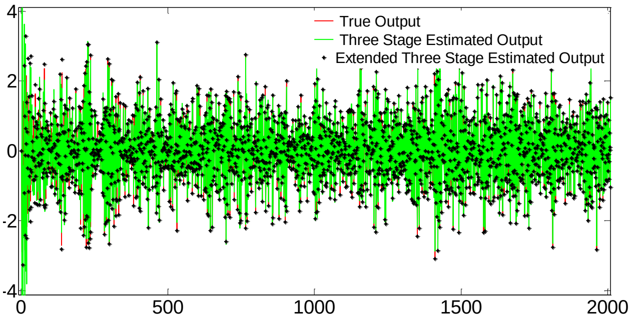


Fig. 4. The true output and the estimated output of three-stage RLS algorithm and extended three-stage RLS algorithm

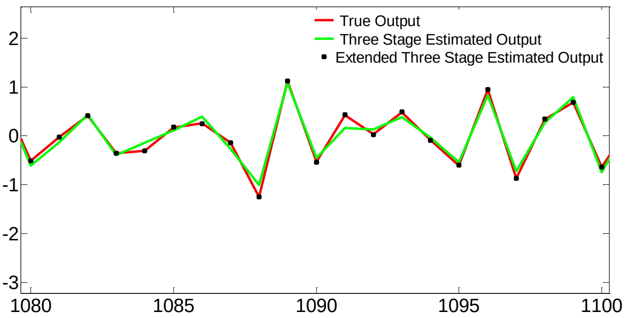


Fig. 5. Window from n =1080 to n =1100 of the true output and the estimated output of three-stage RLS algorithm and extended three-stage RLS algorithm.

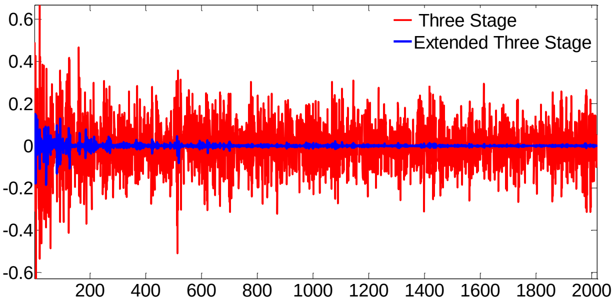


Fig. 6. Residual of three-stage RLS algorithm and extended three-stage RLS algorithm.

FPE is calculated from one to five models as shown in next figure, and as we have supposed, the second order system is the best model order

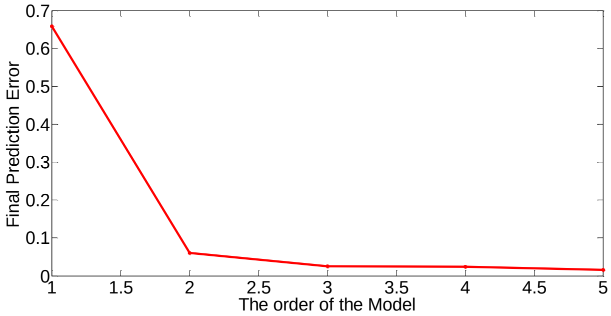


Fig. 7. FPE versus model order

Finally, Root-mean-square error of three-stage RLS algorithm and extended three-stage RLS algorithm versus the sequences is shown in Fig. 8.

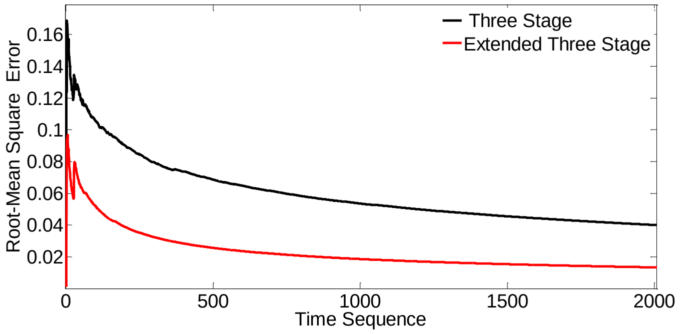


Fig. 8. Root-mean-square error of three-stage RLS algorithm and extended three-stage RLS algorithm versus the sequences.

The figure illustrate that extended three-stage RLS algorithm is more accurate than three-stage RLS algorithm.

VI. CONCLUSIONS

This paper studies extended three-stage recursive identification algorithm for multiple-input single-output (CARARMA) systems. The results show that the proposed extended three-stage RLS algorithm is more accurate and require less computational load compared with three-stage RLS algorithm.

REFERENCES

[1] H. Chen, F. Ding and Y. Xiao, "Decomposition-based Least Squares Parameter Estimation Algorithm for Input Nonlinear Systems using the Key Term Separation Technique", Nonlinear Dynamics vol. 79, pp. 2027-2035, 2015.

[2] Y. Mao, F. Ding. J. W. Ding and X. Wan, “Filtering based least squares parameter estimation algorithms for Hammerstein nonlinear CARMA systems”, American Control Conference, 2017.

[3] H. Ma, J. Pan, L. Lv, G. H. Xu, F. Ding, A. Alsaedi and T. Hayat, “Recursive algorithms for multivariable output-error-like ARMA systems”, Mathematics, vol. 7, pp. 1–18, June 2019.

[4] D. Meng and F. Ding, “Model Equivalence-Based Identification Algorithm for Equation-Error Systems with Colored Noise”, Algorithms. pp. 280-291, 2015.

[5] S. Rachad, B. Nsiri and B. Bensassi, “System identification of inventory system using ARX and ARMAX models,” International Journal of control and automation, December 2015.

[6] J. Ding, “The hierarchical iterative identification algorithm for multi-input-output-error systems with autoregressive noise,” Complexity, 2017.

[7] F. Chen and F. Ding, “Recursive Least Squares Identification Algorithm for Multiple-Input Nonlinear Box-Jenkins Systems Using Maximum Likelihood Principle”, Journal of Computational and Nonlinear Dynamics Vol. 11, 2016.

[8] S. Wang and R. Ding, “Three-stage recursive least squares parameter estimation for controlled autoregressive autoregressive systems”, Applied Mathematical Modelling. pp. 7489-7497, 2013.

[9] B. Bao, Y. Xua, J. Sheng and R. Ding, “Least squares based iterative parameter estimation algorithm for multivariable controlled ARMA system modelling with finite measurement data”, Mathematical and Computer Modelling. pp. 1664–1669, 2011.

[10] L. Chen, J. Li and R. Ding, “Identification for the second-order systems based on the step response”, Mathematical and Computer Modelling. pp. 1074–1083, 2011.

[11] H. Duan, J. Jia and R. Ding, “Two-stage recursive least squares parameter estimation algorithm for output error models,” Mathematical and Computer Modelling, vol. 55, pp. 1151–1159, 2012.

[12] G.Yao and R.Ding, “Two-stage least squares based iterative identification algorithm for controlled autoregressive moving average (CARMA) systems,” Computers and Mathematics with Applications, vol. 63, pp. 975–984, 2012.

[13] D. Wang and F. Ding, “Input-output data filtering based recursive least squares identification for CARARMA systems,” Digital Signal Processing, vol. 20, pp. 991–999, 2010.

[14] X. Wang and F. Ding, “Convergence of the auxiliary model-based multi-innovation generalized extended stochastic gradient algorithm for Box–Jenkins systems,” Nonlinear Dynamics DOI 10.1007/s11071-015-2155-5, June 2015.

[15] F. Ding, “Two-stage least squares based iterative estimation algorithm for CARARMA system modeling”, Applied Mathematical Modelling. vol. 37, pp. 4798–4808, 2013.

[16] C. Yin and A.McKay, “Model Verification and Validation Strategies and Methods: An Application Case Study”, The 8th International Symposium on Computational Intelligence and Industrial Applications (ISCIIA2018), The 12th China-Japan International Workshop on Information Technology and Control Applications (ITCA2018), Tengzhou, Shandong, China, November 2018.

[17] C. Hong, J. Hanan, L. Yan, L. Yong, Y. Yan, Z. wei, L. Jian, S. ying, S. Chun,G. kuo, W. Xiu, Y. An, T. Ping and B. Tai, “Comparison of Crop Model Validation Methods”, Journal of Integrative Agriculture, August 2012.

[18] W. Wang and Y. Lu, “Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model”, Materials Science and Engineering 324 , 2018.

[19] J.Wolberg, “Data Analysis Using the Method of Least Squares”, Technion-Israel Institute of Technology, 2006.

[20] M. Verhaegen and V. Verdult, “Filtering and System Identification”, Cambridge University Press 2007.

[21] R.G.K.M. Aarts, “System Identification and Parameter Estimation”, Course Edition: 2011/2012.

The Demographic Profile Most at Risk of being Disinformed

Kevin Matthe Caramancion
College of Emergency Preparedness, Homeland Security, and Cybersecurity
University at Albany, The State University of New York
Albany, New York, United States
kcaramancion@albany.edu

***Abstract*—This paper explores the relationship between a person's demographic data, age (IV1), gender (IV2), ethnicity (IV3), and political ideology (IV4), and the risk of him/her falling prey to Mis/Disinformation attacks (DV). Participants (n=161) were subjected to the Fake News and deepfake test (15-item). The main data analysis tool employed by this study is multiple linear regression. Important findings of the study include the revelation of the disparity in the performance of the subjects from the underrepresented groups against those who are not. This paper further confirms that an increase in age is a risk for being Disinformed. The predictive model further reveals the profile of a most likely Disinformation victim. The intended target audiences of this paper are policymakers, social scientists, and tech companies.**

Keywords—Fake News, Misinformation, Disinformation, Polarization, Risk

I. Introduction

The pervasiveness of fake news in the digital world has been unparalleled. A number of social networking sites such as Facebook and Twitter allow users and entities alike to create and/or share content in their respective spaces. A particular type of content in these sites is news and journalistic information. More than being able to share content on these sites, information users consume such crucial information from these media, and as such, the integrity of contents is of crucial importance.

On the other side of the spectrum, misleading information had been long existing in the same environment, persisting and mixed with the legitimate news with the intent to deceive users due to several varying reasons and inspirations, primarily political and conspiracy induced. Ultimately, the information consumer has the personal responsibility to decide which of these are factual and which are misleading [1]. Thus, highlighting the ability to detect misinformation with the existing information descriptors, i.e., the contextual clues, is of extreme importance.

With social media feeds consisting of virtually unlimited information, the question comes down to what constitutes a user's judgment of a content's legitimacy. An instrumental set of conditions, i.e., demographic and socioeconomic factors, have long been attributed to the decision-making of users [2][3]. The effects of these on a user in the detection of fake news functions are investigated. The underlying main research question of this paper is "How does a user's demographic attributes influence his ability to detect misinformation?". The basic defense of a user against the risk to the vulnerability of detecting misinformation is condensed into the proper use of using descriptors to recognize fake news when it is spotted, and as such, the contribution to the existing literature of this study will include the methodology used to empirically observe this risk.

The significance of this study is built upon the argument that proper recognition of misinformation is always a precursor phase to combatting the ecosystem of fake news in any space, including social networking sites. All existing solutions be it of any type of technological structure and design, relies first on the detection of an existence of misleading contents [4]. Other pragmatic implications of the results of this study are the increased accuracy in the calculation of risk in humans exposed to fake news campaigns through demographic profiling. From this point, campaigns of awareness to properly recognize fake news can be more focused to the more vulnerable group(s), as revealed by this study. The main output of the study is a predictive model that reveals the projected risk profile of the vulnerable targets of disinformation attacks. This paper will (1) discuss the literature highlighting the disinformation and each demographic factor, (2) discuss propositions and a model for building the predictive model, (3) provide the methodology for conducting the study, (4) share the findings and limitations of the study, and (5) highlight the future research.

II. Literature Review

A. Disinformation and Age

Craik et al. (1986) [5] proposed the reasoning that a user's information perception is dependent on his/her age. This is due to the fact that the mental capacity of a person undergoes many changes and development from infancy to adulthood. However, after reaching a theoretical peak age— this then progresses to decline up to the point of demise. Nelson (1998) [6] further explored this proposition and later revealed that one of the many strands in the development of this constant change is the evolution of a bias based on subjective and personal beliefs. In particular, [5][7] further developed the exploration of belief bias and how it plays a huge part in age and information perception. Two of the findings highlighted two important points: Younger adults, i.e., late teens or early

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

twenties to their thirties, performed more accurately and faster than older adults, i.e., past forty in logical and arithmetic tests. On the contrary, the study revealed that older adults performed better than their younger counterparts on exams involving the establishment of logical inferences from sequential arguments. Ding et al. (2019) [8] explained that logic tends to counteract with beliefs in subjects with varying age groups.

Corollary findings [9] have firmly linked age differences in belief bias, its dynamics, and how it paves the way to misinformation. The study found out that belief bias, more predominantly in older subjects, results to more resistance in a user's position when presented with logical evidence should it be contradicting with their beliefs. The potential for widespread dissemination of misinformation, [10] has examined the individual-level characteristics associated with sharing false articles during the 2016 US presidential campaign and has found a strong influence of age, which persists after controlling for partisanship and ideology: Information consumers over 65 shared nearly seven times as many articles from fake news domains as the youngest age group. Furthermore, this documents that ideology and age were associated with sharing activity as the dependent variable. The independent function of age is strongly observed in: Holding constant ideology, party identification, or both. Age, when separated among the variables, holds a strong influence wherein older consumers were more likely to share fake news than respondents compared with the next-youngest group, 18-29 years of age. This gap in the rate of fake news sharing between those defined by age categorization oldest category and youngest category is notably wide.

Concluding Proposition 1: Existing studies predominantly link stronger correlation to older subjects in the dissemination of misinformation compared to their younger counterparts. Among the age groups, the range of younger adults, 18-29, tends to have lesser documented acts of propagating misleading contents. The prominent cause of these having stemmed from belief bias.

B. Disinformation and Political Ideology

Political leaning is a motivational force to humans. It can inspire or compromise one into performing great atrocities, acts of courage, and kindness. Humans can even sacrifice lives—all in the name of an abstract belief system [11]. When political views function as a predictor, a magnitude of evidence shows that the effect of this to quite a number of users is quite different. The confirmation bias in its form compels users to consider unverified content to be still considered as factual. In extreme cases—even when content is verified to be untruthful is still considered factual by a user with stronger confirmation bias [12].

Generally, there are two partisan divisions, the right, which is highly partisan, while those on the left, which is more objective in their coverage [13]. This results in an asymmetrical media environment [14]. It is in a human's inclination to defend one's ideologies from information that conflicts with his political point of view, the negativity bias. This hypothesis of confirmation bias suggests that conservatives react to threats with greater negativity and motivated information processing than their liberal counterparts [15]. And time has yet again shown that anyone—regardless of political affiliation displays an internal informational processing when confronted with information that contradicts their point of view—resulting in a distinct perception [16]. This biasing effect becomes more obvious when objectionable political information relates to a position that one takes comfort in [17][18]. The negativity bias hypothesis appears such that people will react to threats by defending their ideological in-group and core ideological values when challenged.

Together, these results suggest that a politically natured group will respond to the basic psychological threats with the affirmation of important and salient values to them [15]. Disputes will arise when those holding differing political views collide—which are usually ubiquitous and deep-seated, which can often be followed on common, recognizable lines. Understanding the correlation of distinct political orientations is probably a prerequisite for managing political disputes, which are strongly tagged as a source of the incorrect perception that can then lead to misinformation campaign wars [19]. The cause of this social conflict is backed not just by political dimension but also in a multitude of dimensions in which opposing groups differ from each other in terms of several positions-from tastes in art, explanation of the sciences, up to the tendency to pursue a new type of information, almost always the variability in political ideologies of users may have caused it [20].

Alongside this, an understanding of the cause of its persistence and how it affects information consumption must be central to their themes [21]. Works by [21][22] attempted to first detect if a specific content is politically inclined and challenged the leanings of media contents to better understand fake news in a politically polarized mainstream media by investigating linguistic differences in the discussion of the topic by left- and right-leaning news sources. The study revealed that political leanings tend to employ non-neutral words, with the right inclined to more aggressive while the left inclined to put more emphasis on compassion.

With fake news garnering increasing public attention, a systematic literature review by [4] examined the way broader media environment and people in it may have caused or be damaged by disinformation. Technological studies provide that media sources used discernibly different language and catered for a specific consumer when discussing contents—underscoring a fundamental difference in definitions and portrayals of any issue, and how this ignites a frequent focus on each other, presumably in the spirit of blaming the other for a raised problem [21]. These works are all but the first steps towards identifying the descriptive differences between news sites' varying views in their discussions of a given subject.

Concluding Proposition 2: Existing studies reveal that politics and governance act as a motivator for political acts of a number of information users. Enabled by confirmation bias & negativity bias, this instantiates the creation and spreading of misleading news that highlights the political partisan and discredits the others.

C. Disinformation and Ethnicity

Cross-sectional studies [23][24] argued that ethnicity should be included as a demographic control variable in the assessment of the widespread problems in social sciences. The reason is that present findings with insights from most domains relevant to attitude formation, i.e., stereotyping literature, strongly bridges the potential role of ethnicity and perceived similarity.

Leighley et al. (2009) [25]'s underlying assumption is that the differences in ethnicity between groups may contribute to the homogeneity of culture in organizations. In developing nations, often characterized by a high level of ethnic diversity, concerns arise that groups with heterogeneous values, norms, and attitudes—the broad set of traits that can be referred to as culture—may be unable to agree on policies, the provision of public goods, and the broader goals of society [25]. Each of these traditions reflects a variety of viewpoints on the persistence of ethnic and cultural identities and a wide range of theories on the factors that gave rise to both ethnic and cultural differentiation.

One may not simply look away from the evidence consistent with a synthesis of both views: ethnicity is indeed associated with fundamental differences in values, attitudes, and preferences, including information consumption [26][27][28]. However, the reverse may not be true such that culture is the dependent variable which may not be attributed exclusively to ethnicity. Disinformation campaigns highlight non-specific trajectories on samples that are based on ethnicity variations. This accounts for the economically, and political biases that may help establish the link between culture and ethnicity. The complex relationship between ethnicity and culture in relative terms of fake news propagation had so far remained missing from the economics literature on ethnic heterogeneity.

Nelson (1979) [29] had undertaken a longitudinal study that documented the effects of socioeconomic status (SES), particularly ethnicity, on political participation. As consensus has grown on the role of socioeconomic status, other factors, like ethnicity, have been most of the time prioritized only to secondary in terms of importance. Variations in levels of participation can be traced, in part, to differences in ethnic political culture, and the web is no exception. Ethnicity has a more significant effect than socioeconomic status on levels of participant political culture.

The dynamics in which ethnicity might influence the understanding of social and political information retrieval in a rapidly evolving environment are of emerging concern. Fridkin et al. (2006) [30] recommended that future researchers give more serious attention to this ethnic factor. More studies will fill this gap especially focusing on the quantitative research on the relationship between ethnicity and culture that correlates to an individual's ethnolinguistic identity as a predictor of his norms, values, and preferences [30]. More than looking at the individual participant's ethnicity, the holistic view of the group composition should be taken into consideration—the diversity.

Concluding Proposition 3: The existing studies highlight the degree of overlap between ethnicity through culture and how it creates the attitude of an individual in his social identity in the information environment can all be traced to ethnicity as a potent determinant of civil conflict and public goods, in this case, misinformation. Furthermore, ethnicity is revealed to be a smaller part of a bigger determinant, diversity, that affects information flow on social networks.

D. Disinformation and Gender

Pennycook et al. (2018) [31] & Giacopassi et al. (1986) [24] revealed that weaponized disinformation campaigns and endemic online violence against gender minorities exist. Resources [32] that identify the demographics display consistent affinities to understanding how this phenomenon affects our democratic process to its propagation.

Social media environments are only beginning to grapple with the unintended affordances of their features and are not held accountable by regulatory governing bodies, which should serve as watchdogs and advocates—and gender as a variable, appeared from that. With gender as the issue i.e., variable—international nonprofits, academic institutions, and philanthropic investors see online threats as a form of social injustice and not a concern for democracy and national security. A new kind of authoritarianism seeks to push minority groups aside and halt progress on their rights by controlling social media channels, attacking the press, and limiting freedom of assembly and expression—and gender plays a firm role in this [33].

The support for gender-inclusive social web communication is deteriorating for certain minorities including women, and LGBTQIA+ communities. Belhadjali et al. (2017) [34] shows it's critical to analyze the role gender plays, consciously or unconsciously, for the promotion of more gender-inclusive and harmonious democracies—yet the convergence of gender, democracy, disinformation, and information technology remains understudied. Female politicians who are prolifically wielding social media as a way to overcome underrepresentation and connect with their constituencies—wherein a survey of female politicians from over 100 nations found that more than 85 percent of them use social media and particularly Facebook, with younger legislators being the most active—highlights the need in understanding whether online platforms are a level playing field for political engagement, or replicate the same biases as traditional media outlets [35][36].

Marwick (2017) [37] extracted data from leaders in different industries from the arts and sciences across multiple varying leanings, countries, and regions of the world, reviewed over 100 works of literature, and produced an analysis to identify gender trends in politics and information dissemination on the web in the United States. The research concluded that social media seem to provide some female candidates with an increased ability to promote their "brands" and level the engagement field, exposing that females are able to generate more followers, likes, and user engagement than their male counterparts.

On the other side of spectrum, there is a massive magnitude of evidence [38][39][40] that shows how women leaders and minority activists are often targets of online attacks, harassment aimed at making these personas seem irrelevant. Female politicians around the world expressed the sentiment of how a huge fraction of them having seen demoralizing images of them spread through an information environment or digital medium. The massive spread of gender-charged disinformation campaigns—ranging from mockery to even death threats—transcends into the physical world. Often, attacks come from armies of politically motivated trolls and bots. Stahl (2006) [41] shows that in a political party—female candidates are targeted more often than male candidates by fake news accounts with the explicit intention of halting certain gender-minority from actively engaging in the information medium. When fake news is aimed at certain gender-based communities, the whole information medium produces a bias through an attempt to control the users who can contribute to the ecosystem of information.

Concluding Proposition 4: Gender inclusiveness and equal participation is a precursor towards an actively participatory democracy that allows social media to be utilized effectively to bring all, regardless of gender, closer to government. The need to examine gender as an enabler or otherwise in a participatory environment, including the phenomenon of disinformation, is the very reason why academic institutions, civil society groups, and philanthropists who aim to protect and foster democratic values have a responsibility to look into the gendered dimension of fake news and online violence against gender-minority in politics.

III. Methodology

Figure 1 below describes an overview of the study's methodology. Note that the data collection device is the fake news test.

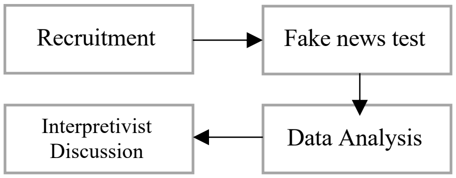


Fig. 1. Block Diagram of Methodology

A. The Population

The survey was opened for an undergraduate class of Introductory in Informatics during the period September 1, 2020, to December 7, 2020. Of the 234 eligible and invited students to take the optional survey, 168 responded. Declared majors and concentrations of subjects are varying, including but not limited to cybersecurity, computer science, informatics, economics, communications, sociology, business administration, accounting, and physics. The age range of subjects is from 18 to 25.

B. Data Collection

A preliminary survey questionnaire that extracts demographic data from each participant was asked. To force the users to select an option, all of these fields are required—by default leaving it empty will prevent one from proceeding to the next item.

To associate the above demographic data to the risk of being disinformed, the users are then later given the fake news test devised by Caramancion (2020) [42] based on the Ecosystem Theory of Disinformation [43]. The test displays a content to a user and explicitly asks for a response if it is legitimate or illegitimate. Furthermore, the fake news test includes screenshots of *deepfakes* which then prompts the users to evaluate if it is legitimate or otherwise.

C. Data Analysis

General descriptive statistics and grade distribution were used to establish the baseline of measures of the results for the dependent variable. This is generally to get an overview of the population's performance. To answer this paper's research questions, the independent variables considered are the demographic data of each participant, i.e., age (in range), political leaning, gender, and ethnicity.

These IVs are then associated with the performance of each participant as their score on the fake news test and will serve as the dependent variable for all of the IVs stated before. To perform this analysis, a multiple linear regression was employed as the main analysis technique of this paper. Dummy coding was employed before loading the IVs to the model. No violation in the assumptions exists, including normality, multicollinearity, and homoscedasticity. Finally, an interpretivist approach was employed by the author to give possible explanations of the results.

D. Controls and Limitations

The survey was deployed in blackboard and is strictly timed so as to mimic the actual stimuli environment of browsing a social media feed. The whole survey was timed 60 minutes and will auto-submit after the timer has elapsed, although, as per the item analysis, the average user typically completed the assessment in 5 to 10 minutes. The range of user attention span per content as revealed by Facebook Research is varying from 1.7 seconds to 12 seconds. The themes of the questions are all relevant to the national affairs within the United States. Backtracking was prohibited to disallow any modification of answers to earlier questions.

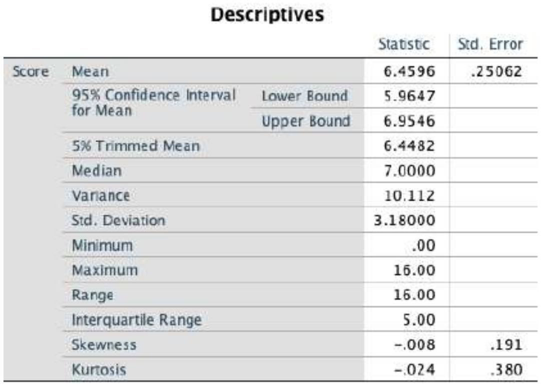
Among all the participant responses, seven attempts were incomplete, i.e., the user opened the assessment but left the survey unanswered, leading the timer to expire and automatically submit their attempts. These responses were removed from the aggregate data.

IV. Results

A. Overview

Table 1 displays the general descriptive statistics of the regression results in aggregate form. Table 2 presents the model information of the predictive model alongside the Durbin-Watson's test. Table 3 highlights the coefficient of the predictors alongside the collinearity diagnostics. Table 4 presents the details of the regression model, including coefficients and their respective statistical significance.

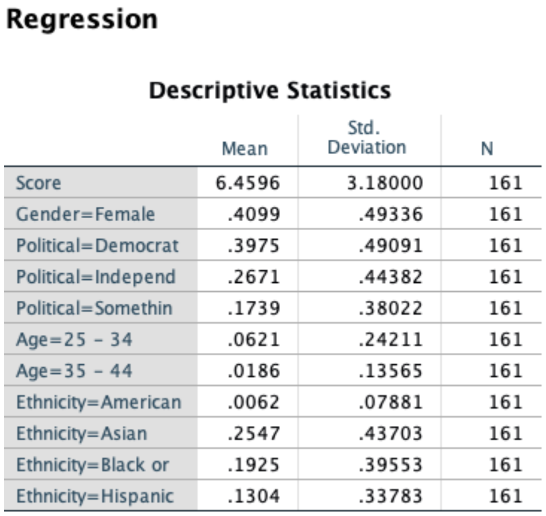
TABLE I. GENERAL STATISTICS



Descriptives

			Statistic	Std. Error
Score	Mean		6.4596	.25062
	95% Confidence Interval for Mean	Lower Bound	5.9647	
		Upper Bound	6.9546	
	5% Trimmed Mean		6.4482	
	Median		7.0000	
	Variance		10.112	
	Std. Deviation		3.18000	
	Minimum		.00	
	Maximum		16.00	
	Range		16.00	
	Interquartile Range		5.00	
	Skewness		-.008	.191
	Kurtosis		-.024	.380

TABLE II. POPULATION STATISTICS

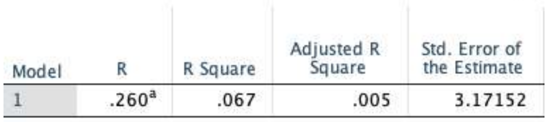


Regression

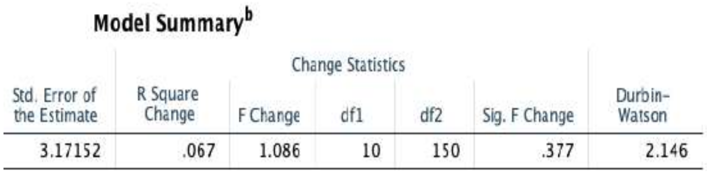
Descriptive Statistics

	Mean	Std. Deviation	N
Score	6.4596	3.18000	161
Gender=Female	.4099	.49336	161
Political=Democrat	.3975	.49091	161
Political=Independ	.2671	.44382	161
Political=Somethin	.1739	.38022	161
Age=25 - 34	.0621	.24211	161
Age=35 - 44	.0186	.13565	161
Ethnicity=American	.0062	.07881	161
Ethnicity=Asian	.2547	.43703	161
Ethnicity=Black or	.1925	.39553	161
Ethnicity=Hispanic	.1304	.33783	161

TABLE III. MODEL SUMMARY



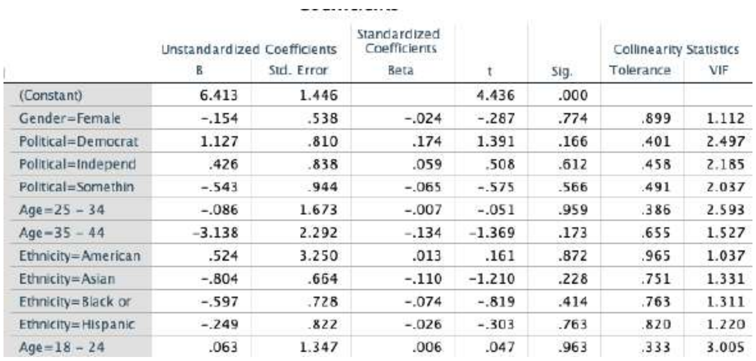
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.260a	.067	.005	3.17152



Model Summaryb

Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
	Change Statistics					
3.17152	.067	1.086	10	150	.377	2.146

TABLE IV. PREDICTOR COEFFICIENTS



	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	6.413	1.446		4.436	.000		
Gender=Female	-.154	.538	-.024	-.287	.774	.899	1.112
Political=Democrat	1.127	.810	.174	1.391	.166	.401	2.497
Political=Independ	.426	.838	.059	.508	.612	.458	2.185
Political=Somethin	-.543	.944	-.065	-.575	.566	.491	2.037
Age=25 - 34	-.086	1.673	-.007	-.051	.959	.386	2.593
Age=35 - 44	-3.138	2.292	-.134	-1.369	.173	.655	1.527
Ethnicity=American	.524	3.250	.013	.161	.872	.965	1.037
Ethnicity=Asian	-.804	.664	-.110	-1.210	.228	.751	1.331
Ethnicity=Black or	-.597	.728	-.074	-.819	.414	.763	1.311
Ethnicity=Hispanic	-.249	.822	-.026	-.303	.763	.820	1.220
Age=18 - 24	.063	1.347	.006	.047	.963	.333	3.005

V. DISCUSSION

Discussion 1: An Increase in Age is the Biggest Risk Factor of Disinformation

The regression model suggests that the higher the age of the subjects, the higher their risk of being disinformed. Subjects with an age range of 25 to 34 score less than their younger counterparts. Furthermore, subjects with age range 35 to 44 scored significantly lower. The optimum age of subjects that scored the highest were 18 to 24, followed by the age group below 18.

Discussion 2: Females Appear to be at More Risk Compared to Males

The regression model suggests that, on average female participants score lower, 1.54 less, compared to the male subjects. Although, among all the demographic factors, gender appears to be the least statistically significant. It should also be further noted that the sample ratio of male-to-female is 95:66.

Discussion 3: Asians, African Americans, Hispanics, and Latinos all have a Bigger Risk of being Disinformed

Interestingly, the predictive model confirmed that the historically underrepresented ethnic group that appeared to garner the lowest score are the Asians, followed by African American subjects, and then the Hispanics and Latinos.

Discussion 4: American Indians and White Subjects have Lower Risk of being Disinformed

The regression model further revealed that American Indians and Alaskan Natives scored the highest, implying their skill in disinformation detection. The ethnic group that follows from this are the Caucasian/White.

Discussion 5: Political Parties Outside the Outside the Major Parties have Bigger Risk of being Disinformed

The regression model suggests that subjects who self-identify as Democrats tend to recognize Disinformation and Deepfakes most accurately. This is followed by those who self-identify as Independent, followed by Republicans. The group who performed the least are the subjects who has Political Affiliation that's Something Else.

VI. Conclusion And Future Works

This study was incepted with the hope of reducing the instances and impact of disinformation on information consumers. The resulting output is a predictive model that uses the demographic factors – a. age, b. gender, c. political ideology, and d. ethnicity as independent variables. The model is based on the interaction among these variables in a given population.

With Caramancion (2020)'s work as inspiration, the first part of the experiment revealed the accuracy of individuals in detecting fake news in a social media feed. The results are then aggregated and analyzed to establish a correlation based on their respective predictors.

Findings are the result of the subject's empirical measure from the experiment mapped out against the four preceding factors. Based on the scores of the subjects, these findings revealed the profile with the weakest ability to detect misinformation, prompting to result in the research question of this study being explicitly answered. The distinct parts of the result, i.e., regression coefficients, are then compared with one another to search for behavioral patterns that revealed the profiles who a. At most risk to be misinformed, b. those who can accurately detect Fake News.

This study lays the foundation in assessing a person's participation in the persistence of Disinformation on the digital space. The ability of technological advances to combat disinformation through detection is learned through training. One practical area for future research is the applications of different style of cyber awareness/education in shielding these individuals against disinformation—therefore highlighting the importance of strengthening the user's capability in recognizing misleading contents against legitimate ones. The variations in the results highlight the need for variations in education. Future studies may enhance the predictive model developed in this paper by injecting more socioeconomic factors as independent variables and through a consistent and incremental validation of the subjects in a longitudinal research. Finally, a population of bigger size may be considered to better reflect the domain attempting to measure.

References

[1] Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.

[2] Kohansal, M. R., & Firoozzare, A. (2013). Applying multinomial logit model for determining socioeconomic factors affecting major choice of consumers in food purchasing: The case of Mashhad. Journal of Agricultural Science and Technology, 15(7), 1307-1317.

[3] Sánchez, M., Beriain, M. J., & Carr, T. R. (2012). Socioeconomic factors affecting consumer behaviour for United States and Spanish beef under different information scenarios. Food Quality and Preference, 24(1), 30-39.

[4] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.

[5] Craik, F. I., & Bialystok, E. (2006). Cognition through the lifespan: mechanisms of change. Trends in cognitive sciences, 10(3), 131-138.

[6] Nelson, K. (1998). Language in cognitive development: The emergence of the mediated mind. Cambridge University Press.

[7] Craik, F. I. (2002). Human memory and aging. In International Congress of Psychology, XXVII, 2000, Stockholm, Sweden; Based on addresses and lectures presented at the aforementioned congress. Psychology Press/Taylor & Francis (UK).

[8] Ding, D., Chen, Y., Lai, J., Chen, X., Han, M., & Zhang, X. (2019). Belief Bias Effect in Older Adults: Roles of Working Memory and Need for Cognition. Frontiers in Psychology, 10.

[9] Fortier A., Burkell J. (2014). Influence of need for cognition and need for cognitive closure on three information behavior orientations. *Proc. Am. Soc. Inform. Sci. Technol.* 51 1–8.

[10] Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science advances, 5(1), eaau4586.

[11] Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. Motivation and Emotion, 36(1), 55-64.

[12] Plous, S. (1993), The Psychology of Judgment and Decision Making, p. 233

[13] Rubin, V. L., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In Proceedings of the second workshop on computational approaches to deception detection (pp. 7-17).

[14] Benkler, Y., Faris, R., Roberts, H., & Zuckerman, E. (2017). Study: Breitbart-led right-wing media ecosystem altered broader media agenda. *Columbia Journalism Review*, *3*, 2017.

[15] Brandt, M. J., Wetherell, G., & Reyna, C. (2014). Liberals and conservatives can show similarities in negativity bias. Behavioral and Brain Sciences, 37(3), 307-308.

[16] Nickerson, R. (1998), "Confirmation bias: A ubiquitous phenomenon in many guises", Review of General Psychology, 2 (2): 175–220, doi:10.1037/1089-2680.2.2.175

[17] Crawford, J. T. (2014). Ideological symmetries and asymmetries in political intolerance and prejudice toward political activist groups. Journal of Experimental Social Psychology, 55, 284–298.

[18] Taber, C. S., Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. American Journal of Political Science, 50, 755–769.

[19] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. Nature communications, 9(1), 1-9.

[20] Raaijmakers, Q. A., & Hoof, A. V. (2006). Does moral reasoning represent sociomoral structure or political ideology? A further exploration of the relations between moral reasoning, political attitudes, consistency of moral thought, and the evaluation of human rights in Dutch young adults. Social Behavior and Personality: an international journal, 34(6), 617-638.

[21] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. Science, 363(6425), 374-378.

[22] Che, X., Metaxa-Kakavouli, D., & Hancock, J. T. (2018, October). Fake News in the News: An Analysis of Partisan Coverage of the Fake News Phenomenon. In Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 289-292).

[23] Ong, J. C., & Cabanes, J. (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. Newton Tech4Dev Network.

[24] Giacopassi, D. J., & Dull, R. T. (1986). Gender and racial differences in the acceptance of rape myths within a college population. Sex Roles, 15(1-2), 63-75. Myths within a college population. Sex Roles, 15(1-2), 63-75.

[25] Leighley, J. E., & Matsubayashi, T. (2009). The implications of class, race, and ethnicity for political networks. American Politics Research, 37(5), 824-855.

[26] Brown, T. N., Sellers, S. L., Brown, K. T., & Jackson, J. S. (1999). Race, ethnicity, and culture in the sociology of mental health. In Handbook of the sociology of mental health (pp. 167-182). Springer, Boston, MA.

[27] Verba, S., Schlozman, K. L., Brady, H., & Nie, N. H. (1993). Race, ethnicity and political resources: Participation in the United States. British Journal of Political Science, 23(4), 453-497.

[28] Austin, E. W., & Nelson, C. L. (1993). Influences of ethnicity, family communication, and media on adolescents' socialization to US politics. Journal of Broadcasting & Electronic Media, 37(4), 419-435.

[29] Nelson, D. C. (1979). Ethnicity and socioeconomic status as sources of participation: The case for ethnic political culture. American Political Science Review, 73(4), 1024-1038.

[30] Fridkin, K. L., Kenney, P. J., & Crittenden, J. (2006). On the margins of democratic life: The impact of race and ethnicity on the political engagement of young people. American Politics Research, 34(5), 605-626.

[31] Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of personality.

[32] Shu, K., Wang, S., & Liu, H. (2018). Understanding user profiles on social media for fake news detection. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 430-435). IEEE.

[33] Michael, K. (2017). Bots trending now: Disinformation and calculated manipulation of the masses. IEEE Technol. Soc. Mag., 36(2), 6-11.

[34] Belhadjali, M., Whaley, G. L., & Abbasi, S. M. (2017). Misinformation online: A preliminary review of survey results on Americans' perceptions by gender, ethnicity, and party affiliation. In Competition Forum (Vol. 15, No. 2, pp. 324-328). American Society for Competitiveness.

[35] Newsom, V. A., & Lengel, L. (2012). Arab Women, Social Media, and the Arab Spring: Applying the framework of digital reflexivity to analyze gender and online activism. Journal of International Women's Studies, 13(5), 31-45.

[36] González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. Social Networks, 38, 16-27.

[37] Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. New York: Data & Society Research Institute.

[38] Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. Berkman Klein Center Research Publication, 6.

[39] Swank, E., & Fahs, B. (2013). An intersectional analysis of gender and race for sexual minorities who engage in gay and lesbian rights activism. Sex Roles, 68(11-12), 660-674.

[40] Olson, C. C., & LaPoe, V. (2018). Combating the digital spiral of silence: academic activists versus social media trolls. In Mediating Misogyny (pp. 271-291). Palgrave Macmillan, Cham.

[41] Stahl, B. C. (2006). On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. Informing Science, 9.

[42] Caramancion, K. M. (2020, September). Understanding the Impact of Contextual Clues in Misinformation Detection. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-6). IEEE.

[43] Caramancion, K. M. (2020, March). An exploration of disinformation as a cybersecurity threat. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 440-444). IEEE.

On error correction performance of LDPC and Polar codes for the 5G Machine Type Communications

Salima Belhadj
Department of Electrical Engineering
University of Tahri Mohammed
Algeria
Belhadjsalima08@gmail.com

Moulay Lakhdar Abdelmounaim
Department of Electrical Engineering
University of Tahri Mohammed
Algeria
moulaylakhdar78@yahoo.com

Abstract—**Channel coding for the 5th generation (5G) wireless communication system has to fulfill diverse requirements arising from new machine type communication (MTC) services. The 5G-MTC applications can be classified into two categories: Ultra-Reliable Low-Latency Communications (URLLC) and massive Machine-Type Communication (mMTC). Polar code and Low-Density Parity Check (LDPC) code are among the most advanced channel coding techniques known today that have the potential to be used in the 5G. The main objective of this paper is to evaluate the error correction performance of Polar and LDPC coding schemes in the case when short to medium information block lengths are transmitted, as it is often in mMTC and URLLC scenarios. These codes are evaluated in terms of both Block Error Rate (BLER) and Bit Error Rate (BER). Simulation results show that Polar codes exhibit a much better error correction capability compared to the LDPC codes for short block lengths, while they have a comparable performance at medium block length.**

Keywords—5G-MTC, LDPC, Polar codes, mMTC, URLLC.

I. Introduction

With the rapid development of wireless communication technology, the 5th generation (5G) has received considerable attention and emerged as a more advanced way in telecommunications [1]. Compared to the fourth-generation (4G) wireless systems, 5G has been designed not only to enhance mobile broadband (MBB) applications but also to support new services centered on machine type communications (MTC) [2], [3].

Machine-type communication (MTC) is one of the main communication paradigms for a wide range of emerging services that enables devices to interconnect wirelessly with minimal or without human intervention, the main usage scenarios of 5G-MTC are [4]: massive MTC (mMTC) and mission-critical MTC (mcMTC) also known as ultra-reliable low-latency communications(URLLC). As the term speaks for itself, mMTC is about providing wireless connectivity to a large number of machine type terminals, with low costs and extremely low energy consumption [5]. While URLLC, is about communication with a high level of reliability (e.g., 99.999%) and low latency in the order of 1ms or less [6]. Recently, numerous researches have been performed to meet the challenging requirements of these scenarios such as research on the field of channel coding.

Channel coding is one of the most prominent and critical subjects in today's wireless communication systems. The goal of channel coding is to protect the information from noise and interference encountered in the transmission through the channel and this is achieved by using a channel encoder and decoder. The former introduces redundancy in a controlled manner in the transmitted information sequence, while the latter exploits this redundancy to detect and correct errors on the receiver side.

The recent interest in channel coding schemes for the 5G has directed to polar codes and LDPC codes [7]. For the Enhanced Mobile Broadband (eMBB) scenario, LDPC code has been chosen for the data channels and Polar code has been adopted for the control channels [8]. As for URLLC and mMTC use cases, research on efficient coding schemes is still ongoing. The candidate coding schemes considered for 5G networks are Turbo code, LDPC code and polar codes [9]. However, Turbo codes are not recommended to use for MTC scenarios since it exhibits a poor performance when information block length is small. Besides, the complexity of Turbo codes is much higher than other modern coding schemes [10]. LDPC has come out as one of the major channel coding contender for 5G [11]. While recent investigations [12] demonstrated that Polar codes outperform modern coding schemes without any sign of error floor, Furthermore, they have the lowest system complexity among the competitive codes that might be used in 5G systems which render them preferred for 5G mMTC/URLLC use cases [13].

In order to achieve the requirements of the mMTC and URLLC, the selected channel coding scheme should have the capability to support short information blocks with low computation complexity, low latency and very good error correction performance to provide higher reliability. Hence, This paper mainly focuses on error correction performance of 5G channel coding schemes namely Polar code and LDPC code for short and medium information block lengths(64<K<1024) in the context of mMTC/URLLC scenarios. However, there are other parameters needed to be taken into account to determine the appropriate coding technique for URLLC and mMTC use cases. The considered codes are evaluated and compared in terms of BLER and BER over an Additive Gaussian Noise (AWGN) channel.

Our paper is organized as follow. In Section II, we briefly introduce the 5G channel coding techniques namely: LDPC codes and Polar codes. BER and BLER simulations results of the discussed codes are provided in SectionIII. Finally, we provide conclusion in Section IV

II. PRELIMINARIES OF CHANNEL CODES

In this section, LDPC and Polar coding schemes are briefly introduced, respectively.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

A. LDPC codes:

In 1962, R. Gallager introduced a family of error correction codes, called Low-Density Parity Check (LDPC) Codes [14]. But they were then largely forgotten. LDPC codes were rediscovered by Mackay in 1996 [15] and, since then, they have been applied in numerous communications systems. As the name implies, they are characterized by a sparse parity check matrix H, where sparse means that most of the entries are zero. An example of H is the following:

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \qquad (1)$$

LDPC code can also be described by a graph, called tanner or bipartite graph. [16], which is composed of Check nodes (CNs) and the Variable nodes (VNs). In the Tanner graph, variable node i is connected to check node j, whenever h_{ij} of H is non-zero, as illustrated in Fig. 1.

Recently, research on LDPC codes has been oriented on quasi-cyclic (QC) LDPC code. The 5G LDPC code belong to the class of QC-LDPC codes, where two base graphs are defined [17].

The LDPC decoding can be implemented by using Sum-Product(SP) algorithm, but it suffers from high complexity. This complexity can be greatly reduced by using the min-Sum algorithm (MSA)[18]. In order to increase the convergence speed of MSA, Layered Min-sum algorithm was proposed [19].

B. Polar codes

Polar code [20], invented by Arikan in 2008, is a special class of error correcting codes that can provably achieves the channel capacity. Polar codes exploit a novel concept called channel polarization, which converts the transmission channel into virtual channels of different capacities. In the limit of infinite block-length, it can be shown that channels become either perfectly noiseless or very noisy. Polar codes are then constructed by only using the noiseless channels for transmitting information while freezing the inputs of all other channels to known values.

The encoding of Polar code with code length N ($=2^n$, $n \geq 1$) is performed using the generator matrix obtained from the polar transform. The generator matrix G_N can be written as:

$$G_N = F^{\otimes n}, \quad F = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad (2)$$

Where $F^{\otimes n}$ denotes the n^{th} Kronecker power of F. The encoding is a process to obtain the encoded bits x $=\{x_1, x_2, \ldots, x_N\}$ through x = u G_N=u $F^{\otimes n}$ for a given source vector u = $\{u_1, u_2, \ldots, u_N\}$. The source vector u consists of the frozen bits and information bits. Fig.2 shows the polar encoding for N=4.

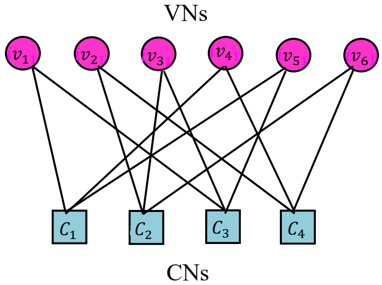


Fig.1. Tanner graph for LDPC Code.

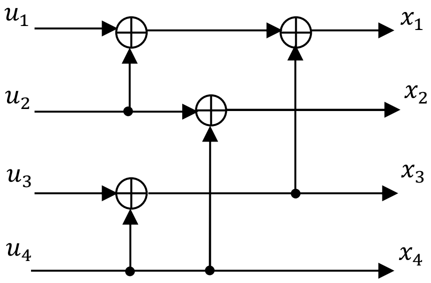


Fig.2. Polar encoder for N=4

The decoding of polar code can be achieved by a Successive Cancellation (SC) algorithm. Unfortunately, Polar code under standard SC algorithms [20] is rather unsatisfactory at short blocklengths. For this reason, Successive Cancellation List (SCL) decoder has been introduced in [21] to mitigate these problems. The performance of the SCL decoder can be further improved by concatenating them with a Cyclic Redundancy Check (CRC) codes (CRC-SCL), where the outer CRC code is used to determine a valid codeword within the list of candidates at the end of the aforementioned decoding process (SCL) [22].

III. SIMULATION RESULTS

The performance of LDPC and polar codes are evaluated for information block lengths of K= 64,128, 256, 512 and 1024 bits with code rate R=1/3. The evaluations are performed in terms of BLER/BER vs. SNR using Binary Phase shift keying (BPSK) for modulation and the channel considered is AWGN. The results are given in below figures from Fig.3 to Fig.7.

In our simulation, we have considered LDPC code based on the 5G specifications [17] and layered min-sum algorithm with 15 iterations was used in the decoder. CRC-SCL algorithm with list size L=8 and CRC of length 16 is used to decode Polar code.

From Fig. 3 and Fig. 4, it is clear that the performance of Polar code is superior to that of LDPC code for short information block lengths (K $\leq$ 128), in both BER and BLER. The performance gain of Polar code against LDPC code is around 1 dB for k=64, and 0.65 dB for k=128 at BLER=10^{-4}.

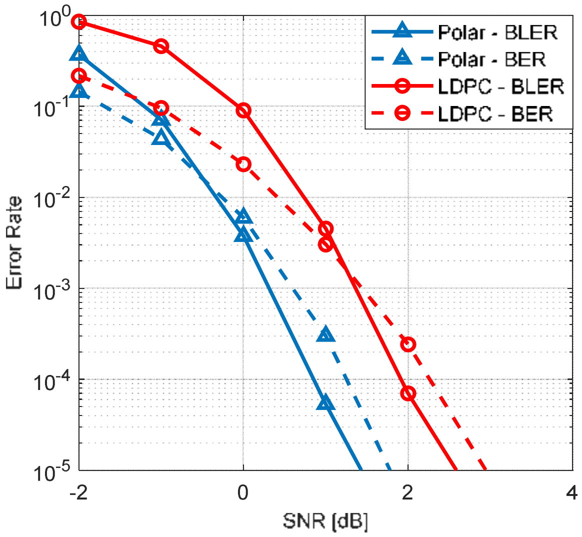


Fig.3. BER and BLER Performance of channel codes for K = 64 bits and R=1/3.

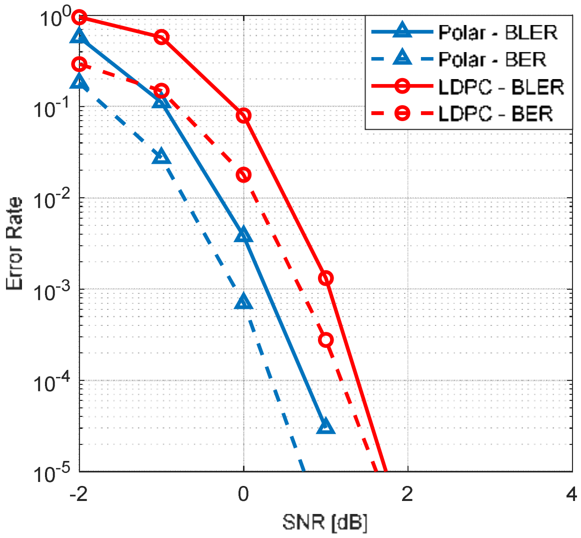


Fig.4. BER and BLER Performance of channel codes for K = 128 bits and R=1/3.

Furthermore, it can be seen from the results that as the code length increase, the performance gain of polar codes over LDPC codes decreases. As shown in Fig.5, Polar codes perform slightly better than LDPC in BER, while the performance of LDPC come close to the performance of the polar code in BLER at this information block length (K=256).

We can see from Fig. 6 and Fig. 7, that LDPC code has comparable performance as that of polar code at information bit length of K=512 and K=1024. We also observe that the performance of LDPC is around 0.15 dB better than the Polar code at BLER=10^{-4} for K=1024. So, it is concluded that at short block lengths polar code offers the best error correction performance, while LDPC and Polar offer similar performance at moderate information block length.

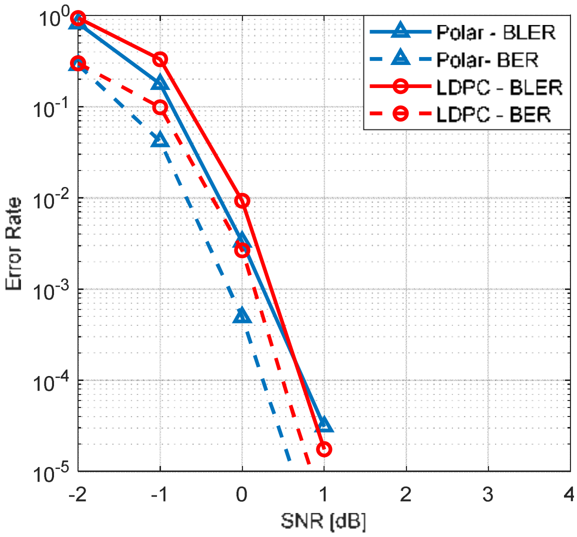


Fig.5. BER and BLER Performance of channel codes for K = 256 bits and R=1/3.

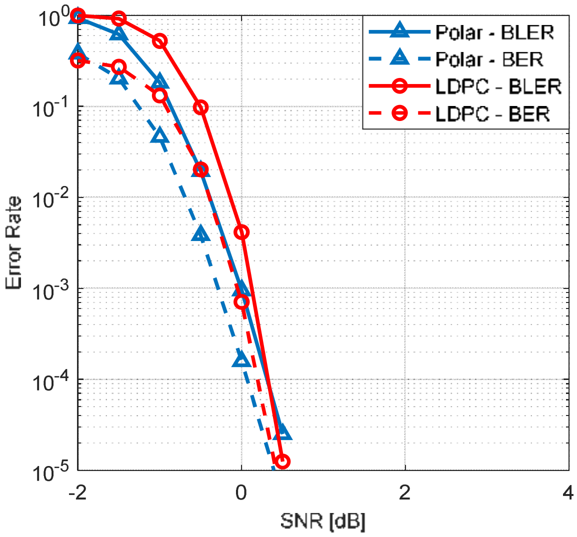


Fig.6. BER and BLER Performance of channel codes for K = 512 bits and R=1/3.

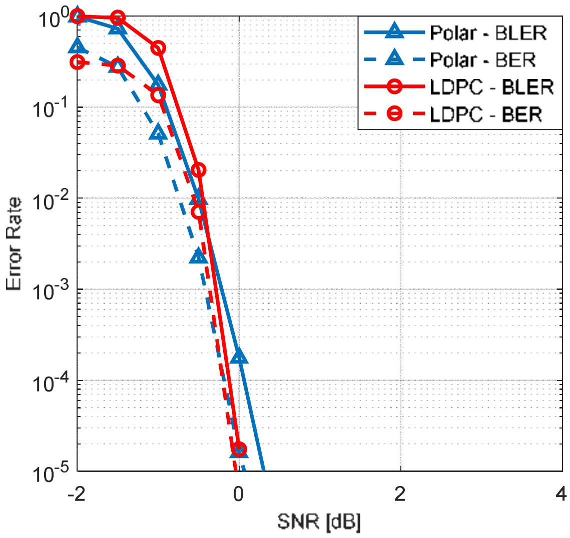


Fig.7. BER and BLER Performance of channel codes for K = 1024 bits and R=1/3.

IV. CONCLUSION

This paper has presented the error correction performance evaluation of LDPC code and polar code for short to medium information block lengths (K ≤ 1024) at a code rate of 1/3 under URLLC and mMTC scenarios. AWGN and BPSK are considered as the communication channel and modulation scheme, respectively. From the results, it is observed that the performance of Polar code with CRC-SCL decoder is better than that of LDPC codes for short information block lengths. While at medium information block length they show a comparable performance. Consequently, it can be said that Polar codes are a promising channel coding scheme in such scenarios.

REFERENCES

[1] J. A. Adebusola, A. A. Ariyo, O. A. Elisha, A. M. Olubunmi and O. O. Julius, "An Overview of 5G Technology," *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*, Ayobo, Nigeria, 2020

[2] A. Ghosh, A. Maeder, M. Baker and D. Chandramouli, "5G Evolution: A View on 5G Cellular Technology Beyond 3GPP Release 15," in *IEEE Access*, vol. 7, pp. 127639-127651, 2019, doi: 10.1109/ACCESS.2019.2939938.

[3] ITU, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," Nov. 2017*, report ITU-R M.*2410-0.

[4] S. Zhang, Y. Wang and W. Zhou, "Towards secure 5G networks: A survey", *Comput. Netw.*, vol. 162, Oct. 2019, [online] Available: https://linkinghub.elsevier.com/retrieve/pii/.

[5] A. Jayawickrama, Y. He, E. Dutkiewicz and M. D. Mueck, "Scalable Spectrum Access System for Massive Machine Type Communication," *in IEEE Network*, vol. 32, no. 3, pp. 154-160, May/June 2018

[6] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, "Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects,*" in IEEE Wireless Communications*, vol. 25, no. 3, pp. 124-130, JUNE 2018.

[7] D. Hui, S. Sandberg, Y. Blankenship, M. Andersson and L. Grosjean, "Channel Coding in 5G New Radio: A Tutorial Overview and Performance Comparison with 4G LTE," in *IEEE Vehicular Technology Magazine*, vol. 13, no. 4, pp. 60-69, Dec. 2018.

[8] N. Yang et al., "Reconfigurable Decoder for LDPC and Polar Codes", *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-5, 2018.

[9] H. Gamage, N. Rajatheva and M. Latva-aho, "Channel coding for enhanced mobile broadband communication in 5G systems," 2017 *European Conference on Networks and Communications* (EuCNC), Oulu, 2017, pp. 1-6

[10] Z. R. M. Hajiyat, et al., "Channel Coding Scheme for 5G Mobile Communication System for Short Length Message Transmission," *Wireless Personal Communications,* vol. 106, no. 2, pp. 377–400, 2019

[11] K. Arora, J. Singh and Y. S. Randhawa, "A survey on channel coding techniques for 5G wireless networks", *Telecommun. Syst.*, vol. 73, pp. 1-27, Nov. 2019.

[12] "R1-1611081 - Final Report of 3GPP TSG RAN WG1 #86bis v1.0.0," 3GPP, *Final Minutes Report*, Nov. 2016.

[13] M. V. Patil, S. Pawar and Z. Saquib, "Coding Techniques for 5G Networks: A Review," 2020 *3rd International Conference on Communication System, Computing and IT Applications* (CSCITA), Mumbai, India, 2020, pp. 208-213,

[14] R. G. Gallager, "Low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 8, pp. 21–28, 1962.

[15] D. MacKay and R. Neal, "Good codes based on very sparse matrices," in *Proc. 5th IMA Conf. Cryptography and Coding*, Oct. 1995, pp. 100–111.

[16] R. Tanner, "A recursive approach to low complexity codes," *IEEE Transactions on Information Theory*, vol. 27, no. 5, pp. 533–547, Sep. 1981.

[17] T. Richardson and S. Kudekar, "Design of Low-Density Parity Check Codes for 5G New Radio," *IEEE Communications Magazine*, vol. 56, pp. 28-34, 2018.

[18] P. Dhanorkar and M. Kalbande, "Design of LDPC decoder using message passing algorithm," *2017 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 2017, pp. 1923-1926

[19] D. E. Hocevar, "A reduced complexity decoder architecture via layered decoding of LDPC codes," *IEEE Workshop on Signal Processing Systems*, pp. 107 - 112, 2004.

[20] E. Arıkan, "Channel Polarization: A Method for Constructing Capacity- Achieving Codes for Symmetric Binary-Input Memoryless Channels," *IEEE Trans. Inf. Theory,* vol. 55, no. 7, pp. 3051–3073, Jul. 2009.

[21] I. Tal and A. Vardy, "List Decoding of Polar Codes," *IEEE Int'l. Symp.Info. Theory* (ISIT), 2011, pp. 1–5.

[22] K. Niu and K. Chen, ''CRC–aided decoding of polar codes,'' *IEEE Commun. Lett.*, vol. 16, no. 10, pp. 1668–1671, Oct. 2012

Mobile Application Recommendation System for Mobile Data Plan Research in progress

Mohamed Imran Mohamed Ariff
Department of Computer Science,
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
moham588@uitm.edu.my

Nor Ainol Mat Yaqin Fadzil
Department of Computer Science,
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
ainolyaqin414@gmail.com

Noreen Izza Arshad
Department of Computer and Information Sciences
Faculty of Science and Information Technology
Universiti TEKNOLOGI PETRONAS
Seri Iskandar, Perak, Malaysia
noreenizza@utp.edu.my

Khairulliza Ahmad Salleh
Department of Computer Science,
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
khair279@uitm.edu.my

Jufiza A.Wahab
Department of Mathematics
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
jufiz279@uitm.edu.my

Anis Zafirah Azmi
Department of Mathematics
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
anis9108@uitm.edu.my

***Abstract*—This paper presents a research in progress mobile application recommendation system for mobile data plan. With the vast amount of mobile data plan available, students must put a lot of effort in finding a mobile data plan that is the best value for money or the best for their heavy usage needs. Thus, paper proposes a recommendation system to aid students in the selection process of a mobile data plan. Further, hybrid filtering will be applied in the development process of this system to recommend a few suitable mobile data plans.**

***Keywords*—Recommendation System, Hybrid Filtering, Mobile Data Plan, MDLC**

I. Introduction

In recent years, telecommunication using mobile phones are entangling to everyone's daily life. The use of mobile phone is not mainly used for telephone conversation, but it has also been used mainly: (1) for text messaging and (2) surfing the net via the use of mobile data provided by the mobile telco operators. The demand for a better use of mobile data is increasing enormously as the usage of a more stable and fast mobile data is increasing due to various use and complex usage the mobile devices among users. Among the usage of mobile phone via the use of the mobile data are internet surfing, video streaming, downloading, and mobile learning mainly among students. Mobile learning is a learning process using various mobile gadgets, and it is transforming the learning environment by allowing students to participate in an asynchronous, and ubiquitous mode [1]. The extensive implementation of mobile learning has resulted in students relying on the usage of their personal mobile data in order to gain access to extensive learning materials[2]. Thus the demand for students in getting the best mobile data plan is vital [3].

The structure of this paper is as follows. The following section will highlight previous studies on the topic of interest. Following that, the recommendation technique relating to the topic of interest will be presented. Next the methodology and system architecture of the project will be presented. Finally, future works of this project will be presented.

II. Literature Review

Previous studies has highlighted that: (1) the use of mobile phone in learning enables student to communicate easily and better with their fellow lecturers, and (2) it also helps students to exchange information and study materials [4-7]. However, other related studies on mobile learning has highlighted that this kind of learning creates financial strains on students as they stressed the need of having a better mobile data plan in order to fully utilize the purpose of mobile learning [8].

The demand for the usage mobile data plan has risen over the years and are driven by several factors, among which: (1) mobile data fixed rates, (2) how easily it connects to laptops (3) online gaming (4) the ability to watch videos while on the go and (5) educational purposes [5, 9]. As for a student, the demand for a good and reliable mobile data plan from a mobile telco provider is vital, as students use extensive mobile data in for their mobile learning process [10]. Furthermore, every student needs: (1) the lowest mobile call rate, (2) uninterrupted mobile service, and (3) mobile data plan at the best price [11].

Over the years, there has been a great amount of mobile data plan available, and students have to put a lot of effort in finding the most suitable mobile data plan. Furthermore, is has also been suggested the implementation of a recommendation mobile data plan system in order to help users (i.e. students) to choose the best available mobile data plan [12, 13]. As such by creating a recommendation system, it would help minimize the information overload in a personalized way (i.e., for students) thus it would be also beneficial not only for the end user, but for the mobile telco provider as the would get more information about the users need. A recommendation system is a specific information

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

filtering system which ranks the existing information base on specific user conditions. Further, the outcomes from a recommendation system would rank the information based on specific filtering technique.

III. Filtering Technique

Recommender systems are software applications which is designed to provide suggestions that could be of interest a User [14, 15]. Further, each recommender application is base on certain technique. The following sub section will highlight and briefly explain three main filtering technique which are widely highlighted in previous research, mainly the: (1) content base, (2) collaborative filtering (CF), and (3) hybrid.

A. Content base filtering

Content base filtering is a technique base on the assumption of a user and several users personal interests. If a user intertest in a particular topic of interest remains the same for a few days, thus it can be said that the user's topic of interest will remain the same for the near future. In a web surfing scenario, a user will tend to search the same topic of interest. Thus, the content base filtering technique will only show those information related to a user's topic of interest and similar users with the same topic of interest. [16-18]. Figure 1 below refers to content base filtering.

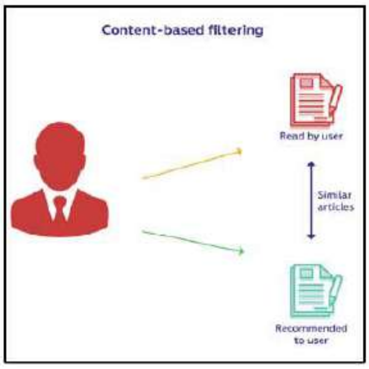


Fig. 1. The content base filtering technique

B. Collaborative filtering

This technique uses the idea of spreading news, word of mouth, people's option, and reviews [11, 15]. This filtering technique then filters and sorts based on similarity them accordingly to help users make decisions[13, 16]

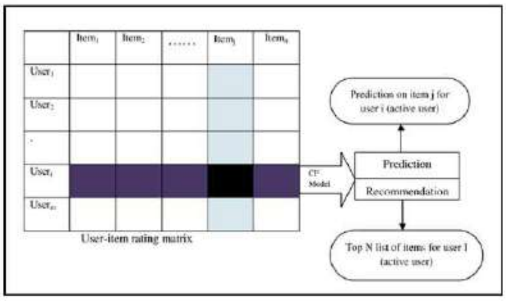


Fig. 2. Collaboarative filtering

C. Hybrid filtering

This technique (Figure 3) combines both filtering techniques mention in the above sub sections. The outcome of this technique is based on the users topic of interest combined with other information such as previous reviews, option and spreading news [17].

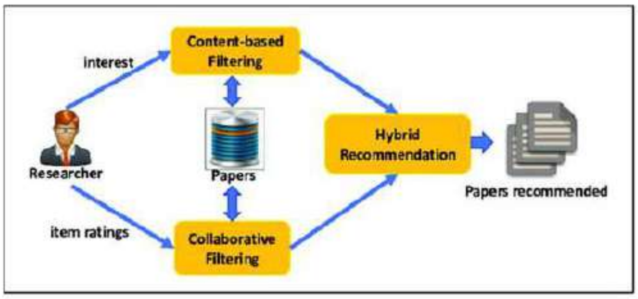


Fig. 3. Hybrid Filtering

IV. Methodology

The proof of concept of the proposed system in this project will be using the Multimedia Development Life Cycle (MDLC) methodology [19] (figure 4). A well-developed methodology plan will save time, money, and multiple modifications. The MDLC activities concentrate on technological aspects of the product development [20]. The MDLC consists of: (1) conceptualization phase, (2) development plan phase, (3) preproduction phase, (4) production phase, (5) postproduction phase and (6) documentation phase. Each of this phase will be briefly explained in the following sub sections.

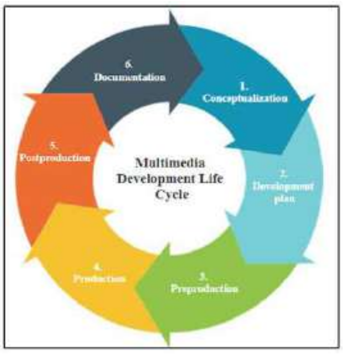


Fig. 4. Multimedia Development Life Cycle (MDLC)

A. Conceptualization

This phase constitutes of the categorization of domain stage. This project will be categorized into three main domain: (1) information – to be given to the users, (2) filtering – base on certain criteria, and (3) outcome – providing the best recommendation to the user.

B. Development

This phase constitutes the development of the system architecture for the mobile application. The system architecture for this system will be explained in the next section.

C. Preproduction

In this phase, the mobile application is coded. This application will be using several tools such as jquery. mobile-1.0 as the platform, Adobe Dreamweaver CS5, and XMAPP control panel v.3.1.0

D. Production

In this phase, all the functional requirements of the prototype mobile application will be analyzed and tested to obtain user feedback.

E. Postproduction

In this phase, further testing will be conducted, using scenario base testing and black box testing.

F. Documentations

In this phase, proper documentation and deployment of the mobile application will be conducted.

V. Proposed System Architecture

The system architecture (figure 5) is the conceptual model that describes a system's structure, behaviors, and more opinions [21]. In the Multimedia Development Life Cycle process, serves as the key milestone. The goal of the activities of system architecture is to define a comprehensive solution based on logically connected and compatible principles, concepts, and properties. Further, the system architecture has features and properties that satisfy, the problem or opportunity expressed by the system specifications and principles of the development life cycle which is then implemented by technology.

The activities presented in Figure 5 defines a thorough project outcome solution based on the logically associated and compatible standards of the proposed project. The outcome of this data plan recommendation system implementation comes from several data plans from any telecommunication company that users would choose to access the internet connection. This project would be using the website's triplet and metadata files as a database in the data plan recommendation system. This data plan recommendation system would be using the hybrid filtering technique. This proposed project implements hybrid filtering technique. This technique combines the user topic of interest and other similar reviews focusing on the same topic of interest. Furthermore, the hybrid technique also utilizes other users reviews to provide the best possible recommendation [17, 21].

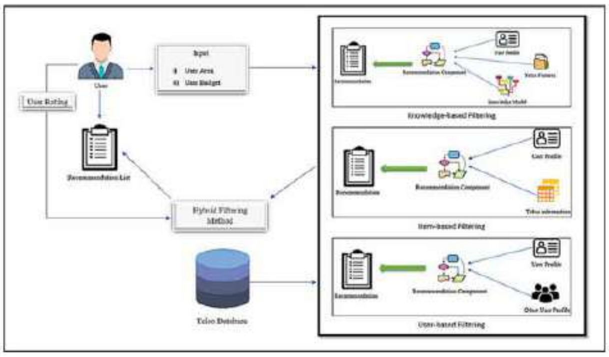


Fig. 5. System architecture

The system architecture will also help to prioritize conflicting goals. In the early stages of a research project system, the system architecture ensures that a design strategy can create an appropriate framework. Furthermore, the design risks and several mitigation plan can also be recognized early especially in the construction process by designing efficient architecture. Functional requirements, including software coding standards, instruments, and platforms, are also determine in the system architecture. Hence, it will ensure success in proposed project and it also gives the right technological solutions.

VI. Future Works

Numerous telco providers offer affordable and reasonable data plans which best suits students. However, for students the selection of a good data plan is vital, as students mostly uses extensive mobile data in their daily usage especially for their mobile learning.

The propose recommendation system opens new ways for customized knowledge and information to be retrieved from the Internet. It also aids to mitigate the issue of information overload, by using the recommendation systems. This is because users (i.e., students) are only interested in those information that are is necessary for their usage.

Further, this work in progress paper has highlighted several filtering techniques which can used by the recommendation system. The proposed mobile data plan recommendation system will be using the hybrid filtering technique, as this technique is a complete technique focusing on a user's interest and other users' reviews. It is hope that this project will inspire and serve as a guide to develop other recommendation system.

Acknowledgement

This work is supported by the Department of Computer Science, Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Perak Branch under the final year project (FYP) 2021.

REFERENCES

[1] Hyman, J.A., M.T. Moser, and L.N. Segala, *Electronic reading and digital library technologies: understanding learner expectation and usage intent for mobile learning.* Educational Technology Research and Development, 2014. 62(1): p. 35-52.

[2] Callo, E.C. and A.D. Yazon, *Exploring the factors influencing the readiness of faculty and students on online teaching and learning as an alternative delivery mode for the new normal.* Universal Journal of Educational Research, 2020. 8(8): p. 3509-3518.

[3] Bulut, E. and B.K. Szymanski. *Understanding user behavior via mobile data analysis*. in *2015 IEEE International Conference on Communication Workshop (ICCW)*. 2015. IEEE.

[4] Patil, R.N., et al., *Attitudes and perceptions of medical undergraduates towards mobile learning (M-learning).* Journal of clinical and diagnostic research: JCDR, 2016. 10(10): p. JC06.

[5] Al-Fahad, F.N., *Students' attitudes and perceptions towards the effectiveness of mobile learning in King Saud University, Saudi Arabia.* Online Submission, 2009. 8(2).

[6] Cavus, N. and D. Ibrahim, *m - Learning: An experiment in using SMS to support learning new English language words.* British journal of educational technology, 2009. 40(1): p. 78-91.

[7] Clarke, P., et al. *Using SMSs to engage students in language learning*. in *EdMedia+ Innovate Learning*. 2008. Association for the Advancement of Computing in Education (AACE).

[8] Venkatesh, B., et al., *Assessing Indian Student's Perceptions Towards M-Learning Some Intial Conclusions.* International Journal of Mobile Marketing, 2006. 1(2).

[9] McQueen, D., *The momentum behind LTE adoption [sGPP LTE].* IEEE Communications Magazine, 2009. 47(2): p. 44-45.

[10] Adebiyi, S.O., H.A. Shitta, and O.P. Olonade, *Determinants of customer preference and satisfaction with Nigerian mobile telecommunication services.* BVIMSR's Journal of Management Research, 2016. 8(1): p. 1.

[11] Ojukwu, N.V., *Determinants of Consumer's choice of telecommunication service provider*. 2017.

[12] Lai, Y.-T., et al. *Mobile data usage prediction system and method*. in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. 2017. IEEE.

[13] Xiao, A., et al. *An in-depth study of commercial MVNO: Measurement and optimization*. in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 2019.

[14] Bancu, C., et al. *ARSYS--Article Recommender System*. in *2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. 2012. IEEE.

[15] Thorat, P.B., R. Goudar, and S. Barve, *Survey on collaborative filtering, content-based filtering and hybrid recommendation system.* International Journal of Computer Applications, 2015. 110(4): p. 31-36.

[16] Felfernig, A., et al., *Basic approaches in recommendation systems*, in *Recommendation Systems in Software Engineering*. 2014, Springer. p. 15-37.

[17] Yang, S., et al., *Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach.* Knowledge-Based Systems, 2017. 136: p. 37-45.

[18] Maidel, V., et al., *Ontological content - based filtering for personalised newspapers: A method and its evaluation.* Online Information Review, 2010.

[19] Rahayu, S.L. and R. Dewi. *Educational Games as A learning media of Character Education by Using Multimedia Development Life Cycle (MDLC)*. in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*. 2018. IEEE.

[20] Aleem, S., L.F. Capretz, and F. Ahmed, *Game development software engineering process life cycle: a systematic review.* Journal of Software Engineering Research and Development, 2016. 4(1): p. 1-30.

[21] Booch, G., *Systems Architecture.* IEEE software, 2010. 27(4): p. 96-96.

Metal Oxide Surge Arrester: A Tool for Lightning Stroke Analysis

M. Shaban
Dept. of Electrical and Electronic Engineering
Loughborough University, UK

***Abstract*— Surge arrester is used to protect the equipment of electrical transmission and distribution systems from the effects of lightning and switching overvoltage. They are very much reliable devices which can work for decades without causing any problems provided they are properly designed, configured, and maintained. This paper examines the effects of lightning strokes in an electrical network and analyzes two different frequency dependent models of metal oxide surge arrester following the IEEE standards. The objective is to create a suitable model of tower which considers the propagation of surge, the footing resistance and the two types of air-gap leaders. It is found that, with the identification of overvoltage, the rate of failure of a substation could be determined. This is done through a comprehensive study that uses different strategies among the one available in literature. It was also found that all the strategies generated a similar transient overvoltage. It was observed that the range of values obtained could be considered as complying with the standards.**

Keywords: Lightening Stroke; Frequency Dependent Model; Electrical Tower; Surge Arresters; Transient Overvoltage

I. INTRODUCTION

In transient studies, the models are usually different from the classic models which are often used in power systems study. All lines, including the span of lines, towers, cables, and surge arresters, are expected to be frequency dependent (FD). A transformer is also expected to be a very peculiar model because in higher frequencies, the most important parameter is the surge capacitance of the primary and the secondary winding of the transformer. Similarly, all the equipment of a sub-station is expected to be modeled using stray capacitance [1]. The lines and spines in a sub-station are very short; therefore, to simulate this system, a precise step should be taken within a short time.

The equipment in the 420 kV system normally has a standard lightning impulse withstand level of 1425 kV. Hence, according to the IEC standards on insulation coordination [2, 3], the highest occurring voltage in the case of a non-self-restoring insulation in operation should stay below this value by a factor of 1.15, i.e., it should not exceed 1239 kV. Nevertheless, the lightning impulse protection level of 823 kV offers enough protection.

Several significant factors can cause the voltage at the terminals of the equipment to be protected to take on a considerably higher value. When inductive voltage drops; it discharges currents higher than the nominal discharge current and provides separation effects through traveling wave processes between the terminals of the arrester and the equipment to be protected. The latter phenomenon must be considered when planning the optimal location of an arrester [4]. It is, therefore, possible to increase the energy absorption capability of the arrester bank to the extreme values by connecting, normally, up to 100 metal oxide columns in parallel [5]. Despite, there are many other options for optimizing the necessary investment against the power supply quality by protecting only part of the towers, part of the phases, or by choosing less expensive arresters of low energy absorption capability based on an arrester failure risk analysis [6].

Lightning faults are of two types: the flashovers, mainly single-phase, following a screen failure and caused by direct hitting to the phase conductors and the back flashovers, which can occur when the lightning strike hits a tower or the ground wire. In this case, the potential at the top of the tower rises significantly and can exceed the dielectric strength of the insulators string [7-9].

Recently, the polymeric ZnO surge arresters have been developed and have been put into operations on transmission lines to limit the overvoltage based on their characteristics. A significant number of surge arresters of lines are in service today and there is a high demand for them at different voltage levels [10-13]. A study, aimed at improving the performance of transmission lines in the face of constraints, requires the modeling of the element of each line considering the effect of the frequency [14].

This paper uses the modified model of surge arresters recommended by the IEEE [15, 16] to examine the phenomenon and effect of lightening stroke on a designed surge arrester. This dynamic model is used to evaluate a lightning overvoltage protection. A lightning overvoltage is generated by using Cigre (International Council on Large Electric Systems) I-surge model on top of the tower. The resulting shape of the overvoltage and its magnitude are computed at two different points. The simulations are repeated with various parameters and with two types of surge arresters (FD model and ZnO model.

The simulations are performed with the popular transient software known as Alternative Transient Program (ATP), which itself is a revised version of the Electromagnetic Transient Program 3.1 (EMTP-RV).

II. SURGE ARRESTER MODELLING

Normally, a surge arrester is modelled using ZnO arrester which is in a nonlinear library of the ATP Draw. In its data function, the rating voltage of a surge arrester and the desired

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

voltage rating are computed together. A surge arrester cannot be simply considered as a nonlinear resistor as it is in the EMTP library. Rather, it must be modified as described in the IEEE working group 3.4.11 [17]. After the modification, it becomes a frequency dependent (FD) model of a surge arrester and gives much better response for lightning transients as compared to the simple model present in the library of the ATP Draw. The frequency dependency is built by the RL and the RLC filters as shown in Fig. 1.

In this case, two surge arresters are used: one at the primary level of a transformer and the other at the secondary of a transformer. Table 1 shows the characteristics of the surge arresters. A surge arrester should be modelled in line with these characteristics. Most of the times, a manufacturer might provide these standard values.

The values of *R0, L0, C0, R1,* and *L1* are given in the IEEE working group. The formulae for calculating these values are given below:

- *L1 = 15 d/n uH*
- *R1 = 65 d/n ohms*
- *L0 = 0.2 d/n uH*
- *R0 = 100 d/n ohms*
- *C = 100 n/d pF*

Where "*d*" is the estimated height of the arrester (in meters), which is 0.56 m. "*n*" is the number of columns of the arresters in parallel which is only one in this case. The circuit shown in Fig. 1 is converted into a sub-circuit, which is shown in the form of horizontal lines in Fig. 2. All the values of the parameters are defined in the mask.

I-surge is taken from the library of sources and connected with the arrester to verify it. The scope is used to measure the voltage. In the I-surge source itself, we can measure the voltage and waveform options of the current. After running the simulation, the current output is generated using the scope view command in the software.

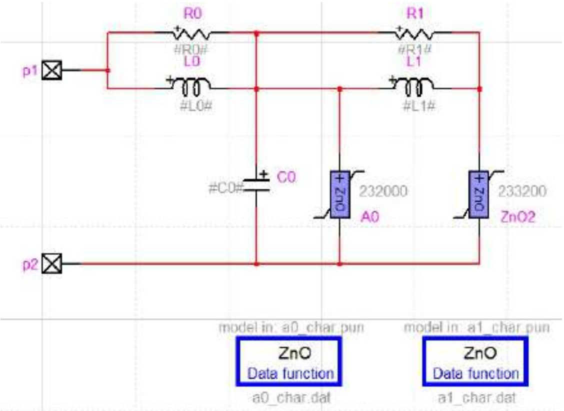


Fig. 1. HV Surge arrester.

TABLE I. SURGE ARRESTER CHARACTERISTICS.

	High Voltage Side	Medium Voltage Side
Rated Voltage	36 kV	12 kV
Continuous operating Voltage	30.04 kV	10.02 kV
Nominal Discharge Current	8/20 us, 10 kA	8/20 us, 10 kA
Residual Voltage at Nominal Current	116.6 kV	40 kV

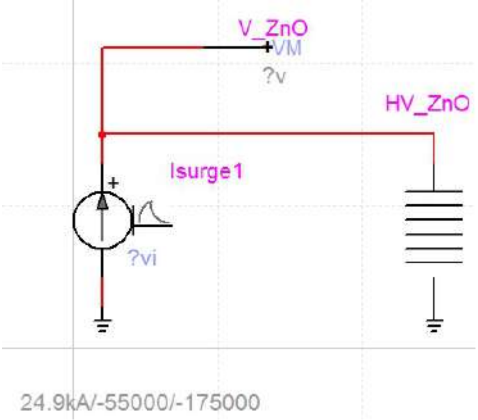


Fig. 2. HV Arrester energization.

It is observed that the pulse is same as defined by the standard shown in Fig. 3 and the voltage generated because of this current and the current itself are compared in Fig. 4. The maximum voltage is identified using the delta cursor command in scope view and calculating the residual voltage at the nominal current, which is the same as described in Table I by the standard.

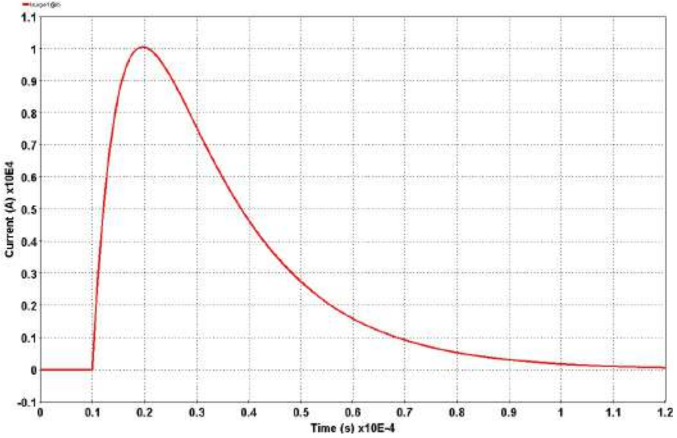


Fig. 3. Impulse defined by standard.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

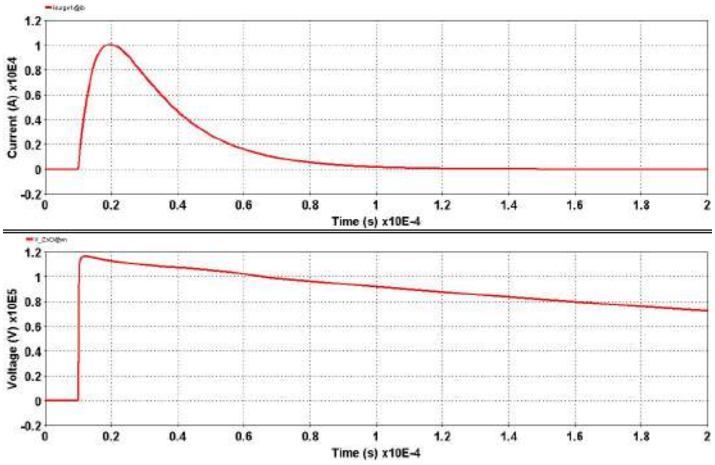
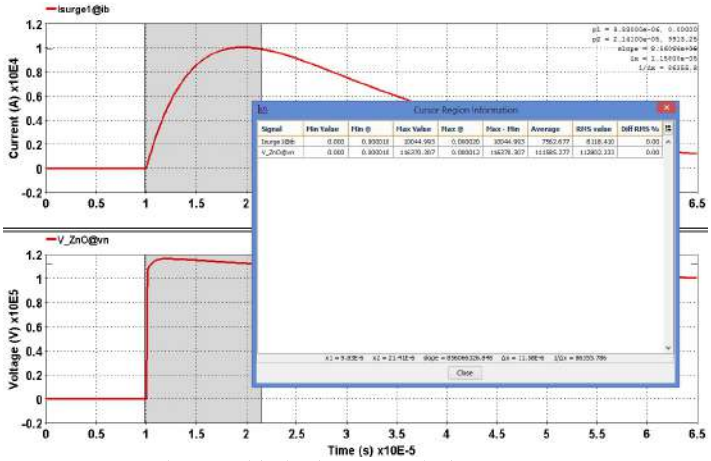


Fig. 4. Impulse and voltage at source.



Signal	Min Value	Min @	Max Value	Max @	Max - Min	Average	RMS value	Diff RMS %
Isurge1@ib	0.000	0.000018	10044.993	0.000020	10044.993	7562.677	8118.430	0.00
V_ZnO@vn	0.000	0.000018	116270.307	0.000012	116278.307	111585.277	112802.333	0.00

Fig. 5. Residual voltage at nominal current.

Fig. 5 shows the maximum value of the residual voltage at the nominal current on the high voltage side of arrester. This maximum voltage is in line with the acceptable standard. Therefore, the desired surge arrester that can be used in the network for the study of lightning is successfully modelled.

III. THE ELECTRICAL TOWER

When lightning strikes on the tower, there will be surge inside the tower with a reflection. Therefore, it is important to model a tower which considers reflection, especially, towers that are taller than 50 m. Towers are generally considered as an ideal single conductor-distributed parameter (CP model in EMTP-RV) [18]. However, more complex models are needed for consideration when towers are taller than 50 m. The surge impedance for the proposed conductor can be calculated using equations (1) and (2).

$$Z_{surge} = 60 \ln\left(\cot\left(\frac{1}{2}\tan^{-1}\left(\frac{R_{avg}}{h_1 + h_2}\right)\right)\right) \tag{1}$$

While

$$R_{avg} = \frac{r_1 \times h_2 + r_2\left(h_1 + h_2\right) + r_3 \times h_1}{h_1 + h_2} \tag{2}$$

R_1 = Tower top radius (m),

R_2 = Tower midsection radius (m),

R_3 = Tower base radius (m),

h_1 = Height from base to midsection (m),

h_2 = Height from midsection to top (m).

The propagation velocity in the towers is estimated at 80 % of the speed of light [19]. The CP line can represent each section of the tower. For the branch, only inductance can be used if they are short in length. The CP line can be used as well; however, it depends on the length of the branch. Fig. 6 shows how the case is going to be as the first step for an electrical network.

For the length of 20 m, 3 m, 3 m, and 4.65 m, CP lines are used for each section of the tower while inductors are used for the branches. The air-gap leader for supporting the phasor and the footing resistance could also be used. As for the resistance, because of the ionization of the soil, the footing resistance changes with the value of the current at a certain point. Thus, in low currents, a constant footing resistance is used and when the current becomes higher, the resistance changes according to the equation (3).

For footing resistance, the control resistance in nonlinear library is used. In this controlled resistance, we input the values of admittance and the current. There are two types of resistances possible here, either the constant or one which depends on the current, thus, at this point, the input selector is used. If the input selector pin is 1, then the resistance will be constant, and the second input is the function of the resistance which depends on the current. This is how the footing resistance is built.

$$R_i = \frac{R_o}{\sqrt{1 + \left(\frac{I}{I_g}\right)}} \tag{3}$$

R_o – footing resistance at low current and low frequency, i.e., 50 or 60 Hz [Ω].

I – Stroke current through the resistance [kA].

$I_g = \dfrac{\rho.E_0}{2.\pi.R_0{}^2}$ - limiting current to initiate sufficient soil ionization [kA].

ρ - Soil resistivity [Ω m].

E_o – is the soil ionization gradient, recommended value: 400 [kV/m].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

After building the footing resistance, the air-gap leader is to be considered, for which there are two different models that can be used.

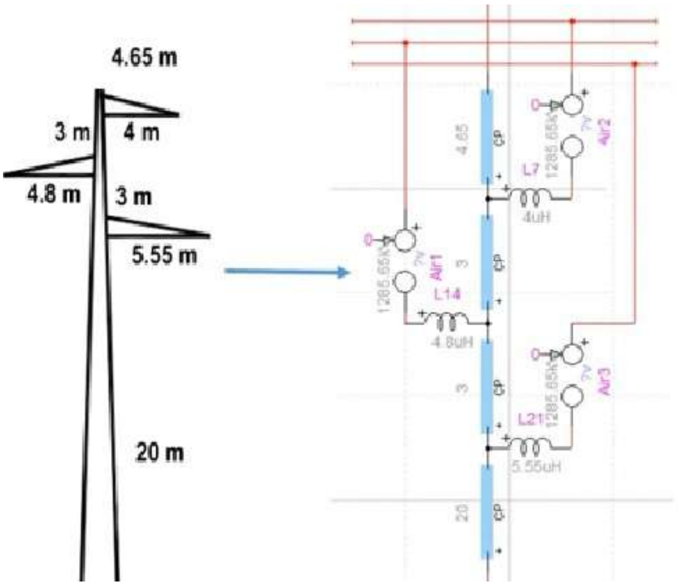


Fig. 6. Electrical tower in EMTP.

A. Disruptive Effect Model

The first one is called disruptive effect model [20]. Flashover occurs when the equation (4) becomes true.

$$\int_{t_o}^{t}\left(\left|V_{gap}(t)\right|-V_0\right)^{k} dt \geq D \qquad (4)$$

-t_o is the time point at which V_{gap} becomes greater than V_o. When the voltage V_{gap} goes below V_o the integral is reset.

The gap is an ideal open switch before the flashover, and it becomes an ideal closed switch after the flashover. The gap remains closed after the flashover until the control signal becomes greater than zero, in which case it will reset (open) the gap. Fig. 7 shows the disruptive effect model along with the ground resistance.

The disruptive effect model with the same equation can be found in the switch library with the name, airgap. The same airgap is used for all three phases. The purpose of the input to the airgap is to reset it after the flashover so we insert zero as an input. This means that it will never reset itself. After this, the stray capacitance between the tower and the phase is used because every stray capacitance has an influence when dealing with high frequencies. The typical value for the capacitor is 80 *pF*. After making all the connections, a sub-circuit was produced creating the tower in EMTP-RV.

B. Air Gap Leader

This is even more advanced model that is why the simulation time will be higher and the simulation will be slow when this type of air-gap is used [21] and it is shown in Fig. 8. This model is not in the library but exists in the examples with the name airgap leader, which is a valid example of this airgap. The same tower is used but only the replacement of the disruptive model with this air-gap model is done as shown in Fig. 9.

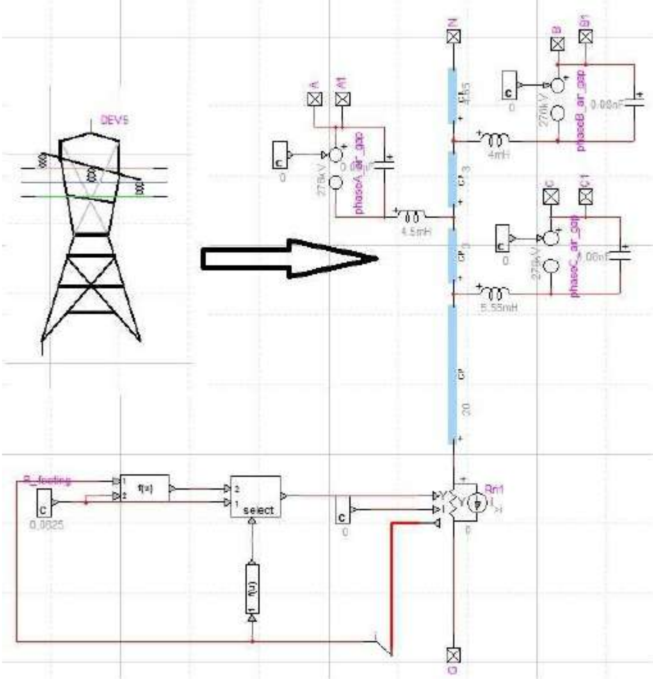


Fig. 7. Disruptive effect model along with ground resistance.

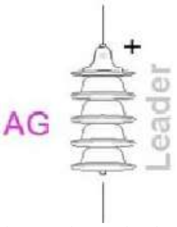


Fig. 8. Air gap leader.

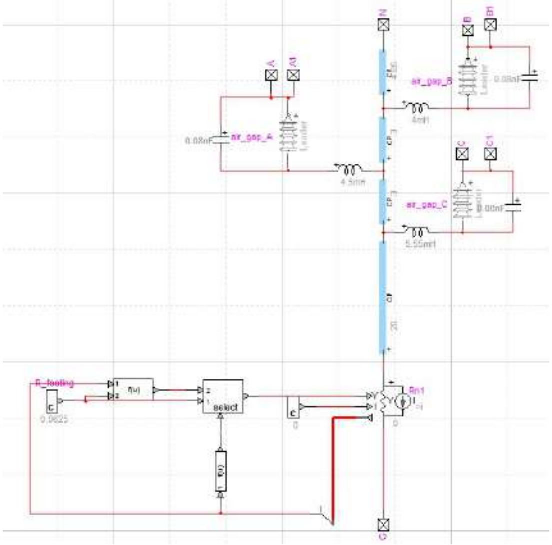


Fig. 9. Airgap model along with ground resistance.

IV. THE ELECTRICAL NETWORK

In this section, each span of the tower where the lightning strike can occur is modelled. The FD model of the line is used and will copy the tower that was built using the first airgap i.e., the disruptive effect model. The line data case is used to enter the date for FD lines. Four conductors (three phases and one neutral) were used with the 300 m length for each. The lightning will strike on the tower and there is going to be a back flashover, so tower should be modelled in a way such that there is no back flashover after the last tower modelled. After the lightning strikes, there will be a reflection between towers. Something should be done after the last tower so that there will be no reflection. Therefore, to get rid of the undesirable reflection, a very long line of 100 km was placed which is presumably transposed. There will then be a voltage source of 64 kV connected using BUS.

To model a substation, every element should be modelled e.g.
Circuit Breakers (CB): 2*50 *pF*
Capacitive voltage transformers (CVT): ~4400 *pF*
Current transformers (CT): 200-800 *pF*
Power transformer: 1-6 *nF*

Fig. 10 shows the detailed circuit diagram drawn in EMTP-RV using cables, lines, and a transformer. Many examples of the typical capacitance values for a lot of equipment are explained in Annex B IEEE C37.0111-2011. However, the preferred data is the one provided by the manufacturer especially, for the transformer. When one uses a short cable to connect the tower with a substation, it can be modelled with RL. One can then connect the surge arrester with all three phases and these surge arresters can be grounded using another small cable.

For the cables going through a substation, the CP model must be used for the cable lengths which are more than 50 m which will reproduce the propagation. Therefore, there is no need to use the FD model for short cable lines because it can create some convergence problems. When the CP model for a 50 m cable was used that connects the phases with a CVT and model CVT with a capacitor with a value of 4400 *pF*, the transformer used is three phase YgD+30º. However, in fact, the transformer is not valid for a high frequency because it has a lot of high value inductance; and in high frequencies, the inductances are like open circuit.

Here, the stray capacitance becomes important, and the values of the stray capacitance were taken from [22] which are 5 *nF*, 15 *nF*, and 20 *nF*. For the lightning, the Cigre current source was used from the sources library and was connected to one of the towers. In the data case of the current source, the data like starting and ending time of the lightning strike was inserted. The subsequent strikes could also be observed by adding current sources in parallel and by changing the starting time.

After running the simulation, the observed parameters were the voltage at the top of the tower, current through the airgap, voltage within the substation, and the energy through the surge arrester. For the simulation, the time-step was small for the lightning, i.e., 0.005 *us*; and the simulation time was 300 *us*. The simulation was very slow, and, during the event, energy was plotted through a deterministic approach.

V. RESULTS AND DISCUSSION

While considering the weather forecast, one can determine the worst lightning strike that could occur. When the overvoltage is identified, the probability of the equipment failing, considering this voltage, can be calculated. This facilitates the determination of the rate of failure of the substation. Fig. 11 shows the waveform of the energy through the surge arresters.

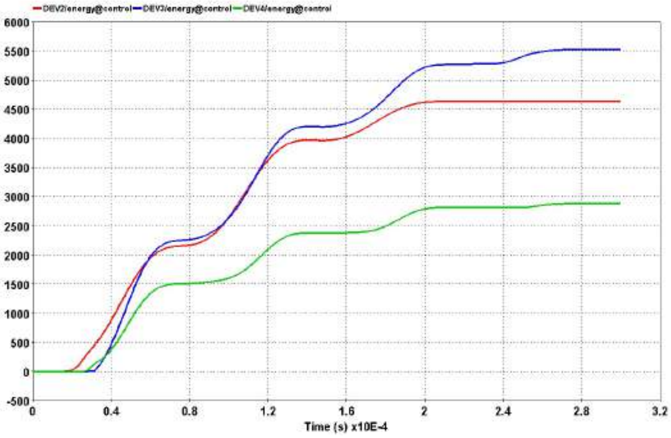


Fig. 11. Energy through surge arresters.

Similarly, the voltage at the top of the tower and on the substation is observed in Fig. 12. It was seen that the voltage was smooth at the substation because of the surge arrester.

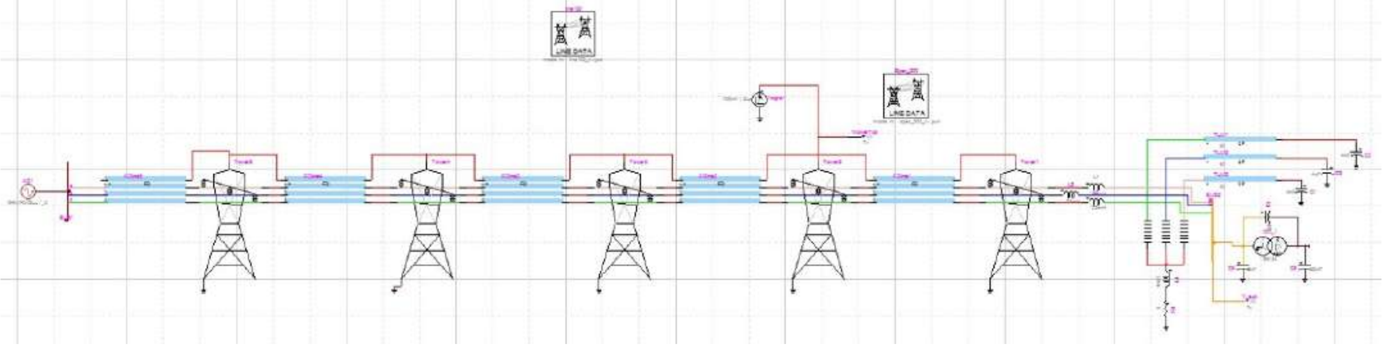


Fig. 10. Electrical network using cables, lines, and towers.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

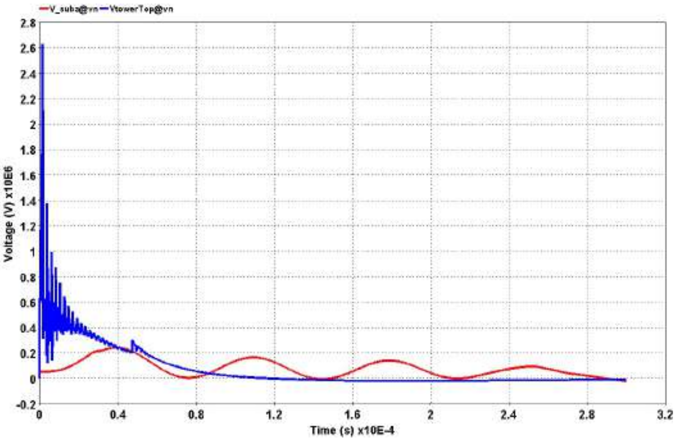


Fig. 12. Voltage at tower top and at substation.

VI. CONCLUSION

This paper modelled the surge arrester according to the IEEE standard to investigate the lightening stroke effect on different surge arresters and is suitable for all types of overvoltage studies. The calculation of energy during the event is a deterministic approach and can be used for weather forecast. Therefore, the worst lightning strike can be determined in advance. It is concluded that with the identification of overvoltage, the rate of failure of a substation could be determined.

Nonetheless, the obtained results indicate that the lightning current (amplitude) and the tower footing resistance must be considered in the study of lightning protection. A major part of the surge current was dissipated through the ground. This study shows that several alternatives that can improve line performances and ensure compromise between these alternatives must be considered. It is recommended that the secondary protections should also be made as the improvement of the transmission line performances alone does not ensure complete protection of a power system.

REFERENCES

[1] IEC 60099-4, Edition 1.1, “Surge arresters – Part 4: Metal-Oxide Surge Arresters without gaps for A.C. Systems”, pp. 19-23, 1998.

[2] IEC 60071-1, Seventh edition, “Insulation Co-ordination – Part 1: Definitions, Principles and Rules”, pp. 09-14, 1993.

[3] IEC 60071-2, Third edition, “Insulation Co-ordination – Part 2: Application Guide”, pp. 28-32, 1996.

[4] IEC 60099-5, First edition, “Surge Arresters – Part 5: Selection and Application Recommendations”, pp. 01-10, 1996.

[5] Mainville J., Rofffon P., Rillin L.P. and Hinrichsen V., “Pressure Relief Tests on Varistors for the Series Compensation Banks Installed at the Montagnais Substation”, IEEE Transactions on Power Delivery, Vol. 9, No. 2, pp.781-787, 1994.

[6] Tarasiewicz Eva J., Finn Rimmer and Atef S. Morched, “Transmission Line Arrester Energy, Cost, and Risk of Failure Analysis for Partially Shielded Transmission Lines” IEEE Transactions on Power Delivery, Vol. 15, No. 3, pp. 919-924, 2000.

[7] J. G. Anderson, “Lightning Performance of Transmissions Lines”, Transmission line reference book 345 kV and above, pp 545-578, 1982.

[8] CIGRE 33.01, “Guide to Procedures for Estimating the Lightning Performance of Transmission Lines”, October 1991.

[9] IEEE Working Group on LPTL,“Estimating Lightning Performance of Transmission Lines II – Updates to Analytical Models”, IEEE Transactions Power Delivery, Vol. 8, No. 3, July 1993.

[10] Gatta F. M., F. Iliceto and S. Lauria, “Lightning Performance of HV Transmission Lines With Grounded or Insulated Shield Wires”, Proceedings of 26th International Conference on Lightning Protection (ICLP 2002), Cracow, Poland, pp. 475-480. 2002.

[11] Tarchini J. A. and W. Gimenez, “Line Surge Arrester Selection to Improve Lightning Performance of Transmission Lines”, IEEE Conference Proceedings on Power Technology, Vol. 2, pp. 6-12, 2003.

[12] Dudurych I. M., T. J. Gallagher, J. Corbett, and M. Val Escudero, “EMTP Analysis of the Lightning Performance of a HV Transmission Line”, IEEE Proceedings on Generation, Transmission and Distribution, Vol. 150, No. 4, pp. 501-506, 2003.

[13] Ito Takamitsu, Toshiaki Ueda, Hideto Watanabe, Toshihisa Funabashi and Akihiro Ametani, “Lightning Flashovers on 77-Kv Systems: Observed Voltage Bias Effects and Analysis”, IEEE Transactions on Power Delivery, Vol. 18, No. 2, pp. 545-550, 2003.

[14] Imece Ali F., Daniel W. Durbak and Hamid Elahi, “Modeling Guidelines for Fast Front Transients”, IEEE Transactions on Power Delivery, Vol. 11, No. CONF-950103, 1996.

[15] A. Bayadi, K. Zehar, S. Semcheddine and R. Kadri, “A Parameter Identification Technique for a Metal-Oxide Surge Arrester Model based on Genetic Algorithm”, WSEAS transactions on Circuits and Systems, Issue 4, Vol 5, pp 549-554, April 2006.

[16] A. Bayadi, “Parameter Identification of ZnO Surge Arrester Models Based on Genetic Algorithms”, Electrical Power Systems, Vol 78, No. 7, pp 1204-1209, July 2008.

[17] IEEE Working Group 3.4.11, “Modeling of Metal Oxide Surge Arresters”, IEEE Transactions on Power Delivery, Vol. 7, No. 1, pp. 302-309, Jan. 1992.

[18] Chisholm W.A., Chow Y. L. and Srivastava K.D., “Lightning Surge Resonse of Transmission Towers”, IEEE Transactions on Power Apparatus and Systems, Vol. 9, pp. 3232-3242, Sept. 1983.

[19] Uglesic Ivo, “Modelling of Transmission Line and Substation for Insulation Coordination Studies” In 3 days training: “Simulation & Analysis of Power System Transients with EMTP-RV” in cooperation with Graz University of Technology, University of Zagreb and University of Sarajevo, Dubrovnik, Hrvatska, 27-29/04/2009.

[20] Rioual Michel, “Measurements and Computer Simulation Of Fast Transients through Indoor and Outdoor Substations”, IEEE Transactions on Power Delivery, Vol. 5, No. 1, pp. 117-123, Jan. 1990.

[21] Shindo Takatoshi, Megumu Miki, Yoshinori Aihara and Atsushi Wada, “Laser-Guided Discharges in Long Gaps”, IEEE Transactions on Power Delivery, Vol. 8, No. 4, pp. 2016-2022, Oct. 1993.

[22] Das J. C., “Transients in Electrical Systems: Analysis, Recognition, and Mitigation”, McGraw Hill Professional, pp. 384-389, 2010.

A Novel Machine Learning Based Screening Method For High-Risk Covid-19 Patients Based On Simple Blood Exams

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Mohit Gupta
Student – AIML
Great Learning
Bengaluru, India
mohit19mahajan@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Mumbai, India
anwesh@greatlearning.in

Richa Agrawal
Student – AIML
Great Learning
Bengaluru, India
epost.richa@gmail.com

Sachin Padasali
Student – AIML
Great Learning
Bengaluru, India
smpadasali@gmail.com

Arti Kumari
Student – AIML
Great Learning
Bengaluru, India
aartikumariarwal@gmail.com

Prabu Purushothaman
Student – AIML
Great Learning
Bengaluru, India
crprabu@gmail.com

***Abstract* - This paper presents a predictive model to potentially identify high-risk COVID-19 infected patients based on easily analyzed circulatory blood markers. These findings can enable effective and efficient care programs for high-risk patients and periodic monitoring for the low-risk ones, thereby easing the hospital flow of patients and can further be utilized for hospital bed utilization assessment. The present machine learning-based SV-LAR model results in a high 87% f1 score, harmonic mean of 91% precision, and 83% recall to classify COVID-19, infected patients, as high-risk patients needing hospitalization.**

Keywords – COVID-19, SARS-CoV-2, pandemic, patient outcomes, machine learning models

I. INTRODUCTION

COVID-19 is a new disease for which effective treatment is still awaited. It was declared a pandemic by World Health Organization (WHO) on March 11, 2020 [1]. As of January 28, 2021, more than 100 million people have been affected by this infection causing more than 2 million fatalities [2]. Global health care now faces unprecedented challenges with the widespread and rapid human-to-human transmission of SARS-CoV-2 and high morbidity and mortality with COVID-19 worldwide [3].

COVID-19 patients get worse quickly and aggressively.
In addition to high transmissibility SARS-CoV-2 infection it is also characterized by fever, dry cough, weakness, headache, dyspnoea, and loss of smell and taste in the early stages, which are common symptom of cold and flu [4]. The early onset of common symptoms can rapidly change to acute respiratory distress syndrome (ARDS), acute cardiac injury, cytokine storm, coagulation dysfunction, and multi-organ failure if the disease is not resolved, resulting in patient death [5]. Early studies showed that COVID-19 patients with comorbidity may lead to poor prognosis, increasing the risk of severe illness from COVID-19. Among laboratory-confirmed cases, patients with any comorbidity yielded poorer clinical outcomes than those without [6]. Several studies have been conducted to find a correlation between pre-existing medical conditions and their impact on COVID-19 prognosis.

In a meta-analysis by Wang et al, Hypertension, diabetes, Chronic obstructive pulmonary disease (COPD), cardiovascular disease, and cerebrovascular disease were found to be the major risk factors for patients with COVID-19 [7]. Several risk factors that led to the progression of COVID-19 pneumonia were identified, including age, history of smoking, maximum body temperature at admission, respiratory failure, albumin, and C-reactive protein [8]. Given the virtually unstoppable global trend of SARS-CoV-2, together with the high prevalence of comorbidities worldwide, the combination of these two conditions poses greater clinical, societal, and economic burdens to healthcare systems [9].

Until now the source of the pathogenesis of the COVID-19 remains unclear, and no specific treatment has been recommended for coronavirus infection except for meticulous care. The world is ready to receive the vaccines as approved worldwide, but the threat continues with mutating strains of the virus. Therefore, the need for a better solution for providing care to those who absolutely need it and to predict the future requirements for better planning and management for better patient outcomes, continues. In

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

several articles, researches have indicated the need for better hospital management by early identification of patients requiring hospitalization and possible further triage [10].

In another attempt to decode the comorbidity-related risks in COVID-19 patients, Zhao et al. developed a logistic regression-based classification model to predict two primary outcomes of admission to the intensive care unit (ICU) and death. The risk score model yielded accuracy with an Area under the curve (AUC) of 0.74 ([95% CI, 0.63–0.85], p = 0.001) for predicting ICU admission and 0.83 ([95% CI, 0.73–0.92], p<0.001) for predicting mortality for the testing dataset. This model was developed and internally validated using data from the COVID-19 persons under investigation (PUI) registry of 4997 patients from a major academic hospital (Stony Brook University Hospital) [11] in New York. Another finding was that the mortality group uniquely contained cardiopulmonary parameters as top predictors.

In another study aimed to clarify high-risk factors for COVID-19, researchers used Multivariate Cox regression to identify risk factors associated with the progression of the disease. Univariate and multivariate analyses showed that comorbidity, older age, lower lymphocyte count, and higher lactate dehydrogenase at presentation were independent high-risk factors for progression. A novel scoring model, named CALL [12], with an area under the receiver operating characteristic curve (ROC) of 0.91 (95% CI, .86–.94) was established to help clinicians better choose a therapeutic strategy.

Elisa Grifoni et al, tested the predictive power of the CALL score in an Italian COVID-19 population admitted to hospital from 12 March to 20 April 2020 and consisting of 210 patients. Their findings concluded that the CALL score is a good prognosticator for in-hospital mortality but not for progression to severe COVID-19 in their settings [13].

In another statistical analysis regarding the associations between increased cardiac injury markers and the risk of 28-day-all-cause death of COVID-19 patients in the Chinese population, the 5 myocardial biomarkers (high-sensitivity cardiac troponin I, creatine phosphokinase)-MB, N-terminal pro-B-type natriuretic peptide, creatine phosphokinase, and myoglobin) were found to be significantly prognostic of COVID-19 mortality [4].

Baseline patient characteristics, laboratory markers, and chest radiography can predict short-term critical illness in hospitalized patients with COVID-19, with an internally validated AUC = 0.77 using a logistic regression-based risk model developed by Steven Schalekamp, et al [14]. In another study, Zhou et al. validated a nomogram including 6 predictors: age, respiratory rate, systolic blood pressure, smoking status, fever, and chronic kidney disease. The model demonstrated a high discriminative ability in the training cohort (C-index = 0.829), which was confirmed in the external validation cohort (C-index = 0.776). In addition, the calibration plots confirmed good concordance for predicting the risk of ICU admission [15].

In another study, a regression analysis showed that C-reactive protein (CRP) was significantly associated with aggravation of non-severe COVID-19 patients, with an area under the curve of 0.844 (95% confidence interval, 0.761–0.926) and an optimal threshold value of 26.9 mg/L [16]. In a Spanish study, COVID-19 patients with normal levels of lymphocytes or mild lymphopenia, imbalanced lymphocyte subpopulations were early markers of in-hospital mortality [26].

Despite several initiatives aimed at containing the spread of the disease, countries are faced with unmanageable increases in the demand for hospitalization and ICU beds [18]. The health care system globally, has been stressed and stretched to its limit. In order to help in patient triage, several attempts have been made to discover early predictors of COVID-19 disease progression and spread. Identification of such factors that predict complications of COVID-19 is pivotal for guiding clinical care, improving patient outcomes, and allocating scarce resources effectively in a pandemic. Medical resource allocation assessments should be based on a risk/benefit approach considering the intensity of transmission, the health system's capacity to respond, other contextual considerations (such as upcoming events which may alter transmission or capacity) and the overall strategic approach to responding to COVID-19 [25] in each specific setting.

We think that it requires agile decision-making based on ongoing situational assessments at the most local administrative level possible. We propose a predictive machine learning model that identifies a potential high-risk patient from the COVID-19 patient population based on blood based circulatory markers. These predictions would help the administrators to make provisions for the scarce 'hospital beds. Consequently, the model can help in providing better public health and social measures to alleviate patient care during the pandemic time thereby improving patient outcomes at large.

II. MATERIALS AND METHODS

A. The Dataset

We have used data published on a public forum , that of Hospital Israelita Albert Einstein, at São Paulo, Brazil [19]. The dataset contains records of patients that were tested for COVID-19 using SARS-CoV-2 Reverse transcription polymerase chain reaction (RT-PCR) and additional blood tests between the 28th of March 2020 and 3rd of April 2020. All data were anonymized following the best international practices and recommendations. The full dataset released included 5,644 individual patients' clinical test results that were standardized to have a mean of zero and a unit standard deviation. It provided information of patient hospitalization into three types of wards in the hospital, such as regular ward, semi-intensive care unit, and intensive care unit as depicted in Fig 1.

The information of patients admission to various wards in the hospital was used to create the target variable for the current problem statement. Hospitalization is needed by patients needing extra care and monitoring due to health

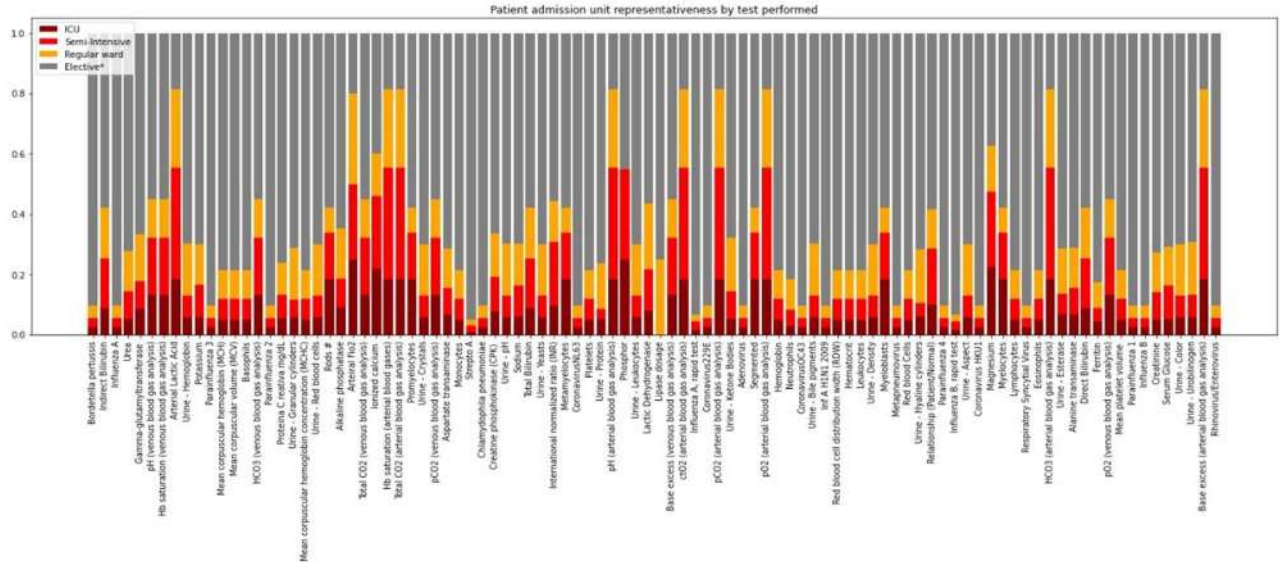


Fig.1. Distribution of patient admission into the three hospital wards across various tests performed

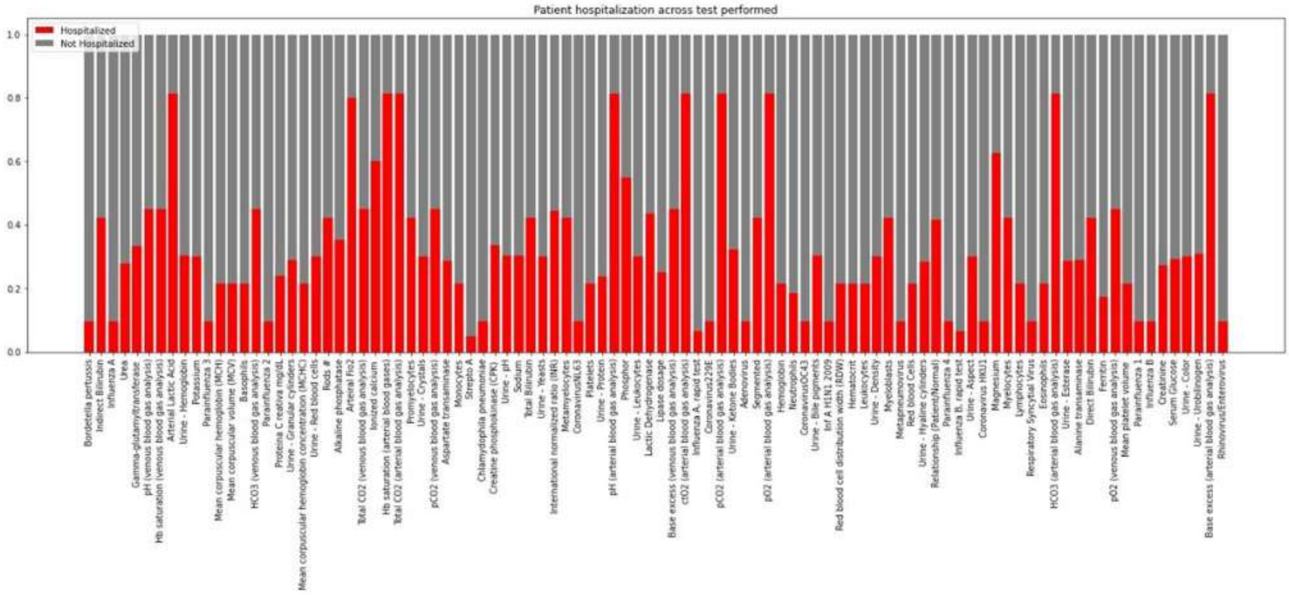


Fig.2. Distribution of hospital admission across various tests performed

who despite being infected do not need hospital admission constitute the low-risk patient population. This formed the basis of the binary classification and the target label for classification of patients {needing hospitalization in any of the hospital wards = 1, no hospitalization needed = 0} for the current objective. The distribution as per figure 1 above was therefore transformed to look like Fig 2 below.

As the current hypothesis is set around blood analysis, we have carefully selected features of routine blood analysis only. Parameter related to patient age was not considered in order to avoid any age related bias in the present analysis. Tests pertaining to viral or bacterial infections other than SARS-CoV-2 were also dropped. It is our objective to find blood based markers to identify high-risk patients and therefore features related to routine urine analysis were also dropped from the current model. Blood gas analysis either on venous blood or arterial blood is also not included in the current analysis. Largely because the blood samples are required to be tested in a 30 minute window or need a cold supply chain [19]. It is our intent to find markers that eases the hospital workload during the pandemic and therefore it is counterintuitive to include tests that need immediate attention and hospital setting to give good results.

It is for this reason that the working dataset for the model building exercise includes test parameters form a simple blood workup, keeping in mind that the sample could be collected form patients’ home environment and not necessarily in the hospital setting. Fig 3 presents the frequency of each test performed amongst the blood analysis related parameters considered for the model building, in the select dataset of 558 patient records tested positive for SARS-CoV-2.

For the purposes of our research we extracted records of patients that were tested positive for the RT-PCR test for the ongoing COVID-19 infection. Our working dataset consisted 558 records of patients infected with COVID-19 from the whole of 5644 patient records. The target variable is the hospital admission class amongst the extracted records. Following our research question we formulated the hypothesis that we were to test -

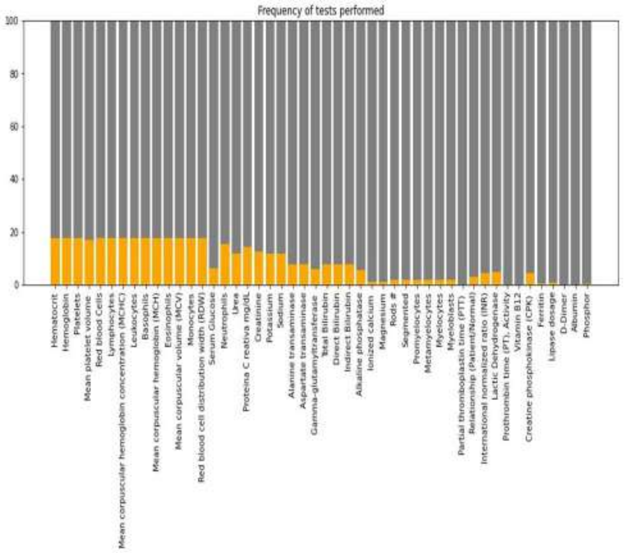


Fig.3.Percent blood analysis done on Covid positive patients

H0: There is no correlation between blood analysis of a COVID positive patient and his/her hospitalization

H1: There is a correlation between blood analysis of a COVID positive patient and his/her hospitalization

B. Model building

We started model building after data processing. The working data had missing values and columns with all null values. Features with more than 95% missing values were dropped and remaining missing values were imputed with the mean. We started with simple linear algorithms such as logistic regression, ridge classification and elastic net, and moved on to non-parametric algorithm such as Nearest Neighbours Classifier and Gaussian Naive Bayes classification. We also used tree based algorithms like decision tree classifier and extra tree classifier. Multiple ensemble techniques like random forest classifier, bagging classifier, adaboost classifier were also used to model the target label with select features of the prepared dataset. We used scikit learn library of machine learning algorithms [20]. Based on our findings, we propose the SV-LAR model for our 2-class (SARS-CoV-2 positive induced hospitalization or not) classification. The proposed model uses voting classifier ensemble based on logistic regression, random forest and adaboost classifier. The working dataset also has class imbalance. Only 10 % labels of the working dataset are positive class of hospitalised COVID positive patients, in the three available hospital wards. We have used SMOTE (Synthetic Minority Oversampling Technique) on the training data to deal with the class imbalance by upsampling the positive class [21].

C. Model performance measures

The performance of the model is expressed in terms F1 score, precision and recall. As we attempt imbalanced classification problem F1 score metric becomes more relevant. It is a measurement that considers both precision and recall to compute the score that can be interpreted as a weighted average of the precision and recall values. High F1 score (closer to 1.0) is desirable in our model. Precision is determined by the number of correctly labeled annotations divided by the total number of annotations added by the machine-learning annotator. It indicates how accurately the model has labelled the two classes. Another metric, recall specifies how many mentions that should have been annotated by a given label were actually annotated with that label. A recall score of 1.0 means that every mention that should have been labeled as entity type A was labeled correctly. In this imbalanced healthcare dataset high scores for precision and recall are desirable for an ultimate high f1 score [23, 24].

III RESULTS AND DISCUSSION

A. Evaluation

The proposed SV-LAR model of the COVID-19 infected patients produces a classification f1-score of 87%. With precision at 91% and recall at 83%. The confusion matrix is presented in the figure below.

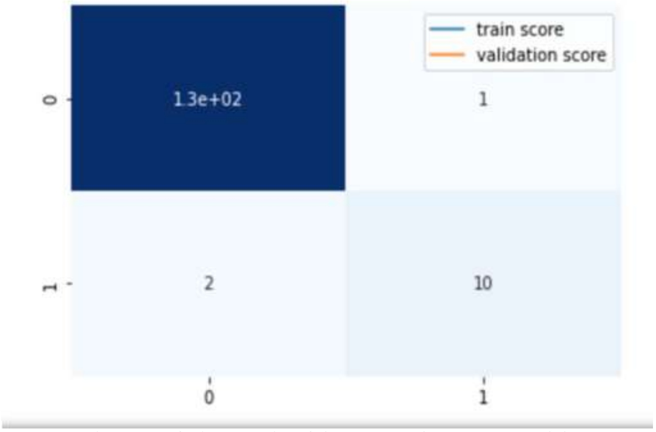


Fig.4. Confusion matrix of the proposed SV-LAR model

To the best of the authors knowledge, it is the first study to report a predictive machine learning model with high precision and recall for the triage of high-risk COVID-19 patients using simple blood exams. As per our findings, it would be possible to identify high-risk COVID-19 patients with more than 83% sensitivity and 91% specificity based on blood analysis alone. While the majority patients are asymptomatic, about 10% patients need hospitalisation. Our solution can enable healthcare professional, segregate potential high-risk patient in need of high degree of hospital care based on simple and cheap blood analysis and monitor them closely.

TABLE 1 CONFUSION MATRIX OF SV-LAR MODEL

	precision	recall	f1-score	support
0	0.98	0.99	0.99	128
1	0.91	0.83	0.87	12
accuracy			0.98	140
macro avg	0.95	0.91	0.93	140
weighted avg	0.98	0.98	0.98	140

If RT-PCR and blood analysis samples were to be taken from residence (as this service is available in India, and possibly in other countries as well), many patients can save a trip to the hospital emergency rooms thus saving time for both the hospital staff and themselves. This model can surely help in managing the pandemic related patient flow effectively and efficiently. Also, this will extend care 'as needed' by each patients' condition, where a high-risk patient is not sent home and a 'not at risk' patient is managed quickly without risking unnecessary hospital visit. Thus our model ensures using available resource where needed and thereby improves patient outcomes and hospital's healthcare burden as well.

B. Proposed solution for patient management

Once our model is deployed in healthcare setting it will enable quick movement of patients and help manage hospital resources more effectively. The suspected patients take a SARS- CoV2-RT-PCR test and simultaneously give blood samples for analysis. If tested positive for COVID-19 infection their blood result is tested through the pretrained SV-LAR model.

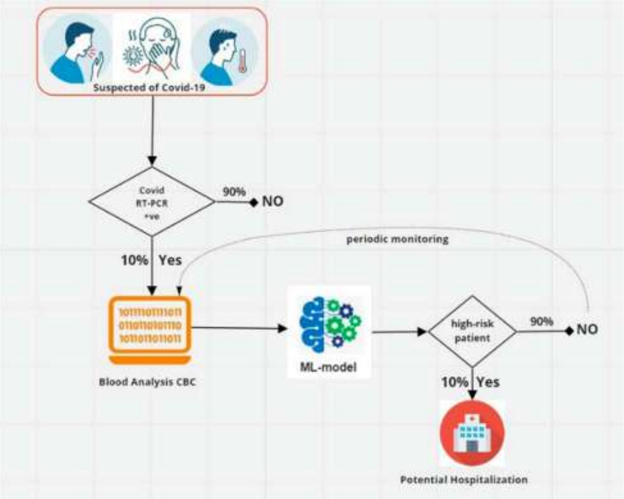


Fig.5. Envisioned patient flow using SV-LAR model

The model predicts their risk of hospitalization. If the patient belongs to high-risk class, the he/she should be tested for other tests like blood gas analysis and should be admitted upon physician approval. If the patient is deemed not at high-risk he/she should be sent back with periodic blood workup. The blood samples should be periodically analysed and tested through the SV-LAR model for risk assessment. This will enable healthcare staff to monitor each patients' development effectively.

IV. CONCLUSION

By the simple intervention of machine learning model (SV-LAR) with f1-score of 87%, the identification of a potential high-risk patient can be performed easily. The differential costs of tests required to the prediction is also underwhelming. SV-LAR model can be used to benefit healthcare workers by identifying about 10% high-risk COVID-19 patients from the increasing COVID-19 patient population. The precision of the model is a high 91% and 83% recall for the positive class. Simply put, 83 out of 100 high-risk patients can be identified correctly using this model and can be taken into hospital care for further treatment.

SV-LAR model is potentially fastest way to triage COVID-19 patients into high-risk and low-risk groups. Not only that, it enables to monitor patients via simple, non-expensive, quick, robust and minimally-invasive blood analysis. The identified high-risk patient population can then be put through more tests and procedures and can be treated accordingly. The low-risk patients, on the other hand, can be remotely monitored for any change in the patients' prognosis via the same model.

Our proposed model can be utilised globally. It relies on basic blood analysis, which is the most simple and established diagnostic service readily available in the healthcare system of any nation. This enables improved management of pandemic agnostic of the socio-economic standing of the nation.

An additional positive impact is related to hospital and patient flow management. It allows patient journey to be monitored from a distance providing better isolation of COVID-19 patients. Given that blood samples could be drawn periodically and analysed away from hospital emergency rooms, it allows ERs to work more efficiently despite the pandemic.
A limitation of our study however, is that not every patient hospitalized needs ICU. Due to the lack of data we could not segment the ICU needing patients from hospitalization requirements. As more data is collected and made available we can further refine the model. We believe that as more data will be incorporated in the model its performance and reliability will increase.

At this point, the model is developed from data available from patient emergency room visits from one hospital only. The model remains to be tested across geographies and hospitals for increased robustness. So far we have utilised only machine learning algorithms and not used deep learning algorithms to train the data. One of the reason was the size of the information available. As the available dataset was small we restricted our approach to include machine learning models. It is known that neural networks and deep learning algorithms work better on large datasets or else they tend to overfit. The performance of the model in larger datasets remains to be seen both using proposed machine learning algorithm and transferring similar approach in deep learning algorithms.

V. REFRENECES

[1] "Archived: WHO Timeline - COVID-19," *Who.int*. [Online]. Available: https://www.who.int/news/item/27-04-2020-who-timeline--covid-19. [Accessed: 14-Feb-2021].

[2] "COVID-19 map - johns Hopkins Coronavirus resource Center," *Jhu.edu*. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 14-Feb-2021].

[3] A. Shander *et al.*, "Essential role of Patient Blood Management in a pandemic: A Call for Action: A call for action," *Anesth. Analg.*, vol. 131, no. 1, pp. 74–85, 2020.

[4] CDC, "COVID-19 and Your Health," *Cdc.gov*, 03-Feb-2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html. [Accessed: 14-Feb-2021].

[5] Silva, and D. L. Guidoni, "Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data," *bioRxiv*, p. 2020.06.26.20140764, 2020.

[6] W.-J. Guan *et al.*, "Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis," *Eur. Respir. J.*, vol. 55, no. 5, p. 2000547, 2020

[7] B. Wang, R. Li, Z. Lu, and Y. Huang, "Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis," *Aging (Albany NY)*, vol. 12, no. 7, pp. 6049–6057, 2020.

[8] W. Liu *et al.*, "Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease," *Chin. Med. J. (Engl.)*, vol. 133, no. 9, pp. 1032–1038, 2020.

[9] Y. Zhou *et al.*, "Comorbidities and the risk of severe or fatal outcomes associated with coronavirus disease 2019: A systematic review and meta-analysis," *Int. J. Infect. Dis.*, vol. 99, pp. 47–56, 2020.

[10] Swiss Society Of Intensive Care Medicine, "Recommendations for the admission of patients with COVID-19 to intensive care and intermediate care units (ICUs and IMCUs)," *Swiss Med. Wkly*, vol. 150, no. 1314, p. w20227, 2020.

[11] Z. Zhao *et al.*, "Prediction model and risk scores of ICU admission and mortality in COVID-19," *PLoS One*, vol. 15, no. 7, p. e0236618, 2020.

[12] D. Ji *et al.*, "Prediction for progression risk in patients with COVID-19 pneumonia: The CALL score," *Clin. Infect. Dis.*, vol. 71, no. 6, pp. 1393–1399, 2020.

[13] E. Grifoni *et al.*, "The CALL score for predicting outcomes in patients with COVID-19," *Clin. Infect. Dis.*, vol. 72, no. 1, pp. 182–183, 2021.

[14] J.-J. Qin *et al.*, "Redefining cardiac biomarkers in predicting mortality of inpatients with COVID-19," *Hypertension*, vol. 76, no. 4, pp. 1104–1112, 2020.

[15] S. Schalekamp *et al.*, "Model-based prediction of critical illness in hospitalized patients with COVID-19," *Radiology*, vol. 298, no. 1, pp. E46–E54, 2021.

[16] Y. Zhou *et al.*, "Exploiting an early warning Nomogram for predicting the risk of ICU admission in patients with COVID-19: a multi-center study in China," *Scand. J. Trauma Resusc. Emerg. Med.*, vol. 28, no. 1, p. 106, 2020

[17] G. Wang *et al.*, "C-reactive protein level may predict the risk of COVID-19 aggravation," *Open Forum Infect. Dis.*, vol. 7, no. 5, p. ofaa153, 2020.

[18] S. Chikode, N. Hindlekar, P. Padhye, N. Darapaneni, and A. R. Paduri, "COVID-19: Prediction of Confirmed cases, active cases and health infrastructure requirements for India," International Journal of Future Generation Communication and Networking, vol. 13, no. 4, pp. 2479–2488–2479–2488, 2020Allen Institute For AI, "COVID-19 Open Research Dataset Challenge (CORD-19)."

[19] Allen Institute For AI, "COVID-19 Open Research Dataset Challenge (CORD-19)." .

[20] P. K. Nigam, "Correct blood sampling for blood gas analysis," *J. Clin. Diagn. Res.*, vol. 10, no. 10, pp. BL01–BL02, 2016.

[21] 1. Supervised learning — scikit-learn 0.24.1 documentation," *Scikit-learn.org*. [Online]. Available: https://scikit-learn.org/stable/supervised_learning.html. [Accessed: 14-Feb-2021].

[22] "Welcome to imbalanced-learn documentation! — imbalanced-learn 0.7.0 documentation," *Imbalanced-learn.org*. [Online]. Available: https://imbalanced-learn.org/stable/index.html. [Accessed: 14-Feb-2021].

[23] K. P. Shung, "Accuracy, precision, recall or F1? - towards data science," *Towards Data Science*, 15-Mar-2018. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9. [Accessed: 14-Feb-2021].

[24] "API Reference — scikit-learn 0.24.1 documentation," *Scikit-learn.org*. [Online]. Available: https://scikit-learn.org/stable/modules/classes.html. [Accessed: 14-Feb-2021].

[25] N. Darapaneni et al., "A machine learning approach to predicting covid-19 cases amongst suspected cases and their category of admission," in 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 375–380

[26] S. Cantenys-Molina, E. Fernández-Cruz, P. Francos, J. C. Lopez Bernaldo de Quirós, P. Muñoz, and J. Gil-Herrera, "Lymphocyte subsets early predict mortality in a large series of hospitalized COVID-19 patients in Spain," *Clin. Exp. Immunol.*, no. cei.13547, 2020

American Sign Language Detection Using Instance-Based Segmentation

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Prasad Gandole
Student – AIML
Great Learning
Mumbai, India
prasadgandole@gmail.com

Sureshkumar Ramasamy
Mentor – AIML
Great Learning
Mumbai, India
rsureshmca@gmail.com

Yashraj Tambe
Student – AIML
Great Learning
Mumbai, India
yashrajtambe54@gmail.com

Anshuman Dwivedi
Student – AIML
Great Learning
Mumbai, India
rjanshuman@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Mumbai, India
anwesh@greatlearning.in

Vihan Parmar
Student – AIML
Great Learning
Mumbai, India
vihanparmar@outlook.com

Kirthi Krishnan Ganeshan
Student – AIML
Great Learning
Mumbai, India
kirthi.krishnang@gmail.com

***Abstract* -- Deaf & mute people communicate usually in Sign Language with each other. Although the sign-language is very well known among these people it's quite unknown by other people. An attempt is made in this project to bridge the communication gap for the people who don't know Sign Language. The sign language used is the American Sign Language Lexicon Video dataset for this project which can be further extended to different languages. The dataset was in video format so a dominant frame extraction algorithm that would extract the dominant images from the video was used to create an as-per-requirement dataset. We found a lot of work had been already done on finger-spellings recognition but since finger-spellings make it much more complicated in actual time we thought of moving ahead with rather identifying words. The work was completed on a Computer Vision's State-of-the-art model based on Deep Learning that helped in achieving the completion of the task at instance level classification making the recognition task to a better identification at pixel-level which in-turn helps in achieving better results. An attempt has been made here to employ the latest method to make the computer learn sign languages that are independent of a person's complexion, lighting conditions, and orientations.**

Keywords-- Key-Frame Extraction, Deep Learning, Computer Vision, Instance Segmentation, American Sign Language Lexicon Video Dataset, Object Detection based SOTA model.

I.INTRODUCTION

This project demonstrates the ability to bridge the communication gap between the autistic-mute individuals and regular individuals in order to bring about a meaningful social impact and foster a sense of unity amongst the disabled by using the power of Artificial Intelligence. The aim was to develop an efficient system that can improve the lifestyle conditions for the autistic mute and hopefully contribute to the greater whole. Disability figures in India have risen about 30% in the previous decade, with about 20.3% of people having movement disabilities, 18.9% having hearing impairments. A major study published in 2018 of five sites in India found that 9.2% of children aged between 2-5 years and 13.6% of the children aged between 6-9 years had at least one of seven neurodevelopmental disorders (vision impairment, epilepsy, neuro motor impairment including cerebral palsy, hearing impairment, speech and language disorders, autism spectrum disorders and intellectual disability[27]. The number of impaired people has recently reached about 400 million and therefore extensive research and studies have been accelerated in order to ease the means for communication amongst the disabled. Currently the communication between the autistic mute community and non-autistic mute takes place with an expert interpreter which becomes inconvenient as well as expensive since human learning and expertise in the subject is involved. Aiming to create a ground of equal opportunity is the foremost objective of this approach. A lot of work has already been carried out in this domain , a lot of them referring to Fingerspelling Recognition.

To a dated back research, the concept of gesture recognition proposed by Junji Yamato et al. [6] included Time-Sequential models which used Hidden Markov Models. HMM increases the computation complexity with a very complex logic to encompass in a program. A variety of visual techniques were discussed which mentioned the use of gloves that simplifies the further processing parts, or skin colour based algorithms which typically aims to bring the hand colour as a feature extraction process or finger detection which tries to detect gestures based on the number. of fingers closed or open, as proposed by Saba Joudaki et. al. [1]. The techniques again have huge computation demands. A Real-time American Sign Language Letters Recognition was proposed by Authors, [2] Xinyun Jiang et. al., which used PCA for feature extraction and then a One-vs-One SVM classifier was used to classify the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

gestures [2]. But only a limited number of gestures can be trained since with larger datasets, SVM takes a longer time to converge. One method proposed using Haar Cascades for sign language feature extraction and then using CNN, proposed by Sanket Kadam et. al. [3]. Again Haar Cascades adds an increasing computation which is a template matching algorithm. Another development made was the use of depth maps in fingerspelling recognition. Depth map was basically created from a depth sensing device which basically used the Microsoft Kinect Camera, proposed by Byeongkeun Kang et. al. [4]. This camera shows up different colour scales depending on the distance from the camera. This helps in providing additional information in feature extraction but again the depth map needs to be extensively created. A similar approach was specified previously as well that used MLRF model thereby classifying the finger spellings which is a very lightweight model used for reducing memory consumption as well as boost up speed, proposed by Alina Kuznetsova et. al. [5]. The technique was implemented on a rotation- position- and scaling- invariant features extracted from the images. A very simple arrangement was proposed by Thad Starner et. al. [7] that established usage of desk and wearables. This requires the signer to be sitting at a desk or using a wearable that actually sees the sign from the same angle as the signer would see it, again the proposed uses HMM for modelling. A different feature extraction was proposed by Ariya Thongtawee et. al. [8] that would include different techniques like number of white pixels at the edge of the image, finger length from the centroid point, angle between fingers of first and last frame. This method exhaustively uses a very strong feature extraction algorithm to gain the idea of a gesture at the very approach possible. A different technique for feature extraction was proposed by M.M. Islam et. al. [9] that uses K-Convex Hull which has multiple complex algorithms which introduces due latency and computational complexities. A paper proposed vision based features extraction using kurtosis and PCA, by M.M. Zaki et. al. [10].The paper proposes the use of a 3D motion sensor which is a palm-sized Leap motion sensor which was then an economical option to Cyberglove or Microsoft Kinect & the features thus extracted were classified using K-NN and SVM. The depth image technology used by C.Dong et.al. in [12] uses depth comparison to extract features and a Random Forest as well as a constrained link angle algorithm to exactly classify the gesture into a finger spelling. Another visual based feature extraction technique that was used was SURF which is an Image Processing algorithm, was proposed by C.M. Jin et. al. [13] for feature extraction which then was forwarded to SVM for classification. A different idea was proposed which converted sign language to text and speech, proposed by V.N.T. Truomg et. al. [14] was based on Haar-Cascade algorithm and pattern matching to classify the gesture. A dynamic Sign language detection technique based on Sign Language Trajectory and Key Frame extraction [15] was proposed by Yufei Yan et. al. It uses Joint Point Weight Assignment of Hand Trajectory that helps in recognition of two hand gestures keeping a view that two points brings two different information at a time. Then DTW (Dynamic Time Warping) is used to classify the gestures. A further refined algorithm was proposed [16] by Murat Tskiran et. al. which converted the image from RGB to YCbCr space which was then forwarded to skin colour detection algorithm and then hand detection using K-convex hull algorithm. This method proves computationally intensive and then uses CNN to classify the gestures. A paper on Indian Sign Language [17] by Sajanraj T D et. al. proposed a technique called CLACHE for feature extraction. This led to extraction of Region of Interest and then using R-CNN, the gestures were classified. A very complex set of algorithms [18], proposed by Archana S. Ghotkar et. al. used HSV and YCbCr colour model, Haar-like features, adaptive skin colour model, location, angle, velocity & motion pattern P2-DHMMS which seems in itself quite tedious to achieve, puts a very complex combinations of algorithms with a requirement of heavy computation power. Use of CNN was done based on a simple Skin Segmentation and Image category classification is proposed in [19] by Shadman Shahriar et. al. which uses bounding box for feature extraction. Another proposal [20] by Helen Cooper et. al. was using the HOG, DTW & Kalman filters to extract features and then uses HMM and CNN which is quite simpler as compared to the previous mentioned algorithms.

In our approach, we propose a simple and yet an effective method of pixel level segmentation, which was the basic idea for using the model. The concept of semantic/instance based segmentation was thus a better solution and it could be understood by asking these questions to classify the approach into 4 major categories:

1. *Semantic Segmentation*:

Given an image, could we classify each pixel as belonging to a particular class?

2. *Classification+Localization:*

Could we also get the location of the said object in that image by drawing a bounding box around the object?

3. *Object Detection:*

In a real-world setting, we don't know how many objects are in the image beforehand. So can we detect all the objects in the image and draw bounding boxes around them?

4. *Instance Segmentation:*

Can we create masks for each individual object in an image? It is different from semantic segmentation. How? Let's say we're looking at an image then we won't be able to distinguish between the two objects of the same class using semantic segmentation procedure as it would sort of merge both the objects together. We hence intend to do this pixel level segmentation by using a State-of-the-art(SOTA) instance based segmentation in a single architecture that throughputs a pixel level accuracy in detecting gestures. Our model gives a high accuracy output as the algorithm makes sure that we not only have a classification of the object of interest with a bounding box but also an instance based segmentation of the image. The dataset used for our approach is the American Sign

Language Lexicon Video Dataset designed by Stan Sclaroff, Carol Neidle, Vassilis Athitsos, J.Nash, A.Stefan, Q.Yuan, A.Thangalli [21]. In the consequent sections, we will map out a step by step procedure and the workings of our architecture.

II. MATERIALS AND METHODS

The detailed model architecture and its procedure are given as follows:

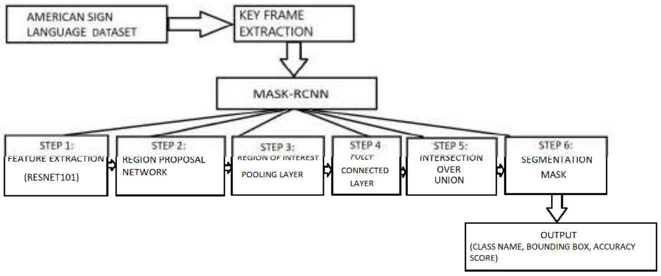

AMERICAN SIGN LANGUAGE DATASET
KEY FRAME EXTRACTION
MASK-RCNN
STEP 1: FEATURE EXTRACTION (RESNET101)
STEP 2: REGION PROPOSAL NETWORK
STEP 3: REGION OF INTEREST POOLING LAYER
STEP 4 FULLY CONNECTED LAYER
STEP 5: INTERSECTION OVER UNION
STEP 6: SEGMENTATION MASK
OUTPUT (CLASS NAME, BOUNDING BOX, ACCURACY SCORE)

Fig.1 Model Architecture

A. *Data Acquisition and Pre-processing*

American Sign Language Lexicon Video[24] data was leveraged for building the model. The dataset used here was in a video (.mov) format, which contains over 3300 signs, each sign represented from a different angle. Some of the lingual annotations included in the dataset are the start and end times of a sign, start and end hand labels, etc. For supporting computer vision tasks and recognition of sign languages, the dataset had numeric IDs for every sign, and every sign is represented in a video sequence with a calibrated camera. Out of the total set of gestures, forty distinct gestures were used to train the model using a total of 667 images, out of which model was trained on 551 images and tested the model on 116 images. All images were extracted to a size of 640*480 pixels since the movements were distinctive enough for this particular pixel size. Using a reasonable pixel size also helped in a faster training procedure with reduced computational complexity.

B. *Key-Frame Extraction*

C. *Mask RCNN Model*

Mask RCNN algorithm has been leveraged for building the model. Features extraction, classification, instance based segmentation is a part of single architecture. Mask RCNN is a widely used Computer Vision algorithm classified as State-of-the-art model for object detection on providing maximum accuracy while at the same time providing pixel level accuracy. The Mask-RCNN model, built on the top of ResNet101 architecture, feeds on images provided by Key-Frame

Key Frame Extraction algorithm was used as a data pre-processing step which generates unique key frames from the American Sign Language dataset that was pre-recorded.[22]. The fact that the Key frame extraction algorithm was used was that it was a pre-processing step, it helped us segregate only the frames of interest and it helped us speed up the process of frame retrieval. Key frame extraction played a pivotal role in order to process real time videos which added up to the efficiency of our model architecture. The following steps detail the Key Frame Extraction algorithm:

Step 1:

To find the difference between two consecutive frames in terms of edges so as to find significantly different frames that define different action

For each video frame k = 1 to N

- Read frame Vk and Vk+1
- Obtain the gray level image for Vk and Vk+1
  - Gk = Gray image of Vk
  - Gk +1 = Gray image of Vk +1
- Find the edge difference between Gk and Gk + 1 using the Canny edge detector.
  - Let diff(k) be their difference.
  - dif(k) = summation of i,j (Gk - Gk + 1) where i, j are row and column index.

Step 2:

Compute the mean and standard deviation

Step 3:

Compute the threshold value

- Threshold = M + a*S , where a = constant

Step 4:

Find the key frames for k = 1 to (N-1)

- If diff(k) > Threshold then,
- Write frame Vk +1 as the output key-frame

extraction, and serves as the core algorithm for our model architecture. Although this algorithm was computationally intensive, the underlying working of being highly selective in choosing the features right from the first stage provided a headway in accuracy and a trade-off to higher resource usage while at the same time solved the problem for vanishing gradients.

Once major features were found having high probable objects within, these were brought on a uniform scale by means of ROI pooling and then further classified into the gesture list available beforehand along with position of the hands. This was refined via IoU to extract even richer images and finally giving the rich extracted images to the segmentation mask in order to throughput an instance based segmentation. This gave the model a high accuracy output as the Mask RCNN model went through a series of steps in order to make sure that classification of the object with a bounding box as well as instance based segmentation of the image.

The workflow of the Mask RCNN model is as follows: -

1. *Feature Extraction*

ResNet (Residual Networks), a classic neural network algorithm, a winner of the ImageNet challenge held in 2015 was used as a backbone for our project. It allowed training on extremely deep neural networks while being able to efficiently solve the problem of vanishing gradient at the same time. It has 101 blocks in it, each block contains a convolution layer, activation layer as Relu, followed by a convolution with a skip connection as the latter adds input directly to the output of the previous convolution layer which is commonly known as a *'Residual Block'*. With Python and Keras library, the Resnet101 architecture which was readily available in pre-trained mode was implemented for creation of the model

2. *Region Proposal Network*

It was basically used to determine whether a specific region has an object in it or not, if it did not find an object in certain features, it eliminated those features and considered only the features which hadan object in it. It returned the candidate bounding boxes.

3. *Region of Interest*

Region of Interest was found in order to extract even more richness to the existing image. Region of interest determined how an image could contain a specific region with more information in them. The region obtained from RPN was of different sizes, thus pooling layers was used to bring them to the same size.

4. *Fully Connected Layer*

These regions were forwarded to a Fully Connected layer to predict the bounding boxes and class labels.

5. *Intersection Over Union*

For all the predicted regions, we computed the intersection/union (intersection over union). Condition: If IoU is >= 0.5, then consider the region as ROI & if not, then exclude the region.

6. *Segmentation Mask*

Once the ROI was found w.r.t. the IoU values, we added a mask branch to the existing architecture. This gave the segmentation mask for each region that contained an object (element wise segmentation). It also returned a mask of size 28*28 for each region which was then scaled up for inference. This was the final step, which gave a class label of the gesture in an image, a bounding box around the gesture in that image, and an instance based segmentation of the objects in that image.

III. RESULTS

Training losses and validation loss is shown below that explains the model's losses which is inclusive of the bounding box loss as well as the classification loss that gives an idea about how the losses are improvised.

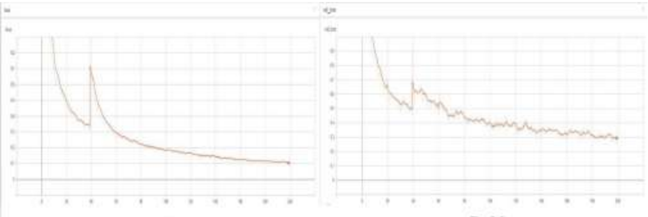


Fig. 2.1

Fig. 2.2

Fig.2.1 is the plot of training loss whereas Fig. 2.2 is the validation loss. The losses graph is loss vs. epoch graph

SGD(Stochastic Gradient Descent) was the optimizer in use because it has faster convergence as compared to adam or other optimizers due to the fact that a paper[26] claims the following — "We observe that the solutions found by adaptive methods generalize worse (often significantly worse) than SGD, even when these solutions have better training performance. Again gradient clipping is used to reduce any issues of exploding gradient.

To ensure the model does not overfit there were some different strategies which were used. One basic strategy was to use weight decay regularizer[25] achieving then state-of-the-art results reported:

"…and weight decay of 0.0005. We found that this small amount of weight decay was important for the model to learn. In other words, weight decay here is not merely a regularizer: it reduces the model's training error"

In terms of hyperparameter tuning the model used was pre-trained on COCO-dataset hence the model was tuned on few hyperparameters which included learning rate & weight decay. For the purpose of optimization the learning rate was tuned from 0.01 to 0.0001 whereas for the purpose of regularization the weight decay was tuned from 0.01 to 0.001. All the hyperparameters tuned where tuned specifically for this use case and are the result of experimental tuning to improve the results.

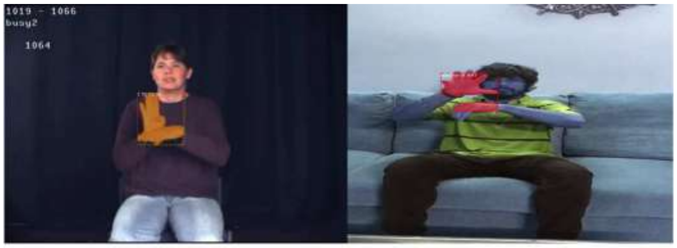


Fig. 3.1 Fig. 3.2

Fig.3.1 is from the validation set which informs us about the validation score of the model and in order to test if the model works accurately with different lighting conditions and orientation. in Fig. 3.2. we tested the model by having a person perform the sign language, in this case 'Advise'.

The basic evaluation procedure of any Object Detection(Semantic Segmentation as well) model proceeds with a term known as mAP(mean Average Precision). The reason behind using this was because Object Detection was localizing multiple objects in a single image. In order to calculate how each object was identified in each image as well as how near to real description the prediction was, this was not used. As per the use case, the majority of the signs were single handed signs which is why the concept of object detection stood redundant. Hence, the idea was to use the same metric of Object Localization which is IoU but averaging over all the images IoU's helped in detecting how good is the performance of the model on the entire validation set, hence introduced average IoU. An average IoU of ~95% was achieved on the train set whereas ~90% was achieved on the validation set

IV. DISCUSSION AND CONCLUSION

Algorithms used before like Skin Detection Algorithm: Skin Region Detection, Feature Extraction: K-Convex Hull Algorithm, Classification: Convolutional Neural Network are not used in place of Key-Frame extraction because former mentioned techniques had multiple disadvantages. Skin Detection Algorithm would not intelligently detect skin regions according to dynamic lighting conditions and was only as good as it is trained. The other, otherwise widely used, K Convex Hull Algorithm suffered from slow recursion, overhead of the repeated subroutine calls, required excessive memory for storing intermediate results of sub- convex hulls, unable to control and guarantee sub-problem size resulting in sub-optimum worst-case time performance. Also the underlying 'Divide and Conquer' algorithm was not ideal if the points to be considered were too close to each other. Instead the Key-Frame extraction algorithm retained information with lesser images and was independent of the input condition like light, skin region, cloth color etc. which were all taken care of as these would remain the same through the gesture hence these would sbe summarily neglected for extracting the features.

The Mask R-CNN takes care of various steps inside its architecture all the way from extracting features in the first step, determining if that extracted feature consists of an object or not in the second step in order to ensure that only those features with objects in it are processed for the next step and then it brings all the features consisting of objects to the same size in the third step by using Region of Interest pooling layer and further classifying the image with a bounding box in the fourth step. It then passes on the classified images along with their bounding boxes to intersection over union to extract instance based segmentation.

To prepare the training data quickly, A better tool is required for automating the frame annotation process, which if fed with initial geometry of a target object, may decide what possible shapes the target object can have, i.e. for fingers, gaps (hull) between the fingers having relation between their lengths/breadth and with that of palm as given, then the auto-annotation will correctly contour the hands in any posture and can give additional output of hand rotation angles which may serve as additional input feature.

The first version of our model detects ASL from recorded video as real time video could not be processed due to technical constraints like limited GPU, free memory on Google Colab. Also its code intensive being on TensorFlow 1.0x platform. VMWares give us sufficient access to resources but even that has certain limits. YOLO can be tried in place of Mask-RCNN provided the constraints are met.

Since the project does not serve the masses, getting investors to up-scale the project can get tricky. However, the promising nature of the project can rope in players soon if a solution which integrates a full end to end product comes up which has a well-designed, user-friendly interface and runs on slower hardware and can work even without internet connection.

An attempt can be made to train the model on a Mobile NET architecture with the features learnt from the current model. The new Mobile app size model will do the same task. Secondly a bi-directional LSTM model can be built to frame multiple gestures into a grammatically correct sentence which then can be converted to speech and in another form as per the customer requirement – this will increase the business scope of the model and its application.

REFERENCES

[1] Saba Joudaki, Dzulkifli bin Mohamad, Tanzila Saba, Amjad Rehan, Mznah Al-Rodhaan & Abdullah Al- Dhelaan (2014) Vision-Based Sign Language Classification: A Directional Review, IETE Technical Review, 31:5, 383-391, DOI: 10.1080/02564602.2014.96157

[2] Xinyun Jiang and Wasim Ahmed, Hand Gesture Detection based Real-Time American Sign Language Letters Recognition Using Support Vector Machine (2019)

[3] Sanket Kadam, Aakash Ghodke, Sumitra Sadhukhan, Hand Gesture Recognition Software Based on Indian Sign Language (2019)

[4] Byeongkeun Kang, Subarna Tripathi, Truong Q. Nguyen, Real-time Sign Language Fingerspelling Recognition using Convolutional Neural Networks from Depth map (2015)

[5] Alina Kuznetsova, Laura Leal-Taix´e, Bodo Rosenhahn, Real-time sign language recognition using a consumer depth camera (2013)

[6] Junji Yamato,Jun Ohya,Kenichiro Ishii, Recognizing Human Action in Time-Sequential Images using Hidden Markov Model (1992)

[7] Thad Starner and Alex Pentland, Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video (1998)

[8] Ariya Thongtawee, Onamon Pinsanoh, Yuttana Kitjaidure, A Novel Feature Extraction for American Sign Language Recognition Using Webcam (2018)

[9] M.M. Islam,S.Siddiqua,J.Arfan, Real Time Hand Gesture Recognition using different algorithms based on ASL (2018)

[10] M.M. Zaki, S.I. Shahee,Sign language recognition using a combination of new vision based features (2011)

[11] C.Chuan, E.Regina, C.Guardino, American Sign Language Recognition using Leap Motion Sensor (2014)

[12] C.Dong,M.C.Leu,Z.Yin,American Sign Language using Microsoft Kinect (2015)

[13] C.M. Jin, Z.Omar, M.H.Jaward, A Mobile Application of American Sign Language Translation via Image processing algorithms (2016)

[14] V.N.T. Truomg, C.Yang, Q.Tran, A Translator to American Sign Language to Text and Speech (2016)

[15] Yufei Yan, Zhi jun Li,Qunzhu Tao, Chenyu Liu, Research on Dynamic Sign Language Algorithm Based on Sign Language Trajectory and Key Frame Extraction (2019)

[16] Murat Tskiran, Mehmet Killioglu , Nihan Kahraman, A Real-Time System For Recognition Of American Sign Language By Using Deep Learning (2018)

[17] Sajanraj T D, Beena M V, Indian Sign Language Numeral Recognition Using Region of Interest Convolutional Neural Network (2018)

[18] Archana S. Ghotkar , Gajanan K. Kharate, Study of vision based hand gesture recognition using indian sign language (2014)

[19] Shadman Shahriar, Ashraf Siddiquee, Tanveerul Islam, Abesh Ghosh, Rajat Chakraborty, Asir Intisar Khan, Celia Shahnaz, Shaikh Anowarul Fattah, Real-Time American Sign Language Recognition Using Skin Segmentation and Image Category Classification with Convolutional Neural Network and Deep Learning (2018)

[20] Helen Cooper,Brian Holt and Richard Bowden, Sign Language Recognition(2011)

[21] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan and A. Thangali, *The ASL Lexicon Video Dataset,* CVPR 2008 Workshop on Human Communicative Behaviour Analysis (CVPR4HB'08)

[22] S. K. Ramasamy, Multi-Scale Data Fusion for Surface Metrology. Proquest, Umi Dissertation Publishing, 2012.

[23] “Detection and Segmentation,” Stanford.edu. [Online]. Available: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf. [Accessed: 05-Apr-2021]http://vlm1.uta.edu/~athitsos/asl_lexicon/

[24] Alex Krizhevsky, et al. “*ImageNet Classification with Deep Convolutional Neural Networks*”(2012)

[25] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, Benjamin Recht,*The Marginal Value of Machine Learning*, NeurIPS(2017

[26] Neurodevelopmental disorders in children aged 2–9 years: Population-based burden estimates across five regions in India. (218 C.E.). PLOS MEDICINE, 1. https://doi.org/10.1371/journal.pmed.1002615

Traffic Monitoring and Analysis At Toll Plaza

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Umang Maheshwari
Student – AIML
Great Learning
Pune, India
umang2809@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Pune, India
anwesh@greatlearning.in

Parikshit Bangade
Student – AIML
Great Learning
Pune, India
parikshitbangde@gmail.com

Sushilkumar C Thorawade
Student – AIML
Great Learning
Pune, India
thorawadesushilk@gmail.com

Amit Mane
Student – AIML
Great Learning
Pune, India
meet9426@gmail.com

Rushikesh Borse
Mentor – AIML
Great Learning
Pune, India
rpborse@gmail.com

***Abstract*--This paper presents a solution to Traffic monitoring and analysis at Indian toll plazas. In some of the areas, the work is done on Vehicle detection and localization, vehicle registration detection, and character recognition and vehicle make classifier (Currently concentrated on Indian Cars only). Accurate detecting and localizing of objects in computer vision has always been a core problem, to the rescue of which Tensor Flow Object detection API comes with implementations provided by RCNN family (Basic RCNN, Fast RCNN, and Faster RCNN) and Single-shot detection models. Experimenting over a range of models from Faster RCNN with inception v2, SSD MobileNet, and SSD inception v2 for vehicle and vehicle registration number detection problem we got higher accuracy with Faster RCNN with inception v2 for RMS Prop optimizer. For the detected vehicle registration number, we used an SVM-based multiclass classifier. A custom image dataset is used to train the model. The vehicle make classifier is implemented with VGG16 and the dataset was selected to contain only Indian cars.**

Keywords—Vehicle Detection, Vehicle registration number, Faster RCNN, VGG16 and SVM

I. Introduction

India has a huge web of highways all over the country connecting cities and states together. Road transportation being the most widely used mode of commuting and transporting goods from one city or state to other city and state. A need of monitoring and analysis of traffic at Indian roads is of paramount importance to civil and military surveillance for use of development and security-based decisions. Also, the derived insights could be a source of input to help find solutions to wide variety of problems a commuter face.

Indian traffic at toll plazas is very complex and critical with a diversified volume of vehicles likes of bikes, cars, trucks, and buses running on road at any given time of the day. Causes a wide variety problem of which congestion being most common on Indian toll plazas. The maintenance of these highways is very important to keep the connectivity intact.

Toll Plazas or toll booths are special counters build on highways to collect tax from the general population using the highway for commuting. This tax is basically to retrieve the cost to build and maintain the highway. So, it a usual scenario of getting a huge congestion of traffic where everybody is lined up to pay the taxes before they move ahead.

As a result of these congestion, it not only causes delays to commuters but also adds to loss of fuel and productive time. On a larger scale to observe the impact can be as large as impacting the economy negatively. Also, the kind of pollution caused in scenarios of traffic congestion is tremendous impacting environment.

II. Motivation of Work

Spending hours at toll plaza in heavy traffic is very common in India and people do complain of waiting long before they are able to pay and pass the toll booth. As per the toll rules of NHAI [1] there is a max of 3 minutes waiting period for every vehicle to be able to pay and pass the toll booth and if this time exceeds beyond 3 minutes due to congestion, delay by toll booth authority or any other similar reasons should be allowed a free passage without a fee.

A necessity of monitoring of the toll booth functioning and scenario identification of congestion help us find better solutions of several problems at toll booths.

III. Related Work

Our proposed solution targeted few of the milestones on which past work done. A vision-based vehicle identification system providing solution of object extraction [16], object tracking, occlusion detection and segmentation, and vehicle

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

classification. In situations where vehicles on the road may occlude each other, their trajectories may merge or split and to handle this they developed three processes: occlusion detection, motion vector calibration, and motion field clustering. Finally, the segmented objects were classified into seven different categorized vehicles [2].

Object Detection solution based on TensorFlow's Object Detection API being a promising technique providing ability to build and deploy image recognition software. Object detection technique not only classifies and recognizes objects of interest in images but also localizes them and marks them with bounding boxed around them. Solution mostly focuses on detecting harmful objects like threatening objects for which they got Tensor flow Object Detection API to train model and have used Faster R-CNN algorithm for implementation [3].

Region-based convolutional neural network for real-time hand gesture recognition. A Faster region-based convolutional neural network (Faster-RCNN) with Inception V2 architecture was used. Their solution observed average precision, average recall, and F1- score by training the model with a learning rate of 0.0002 for Adaptive Moment Estimation (ADAM) and Momentum optimizer, 0.004 for RMSprop optimizer. A better result of precision recall and F1-score values were attained with ADAM optimization algorithm after evaluating over custom test data [5].

IV. Our Proposed system

Proposed solution is an integration of models trained to help achieve objectives like vehicle detection on live video streams, Vehicle counting, Vehicle model type classification and vehicle registration number detection.

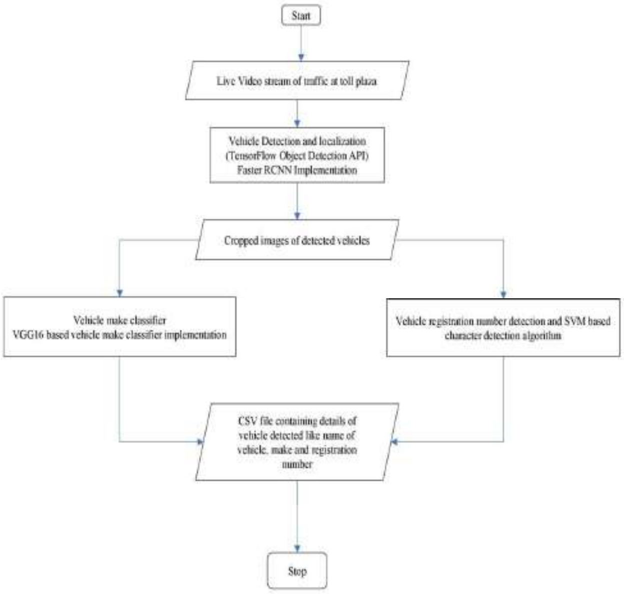

Start
Live Video stream of traffic at toll plaza
Vehicle Detection and localization (TensorFlow Object Detection API) Faster RCNN Implementation
Cropped images of detected vehicles
Vehicle make classifier VGG16 based vehicle make classifier implementation
Vehicle registration number detection and SVM based character detection algorithm
CSV file containing details of vehicle detected like name of vehicle, make and registration number
Stop

Fig 1. Solution Workflow

The above figure shows the workflow followed in our solution. We started with detection and localization of our desired objects which are different types of vehicles and form distinguishing boundaries for the located objects into video frames and cropping the detected objects to be stored locally and used for further processing in outlaid milestones.

In Object Localization we detect where the object is situated, and Object detection involves detecting multiple objects in an image. In Object detection, we break down the original image into multiple images and perform Object Localization on each part. Extracted image parts are called ROI (Region of Interest) / Anchor Boxes/ Priors. To solve the problem of classification and regression (localization), we have different models which uses different approaches such as R-CNN uses 'Selective Search' to extract ROI's. Fast-RCNN also uses Selective Search but extract ROIs from Feature Map. Faster R-CNN uses Region Proposal Network (RPN) to extract ROIs and in SSD, we (humans) provide the ROIs using multiple grids and anchor boxes (or priors). Allow CNN to look at different ROIs to see if there is an object in the ROI and what is their bounding box.

Constructing a true Deep Learning model with ability to precisely localize and detect multiple objects in same image is always a core challenge with traditional computer vision and hence TensorFlow Object detection API an open-source solution built on top of TensorFlow makes it bit easy to construct, train and deploy Object detection models. So, for object detection we used TensorFlow Object detection API's one of the powerful tools to help us achieve the objective detection milestone and use algorithm for implementation.

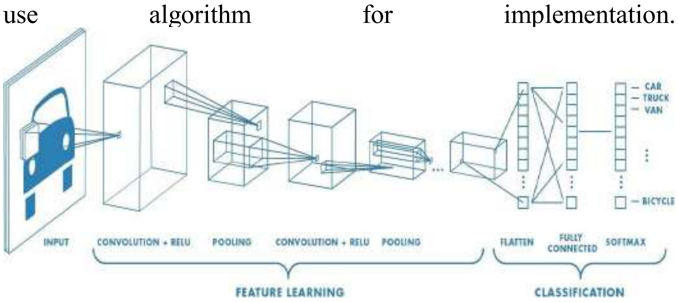


Fig 2. CNN Architecture

Input image applied with a layers of convolution layers (with RELU activation function) Max pooling layer to concise the derived feature maps by extracting the max values of the feature map values to concentrate on the main features of detected objects. The output of pooling layer then flattened feature map then passed through fully connected Neural network layers. SoftMax activation enables to do multi-class classification.

For the detected vehicles in multiple video frames using Faster RCNN implementation it became a challenge to monitor the moving vehicles and hence we applied a naming logic to address the detected objects and monitor them in subsequent frames. A combination of location directory, a counter to track number of cropped images, class index id, class index name and object id were used as components to form a name for the detected vehicles.

A unique object id talked about is derived via Centroid tracking algorithm are given in the derived name space which is then saved in csv format and finally merging all csv for detailed analysis. such as counting number of vehicles etc. Centroid tracking algorithm relies on the Euclidean distance between:

- Existing object centroids (i.e., objects the centroid tracker has already seen before).
- New object centroids between subsequent frames in a video.

Further to the processing in the solution the detected vehicles with bounding boxes are cropped and used for vehicle make classification and vehicle registration number detection. K. Simonyan and A. Zisserman [4] from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" proposed VGG16 convolutional neural network. We used VGG16 based image classifier for identifying the make of detected vehicles in first stage of our system.

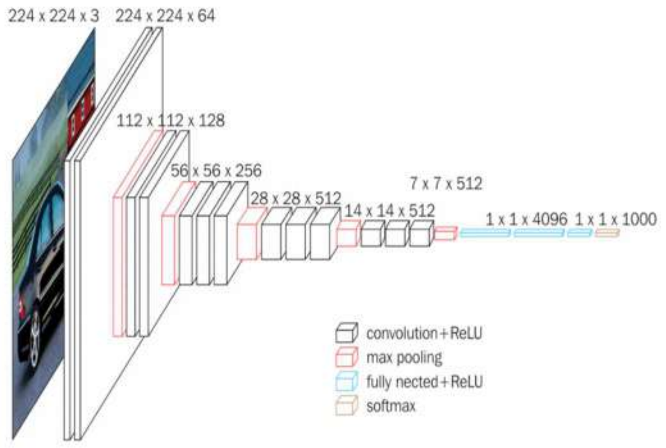


Fig 3. VGG16 CNN

The same cropped image of object detection stage of our system was used as an input for vehicle registration number detection. A background subtraction achieved with Otsu Thresholding. Background subtracted Images were then subjected to segmentation where each character was further predicted with an SVM based character recognition algorithm.

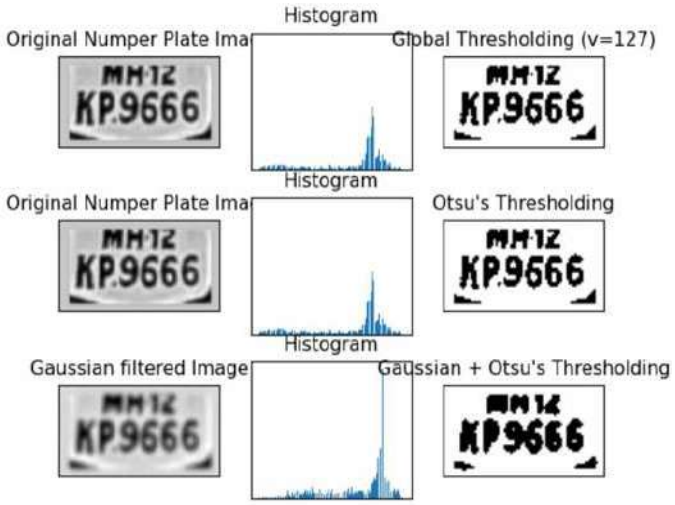

Original Numper Plate Ima
Histogram
Global Thresholding (v=127)
Original Numper Plate Ima
Histogram
Otsu's Thresholding
Gaussian filtered Image
Histogram
Gaussian + Otsu's Thresholding

Fig 4. Background Subtraction (Otsu Thresholding)

Finally, the output of all these models is clubbed to form a CSV file with insights like Vehicle counts, Vehicle registration number, Vehicle make classification description along with timestamp showing the date and time. This insight then redirected over dashboard for live monitoring and further analysis.

With all the data collected via the above workflow can be conducted to help in different domains of interest.

- Traffic inflow at toll plazas to extract busiest day, week of the month, busies hour of the day etc.
- Toll collections stats and each vehicle service time monitoring.
- Security insights with Vehicle registration number detection and vehicle model type detection.

V. Algorithms

A. SVM based Character Recognition

The hypothesis function h is defined as:

$$h(x_i) = \begin{cases} +1 & if\ w.x+b\geq 0 \\ -1 & if\ w.x+\ b<0 \end{cases} \qquad (1)$$

The point above or on the hyperplane will be classified as class +1, and the point below the hyperplane will be classified as class -1. Computing the (soft margin) SVM classifier amounts to minimizing an expression of the form,

$$\left[\frac{1}{n}\sum_{i=1}^{n} max\left(0,1-y_i(w.x_i-b)\right)\right]+\lambda\left|\left|w\right|\right|^2 \qquad (2)$$

We focus on the soft-margin classifier since choosing a sufficiently small value for lambda yields the hard-margin classifier for linearly classifiable input data.

B. Faster R-CNN

Replaces the selective search method with region proposal network (RPN) which makes the algorithm faster time/Image – 0.2 second.

$$\text{IoU} = \frac{\text{Anchor} \cap \text{GroundTruth}}{\text{Anchor} \cup \text{GroundTruth}} \qquad (3)$$

$$\text{IoU} = \frac{\text{A} \cap \text{Gt}}{\text{A} \cup \text{Gt}} \begin{cases} > 0.5 = \text{Object} \\ < 0.5 = \text{No Object} \end{cases} \qquad (4)$$

Computing the Intersection over Union is dividing the area of overlap between the bounding boxes by the area of union.

In the numerator we compute the *area of overlap* between the *predicted* bounding box and the *ground-truth* bounding box. We are keeping the threshold value of 0.5.

The denominator is the *area of union*, or more simply, the area encompassed by *both* the predicted bounding box and the ground-truth bounding box.

Dividing the area of overlap by the area of union yields our final score — *the Intersection over Union.*

$$\text{IoU} = \frac{\text{A} \ \cap \text{Gt}}{\text{A} \ \cup \text{Gt}} \begin{cases} > 0.5 = \text{Object} \\ < 0.5 = \text{No Object} \end{cases}$$

VI. System Environment setup

A. Hardware

TABLE.1. HARDWARE COMPARISON

Machine Type	CPU Processor	GPU	RAM	Clock speed
Machine 1 - Asus vivo book-max A541uv	Intel® Core TM i3 7100U	Nvidia 920MX	12GB	2.4GHz
Machine 2 - Acer Predator Helios 300	Intel® Hexacore Intel i7 9750	Nvidia RTX 2060	16GB	2.6Ghz

B. Dataset

TABLE.2. DATASET DETAILS

Models	Dataset	Train Images	Testing Images	Number of classes	images /classes
Vehicle detection Model	Google image dataset v4, Custom images	701	180	6	145 approx.
Vehicle registration number detection model	Google image dataset v4, Custom images	2000	376	1	2376
Vehicle Make classifier	Google Custom images using labelImg	689	272	17	60 approx.
SVM Character recognition (0-9 digits, 26 alphabets)	Google Custom images using labelImg	1260	180	36	40 per character

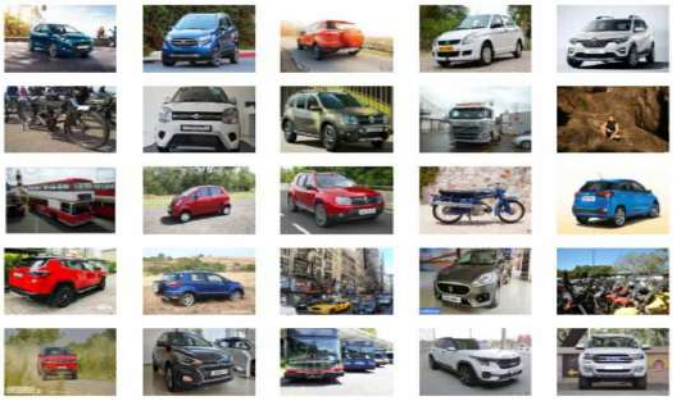


Fig 5. Actual dataset for Object detection and make classifier

These images for object detection were taken to ensure that they have the desired vehicle class covering major portion of the image with a wide variety of vehicle orientation like middle/close range in front of the camera, middle/close range in the left, close/middle range in the right, and far range to enable the model to learn better. The images sizes vary from 300*300 to 1000*1200 pixels. Several instances of a same vehicle are included with different bounding hypotheses. Also, few images of multiple classes like Car, Bus or Truck and person was included to enable the model to learn various weights from such perspective. We included around 120-150 images per class and the classes were – Car, Truck, Bus, Rickshaw, bike and Person.

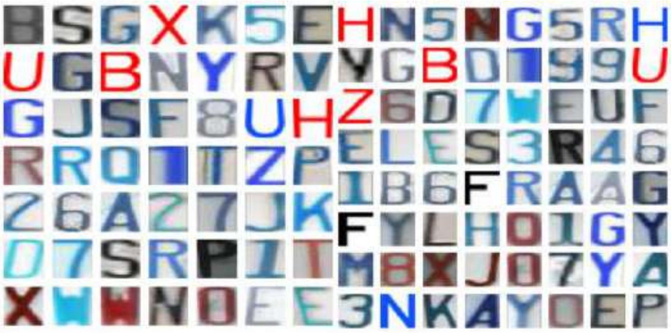


Fig 6. Character recognition dataset

For vehicle registration detection we used images with vehicle registration numbers covering the entire image size. To add diversity, we included Indian plus foreign registration number plates and multiline number plates to enable our model to learn them better. The selected images were of size 20*20 pixels for SVM based character recognition model.

At this moment our solution only considers the class Cars for Vehicle make classification for which we used google images of Indian car models like Honda Amaze, Hyundai i10, Suzuki Wagon R, Suzuki Swift etc. Images ranging from 300*300 to 1000*1200 pixels majorly coving the entire image size. The Cars were selected to have wide variety of vehicle orientation like middle/close range in front of the camera, middle/close range in the left, close/middle range in the right, and far range to enable the model learn better.

VII. Results & Discussion

A. False detection results

For vehicle registration detection the model detected the flashlights as vehicle registration number.

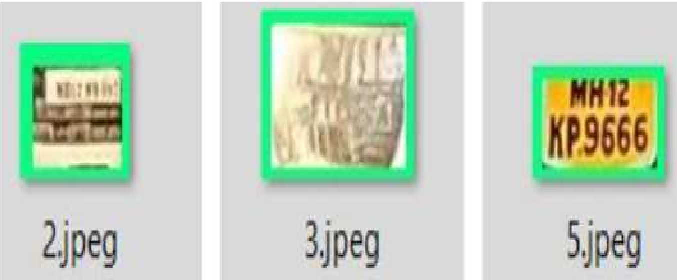


Fig 7. False detection of vehicle registration number

B. Results of Evaluation

For object detection we use the concept of **Intersection over Union** (IoU). IoU a Jaccard index-based computes intersection over the union of the two bounding boxes; the bounding box for the ground truth and the predicted bounding box. So, a value of 1 would be perfect overlap.

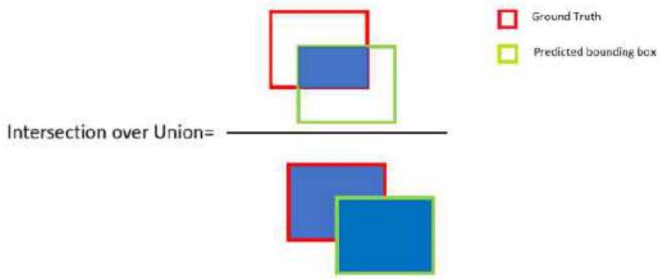


Fig 8. Intersection over Union

With a threshold value the prediction can be expressed as positive or negative. Let us say IoU threshold is set to 0.5, in that case:

- if IoU ≥ 0.5, classify the object detection as True Positive (TP).
- if IoU < 0.5, then it is a wrong detection and classify it as False Positive (FP).
- When a ground truth is present in the image and model failed to detect the object, classify it as False Negative (FN).
- True Negative (TN): TN is every part of the image where we did not predict an object. This metrics is not useful for object detection, hence we ignore TN.

Set IoU threshold value to 0.5 or greater. It can be set to 0.5, 0.75. 0.9 or 0.95 etc.

C. Losses

$$L(\{\text{pi}\},\{\text{ti}\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \qquad \textbf{(5)}$$

i *= index of an anchor box in an mini batch*
pi *= probability of object being present in given ith index anchor box predicted by model*
p*i *= object being present in given ith index anchor box specified in the Label (1= positive anchor, 0=negative anchor)*
ti *= predicted bounding box for object being present (only for positive anchor and do not care for negative anchor)*
t*i *= ground truth of bounding box for ith positive anchor*
Lclassification *= classification loss for object present or not*
Lregression *= regression loss for bounding*
{pi} *= output of classification layers*
{ti} *= output of regression layers*

As referred in earlier section object detection deals with two objective viz classification and localization classification deals with whether object is present or not whose loss function is given by 1/Nclases * Lcalsification (pi,p*i) . Localization deals with regression loss given by λ 1/Nregression*Lregression(ti,t*i).Here , 1/Nclases and 1/Nregression is used to normalize and is weighted by a balancing parameter λ used for balancing the trade of between the two losses. Nclasses is equal to mini batch size here we have taken 64, normalized by the number of anchor locations (i.e., Nregression ~ 2, 400). Regression box losses are only calculated against the positive anchor boxes.

D. Final Model outcome

The trained models are tested over the test samples and their performance is evaluated for a few parameters such as precision, recall and IoU.

TABLE.3. EVALUATION METRICES

Objectives	**Models used**	**Metrices**		
		Avg. Precision	**Avg. Recal**	**Avg. F1 Score**
Object detection	FasterRCNN-InceptionV2	0.453	0.678	0.543
	SSD_MobileNet_V2	0.352	0.453	0.3961
	SSD_InceptionV2	0.384	0.461	0.418
Vehicle number plate detection	FasterRCNN-InceptionV2	0.749302	0.689	0.7177
Vehicle Model type detection	VGG16	0.481	0.514	0.496

VIII. CONCLUSIONS

While experimenting with a few models used for vehicle detection and vehicle registration number detection and make classifier the results observed as Faster RCNN with Inception v2 and VGG16 proved to having considerably better precision, recall and F1 scores and higher accuracy against other counterparts. Table 3 gives a tabular description of all the evaluation metrices.

Faster RCNN with inception v2 model is trained for 12000 epochs with momentum optimizer 0.09 and learning rate 0.00002.

IX. REFERENCES

[1] MYADVO TECHSERVE PRIVATE LIMITED, “3-minute waiting rule at tolls plazas on national highways,” *Myadvo.in*. [Online]. Available: https://www.myadvo.in/blog/waiting-rule-at-tolls-plazas-on-national-highways/. [Accessed: 24-Feb-2021].

[2] C.-L. Huang and W.-C. Liao, “A vision-based vehicle identification system,” in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, 2004.

[3] A. Sethi, “Object Detection using the TensorFlow API,” *Analyticsvidhya.com*, 07-Apr-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/build-your-own-object-detection-model-using-tensorflow-api/. [Accessed: 24-Feb-2021].

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv [cs.CV]*, 2014.

[5] R. Bose and S. Kumar, “Hand gesture recognition using faster R-CNN inception V2 model,” in *Proceedings of the Advances in Robotics 2019*, 2019.

[6] “Dr. Rushikesh P Borse,” *Mitaoe.ac.in*. [Online]. Available: https://www.mitaoe.ac.in/school-of-electrical-engineering-Dr-Rushikesh-Borse.php. [Accessed: 24-Feb-2021].

[7] N. Darapaneni *et al.*, “Computer vision based license plate detection for automated vehicle parking management system,” in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2020, pp. 0800–0805.

[8] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv [cs.LG]*, 2014.

[9] V. Bushaev, “Understanding RMSprop — faster neural network learning,” *Towards Data Science*, 02-Sep-2018. [Online]. Available: https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a. [Accessed: 24-Feb-2021].

[10] Keras Team, “Keras: the Python deep learning API,” *Keras.io*. [Online]. Available: https://keras.io/. [Accessed: 24-Feb-2021].

[11] “Tensorflow api - Google Search,” *Google.com*. [Online]. Available: https://www.google.com/search?q=Tensorflow+api&oq=Tensorflow+api&aqs=chrome..69i57j0l4j69i60l3.2733j0j4&sourceid=chrome&ie=UTF-8. [Accessed: 24-Feb-2021].

[12] “VGG16 - convolutional network for classification and detection,” *Neurohive.io*, 20-Nov-2018. [Online]. Available: https://neurohive.io/en/popular-networks/vgg16. [Accessed: 24-Feb-2021].

[13] S. Sharma, A. Sasi, and A. N. Cheeran, “A SVM based character recognition system,” in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2017, pp. 1703–1707.

[14] J. C. Nascimento, A. J. Abrantes, and J. S. Marques, “An algorithm for centroid-based tracking of moving objects,” in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, 1999.

[15] “OpenCV: Image Thresholding,” *Opencv.org*. [Online]. Available: https://docs.opencv.org/master/d7/d4d/tutorial_py_thresholding.html. [Accessed: 24-Feb-2021].

[16] N. Darapaneni, B. Krishnamurthy, and A. R. Paduri, “Convolution Neural Networks: A Comparative Study for Image Classification,” in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 327–332.

A Novel Dataset Based Algorithm for the Fastest Shipping by Allocating Nearest Supplier

Md Imtiaz Ahmed
Computer Science and Engineering
Daffodil Institute of IT
Dhaka, Bangladesh
imtiaz@diit.edu.bd

***Abstract*—Algorithm is an efficient way to solve issues or minimizing efforts in an existing system by imposing new technique or rules on the existing system. Technology creates new platform or elements by which a vast majority of business growing and new idea or algorithm can be more effective if one can attain the algorithm on its system. Online e-commerce platform is increasing and due to COVID-19 the world has been mostly depending on it. Customer always expects fastest delivery when customer uses online platform and the eagerness of customer's expectation always impacts on online platform. In the thought of fastest delivery for the manufacturer or B2C oriented business a fastest shipping algorithm has been proposed where the algorithm will use the dataset of its existing system and will calculate the shortest path to find the nearest located outlets. It will gradually support the manufacturers or brands by identifying the nearest distance outlets for fastest delivery of products and at the same time it will minimize the efforts of work, time and cost.**

Keywords—Dataset based Algorithm, Dijkstra's Algorithm, E-commerce, B2C, Fastest shipping, Allocating nearest supplier.

I. Introduction

Strategies changes in accordance with the technological innovation where new methods always get the priority for solving critical situations. Algorithms are the way where the new possible solution for a problem can be stated. E-commerce business is developing rapidly and due to the COVID-19 people mostly depends on e-commerce. As people more depends on e-commerce so the shipping issues also arise at the same time. How to ship a product in the fastest way from the seller that are mostly identical issues nowadays for getting good reviews of the system.

Multiple sellers can sell the same product as the brand can allow multiple sellers but multiple sellers won't be in the same address or districts or states. Amazon is one of the top e-commerce platforms where multiple sellers can sell the same product as sellers have the authorization to sell the same product and for that, they normally get the approval from the brand and use the brand or manufacturer labelled UPC to list their items. Ali Express is another popular platform where the same products can be sold by multiple sellers.

A product price is distinct for all customers where the brand has rules to sell the same product. How a seller can understand how they will get the order as many of the seller using the same links and same prices. Sellers draft thinking is that they need to put discounts so that they can generate sales. Although sellers can think we can get sales by getting positive reviews and feedbacks. Algorithms will help the platform owner or the brand to find out who will ship the products and who will get the customer order.

Fastest shipping is one of the important factors to keep customers faith in e-commerce. In the fastest shipping way how a seller will be selected from the pool as multiple sellers selling the same products are the key concerns in this paper. As e-commerce has the priority of giving the opportunity to have multiple sellers to use the same platform and the same manufacturer or brand can have a different physical warehouse or showroom.

Algorithms implementation is a possible solution for the fastest shipping and allocating who will ship the products means from whom the product will be purchased. If one brand has multiple storefronts in different places and if the brand uses one e-commerce for its online platform then the algorithm will help them by defining from which place the product will be shipped. At the same time the algorithm if successfully implemented in software or in machine learning it will automatically predict the fastest warehouse or storefronts from where the fastest shipment is possible.

Building an efficient algorithm is necessary to sorting out these issues as it can help the platforms owner or brand owner to achieve more success in regards to the fastest shipment. The main aim of this paper is to create an algorithm which can be the best fit for both popular e-commerce platform like Amazon, Ali express, eBay etc. as well as for the manufacturer or brands. After implementing these algorithms they can have faith that their system or procedures offer the fastest shipment to their customers.

II. Background Analysis

An algorithm is a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation [1][2]. Algorithms are always used for giving solution for specific issues which can be implemented with any languages. It can be used for calculating, data processing, sorting complex problems etc. An algorithm can be expressed or can be executed in a finite set of time or space and it is a well-defined formal language [3].

Shipping collision issues for multiple shipments with four simulation schemas have been proposed [4]. Multiple researchers focus on how to design an effective model for the marine shipment procedure. A Morphing Evolutionary Algorithm has been proposed which will effectively subside the manual works and the performance improvement is 11.67% [5]. Shipping collision and manual works always delay the shipping performance and hence customer gets dissatisfied.

Consumers are the key to e-business and the fastest shipping is necessary to get the reliability or faith of consumers. Consumer adaption of online shipping a vital issue for study and consumers intention depends on it to buy online [6]. Consumer's faithfulness depends on the timely shipping or delivery of a product. For that many models have been imposed or proposed for the fastest delivery [7]. Shipping methods or timely shipping is a great key to become successful in the e-commerce sectors.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

In Amazon, a product can be sold by multiple sellers where they can use the same product links for selling [8]. A product of Amazon has been checked and finds the similarity that it has multiple sellers who are using the same product link for selling their products [9]. As multiple sellers use the same product link for selling their products so that an algorithm needs to identify which seller will get the priority for getting an order. The algorithm will help the platform owner to detect the identical and logical seller who will get the priority.

An observation has been taken into account of a renowned brand named "UGG". It has multiple stores in the USA and Canada where it uses the same online platform to get online sales and running its e-commerce sector [10]. UGG has multiple stores so that they can do the fastest shipment in the USA and Canada based on customer addresses. Definitely, an algorithm can help them find the fastest delivery to its customers so that no manual works needs to notify the outlets manually.

III. Gap or Current Methodology Analysis

Giant platforms like Amazon have their own recommendation or learning algorithms by which they can predict or assign sellers for a specific order. As multiple sellers sell the same product, they must have their own way to determine the sellers with some specific criterion. For brands or manufacturers who don't have any specific parameter to select the method for those, an algorithm can be introduced and by observing these issues a novel algorithm approach has come to a solution that others can implement on their portal.

UGG brands have many outlets in the USA but not in all the states [10]. New York, California, New Jersey etc. states have their outlets but Colorado State doesn't have their outlets. When a customer from Colorado states tries to purchase a product from them then it will be time-consuming for an admin to find the nearest outlets to ship that customer's product. In this parameter, if an algorithm can be implemented properly on its website like Amazon so that the manufacturer can easily notify the outlets that are located in the nearest state. After notifying the outlets will start shipping the product on their own and shipment will be fastest.

There is a number of consecutive ways to ship a product to the customer like by Air, by Parcel or Transport services. One can say that each state has its own Airports and Air shipment services then why should one need to implement the algorithms. The main focus of the proposed novel algorithm is to ship the products to the customer with less time and cost minimization on the shipping. Air shipment cost is pretty high but if the system with the implemented algorithm can show the nearest possible outlets with fewer charges and the fastest shipment then customers will be satisfied as well as the physical works will be reduced.

The manufacturer can have their own warehouse for the online orders, where they likely to keep all the products that they are intended to ship to the customers from online. Expenditure on the warehouse normally high and the shipment procedure from the same warehouse can rigorously follow the same rules like shipment time to consumers, shipment costs to consumers, return policies to the consumer etc. The approached algorithm can help the manufacturers to maximize their sales through each outlet and warehouse expenditure will be low and manpower will be less than in previous.

IV. Proposed Algorithm

UGG brand has a list of outlets in nineteen states in the USA [10] where the country has in total of 50 states [11] [12]. Customers from other states can use the UGG vendor website as a medium for their purchase and customers always expect to get the order as soonest as possible. It must be noted that customers from the same states where the stores are located can order through the portal as well and at that time customers order will directly check their state located outlets first.

In the proposal, a platform needs to be introduced or need to be implemented or if the manufacturer already has an online ERP (Enterprise resource planning) [13] software they can use it. Cloud inventory management software tool like the Finale Inventory software [14] can be used by the manufacturer or the manufacturer can build their own portal where they can easily locate which outlets has which products, colors or sizes. As the tool finale inventory has the features to obtain the store name for each different quantities or items [14] so that the manufacturer can easily locate and can grab their updated store database from the software each time when an order normally placed by the customer. Finale Inventory has the option to download the manufacturer's products product database and it's normally come as a CSV file [15].

Small manufacturers who are still in the development process can use the software for their products inventory, POS, accounting etc. and at the same time they can use the proposed algorithm for their fastest and automated tasks to assign outlets for shipping to the online customers. Manufacturers like UGG, Ray-ban, CK etc. can have their own tool but their tools need to have the product's location so that the products lookup database can be easily trained in the algorithm process. An important factor that needs to be noted that the finale inventory has the POS (Point of Sales) [16] so that when an item physically sold by the outlets the system will automatically deduct that quantity from their database so that the proposed system will always get the accurate score for pointing the exact outlets for the shipment process.

A. Datasets for the proposed Algorithm

The algorithm needs data for the successful operation and the structure of an algorithm depends on which data algorithm will work. For the proposed algorithm two datasets will be needed for the successful operation.

a) Imported Products Dataset

"Products imported dataset" can be imported from the ERP (Enterprise resource planning) software [13] or can be imported from the Finale Inventory web tools [15]. It must have the products name, style no, color name, sizes, availability, location of the product etc. The location of the product is mandatory as the outlet's names normally are used in this column. This dataset can be downloaded in a timely manner so that each time when order placed it will be automatically downloaded for utilizing through the system using the algorithm. The orientation of the dataset table can look like below where one product's secondary data has been listed in table I:

TABLE I.

Item ID	Style no#	Product Name	Color	Size	In stock	Location	State Name
120001	1016222	CLASSIC MINI II BOOT	Black	7 US	10	Citadel, Tanger	California
120001	1016222	CLASSIC MINI II BOOT	Black	7 US	5	Orlando	Florida
120001	1016222	CLASSIC MINI II BOOT	Black	7 US	2	Riverhead	New York
120002	1016222	CLASSIC MINI II BOOT	Black	8 US	10	Chicago	Illinois
120002	1016222	CLASSIC MINI II BOOT	Black	8 US	5	Premium	Nevada
120003	1016222	CLASSIC MINI II BOOT	Chestnut	7 US	2	Seattle	Washington

TABLE II.

State name	California	Connecticut	Florida	Georgia	Hawaii	Illinois	.....	N^{19th}
Alabama	1381.45	1851.01	7910.78	10661.75	7655.61	1252.03	....	-
Alaska	909.83	1057.84	9057.76	9380.39	7545.86	474.66	....	-
Arizona	1614.01	2035.94	8372.03	10680.27	7137.65	977.77	....	-
Arkansas	1808.23	2139.70	8970.02	10438.87	6691.20	715.39	....	-
California	1	478.82	8317.62	9320.27	8411.70	1273.96	....	-
Colorado	2452.26	2724.71	9529.71	10547.76	5992.89	1216.34	....	-
…..	....	....	....	....	....	....	....	-
N^{50th}	-	-	-	-	-	-	-	-

The products name on the website must be the same as the system software has. The table above has the product that is used as an example of the product table dataset [17]. Two different outlet's names have been addressed in the table to give an explanation that those outlets have the same product but they can have quantities in a random order like Citadel has 7pcs and Tanger has 3pcs. When any of those outlets have been sold out then one outlet name will show in the location column. It has been also noted that when an online order comes to an outlet then the outlet will deduct the order quantity from that outlet's POS in-stock quantities. The idea behind the above table is mostly appropriated with the Finale Inventory Software [15] as the dataset of the software are same. If any state outlets don’t have quantities for a specific product it won't show available in this dataset whether if other state outlets have products for the same product then their available stock will show in the dataset.

b) Dataset of distance from each state

Each of the states has a specific distance from the other states in the same country and very rarely the distance from each other matches exactly. But each state contains a number of cities, where the cities have different distances between each other. In this paper, the distance between the states only measured and the dataset has been prepared to have one(1km) distance when the distance of the same state will be calculated. Let's say "X" state distance from the 'X" state distance is one(1km). One can put the distance from the same state as zero (0km) but if one wants to train the dataset to different AI or Machine Learning algorithms then value one (1km) will be good for finding the shortest path. The dataset table will contain 19 states name on the column and on the row, there will be 50 states name. As the manufacturer UGG has outlets in 19 different states in the USA [10] and the country USA has in total 50 states [11][12]. Distance from each state is calculated in kilometers (km) and the distance was found from the web portal named "Distance from to" [18].

B. Algorithm

- Order id and the details of the order will be acknowledged
- From the order the "customer's lived state name" and "item/product id" will be picked out
- Whether the customer's state name matches with any of the states that are on the 1st column of the dataset of distance from each state which must be checked
- If the customer state name matches with the state name that is on the dataset of distance from each state, then an array will be created for the shortest path order to each state from the customer state name by using the "Dijkstra's algorithm" which will use the dataset of distance from each state.
- It will check whether the located nearest state has the ordered product id or not by checking the in-stock quantity column of the dataset of imported products.
- If the located state outlet has quantities of the ordered item/product id, then the outlet will be notified and an order will be placed for that outlet.

If more than one store located in that state outlet has the same product then it will randomly choose one outlet from the imported products datasets where normally the outlet names are listed.

- If the mapped state doesn't have the product id or quantities, then it will look into the shortest path state array for the next state name and will move to step 4.
- If the state name doesn't match with the 1st column of the dataset of distance from each state then it will be marked as an international address and the international shipping team will be notified.

C. Pseudocode

The Pseudocode of the proposed algorithm are listed below:
Let's say the order dataset has the below properties,

```
Customer Order Dataset {
    Customer-name ();
    Customer-address ();
    Product details ();
    Paid amount ();
}
Let's take,
Imported Product dataset = x;
Dataset of Distance from each state = y;
State name from customer-address () = s;
Item-id from customer ordered product details () = p;
s ←customer-address ();
p ← product-details ();
algorithm (x, y) {
    If {
        s → y;
        For each state in y, create the ordered shortest path array from s using Dijkstra's Algorithm:
        a = [Sn]
        for each state Sn of a[Sn]:
            p → x;
            notify the ordered first Sn →x;
    }
    else{
        notify the international shipping team.
    }
}
```

D. Flowchart

The flowchart of the proposed algorithm are given below:

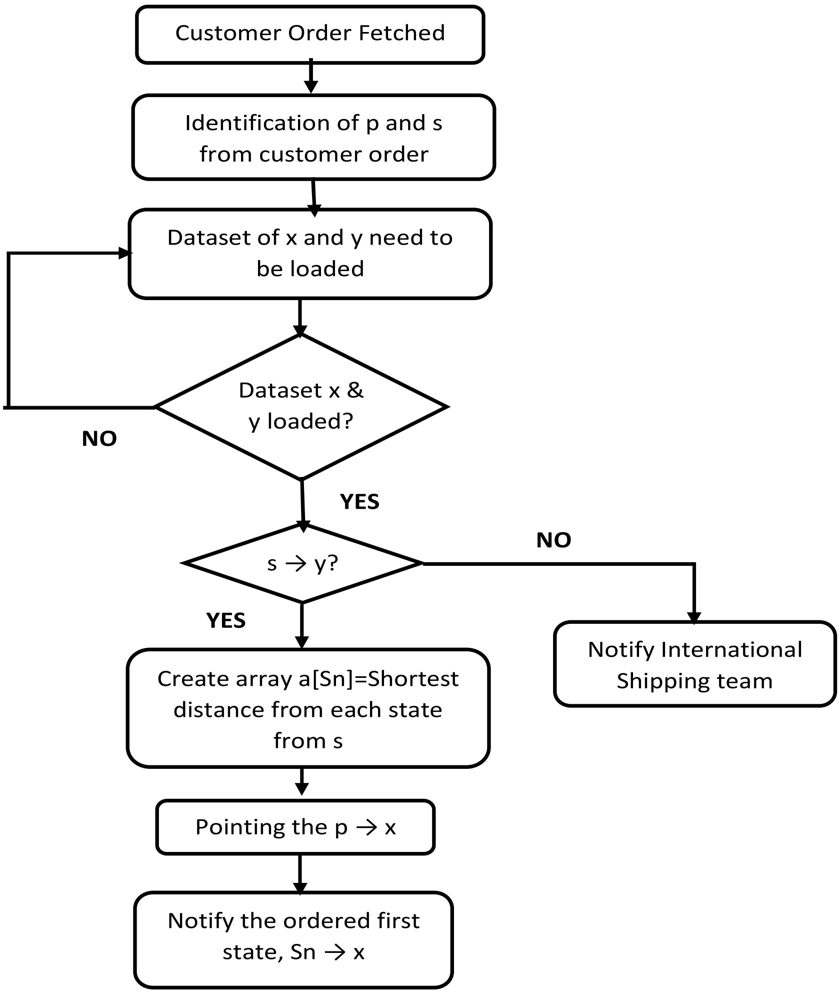

Customer Order Fetched
Identification of p and s from customer order
Dataset of x and y need to be loaded
Dataset x & y loaded?
NO
YES
s → y?
NO
YES
Create array a[Sn]=Shortest distance from each state from s
Notify International Shipping team
Pointing the p → x
Notify the ordered first state, Sn → x

Fig. 1. Flowchart of the proposed algorithm

V. The complexity of the Proposed Algorithm

The proposed algorithm has complexity in some criterion as it depends on the dataset that needs to be fetched by the algorithm. Some of the errors that one can face when trying with the algorithm are listed below:

A. *Outlet allocation*

By observing the dataset of imported products, it shows that one specific state's multiple outlets can have the same product and there are not any specifications on how many quantities each outlet owns. For this issue, the outlets will be chosen randomly but if a customer order a big amount like 10pcs but the one state's none of the outlets has that amount of quantity, then there will be bugs. The algorithm will be well suited for customers like Amazon where normally a customer doesn't order more than 2-3pcs at a time.

B. *Multiple items of an Order*

A customer placed an order which has multiple items like the customer ordered a shoe and a top. The algorithm can solve the multiple items of that order by taking each of the ordered item id's at once to use them through the algorithm. But still, there is one order id for that order so that multiple different outlets will be notified for the same order. As the algorithm needs to do the same tasks twice so that it will be time-consuming as well.

C. *Cloud Service*

The algorithm will work when an order will be placed on the website or in the system so that from the ordered data the algorithm will start manipulating to identify the potential outlets for shipping the order. Cloud server needs to use for the ideal operation of the algorithm. One can do that in their desktop system but for the fastest manipulation cloud will be much better.

D. *Dataset Loading Issues*

The algorithm mainly depends on two specific datasets, those are the "imported products dataset" and "the dataset of distance from each state". The dataset of distance from each state can be loaded once as this dataset won't having any changes but the "imported products dataset" always changed in a manner that the outlet's physical sales will be deducted and new items can be adjusted to the outlets so that if dataset won't be loaded properly every time then wrong outlet's will be notified.

VI. Open Challenges

The proposed algorithm has built with the idea for fastest shipping to the consumers when a manufacturer can use the same website for selling its products where there can be a number of outlets in different states in a country. It will be very identical if one manufacturer can implement this algorithm to their web portal where a cloud server will be the most appropriate for executing the algorithm. Though it has some limitation or challenges for implementing or issuing the algorithm, it will be good for practicing the algorithm and necessary steps can be taken when facing more issues or the issues that listed in this paper. The machine learning approach will be the best fit for the proposed algorithm, as the idea behind the algorithm has the initialization to work with two specific datasets. Machine learning techniques will be a challenge to implement and if someone can build any other edition for this algorithm, it will be taken into account. As the algorithm proposed by observing the complexity, cost, time for a manufacturer for running an e-commerce portal, more new features can be taken into account for more effectiveness of the proposed algorithm.

VII. Conclusion

The goal of the algorithm normally minimizes the tasks or time that take earlier a good amount of time and algorithm performs or normally creates for some specific issues. In this paper the algorithm that proposed are created mainly for the B2C platforms where a manufacturer can easily flourish their online business by allocating different owned outlets for the shipping. Customer always wait for the products arrival and in that circumstance the algorithm will find the shortest distance outlets and that allocated outlets will be the candidate for shipping the products if only it has the available stock. By implementing the proposed algorithm manufacturer won't then have to think for the physical warehouse for keeping the online orders quantities and it will help them by minimizing the cost and time. As outlet's person can do the shipment and can maintain the whole system so that manpower will be less. The manufacturer can also notice how their physical outlets sales and how much outlets have in stock so that supply chain management will be more accurately.

References

[1] "The Definitive Glossary of Higher Mathematical Jargon — Algorithm". Math Vault. August 1, 2019. Archived from the original on February 28, 2020. Retrieved November 14, 2019.

[2] "Definition of ALGORITHM". Merriam-Webster Online Dictionary. Archived from the original on February 14, 2020. Retrieved November 14, 2019.

[3] "Any classical mathematical algorithm, for example, can be described in a finite number of English words" (Rogers 1987:2).

[4] "Simulation of multi-ship collision avoidance with avoidance key ship algorithm" by Y.-Z. Xue, Y. Wei and M. Sun, published on January 2014.

[5] "Ship Design with a Morphing Evolutionary Algorithm" by Ciel Thaddeus Choo; Joo Hock Ang; Simon Kuik; Louis Choo Ming Hui; Yun Li; Cindy Goh, published on July 2020

[6] "An Exploratory Study of Consumer Adoption of Online Shipping" by Songpol Kulviwat, Ramendra Thaku, Chiquan Guo published on July 2008.

[7] "Online ship rolling prediction using an improved OS-ELM" by Chao Yu; Jianchuan Yin; Jiangqiang Hu; Anran Zhang, published on July 2014.

[8] "How to sell the same product that another seller is already selling (same UPC code)?" by Amazon forum on August 2020. Link: https://sellercentral.amazon.com/forums/t/how-to-sell-the-same-product-that-another-seller-is-already-selling-same-upc-code/680979.

[9] "UGG Women's Classic Clear Mini Ankle Boot" by Amazon,Link:https://www.amazon.com/dp/B082HHWHW1/

[10] "UGG Outlet stores in USA and Canada", Link: https://www.ugg.com/outlet-stores.html

[11] "Common Core Document of the United States of America: Submitted With the Fourth Periodic Report of the United States of America to the United Nations Committee on Human Rights concerning the International Covenant on Civil and Political Rights". U.S. Department of State, via The Office of Website Management, Bureau of Public Affairs, on July 9, 2017. Link: https://2009-2017.state.gov/documents/organization/251864.pdf

[12] "U.S. Insular Areas: application of the U.S. Constitution" (PDF). Government Accountability Office. on November 1997. Link: https://www.gao.gov/archive/1998/og98005.pdf

[13] "Applications or defination of Enterprise Resource Planning (ERP)" by Oracle. Link: https://www.oracle.com/erp/what-is-erp/#link1

[14] "Cloud Inventory Management Software for your Growing Business" by Finale Inventory, Link: https://www.finaleinventory.com/cloud-inventory-management-2

[15] "Bulk Import Initial Stock Quantities" by Finale Inventory. Link: https://www.finaleinventory.com/basic-stock-import

[16] "Integrations for Point of Sale (POS) Systems With Finale Inventory" by Finale Inventory, Link: https://www.finaleinventory.com/integrations-pos

[17] "CLASSIC MINI II BOOT" by UGG, Link: https://www.ugg.com/women-boots-classic-boots/classic-mini-ii-boot/1016222.html

[18] Distance from each state measured by the Distance from to website. Link: https://www.distancefromto.net/

Predicting Hospital Beds Utilization For COVID-19 In Telangana, India

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern University Illinois, USA
darapaneni@gmail.com

Mahita GM
Student – AIML
Great Learning
Hyderabad, India
mahigm7@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Pune, India
anwesh@greatlearning.in

Sateesh Kumar Talupuri
Student – AIML
Great Learning
Hyderabad, India
satishtvs@diagnosmart.com

Vasundhara Konanki
Student – AIML
Great Learning
Hyderabad, India
vsndhr.vasu@gmail.com

Shruti Galande
Student – AIML
Great Learning
Hyderabad, India
shrutigalande264@gmail.com

Chiru Hasini Tondapu
Student – AIML
Great Learning
Hyderabad, India
chiruhasini1996@gmail.com

***Abstract*— Our study aims to predict the requirement of isolation beds, oxygen beds, and Intensive Care Unit (ICU) beds in both government and private hospitals for COVID-19 patients in Telangana, India. Using epidemic modeling and time series analysis techniques, we can predict the future daily COVID-19 cases and active cases from historical data of confirmed COVID-19 cases. From the historical data analysis, we can also establish the pattern of hospitalized cases from active cases and also the pattern of various categories of bed occupancy numbers. The pattern is represented by minimum, maximum and average percentages. We have observed the second wave of COVID-19 in many European countries and is spreading more rapidly than the first wave. In India, we must be prepared to address the second wave as it might be more rampant since the lockdown is eased in most of the cities. Since traveling and gatherings are getting back to normal, significantly larger population might be susceptible and would need immediate hospitalization. We presented a process/methodology to estimate bed occupancy numbers from the prediction of COVID-19 positive and active cases using modified Susceptible-Infected-Recovered (SIR) model for the contagion and FB-Prophet model for time series analysis. We used official data from the Government of Telangana bulletins for the COVID-19 pandemic up to December 31st, 2020 for our process/methodology. SIR modeling is more intuitive and explainable but requires a lot of trial and error and assumptions. The FB-Prophet prediction process is simple and accuracy is also better compared to SIR modeling.**

Keywords*— *COVID-19, hospitalization, beds, prediction, models, Telangana, SIR model, ARIMA, Facebook (FB) Prophet,

I. INTRODUCTION

COVID-19 is an infectious disease known to cause severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). The World Health Organization (WHO) has declared the novel Coronavirus (COVID-19) as a global pandemic on 11th March 2020.

In India, the first positive case of COVID-19 was detected on 30th January 2020, in Kerala. The Government of India implemented a complete nationwide lockdown on 24th March 2020 for 21 days, and was continued until 31st May 2020 [12]. The purpose of the nationwide lockdown was to contain the spread of the COVID-19 disease so that the Government could scale up the production of the testing kits for COVID-19 and personal protection equipment (PPE) for the health workers and build extended health infrastructure to accommodate more COVID-19 patients [1][6]. On 3rd June, in India, 952 dedicated COVID hospitals with 1,66,332 Isolation beds, 21, 393 ICU beds, and 72,762 Oxygen supported beds were available. 2,391 dedicated COVID Health Centers with 1,34,945 Isolation beds; 11,027 ICU beds and 46,875 Oxygen supported beds have been operationalized [1].

In the state of Telangana, the first positive case of COVID-19 was detected on 5th March, 2020. As on 11th January 2021, the total positive cases in Telangana are 2,90,008 wherein 60.64% of males and 39.37% of females were affected. There are 281 dedicated COVID hospitals consisting of 4,330 ICU beds and 8,428 Oxygen beds [7].

By now, the world has been made aware of the impact COVID-19 had on the societies and their healthcare systems. Already at the start of the first wave of the pandemic, it became apparent that the capacity of hospital beds could come under great pressure. The morbidity of COVID-19 drastically increased the demand for hospital beds.

Disasters typically come with a sudden influx of unforeseen patients, which almost instantly pushes the boundaries of a hospital's capacity. Lack of bed capacity further increases that burden [2][3][4].

In order to prevent such overflow, in India, on 24th March 2020, a strict lockdown was imposed due to COVID-19 wherein there were only 519 confirmed cases across the country. After 53 days of nationwide lockdown, on 16th May 2020, the confirmed cases shot up to 85,950 in India [13]. In this timeframe, it was observed that the growth rate of the cases had slowed down by 6th April, to the rate doubling every 6 days, and by 18th April [15], to a rate doubling every 8 days. Lockdown 2.0 was imposed between 15th April 2020 till 3rd May 2020. During Lockdown 3.0 and 4.0 until 30th May 2020, the Government segregated all the districts into three zones based on the spread of the virus- green, red and

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

orange - with applied relaxations [13]. Due to this nationwide lockdown over a period of time by means of quarantine, social distancing, hand washing, closure of educational institutions, curfews, wearing a mask etc, the influx in hospitals had decreased comparatively. On 30th May 2020, the Government announced that the lockdown would be lifted from then on, except for the containment zones wherein the lockdown continued till 30th June 2020. From 1st June 2020 till 31st January 2021, there were 8 unlock measures imposed each month which were focused on how restrictions should be eased in a phased manner.

A fragile equilibrium needs to be found between reserving a sufficient number of beds for COVID-19 cases, while also providing sufficient beds for regular, necessary care which cannot be delayed. In order to achieve such balance, predictive models can play an important role, to predict the number of needed beds that should be allocated to the pandemic. We therefore set up forecasting models to predict on each day the needed capacity for different bed types using cases and beds data of Telangana State from Aug 2020 to Nov 2020.

II. MATERIAL

The data is obtained from Telangana Covid19 Healthcare Bulletins [7] from Telangana Government’s website. The daily bulletins available from 1st August, 2020 onwards have information about the number of daily positive cases, total accumulated cases, recovered cases, deaths, active cases, home quarantined cases, total beds in government and private hospitals in the states and how many beds are occupied daily in various categories of beds: regular, oxygen and ICU. From these bulletins, we have collected data from 1st, August, 2020 to 30th, November, 2020 and did analysis on the data.

III. ANALYSIS

A. *Positive, Recovered and Active cases*

Fig.1 shows the daily positive cases and recovered/deceased cases between 1st August, 2020 to 30th November, 2020. The dips in the numbers were observed during Sundays and holidays where the number of tests done were low. The peak number of positive cases around 3000 were during the 4th week of August, 2020.

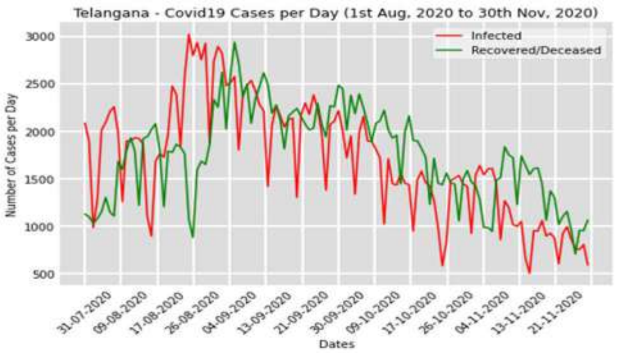


Fig. 1. Daily positive and recovered/deceased cases

Fig.2 shows total accumulated positive cases and recovered/deceased cases. The total positive cases were around 270000 and total recovered cases were around 260000 on the last day of November. There were a total 1460 deaths during the same time [7]. The total active cases reached a maximum of around 32994 during last week of Aug, 2020.

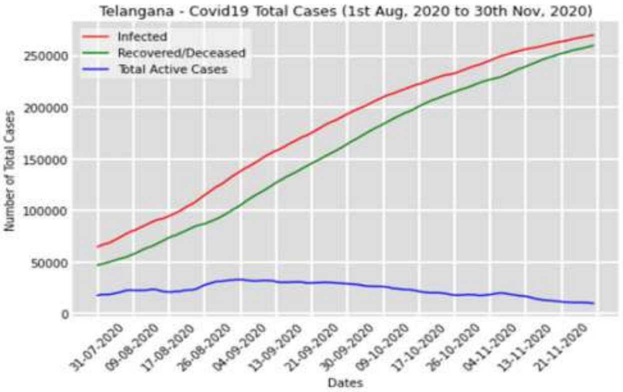


Fig. 2. Cumulative positive, recovered/deceased and active cases

B. *Primary and Secondary contacts testing*

Fig.3 shows the number of primary contacts and secondary contacts tested daily. More numbers of primary contacts are tested. The testing numbers increased during August.

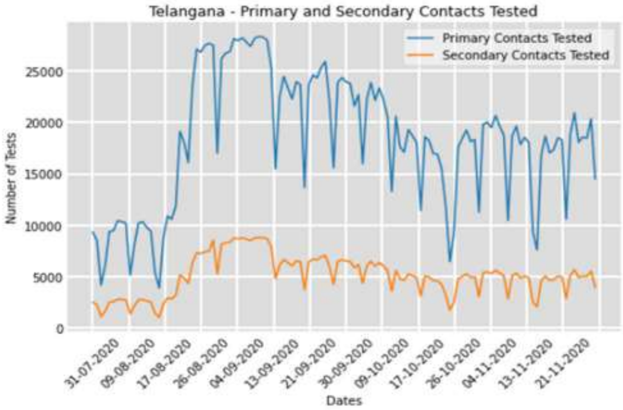


Fig. 3. Primary and Secondary Contacts Testing

C. *Active, Home Isolation and Hospitalized cases*

Fig.4 shows the graph total active cases, total home isolation cases and total hospitalized cases.

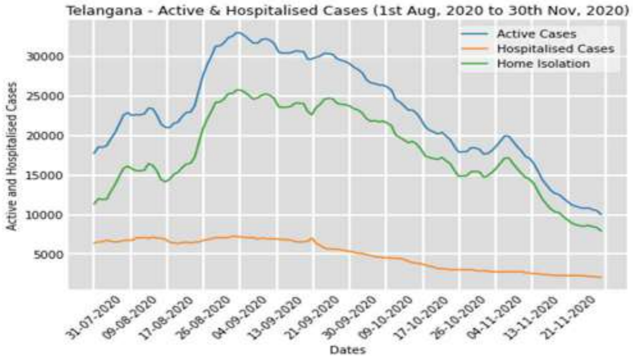


Fig. 4. Total active, home isolated and hospitalized cases

Hospitalized cases are obtained from subtracting the home isolated cases from total active cases. Maximum hospitalized cases were around 7250 during the first week of Sep, 2020.

D. *Various categories beds occupancy numbers*

Fig.5 shows the graph of total beds occupancy numbers of various kinds of beds. The beds are categorised as regular, oxygen and ICU beds. These beds are available in both government and private hospitals. There were around 17000+ beds in both private and government hospitals as on the last day of Nov, 2020. Out of those beds 8500+ were of government and 8500+ were of private. There were 3600+ regular beds, 8500+ oxygen beds, 4800+ ICU beds. Private hospital oxygen beds are occupied most followed by government hospital oxygen beds. As per the protocol/general nature of precautions to be followed, most asymptomatic cases are home quarantined. Symptomatic cases or cases with comorbidities will be referred for hospitalization. Out of all the hospitalized cases, mild symptomatic cases without comorbidities and young age are given regular beds. Old age and cases with comorbidities are given oxygen beds. Very severe cases will be allotted to ICU beds.

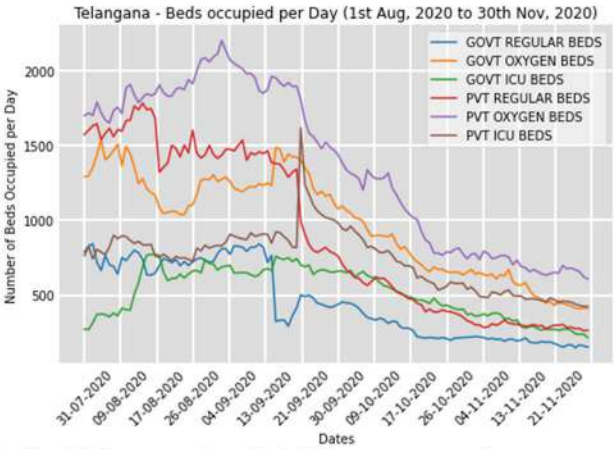


Fig. 5. Various categories of beds daily occupancy numbers

Fig.6 shows the percentage of beds occupancy with respect to total beds.

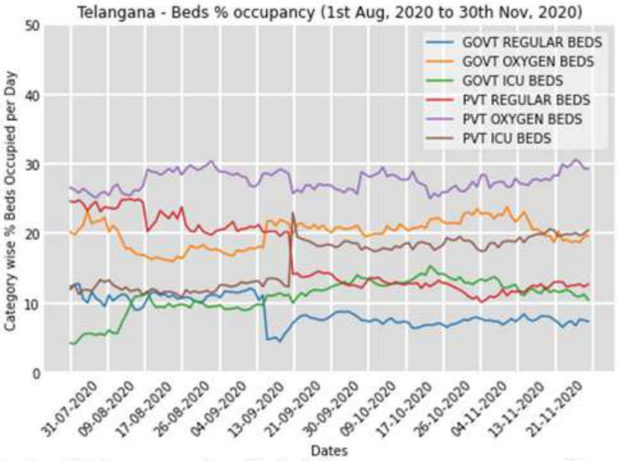


Fig. 6. Various categories of beds daily occupancy percentages with respect to total beds occupied

Private oxygen beds occupancy percentage was between 25 to 30%. Initially private regular beds occupancy percentage was between 20 and 25%, but later that percentage came down to around 10%. The government oxygen beds percentage was varying between 15% to 25%. All other categories of beds occupancy percentages were below 15%.

E. Relationship between active cases and beds occupancy

Fig. 7 shows the approximate relationship (which is estimated from the beds occupancy data) between number of active cases and the beds occupancy pattern.

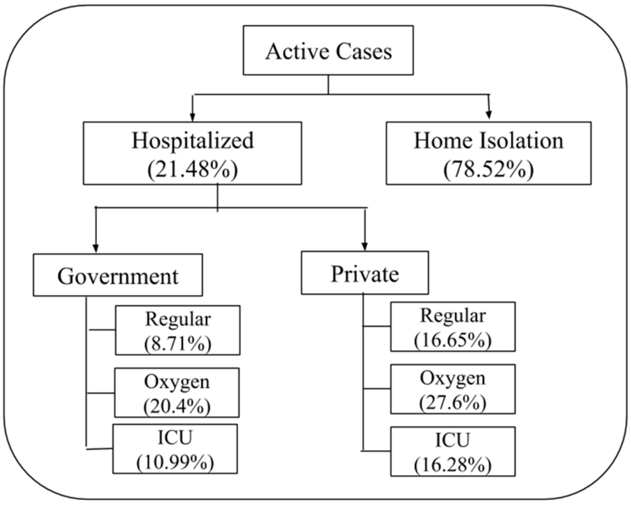


Fig. 7. Estimation Flow chart of Average beds occupancy numbers from number of active cases

Maximum, minimum, average percentages of home isolation, hospitalization and various categories of beds information are given in Table.I.

TABLE I. BEDS OCCUPANCY PERCENTAGE AND THEIR RELATIONSHIP

Percentages	Min	Max	Avg
% Home isolation cases wrt Active Cases	63.8%	86.15%	78.52%
% Hospitalized cases wrt Active Cases	13.85%	36.20%	21.48%
% Govt regular beds wrt Hospitalized Cases	4.46%	12.79%	8.71%
% Govt oxygen beds wrt Hospitalized cases	15.97%	23.85%	20.4%
% Govt ICU beds wrt Hospitalized cases	4.12%	15.35%	10.99%
% Pvt regular beds wrt Hospitalized cases	10.14%	24.95%	16.65%
% Pvt oxygen beds wrt Hospitalized cases	25%	30.62%	27.60%
% Pvt ICU beds wrt Hospitalized cases	10.80%	35.53%	16.28%

IV. METHODOLOGY

From the above analysis, we can conclude that if we have information about the number of active cases, then we can estimate the number of hospitalized cases and further estimate the number of various categories of beds required.

A. Positive, Recovered and Active cases

To estimate the number of beds we shall estimate the number of active cases. The number of active cases are estimated from the number of cumulative positive cases and the number of cumulative recovered cases plus the number of deceased. According to protocol each confirmed case is considered active for 14 days and is quarantined and declared recovered after 14 days unless it is a critical case. As recovery rates are very high, we can consider that most of all confirmed cases will get recovered after 14 days. This gives us the formula for estimating the active cases. Number of active cases is derived by subtracting the number of recovered cases (including number of deceased cases) from the number of cumulative positive cases 14 days ago.

So we formulated the following method for estimating the number of various categories of beds' requirement in future by predicting the number of positive cases and number of active cases.

1) Using the SIR/SEIR model, predict the number of positive cases daily P(t) for the next n number of days from existing data.
2) Using the SIR/SEIR model, predict the number of removed (recovered/deceased) cases daily R(t) for next n number of days from existing data.
3) Accumulate the number of positive cases and removed cases daily PC(t), RC(t) for next n number of days from existing data.
4) Calculate the number of active cases using the formula : AC(t) = PC(t-13) - RC(t). This formula is verified using existing data.
5) Calculate the number of hospitalized cases HC(t) as percentage of Active Cases AC(t). All hospitalized cases require beds. So the total beds requirement is HC(t).
6) Calculate various categories (regular, oxygen, icu (govt, pvt)) of beds as percentage of Total Beds HC(t).

We used the SIR model in our case to predict the number of positive cases and number of recovered cases. We also experimented with Time Series analysis and prediction models like ARIMA and Facebook (FB) Prophet.

B. SIR modeling

In the SIR model [8], there are 3 compartments: Susceptible(S), Infected(I) and Recovered or Removed(R). Susceptible individuals (S) become infected and move into the infected class (I). After some period of time [16] infected individuals recover and move into the recovered class (R). The corresponding equations are given by

$$\frac{dS}{dt} = -\beta . I . \frac{S}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta . I . \frac{S}{N} - \gamma . I \tag{2}$$

$$\frac{dR}{dt} = \gamma . I \tag{3}$$

where S, I and R, are the numbers of susceptible, infected, and recovered individuals in the population. Suppose the unit of time we are considering is days, then

β is the transmission/infection rate and βSI represents the number of susceptible individuals that become infected per day.

γ is the recovery rate and γI is the number of infected individuals that recover per day;

1/γ is the infectious period i.e. the average duration of time an individual remains infected.

An important quantity of any disease model is the reproductive number, R0, which represents the average number of secondary infections generated from one infectious individual. For the SIR model,

$$R0 = \beta N / \gamma \tag{4}$$

where N=S+I+R is the total (constant) population size.

C. ARIMA - Time Series Analysis

Autoregressive Integrated Moving Average (ARIMA) is a simple, flexible and performs well in forecasting tasks [10]. Individual components are AR, I and MA. The AR part simply speaks about autocorrelation using the past data to explain current data. Autocorrelation tries to find out how many lags/periods of time will best predict the current numbers. The 'I' refers to using differencing to attain stationary. Stationary means the statistical properties of the time series such as the median, variance and correlation remain stationary over time. When forecasting a value using a forecasting model, there will be a delta between prediction and the actual value. The MA part can be thought of as a collection of these error terms. These error terms can be incorporated to factor in random fluctuations/irregularities when making our forecast [10].

D. Facebook (FB) Prophet - Time series analysis

An interesting alternative to forecasting problems is using fb-prophet [9] that makes the task of forecasting more accessible and easier to carry out. The great part of fb-prophet is that it automatically detects change points in time series and allows us to factor in hourly, daily, weekly, monthly, yearly trends and even allows us to factor in changes of trends during pre-defined holidays to make our forecast more accurate. Best model can be obtained from turning on weekly seasonality and setting seasonality mode as multiplicative [10]. The forecasted value reveals both the trends, the upper and lower bound forecasts. Prophet is based on decomposable (trend+seasonality+holidays) models. This works best with the time series data that have clear seasonal effects [11]. FB Prophet is strong in dealing with missing data, capturing the shifts in the trend and large outliers. The Prophet framework has its own special data frame to handle time series and seasonality data easily [11],

E. SIR Model - Data fitting

The data collected from Telangana dataset are organized in the form of time-series where the rows are recorded in time (from August to November, 2020), and the three columns are, the total cases I_t, number of infected individuals I and deaths D. Consequently, the number of removals R can be estimated from the data by R = I_t - I - D.

We tried to fit the data to the SIR model by trial and error and visual inspection. The parameters: initial susceptible population, infection rate (beta) and recovery rate (gamma) are adjusted manually through trial and error to fit the data as best as possible to obtain the result graph as shown in Fig.8.

In the graph shown in Fig.8, the red line is actual active cases following the infection curve obtained from the SIR model approximately. The SIR model is used to simulate the susceptible, infection and removed cases from 1st July, 2020 for next 500 days until mid-December, 2021. From Fig.8 we can observe that the infection cases will reduce to zero by March-April, 2021, if there are no second wave of infections.

If there are any new waves of infections, with new data we can adjust the model parameters and fit the data to get a new SIR model, which we can use for predictions.

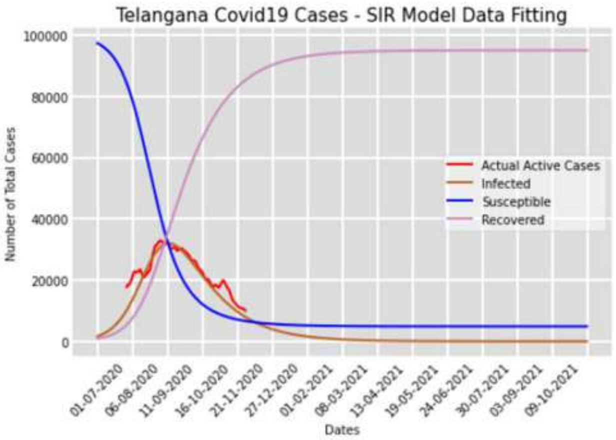


Fig. 8. Active cases data fitting with SIR modeling

The beta (infection/transmission rate) and gamma (recovery/removal rate) varies as shown in Fig.9.

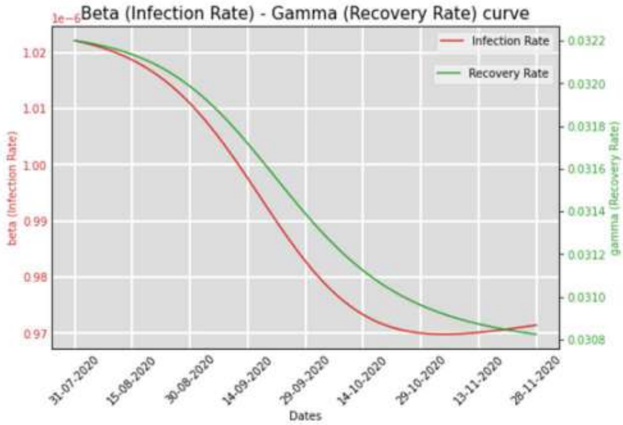


Fig. 9. Telangana Infection and Recovery Rates of SIR Model

V. RESULTS

We can use the SIR model to predict the number of active cases for every day for the next n number of days and from the number of active cases, we can predict the number of hospitalized cases and total number of beds and in turn predict the number of various categories of beds required using the average percentages obtained from existing data which was given in Table. I.

A. *Predicting Active and Hospitalized cases*

Fig.10 shows the number of active cases predicted from the SIR model from 1st December, 2020 to 30th April, 2021. The number active cases predicted for 1st December, 2020 was around 7500+ decreased to around 150+ for 30th April, 2021. Hospitalized cases which are calculated as the percentage of active cases, decreased from 1600+ to 30+ during the same time.

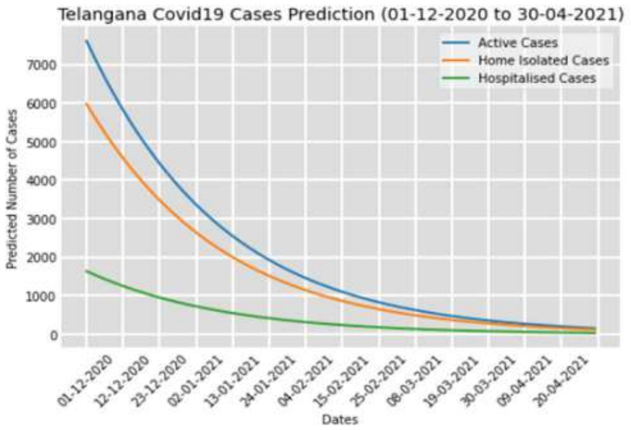


Fig. 10. Active and Hospitalized Cases Prediction from SIR Model

B. *Predicting various categories of beds*

Fig.11 shows the predicted number of various categories of beds from 1st December, 2020 to 30th April, 2021. Private Oxygen beds are generally utilised most and their number has decreased from 450+ beds to single digits.

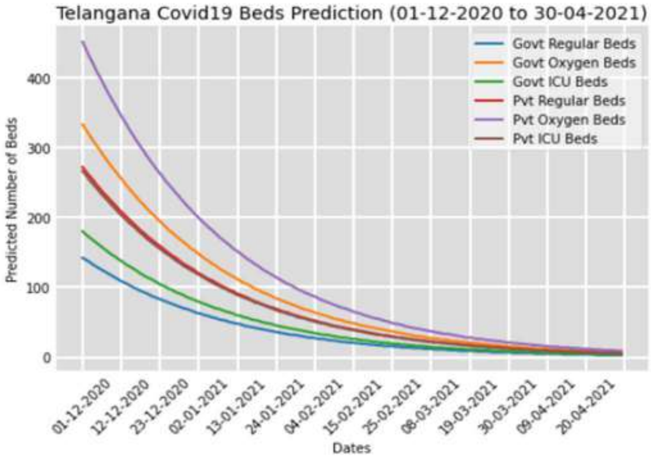


Fig. 11. Number of various categories of beds prediction from total hospitalized cases obtained from SIR Model

C. *Predicting daily positive cases using FB-Prophet*

We also used FB-Prophet for predicting daily positive cases and their trend along with their lower and upper bound as shown in Fig.12.

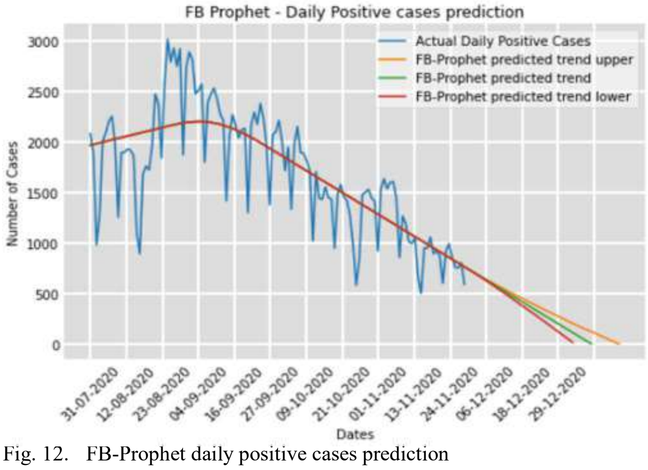


Fig. 12. FB-Prophet daily positive cases prediction

D. *Predicting total active cases using FB-Prophet*

Fig.13 shows the FB-prophet prediction of total active cases and their trend along with their lower and upper bounds

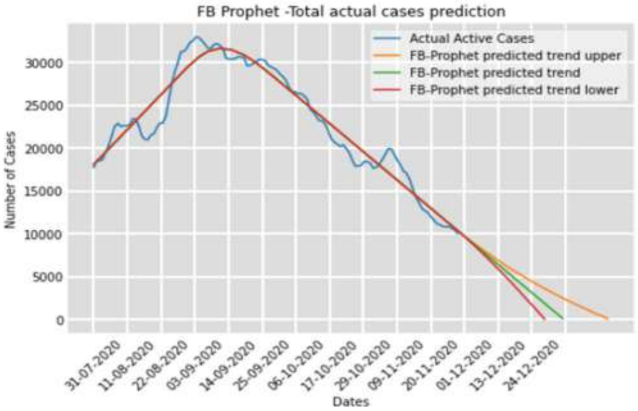


Fig. 13. FB-Prophet total active cases prediction

E. *Accuracy metrics for Predictions*

We used data from 1st Aug, 2020 to 30th Nov, 2020 as training dataset and 1st Dec, 2020 to 23rd Dec, 2020 as test dataset.

SIR Model, ARIMA and FB-Prophet active case predictions and their RMSE values for train and test datasets are shown in Table II.

TABLE II. ACTIVE CASES PREDICTIONS - RMSE VALUES FOR SIR AND FB-PROPHET

Data Set	***SIR RMSE***	***ARIMA RMSE***	***FB PROPHET RMSE***
Train Dataset	3126.61	1887.60	1512.85
Test Dataset	1773.69	14441.98	1611.52

SIR modeling is more intuitive and explainable, but requires a lot of trial and error and assumptions. The FB-Prophet prediction process is simple. The root mean squared error (RMSE) of FB-Prophet predictions is less and more accurate. So FB-Prophet predictions can be used for quick and accurate prediction of daily COVID-19 positive and active cases.

VI. DISCUSSION AND CONCLUSION

Based on Historical data and predictive analysis done so far we have observed that existing government and private hospital beds are sufficient to handle all hospitalized COVID19 cases. Highest occupancy is observed in the Oxygen bed category in both government and private set up.

We have observed second waves in many European countries [14] and the proposed predictive analysis can be used to understand hospital beds availability in future if a second wave occurs in India. This analysis can be applied to analyse hospital beds availability in other states in India.

SIR modeling can be used with extra parameters to simulate lockdown effects, surges, comorbid conditions, festivals/congregations, elections, environmental factors like pollution and weather conditions, air travel restrictions, food habits etc., to approximately predict the infection spread scenarios and be prepared with adequate healthcare infrastructure.

The FB-prophet model also can be used for time series analysis for predicting positive cases more accurately with inclusion holidays, special days etc.

As the vaccination started in India and there is no second wave observed except in a few cities, let us hope present healthcare infrastructure in the country is enough.

VII. REFERENCES

[1] R. Tyagi, M. Bramhankar, M. Pandey, and M. Kishore, “COVID-19: Real-time Forecasts of confirmed cases, active cases, and health infrastructure requirements for India and its states using the ARIMA model,” bioRxiv, p. 2020.05.17.20104588, 2020.

[2] S. M. Moghadas et al., “Projecting hospital utilization during the COVID-19 outbreaks in the United States,” Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 16, pp. 9122–9126, 2020.

[3] V. R. Verma, A. Saini, S. Gandhi, U. Dash, and M. S. F. Koya, “Projecting Demand-Supply Gap of Hospital Capacity in India in the face of COVID-19 pandemic using Age-Structured deterministic SEIR model,” *bioRxiv*, p. 2020.05.14.20100537, 2020.

[4] V. R. Verma, A. Saini, S. Gandhi, U. Dash, and S. F. Koya, “Capacity-need gap in hospital resources for varying mitigation and containment strategies in India in the face of COVID-19 pandemic,” Infect. Dis. Model., vol. 5, pp. 608–621, 2020.

[5] C. Massonnaud, J. Roux, and P. Crépey, “COVID-19: Forecasting short term hospital needs in France,” bioRxiv, p. 2020.03.16.20036939, 2020.

[6] S. Mandal et al., “Prudent public health intervention strategies to control the coronavirus disease 2019 transmission in India: A mathematical model-based approach,” Indian J. Med. Res., vol. 151, no. 2 & 3, pp. 190–199, 2020.

[7] “Media Bulletins,” Gov.in. [Online]. Available: https://covid19.telangana.gov.in/announcements/media-bulletins/

[8] W. N. Arifin, W. H. Chan, S. Amaran, and K. I. Musa, “A Susceptible-Infected-Removed (SIR) model of COVID-19 epidemic trend in Malaysia under Movement Control Order (MCO) using a data fitting approach,” bioRxiv, p. 2020.05.01.20084384, 2020.

[9] Sakib Mahmud, “Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model”. Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/343306716.

[10] Dr. Shikha Gaur, “Global forecasting of covid-19 using ARIMA based FB-Prophet”, International Journal of Engineering Applied Sciences and Technology, 2020, Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467

[11] Indhuja M, Sindhuja PP, “Prediction of covid-19 cases in India using prophet”, International Journal of Statistics and Applied Mathematics 2020; 5(4): 103-106

[12] Sanjana Krishnan, Sahil Deo, Shardul Manurkar, “50 days of lockdown: Measuring India’s success in arresting COVID-19,” *Observational Research Foundation*, 19-May-2020. [Online]. Available: https://www.orfonline.org/research/50-days-of-lockdown-measuring-indias-success-in-arresting-covid-19-66336/.

[13] Wikipedia contributors, “COVID-19 lockdown in India,” *Wikipedia, The Free Encyclopedia*, 15-Jan-2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=COVID-19_lockdown_in_India&oldid=1000471654.

[14] G. J. Milne, S. Xie, D. Poklepovich, D. O’Halloran, M. Yap, and D. Whyatt, “Effectiveness of second wave COVID-19 response strategies in Australia,” bioRxiv, p. 2020.11.16.20232843, 2020.

[15] S. Ramesh and M. Basu, “R0 data shows India’s coronavirus infection rate has slowed, gives lockdown a thumbs up,” Theprint.in, 14-Apr-2020. [Online]. Available: https://theprint.in/science/r0-data-shows-indias-coronavirus-infection-rate-has-slowed-gives-lockdown-a-thumbs-up/399734/.

[16] N. Darapaneni *et al.*, “Coronavirus outburst prediction in India using SEIRD, logistic regression and ARIMA model,” in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2020, pp. 0649–0655.

Dynamic Modeling and Feedback Linearization Control of a 3-D Overhead Gantry Crane System

1st Amin A.M. Fadlalla[1]
Mechanical Engineering Department
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia
https://orcid.org/0000-0002-8444-5052

2nd Mohamed Hassan
Systems Engineering Department
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia
abdeen009@hotmail.com

***Abstract*—The overhead crane systems are important heavy-duty machines that are widely used, especially in construction and industry sectors to transport and lift heavy loads. This paper introduces the mathematical modeling of a 3-D overhead crane. The Euler-Lagrange equation is utilized to extract the dynamic equations of the crane system. The obtained nonlinear crane model has been linearized by applying the feedback linearization technique. Then, a PID controller has been developed and implemented to the obtained linearized crane model. The parameters of the designed PID controller have been tuned by implementing the Differential Evolution (DE) algorithm which significantly improved the performance of the PID controller. The simulation results have been performed using MATLAB/SIMULINK platform. The obtained simulation results clearly showed the impact of applying the Differential Evolution (DE) technique to tune the PID controller gains which provided better performance compared to the manually tuned PID controller gains.**

***Index Terms*—overhead crane, dynamics, PID control, feedback linearization, differential evolution**

I. Introduction

Overhead cranes have extensive use in industry for transportation of heavy loads. To guarantee safe and systematic functioning of the gantry crane, proper design and control is a must. Therefore, several studies were conducted to study the dynamic behavior and control of such system. In order to design a proper control scheme for the crane, the correct dynamic model must be in hand. Many researchers have studied the over head crane system in attempts to optimize it's operation. Among the previous studies, modeling and control of a three-dimensional overhead crane is introduced in [1], where, a new dynamic model of the crane is derived based on a newly defined two-degree-of-freedom swing angles. A decoupled control scheme based on the dynamic model, linearized around the stable equilibrium, is proposed for anti-swing control. It is reported that, the decoupled scheme guarantees not only rapid damping of load swing but also accurate control of crane position and load hoisting for the practical case of simultaneous traveling, traversing, and slow hoisting motions, which was also proven by experiments. Other modeling techniques were carried out such as solid modeling and finite element analysis [2]. Linear controller such as Proportional Derivative (PD) was proposed in [3], the uncertainty in the crane dynamics was compensated by radial basis function neural networks (RBF). Many nonlinear control techniques were attempted, in [4] saturation control approach was used to control a two degree of freedom overhead crane as to control the horizontal position. Sliding mode control [5], [6] was also used to control overhead grane. In addition, partial feedback linearization technique was used to improve the response of a nonlinear controller [7]. LasSalle's invariant set theorem was utilized to develop nonlinear coupling control laws to overhead crane [8]. In [9] Backstepping with sliding mode control was designed for 3D overhead crane, the mathematical model was developed by lagrangian principal. A fuzzy based control laws using inertia theorem was implemented in [10], another enhance fuzzy method was used to minimize the dead zone problem effect. A fuzzy technique in conjunction with proportional derivative (PD) controller was adopted [11]. Fuzzy logic with feedforward-feedback controller was proposed in [12].

In [13], a 3-D crane modeling and control utilizing Euler-Lagrange state-space method and anti-swing fuzzy logic is proposed, where, the mathematical model of the system is deduced. Using laboratory stand and Simulink® environment the state-space representation for that model is introduced and explored. The acquired control design was analyzed & simulated, and the performance was compared with available encoder-based system supplied by the 3-D crane maker Inteco®. In addition, an anti-swing fuzzy logic control has been built and tested. Achieved control strategy is compared with the available anti-swing PI controller planned by the three-dimensional crane producer Inteco®. 5 DOF control strategies are designed, assessed and compared with the various load masses.

Also Artificial intelligence techniques have its share in controlling and stabilizing overhead crane, A particle swarm optimization technique was used to tune the parameters of model predictive control (MPC) [14]. A cuckoo search algorithm was used to model overhead crane using radial basis function neural networks with aid of membrane communication mechanism [15].

II. The 3-D Overhead Gantry Crane System

The three-dimensional overhead crane under investigation with its payload is pictured in Fig. 1, where L represent the

[1] Earlier known as Amin A. Mohammed

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

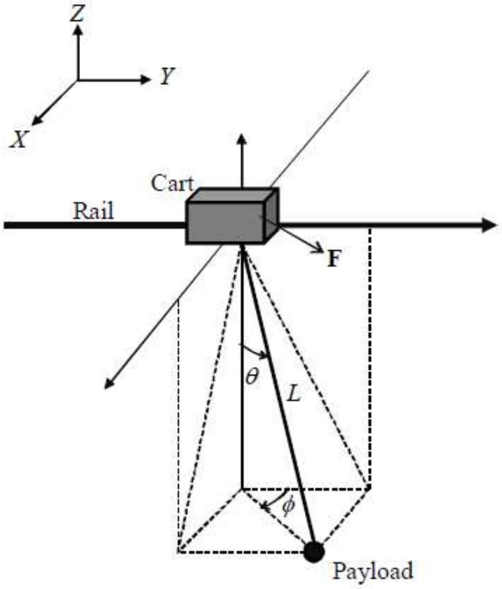


Fig. 1. Representation of the three-dimensional overhead crane.

length of the rope, θ & ϕ the swing angles of the rope, and F is the cart drive force. To simplify the analysis, the payload and the cart are regarded as point masses. The cart friction force and the tension force that may lead to an elongation in the rope are ignored.

III. Crane Dynamic Model

The dynamic model of the 3-D overhead crane system which is essential for control is formulated based on Euler-Lagrange equation [16]. To apply Lagrange equation, first the kinetic energy and the potential energy associated with the system must expressed [17]. As shown in Fig. 1, the position vectors for rail, cart and payload are given by:

$$r_r = [x, 0, 0]$$

$$r_c = [x, y, 0]$$

$$r_p = [x + L\sin\theta\sin\phi, \quad y + L\sin\theta\cos\phi, \quad -L\cos\theta]$$

where y and x denote the cart locations in Y- and X-directions correspondingly.
The system entire kinetic energy is a contribution from the cart, rail, and payload, as given bellow:

$$T = T_{rail} + T_{cart} + T_{payload} \tag{1}$$

Where

$$K_{rail} = \frac{1}{2}m_r\dot{r}_r^2 = \frac{1}{2}m_r\dot{x}^2$$

$$K_{cart} = \frac{1}{2}m_c\dot{r}_c^2 = \frac{1}{2}m_c(\dot{x}^2 + \dot{y}^2)$$

$$K_{payload} = \frac{1}{2}m_p\dot{r}_p^2 + \frac{1}{2}J\dot{\phi}^2 + \frac{1}{2}J\dot{\theta}^2$$

After substituting for r_p and simplifying, the total kinetic energy becomes:

$$\begin{aligned} T =& \frac{1}{2}(m_r + m_c + m_p)\dot{x}^2 + \frac{1}{2}(m_c + m_p)\dot{y}^2 + \\ & m_pL\cos\theta\sin\phi\dot{x}\dot{\theta} + m_pL\sin\theta\cos\phi\dot{x}\dot{\phi} \\ & + m_pL\cos\theta\cos\phi\dot{y}\dot{\theta} - m_pL\sin\theta\sin\phi\dot{y}\dot{\phi} \\ & + \frac{1}{2}(m_pL^2 + J)\dot{\theta}^2 + \frac{1}{2}(m_pL^2\sin^2\theta + J)\dot{\phi}^2 \end{aligned} \tag{2}$$

where g is the gravitational acceleration and J symbolizes the payload's moment of inertia. The potential energy on other hand, is only resulting from the payload, and is given by:

$$V = v_{paylaod} = -m_pgL\cos\theta \tag{3}$$

The Euler-Lagrange equation is given by [18]:

$$\frac{d}{dt}\left[\frac{\partial K}{\partial \dot{q}_i}\right] - \frac{\partial K}{\partial q_i} + \frac{\partial V}{\partial q_i} = Q_i, \quad i = 1, 2, 3, 4 \tag{4}$$

Where q_i and Q_i are the generalized coordinate and the associated generalized force.
The state and control vectors are defined respectively, as:

$$q = \begin{bmatrix} x & y & \theta & \phi \end{bmatrix}^T$$

$$Q = \begin{bmatrix} f_x & f_y & 0 & 0 \end{bmatrix}^T$$

In which f_x and f_y denote the control inputs for the motion in the X- and Y-direction. Out of Equation (4) a set having four equations is achieved for $i = 1, 2, 3, 4$, that is:

$$\begin{aligned} \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{x}}\right] - \frac{\partial K}{\partial x} + \frac{\partial V}{\partial x} &= f_x \\ \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{y}}\right] - \frac{\partial K}{\partial y} + \frac{\partial V}{\partial y} &= f_y \\ \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{\theta}}\right] - \frac{\partial K}{\partial \theta} + \frac{\partial V}{\partial \theta} &= 0 \\ \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{\phi}}\right] - \frac{\partial K}{\partial \phi} + \frac{\partial V}{\partial \phi} &= 0 \end{aligned} \tag{5}$$

From the equation set (5), the dynamic model of the crane can be expressed in matrix (compact) form [19]:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) = Q \tag{6}$$

where the matrices $M(q) \in R^{4\times4}$, $C(q, \dot{q}) \in R^{4\times4}$ symbolize the inertia and Centripetal-Coriolis terms as shown in equations (7, 8).
And $G(q) \in R^4$ is the gravity vector defined as:

$$G = \begin{bmatrix} 0 & 0 & m_pL\sin\theta & 0 \end{bmatrix}^T$$

Since the inertia matrix $M(q)$ is symmetric and positive definite, it's inverse must always exist [20], thus, from the matrix equation in (4), we can write:

$$\ddot{q} = M(q)^{-1}(Q - C(q, \dot{q})\dot{q} - G(q)) \tag{9}$$

$$M(q) = \begin{bmatrix} m_r + m_p + m_c & 0 & m_p L \cos\theta \sin\phi & m_p L \sin\theta \cos\phi \\ 0 & m_p + m_c & m_p L \cos\theta \cos\phi & -m_p L \sin\theta \sin\phi \\ m_p L \cos\theta \sin\phi & m_p L \cos\theta \cos\phi & m_p L^2 + J & 0 \\ m_p L \sin\theta \cos\phi & -m_p L \sin\theta \sin\phi & 0 & m_p L^2 \sin^2\theta \end{bmatrix} \tag{7}$$

$$C(q,\dot{q}) = \begin{bmatrix} 0 & 0 & -m_p L \sin\theta \sin\phi\dot{\theta} + m_p L \cos\theta \cos\phi\dot{\phi} & m_p L \cos\theta \cos\phi\dot{\theta} - m_p L \sin\theta \sin\phi\dot{\phi} \\ 0 & 0 & -m_p L \sin\theta \cos\phi\dot{\theta} - m_p L \cos\theta \sin\phi\dot{\phi} & -m_p L \sin\theta \cos\phi\dot{\phi} - m_p L \cos\theta \sin\phi\dot{\theta} \\ 0 & 0 & 0 & -m_p L^2 \sin\theta \cos\theta\dot{\phi} \\ 0 & 0 & m_p L^2 \sin\theta \cos\theta\dot{\phi} & m_p L^2 \sin\theta \cos\phi\dot{\theta} \end{bmatrix} \tag{8}$$

Moreover, the payload swing angles with respect to XZ-plane, θ_y and with respect to YZ-plane, θ_x can be expressed as follows, by direct project of the appropriate side of Fig. 1:

$$\theta_x = \tan^{-1}(\tan\theta \sin\phi), \qquad \theta_y = \sin^{-1}(\sin\theta \cos\phi) \tag{10}$$

IV. Controller Design

A. Feedback Linearization and PID Control

The essential idea of feedback linearization technique is it to define a control law in such a way that we cancel the nonlinearities in the model and end up with a linear differential equation for error command, and then use the linear control design techniques. This technique sometimes called computed torque control [18], [19], and it has been applied successfully in robotics [17].

Rewrite the model given in Equation (6):

$$M(\theta)\ddot{\theta} + C(\theta,\dot{\theta})\dot{\theta} + g(\theta) = \tau \tag{11}$$

The next task is to develop a PID controller as shown in equation (12), to the dynamic model. A supplementary state variable is required to address the integral part of the PID control. It is defined by ξ, and its derivative with respect to time is $\dot{\xi} = \tilde{\theta}$.

$$\tau = K_P\tilde{\theta} + K_I \int \tilde{\theta} dt + K_D\dot{\tilde{\theta}} \tag{12}$$

Thus, the PID control law could be represented by the below system [20]:

$$\tau = K_P\tilde{\theta} + K_I\xi + K_D\dot{\tilde{\theta}} \tag{13}$$

$$\dot{\xi} = \tilde{\theta} \tag{14}$$

The closed-loop equation's is found by introducing the control law τ of (13) to the crane dynamics (11), i.e.

$$M(\theta)\ddot{\theta} + C(\theta,\dot{\theta})\dot{\theta} + g(\theta) = K_P\tilde{\theta} + K_I\xi + K_D\tilde{\theta} \tag{15}$$

$$\dot{\xi} = \tilde{\theta} \tag{16}$$

As a function of state vector $\begin{bmatrix}\xi^T & \tilde{\theta}^T & \dot{\tilde{\theta}}^T\end{bmatrix}^T$, equations (15-16) could be expressed as:

$$\frac{d}{dt}\begin{bmatrix}\xi \\ \tilde{\theta} \\ \dot{\tilde{\theta}}\end{bmatrix} = \begin{bmatrix} \tilde{\theta} \\ \dot{\tilde{\theta}} \\ \ddot{\theta}_d - M(\theta)^{-1}[K_P\tilde{\theta} + K_I\xi + K_D\dot{\tilde{\theta}} - C(\theta,\dot{\theta})\dot{\theta} - g(\theta)] \end{bmatrix} \tag{17}$$

$\tilde{\theta} = \dot{\tilde{\theta}} = 0$ at the equilibrium, hence $\theta = \theta_d$.
Therefore equation (17) gives:
$[\xi^T \quad \tilde{\theta}^T \quad \dot{\tilde{\theta}}^T] = [\xi^* \quad 0^T \quad 0^T]$, where

$$\xi^* = K_I^{-1}[M(\theta_d)\ddot{\theta}_d + C(\theta_d,\dot{\theta}_d)\dot{\theta}_d + g(\theta_d)]$$

$$\frac{d}{dt}\begin{bmatrix}\theta_d \\ \dot{\theta}_d\end{bmatrix} = \begin{bmatrix} \dot{\theta}_d \\ M(\theta_d)^{-1}[\tau - C(\theta,\dot{\theta})\dot{\theta} - g(\theta)] \end{bmatrix} \tag{18}$$

B. Differential Evolution

Differential Evolution (DE) is renowned as an effective global optimizer. In this study the main objective is to reduce the settling time and the overshoot. DE starts with a population of NP possible solutions labeled $X_{i,G}, \quad i = 1, ..., NP$, where i an index represents the population of a given generation and G represent such generation. The performance of DE relays on three major operations, namely, mutation, followed by generation or reproduction and selection.

The central idea of DE is a new strategy for producing vectors of trial parameter . By summing the weighted difference vector among two population members to a third one, DE produces new parameter vectors. If the obtained vector gives a lower objective function than an old population member, the newly produced vector substitutes the vector that was compared with. Moreover, the best parameter vector is evaluated for every generation as to keep track of the advance of the procedure [22].

In this case the initial population is indiscriminately selected, in a range of 0 to 200, for the three gains of the controller $K_P, K_I, \quad$ andK_D. With original population (Np) and generation $(Ng\ )$ sizes of 20, cross over (CR) and mutation (F) factors of 0.5. The (gross) objective function is constructed as follows:
$OF = 0.5 \times (OF1 + OF2)$ where OF_1 and OF_2 represent the individualistic objective functions for maximum overshoot and settling time.

V. Results and Discussion

In the present analysis, the nominal values of the parameters are taken as $m_p = 0.73kg$, $m_c = 1.06kg$, $m_r = 6.4kg$, $L = 0.7m$ and $J = 0.005kg.m^2$ [21]. The acceleration gravity g is taken as $9.8m/s^2$.

The optimized PID controller parameters obtained using DE are: $K_p = diag(200, 200, 200, 200)$, $K_D = diag(22.28, 22.28, 22.28, 22.28)$, and $K_I = diag(111, 111, 111, 111)$, the achieved results are discussed below.

The obtained results with and without DE algorithm are compared as detailed below. There is a high overshoot and settling time for the position of the cart as shown in Figs. 2-3 in case of manual tuning. However, those high overshoot and settling time have been reduced significantly with the turned PID parameters using DE algorithm. Besides, the rise time has been significantly reduced as detailed in Table I and Table II with excllent improvement in the (absolute) peaks. This clearly highlight the usefulness of applying DE algorithm to turn the parameters of the PID controller.

On the other hand, the settling time for angular displacements is relatively high for the case of manual turning as displayed in Figs. 4-5. Yet, with the help of DE algorithm, the settling time has been dramatically reduced which once again confirm the superior of DE algorithm over the manual approach. Detailed response characteristics for the angular displacement (θ) are provided in Table III, where the rise time was reduced by around 10 times, overshoot was totally removed, and the peak was reduced by around 20%. Similar info are shown in Table IV for the angular displacement (ϕ), where; the rise time, settling time, overshoot, and peak were reduced.

Finally the control signals for the case of manual tuning and optimized tuning are shown in Figs. 6-9 (for comparison only as they're predetermined already). High control signals are required in case of DE algorithm as to achieve rapid responses, which could be cited as a pitfall against DE, however, it is realistic in light of the immediate response that was obtained. On the other hand, very small control inputs are needed in case of manual turning which is penalized by poor and late responses.

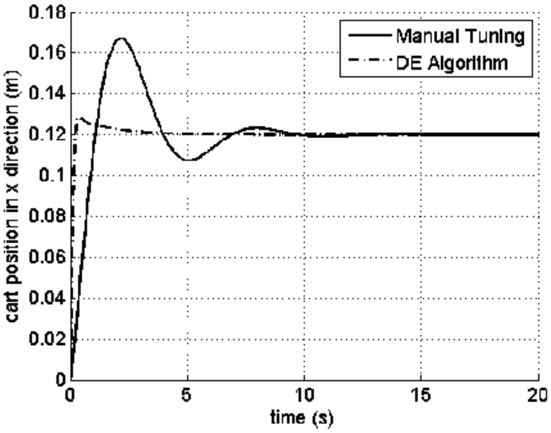


Fig. 2. Position of the cart in X-direction for FBL

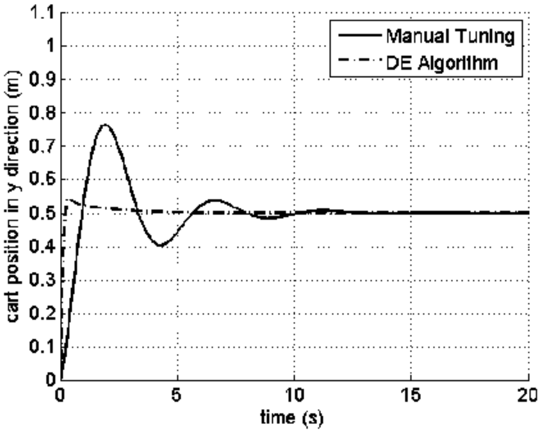


Fig. 3. Position of the cart in y-direction for FBL

TABLE I
COMPARISON OF STEP INFO FOR CART POSITION IN X-DIRECTION

Step Info	PID controller	
	Manual Tuning	*DE algorithm*
Rise Time (s)	0.8068	0.1576
Settling Time (s)	0.1075	2.2223
Overshoot (%)	39.3112	7.7691
Peak (m)	0.1672	0.1293

TABLE II
COMPARISON OF STEP INFO FOR CART POSITION IN Y-DIRECTION

Step Info	PID controller	
	Manual Tuning	*DE algorithm*
Rise Time (s)	0.6925	0.1561
Settling Time (s)	9.6079	2.4303
Overshoot (%)	52.5639	8.4952
Peak (m)	0.7630	0.5425

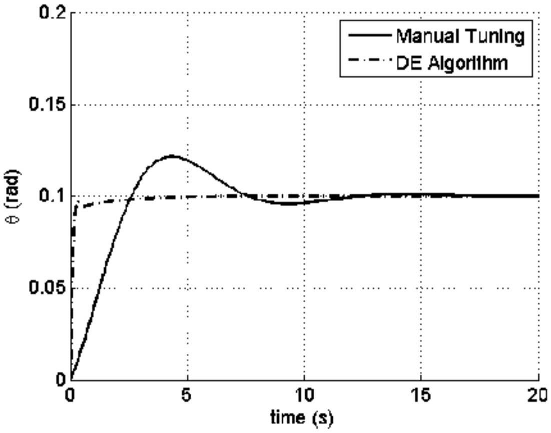


Fig. 4. Angular displacement θ (rad), FBL

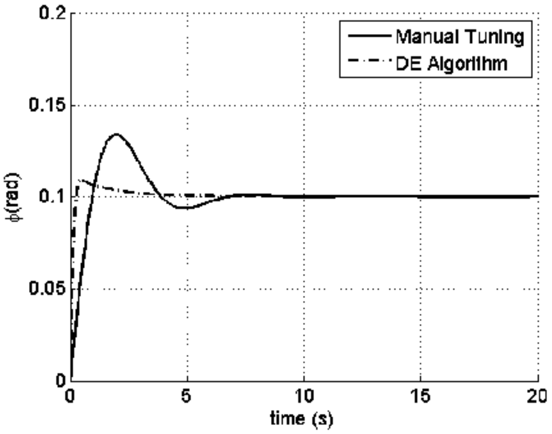


Fig. 5. Angular displacement ϕ (rad), FBL

TABLE III
Comparison of step info for angular displacement (θ)

Step Info	PID controller	
	Manual Tuning	*DE algorithm*
Rise Time (s)	1.9577	0.1826
Settling Time (s)	11.0925	2.5159
Overshoot (%)	21.6526	0.0000
Peak (rad)	0.1215	0.1000

TABLE IV
Comparison of step info for angular displacement (ϕ)

Step Info	PID controller	
	Manual Tuning	*DE algorithm*
Rise Time (s)	0.7290	0.1776
Settling Time (s)	6.1622	2.9315
Overshoot (%)	33.7020	9.5036
Peak (rad)	0.1337	0.1095

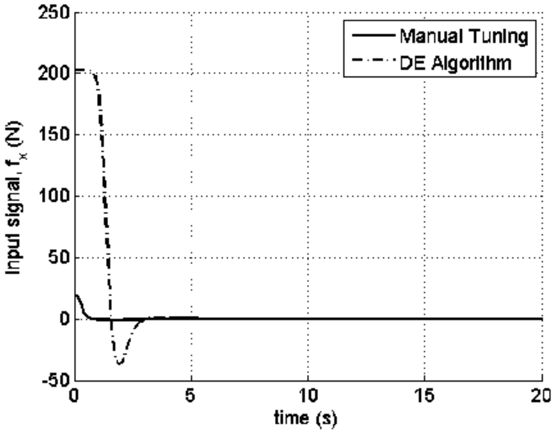


Fig. 6. Control signal f_x using FBL

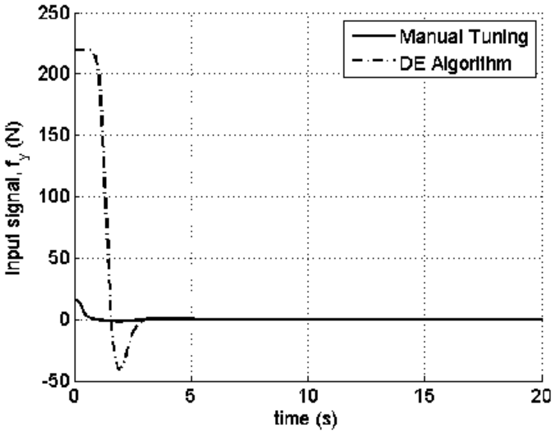


Fig. 7. Control signal f_y using FB

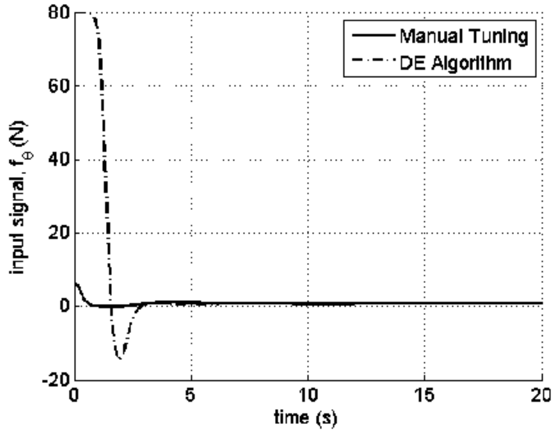


Fig. 8. Control signal f_θ using FB

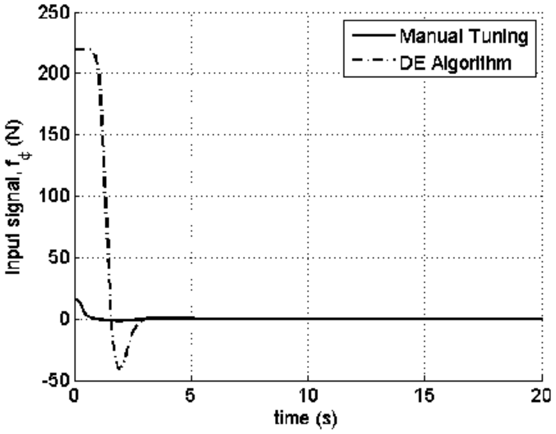


Fig. 9. Control signal f_{phi} using FB

VI. Conclusion

The dynamic model of a 3-D gantry crane including rope length and payload has been developed using the Euler-Lagrange method.

A nonlinear PID control, base on the feedback linearization technique is proposed. The obtained results, include some oscillation, however, by using DE to tune the controller parameters, better results are achieved. The achieved results with DE implementation have greatly reduced the settling time and the maximum overshoot compared with the manual tuning case. Conversely, the control associated with DE algorithm are relatively high compared with the other case.

Acknowledgment

The authors greatly thankful the fund supplied by KFUPM Endowment through Mr. Luhaidan Scholarship Program. Beside, the lead author is thankful to the funding delivered by K.A.CARE Energy Center at KFUPM via K.A.CARE Graduate Students Grant Program, grant number: KACARE182-GSGP-10.

Mohamed Hassan is a PhD scholar at system engineering department. He received his B.Sc degree in Electrical and Electronic Engineering Majoring in Control Systems and instrumentations from the University of khartoum in 2008, and the master degree from KFUPM in 2015. His research interests are in systems modeling and identification, nonlinear control, intelligent control , study of valve stiction phenomena ,Microgrids and power systems optimzation.

Amin A.M. Fadlalla ”was born in Alhedor, East of Gezira, Sudan on 1st January 1988. He obtained his B.S. degree in Mechanical Engineering from University of Khartoum (UofK) in 2011, and the M.S. degree in Mechanical Engineering from King Fahd University of Petroleum & Minerals (KFUPM), Saudi Arabia in 2015 where he has recently completed all the requirements for the PhD degree and will be graduated soon. His research encompass work across the fields of dynamics, control, robotics, and wind turbines”.

References

[1] H.-H. Lee, “Modeling and control of a three-dimensional overhead crane,” Journal of Dynamic Systems, Measurement, and Control, vol. 120, no. 4, pp. 471–476, 1998.

[2] Alkin, C., C. E. Imrak, and H. Kocabas. ”Solid modeling and finite element analysis of an overhead crane bridge.” Acta Polytechnica 45.3 (2005).

[3] Toxqui, Rigoberto Toxqui, Wen Yu, and Xiaoou Li. ”PD control of overhead crane systems with neural compensation.” International Symposium on Neural Networks. Springer, Berlin, Heidelberg, 2006.

[4] Burg, T., et al. ”Nonlinear control of an overhead crane via the saturating control approach of Teel.” Proceedings of IEEE International Conference on Robotics and Automation. Vol. 4. IEEE, 1996.

[5] Wang, Wei, et al. ”Anti-swing control of overhead cranes based on sliding-mode method.” Control and decision 19 (2004): 1013-1016.

[6] Hussein, Emad Q., Ayad Q. Al-Dujaili, and Ahmed R. Ajel. ”Design of Sliding Mode Control for Overhead Crane Systems.” IOP Conference Series: Materials Science and Engineering. Vol. 881. No. 1. IOP Publishing, 2020.

[7] Lee, Soon-Geul, et al. ”Partial feedback linearization control of a three-dimensional overhead crane.” International Journal of Control, Automation and Systems 11.4 (2013): 718-727.

[8] Fang, Yongchun, et al. ”Nonlinear coupling control laws for an underactuated overhead crane system.” IEEE/ASME transactions on mechatronics 8.3 (2003): 418-423.

[9] Tsai, Ching-Chih, Hsiao Lang Wu, and Kun-Hsien Chuang. ”Backstepping aggregated sliding-mode motion control for automatic 3D overhead cranes.” 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2012.

[10] Chang, Cheng-Yuan, and Kuo-Hung Chiang. ”Fuzzy projection control law and its application to the overhead crane.” Mechatronics 18.10 (2008): 607-615.

[11] Chang, Cheng-Yuan, and Kuo-Hung Chiang. ”Fuzzy projection control law and its application to the overhead crane.” Mechatronics 18.10 (2008): 607-615.

[12] Al-saedi, Mazin I. ”Enhancing the Feedforward- Feedback Controller for Nonlinear Overhead Crane Using Fuzzy logic controller.” MS&E 745.1 (2020): 012074.

[13] A. Aksjonov, V. Vodovozov, and E. Petlenkov, “Three-Dimensional Crane Modelling and Control Using Euler-Lagrange State-Space Approach and Anti-Swing Fuzzy Logic,” The Scientific Journal of Riga Technical University-Electrical, Control and Communication Engineering, vol. 9, no. 1, pp. 5–13, 2015.

[14] Smoczek, Jaroslaw, and Janusz Szpytko. ”Particle swarm optimization-based multivariable generalized predictive control for an overhead crane.” IEEE/ASME Transactions on Mechatronics 22.1 (2016): 258-268.

[15] Zhu, Xiaohua, and Ning Wang. ”Cuckoo search algorithm with membrane communication mechanism for modeling overhead crane systems using RBF neural networks.” Applied Soft Computing 56 (2017): 458-471.

[16] M. D. Ardema, “Lagrange’s Equations,” Analytical Dynamics: Theory and Applications, pp. 109–130, 2005.

[17] M. W. Spong and M. Vidyasagar, Robot dynamics and control. John Wiley & Sons, 2008.

[18] R. N. Jazar, Theory of Applied Robotics. Boston, MA: Springer US, 2010.

[19] R. M. Murray, Z. Li, S. S. Sastry, and S. S. Sastry, A mathematical introduction to robotic manipulation. CRC press, 1994.

[20] R. Kelly, V. Santibanez, and A. Loriá, Control of robot manipulators in joint space. London: Springer, 2005.

[21] R. M. T. Raja Ismail, M. A. Ahmad, M. S. Ramli, and F. R. M. Rashidi, “Nonlinear Dynamic Modelling and Analysis of a 3-D Overhead Gantry Crane System with System Parameters Variation.” International Journal of Simulation–Systems, Science & Technology, vol. 11, no. 2, 2010.

[22] Musrrat Ali, M. Pant, and A. Abraham, “Simplex differential evolution,” Acta Polytechnica Hungarica, vol. 6, no. 5, pp. 95–115, 2009.

IR Aware Cell Placement and Clock Tree Performance Optimization in FPGA Memories

Sourabh Aditya Swarnkar, Mohammad Anees, Kumar Rahul, Santosh Yachareni,
Xilinx, Hyderabad, India.
E-mails: (sourabh, manees, kumarr, santoshy) @xilinx.com.

Abstract— Field Programmable Gate Array (FPGA) Memories are synchronous pipelined circuits that rely on the distribution of the clock to achieve the timing requirements. The density of FlipFlops (FFs) are higher in such designs. This paper discusses a novel method and algorithm for reducing the simultaneous switching current demand to address IR hotspots by the staggered placement of these FFs. The higher current demands from a small section of high resistive Power Distribution Network (PDN), is redistributed to other unused sections of the Network. This approach improves the performance of clock network distribution by reducing the clock insertion delay and max skews. Application of proposed algorithm reduces IR-drop by 32.5% and improves clock insertion delay by 36.5%.

Keywords— Dynamic IR- drop, Power Distribution Network (PDN), dynamic current reduction, scan chain, Place and Route (PNR), Clock Tree Synthesis (CTS), FlipFlops (FFs)

I. Introduction

As the complexities of integrated circuits have increased, the power consumption during the circuit test also has grown drastically. Also, the power consumption during the scan testing is higher than the power consumed in normal operation due to the increased switching activity. Excessive test power consumption can lead to dynamic IR-drop which in turn degrades the circuit performance and may lead to functional failure.

For system-on-chip (SoC) designs, FFs are the major building blocks in terms of PPA (Power, Performance and Area). For test purposes, these FFs are hooked into scan chains. Several studies are done on Hold time failures caused by clock skews [1-4]. Traditionally, hold time failures are caused by clock skew from different clock branches, different clock domains or clock jitter, drift, etc. [5].

To improve test run time, the scan chain is partitioned into smaller chains [6]. There are several scan-FFs in each of these chains. During test mode, all scan FFs change state or toggle simultaneously. At each active edge of the clock, every FF captures the incoming data and sends it to the next stage. Also, the CTS targets all the FFs for zero clock skew, so all FFs would switch at the same time. Every active edge of the clock triggers the switching current of all the FFs. Therefore, the peak current demands are synchronous to the clock for the active edges which attribute to a higher on-die IR-drop.

There are mainly two types of supply voltage degradation: (1) Static IR-drop, which is observed when a Power/Ground supply network is considered as a resistive network with DC supply current sources; (2) Dynamic IR-drop, which is caused by switching current in Power/Ground network.

Various cell placement and design floor-planning-related methods have been proposed to restrict supply voltage reduction. Prior works demonstrated different approaches for IR-drop modeling and optimization. IR-drop modeling is very beneficial, particularly for the larger, chip-level designs, which usually have a longer turnaround time to complete the IR-drop analysis.

We compared recent works [7,8,9] in Table 1. Bhamidipati et al. proposed an algorithm to simultaneously reduce the peak current and the intensity of IR hotspots by scheduling useful skews [9]. This technique is useful for reducing peak current demands of timing non-critical cells. Also, this method does not consider clock insertion delays of targeted FFs. Insertion of power staples is a new technique for improving PDN robustness in sub-10nm technologies. Heo et al. introduce an approach to improve power staple insertion at post-placement in [8]. Power staples are short pieces of metal shorting multiple adjacent Power rails, to mitigate the IR-drop. This approach is suitable for designs with less routing congestion. Kahng et al. introduce the first IR-drop aware placement engine in [7]. They present a greedy algorithm to find optimum NP-hard placement of cells to reduce IR-drop. They formulated IR-drop (Voltage) as a function of supply currents and study the distribution of current to each cell to locate the ideal placement to reduce IR-drop. For each cell's peak current, the worst voltage degradation is approximated by a log-sum-exp function and integrated into the objective function. The solution to the NP-hard problem is large in numbers and time-consuming process as greedy search algorithms are not linear but polynomial in time. The polynomial search for a big design at advanced technology nodes can increase the run time almost exponentially.

TABLE 1. IR-Drop Optimization: Previous Works Vs Our Work

Ref. Work	*Design Step*	*Target*	*Method*	*Improve Clock insertion delay*
[7]	Place-Opt	Peak Current Reduction	Cell Movement	NO
[8]	Place-Opt	PDN Resistance Reduction	Staple Insertion	NO
[9]	CTS	Peak Current Reduction	Useful Skew Scheduling	NO
Our Work	Place-Opt	Dynamic IR-Drop	Cell Movement	YES

These methods [7,9] suppress the peak current demands, but they cannot avoid local IR-drops, particularly in the scan test modes, where the peak current demands are

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

synchronous to the clock for the active edges which attribute to a higher on-die IR-drop. Such demands can be very high if the successive FFs have the opposite data stored and placed together in close vicinity.

In our previous work [10], we demonstrated a novel method of FFs clustering to improve the clock insertion delay of FPGA memories. These delays are highly impacted by the big wirelengths inside the Semi-Custom FPGA memories, which limit the operating frequency of the designs. Dynamic IR-drop hotspots result from instantaneous large current drawn in a local region. For test purposes, these clustered FFs are hooked into scan chains.

In this paper, we demonstrate a deterministic linear search algorithm to minimize the local dynamic IR-drop caused by high-density FFs clustering. Unlike NP-hard, our approach is to simplify the search by targeting only a finite number of searches to meet the specifications of design, rather than finding the most optimum solutions. Our proposed approach is to distribute the high current demand from a small section of the Power Distribution Network to the unused [less utilized] sections without disturbing the clustering approach discussed above. This technique not only reduces the formation of local IR hotspots but also proves to be very effective in maintaining the clock skews. This solution is easy to plug into the mainstream physical design Clock Tree Synthesis (CTS) step. Application of proposed algorithm reduces IR-drop by 32.5% and improves clock insertion delay by 36.5%.

II. Power Distribution Network & Dynamic IR-Drop

A Power Distribution Network (PDN) is a stitched network of Power/Ground metal straps to deliver stable supply voltages from PG pads to all the circuit devices in a design. Fig. 1 shows a typical PDN with (a) PG pads, providing the ideal reference voltages (VDD and GND), (b) PG Metal Straps, connecting to PG pads through high metal layer, (3) via stacks, connecting PG meshes with lower-level metal, and (d) Standard Cell PG rails, connecting to PG pins of standard cells in low metal layers from which Standard Cell pins can be accessed.

Construction of a typical PDN in advanced geometry nodes where 16 or more metal layers exist, starts with building M0 rails to which the power and ground pins of standard cells are connected. M0 rails are altered between Standard cell rows, hence these are separated by the height of standard cells. Due to an increase in the current and power density at the advanced technology nodes, the M0 rail alone may not be strong enough to meet IR-drop and Electromigration (EM) requirements.

So, to mitigate these limitations, M2 rails are routed in parallel to M0 rails. The higher-level metal straps are routed perpendicular and parallel alternatively to distribute the current evenly. These straps are connected to the immediate lower metal layer at each cross-intersections by using a via-stack. The size of the via-stack needs to be chosen carefully; larger via-stacks reduce the resistive IR-drop, but consume more routing resources and cause congestion. Upper Metal [M8 and above] layers are progressively made wider to evenly distribute the current across the design. These upper metal layers have less resistivity and introduce very little impact on IR-drop. Lower metal layers in 7nm such as M0/M1 rails and corresponding vias have a higher resistivity.

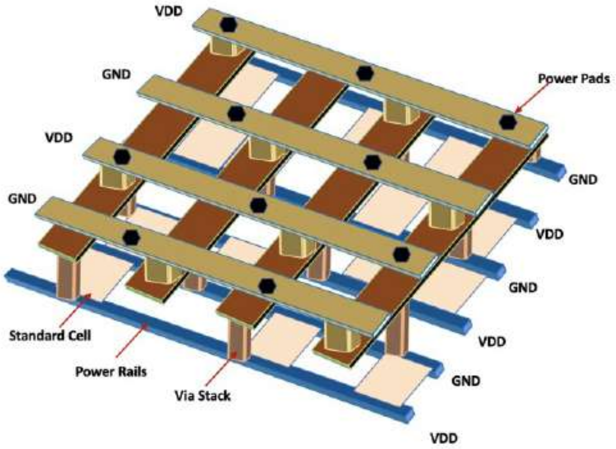


Fig. 1. Typical PDN construction

IR-losses are the ohmic voltage drop on the power grid which depends on the metal layers sheet resistances [11]. In 7 nm technology, the spectrum of metal layer resistivity is large, from the top lesser resistive layers to the bottom higher resistive layers (M0-M3). The chart in Fig. 2 shows the normalized sheet resistance with respect to M0 for each layer of 7 nm node technology. Metal 0 and Metal 1 are more frequently used by standard cells to connect signals and power supplies. The section of PDN which connects the power supply to these cells with M0 and M1 is the major contributor of IR-drop.

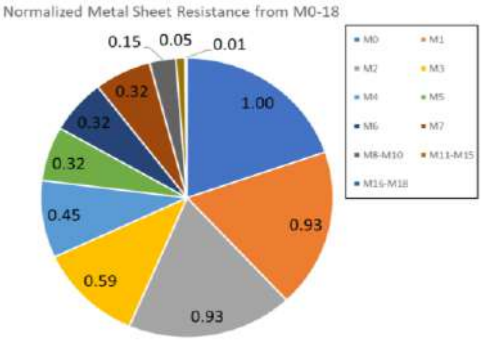


Fig. 2. Normalized distribution chart of Sheet Resistance

III. Problem Statement

The goal of our work can be defined in 2 steps:
(1) Automate the clustering of critical path FFs within a physical bound: The task is to identify the critical timing paths based on the functionality of the sub-blocks. The placement bounds are estimated and defined based on the number of FFs associated with the sub-blocks. This approach reduces the clock insertion delays of targeted FFs.
(2) Dynamic IR-aware placement adjustment of these FFs to minimize the Engineering Change Order (ECO) during high current switching operation such as scan testing: Defining the target IR-drop constraints, max displacement constraint, dynamic IR-report for every targeted sub-block, with the objective of cell moment to reduce IR-hot spots without degrading the clock insertion delay. These displacements also reduce the local routing congestions.

IV. Our Approach

Placing cells to minimize static IR-drop has been proven to be NP-hard [Ref- Supply voltage degradation aware analytical placement]. The solutions to the NP-hard problems are large in numbers. NP-hard problems are solved using a greedy search algorithm that is not linear but polynomial in time. Usually, PDN is designed with a nominal set of wire widths and pitches for different metal layers and then optimized to meet the allowed **IR-drop budget** based on the targeted designs. Our proposed approach is to convert the cell placement search into a deterministic linear search. We only target to achieve the specified IR-drop budget, rather than using greedy search to find the most optimal placement. We focus the optimal placement search within the targeted bounds which is broken into smaller unit bins. We further constrained our cell placement search within standard cell rows of these placement bounds. Every bin is also targeted to have a predefined checker-board patterned FF placement for staggering the FFs. We also derived a predefined IR-drop target budget. This approach transforms the search algorithm to be deterministic and linear rather than greedy.

A. *Automation of placement bounds for critical sub-blocks and CTS control options*

Using placement bound constraints, we physically cluster the placement of all the FFs of the timing critical sub-block. Our previous work [10] shows an improvement in clock-to-out for the targeted sub-block by 36% and in max skew by 39%. In this paper, we propose an automated method to generate the physical bounds. Clock grouping options for related FFs of sub-blocks within the bounds were also applied for reducing clock insertion delays and clock skews.

B. *Defining the target Bin*

To simultaneously reduce the peak of dynamic current surge and intensity of IR hotspots, we break the placement bounds into smaller bins as Placement Units [PU] Fig. 3. The most resistive section of the PDN is the M0-rail and the via stack that connects the M1 metal straps. Considering that the M0-rails and Via0 are highly resistive, they are the highest contributing elements to the formation of high-IR regions. Therefore, if the current, which is demanded through each lower level via stack could be lowered, the intensity of IR hotspots could be greatly mitigated. To mitigate the higher resistivity of lower-level metal layers, we define a single bin to cover vertical M1 pitch (VDD to VDD). This is the minimum distance between Via0 in M0 - VDD or GND power rails. The current drawn from any section of PDN is directly proportional to the number of cells connected.

C. *Acceptable IR-drop target budget calculation*

The calculations for determining an acceptable IR-drop target is based on the following assumptions [12] [for typical 7 nm SoC and nominal Voltage ($V_{nominal}$)= 0.925V power supply]:

i. A standard cell library characterized to $V_{nominal}$ ±10%

ii. Variation within Voltage Regulator $V_{nominal}$ ±2%

iii. 18.5 mV of IR-drop from the Voltage Regulator to the SoC.

Based on the .LIB characterization data, it is clear that the SoC PDN must maintain at least 0.8325 V (0.925V – 10%) to all the standard cells. Also, the worst-case voltage delivered from the Voltage Regulator is 0.9065V (0.925V – 2%). So, the acceptable IR-drop target can be easily determined by subtracting the voltage requirement from the voltage supplied:

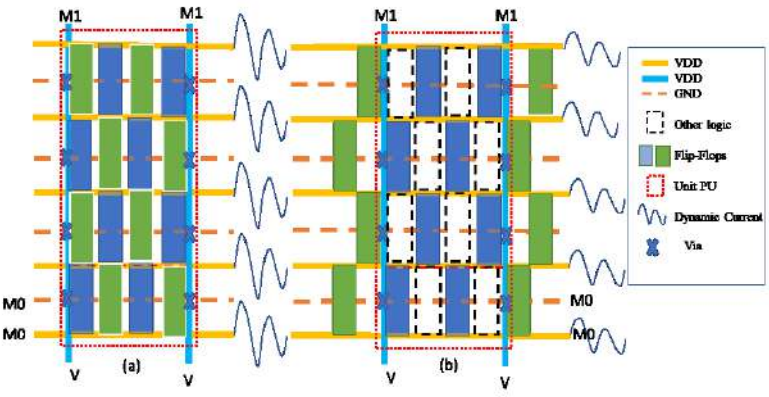


Fig. 3. Unit Placement Bins (a) before, (b) checkerboard pattern

$$\delta = (V_{Supply\_Regulator} – 2\%) – (V_{nominal} –10\%) – 18.5mV$$

$$= 0.9065V – 0.8325V– 18.5mV = 55.5 mV$$

[δ is Total IR voltage drop target budget on VDD and VSS]

D. *The checkerboard pattern for FF placement and Dynamic current simulations*

We propose a checker-board pattern for spacing and placing the cells to reduce the number of cells connected to M0 between M1 pitch Fig. 3. In 7nm, scannable FFs are usually double-height cells to provide easy access to additional scan-related pins, so our targeted bin is extended to cover several rows of scannable FFs. For proper spacing and placement of the cells in a checkerboard pattern, the number of FFs within the M1 pitch is kept equal to the number of rows of FFs inside the bin.

We also constrained the maximum displacement distance of the FFs to be within ±M1 pitch to minimize the impact on clock insertion delay of FFs due to displacement. [Horizontal displacement range of the Unit Scan [-x_{Δ},x_{Δ}] = ±M1 pitch]. Circuit level simulations including 4x4 FFs and RC modeling of the M0/M1 inside the bin before and after the displacement shows a 41% reduction in dynamic current fig. 4. With this reduction, we concluded that the constrained $\pm x_{\Delta}$ is sufficient enough to achieve the IR-drop target budget δ.

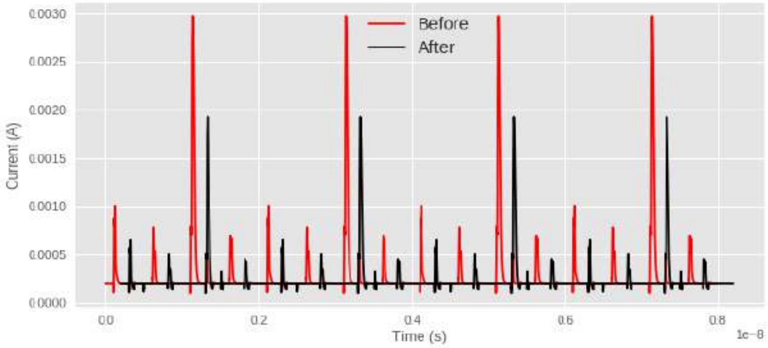


Fig. 4. Dynamic current simulations before and after displacement

E. Explanation of Algorithm

The following pseudo-code provides a detailed description of the algorithm used to implement the proposed solution. The flow can be defined in these steps:

(1) Defining a targeted list of functional modules for improved clock insertion delay. This list is fed into the **prepare_bound** function. For each functional module, a list of origins in the form of Cartesian (X, Y) locations is defined to place placement bounds.

(2)For each functional module, the **commit_bound** function is called. This function has multiple tasks: (a)Bound area is defined as a function of the total number of FFs with 50% extra margin to compensate the push out for checker-board patterned placement. (b) These bounds are further divided into an equal number of rows and columns of the unit bin (PU), (c) Deriving Bbox (Bounding box) of placement bounds based on the number of rows and columns of unit bins, (d) Extracting coordinates of each unit bins for traversing within the bounds, (e) Each unit bin is also further divided into the placeable FF units of 4x4, (f) Co-ordinates of these placeable FF units are also extracted, (g) Create a collection of each placeable units of every bin(PU) in the placement bound.

(3) Placement bound is created from (2) (a),(c)

(4) place opt and legalization

(5) (a)For each row of the collection created in (2)(g), the rows are traversed and placed FFs are staggered in a checkerboard pattern. (b) In case of all the placeable FF units are occupied, row over-flow is applied to continue the displacement of the FFs to the next row.

(6) Legalize placement.

TABLE 2. NOTATIONS

Term	Definition
B	Set of placement Bounds in design
b_k	K^{th} placement Bound of set B
b_{kW}/b_{kH}	Width & Height of b_k
PU	Unit Placement Box
PU_w /PU_H	Width & Height of one PU PU_w =P_{M1} & PU_H = 4*2*S_h
S_h	Site Row (Standard cell Row) Height
FF_W	Width of Unit Scan FF
FFh	Height of Unit Scan FF (Double Height Cells) =2* S_h
N_{FF}	Number of Unit Scan FF inside b_k
P_{M1}	Horizontal Pitch of M1 (Vertical) Power track inside PDN P_{M1} is an Integer multiple of FF_W
a, a_loc	Matrix [N_R x N_C] of PUs inside Bound, corresponding locations
b,b_loc	Matrix [r_p x r_q] of Placeable Unit Scan FF inside PU, corresponding locations
[-x_Δ, x_Δ]	Horizontal displacement range of the Unit Scan FF. χ_Δ= f(P_{M1}) or χ_Δ= f(FF_W)
Ab_k	Area Needed of b_k to have 50% Utilization = 2*[N_{FF}* FF_W*2* S_h] = N_R*N_C
MODE[i]/Origin[i]	List of functional modules/List or targeted origins of Bounds
ab	Collection of origins of each placeable FF units row-wise throughout a placement bound

```
prepare_bound () {
//List Functional Modules targeted for improved clock insertion delay
|    set MODE_[i] ← [READ_path1, READ_path2, Write_path] //defining
|     set Origin_[i] ← [{ll_x1, ll_y1}, {ll_x2, ll_y2}, {ll_x3, ll_y3}]
|      For [i=1; i<sizeof [MODE_[i]]; i++],
|      |      commit_bound (MODE[i] Origin[i])

commit_bounds (arg1, arg2[a b]) {
1: create collection ∈[FF] ← get cells FF of (arg1)
2: get Size of collection ∈[FF] → $N_{FF}$ // collection of $N_{FF}$ inside the Module
3: $Ab_k$ ← 2*[$N_{FF}$* $FF_W$*2* $S_h$] // Area of bound to keep 50% Utilization
4: $N_R$, $N_C$ ← [$INT(\surd(Ab_k)$+1] //Breaking Bound into equal number of
                                       //rows cols of PU
                                       // create matrix containing PU:
5: create_matrix =a; dimension = { $N_R$, $N_C$} → $\begin{pmatrix} R_N C_0 & \cdots & R_N C_N \\ \vdots & \ddots & \vdots \\ R_0 C_0 & \cdots & R_0 C_N \end{pmatrix}$
6: set $\{\chi_{llb}, y_{llb}\}$ ← arg2[0,1]
   set $\{\chi_{urb}, y_{urb}\}$ ← $\{(\chi_{llb} + N_C * PU_W), (y_{llb} + N_R * PU_H)\}$
   set Bbox for $b_k$, bbx1= $[\{\chi_{llb}, y_{llb}\}, \{\chi_{urb}, y_{urb}\}]$
      //Setting Bbox for bound $b_k$, deriving Cartesian X and Y locations
7:$\sum Orig_{PU}(R_i, C_j)$ ← $\sum_{\substack{0\leq R_i \leq N_R \\ 0\leq C_j < N_C}} [\{\chi_{llb} + C_j * PU_w\}, \{y_{llb} + R_i * PU_H\}]$
      //Derive origin for each element of matrix a, as $f(PU_H)$
      //Cartesian X location and $f(PU_W)$ for Cartesian Y location
8: create_matrix =a_loc; dimension = {$N_R$, $N_C$}
      // Assign all the Cartesian origin of each PU to another matrix
9: create_matrix =b; dimension = { $r_p$, $r_q$ } → $\begin{pmatrix} p_3 q_0 & \cdots & p_3 q_3 \\ \vdots & \ddots & \vdots \\ p_0 q_0 & \cdots & p_0 q_3 \end{pmatrix}$
     // Each element of matrix a is further divided into placeable FF units
     //(4x4)

10: create_matrix =b_loc; dimension = {$r_p$, $r_q$}
     // Assign all the Cartesian origin of each PU to another matrix
11: create a collection
        $set\ ab = \begin{pmatrix} R_N C_0 \begin{pmatrix} p_3 q_0 & \cdots & p_3 q_3 \\ \vdots & \ddots & \vdots \\ p_0 q_0 & \cdots & p_0 q_3 \end{pmatrix} & \cdots & R_N C_N \begin{pmatrix} p_3 q_0 & \cdots & p_3 q_3 \\ \vdots & \ddots & \vdots \\ p_0 q_0 & \cdots & p_0 q_3 \end{pmatrix} \\ \vdots & \vdots & \vdots \\ R_0 C_0 \begin{pmatrix} p_3 q_0 & \cdots & p_3 q_3 \\ \vdots & \ddots & \vdots \\ p_0 q_0 & \cdots & p_0 q_3 \end{pmatrix} & \cdots & R_0 C_N \begin{pmatrix} p_3 q_0 & \cdots & p_3 q_3 \\ \vdots & \ddots & \vdots \\ p_0 q_0 & \cdots & p_0 q_3 \end{pmatrix} \end{pmatrix}$

  //get the origin of each placeable FF unit row-wise throughout bound $b_k$
12: a) create bound $b_k$ ← $\binom{\in[FF]}{bbx1}$
    //create_placement bound
    b) apply CTS grouping options
13: run incremental place_opt & legalization

14: For each row of $R_{ab}$ {
        get the origin of all placed cells

     $rearrange\ FF$
     $\rightarrow \begin{cases} \text{If } R_{ab} \text{ is even} \rightarrow (C_0 q_0, C_0 q_2, \dots, C_N q_0, C_N q_2) \\ \text{If } R_{ab} \text{ is odd} \rightarrow (C_0 q_1, C_0 q_3, \dots, C_N q_1, C_N q_3) \\ \text{displacement range} \rightarrow [-x_\Delta, x_\Delta] \end{cases}$
15: for a row overflow

$rearrange\ FF\ ext \begin{cases} R_{ab} + 1\ (If\ R_{ab} \text{ is even} \rightarrow (C_0 q_0, C_0 q_2, \dots, C_N q_0, C_N q_2) \\ R_{ab} + 1\ (R_{ab} \text{ is odd} \rightarrow (C_0 q_1, C_0 q_3, \dots, C_N q_1, C_N q_3) \end{cases}$

16: legalize placement
}
```

V. EXPERIMENT RESULTS

Our proposed dynamic IR hot-spot and clock insertion delay reduction technique was tested on selected sub-blocks of the FPGA Semi-custom Memory design at 7nm technology node with 0.925V power supply. The following EDA tools are adopted to construct our experimental flow: The Synopsys IC Compiler II (ICC-II) [13] was utilized for Design planning, PDN construction, cell placement, CTS, and routing. The proposed flow script is written in TCL language, which is the native scripting language in ICC-II's shell. The IR-drop is calculated using Ansys Totem [14]. Static timing analysis is reported by PrimeTime [15].

A. IR hot spot reduction

Fig. 5 reflects the effectiveness of our proposed algorithm in reducing IR hot spots. In this figure, the IR map of FPGA Semi-custom Memory design before and after the application of the proposed algorithm is illustrated. We targeted to apply the algorithm on 3 functional modules of the design. As demonstrated, the IR hot- spots of Placement Bounds (b1, b2, and b3) are completely mitigated. In the map, the Red color denotes voltage drops above 60mV and the Yellow color denotes voltage drops below 60mV.

In Fig. 6, the top 10 bins are sorted based on the voltage degradation within each placement bound. The 3D plot contains the normalized Cartesian location of bins in (X, Y), and Z represents the IR-drop in mV. Voltage drops before and after applying the proposed algorithm are captured. The worst IR-drop is improved by ~32.50%. For all 3 Placement Bounds, we achieved the IR-drop target budget [δ] of 55.5 mV. Table 3 captures the IR-drop improvement for all the targeted sub-blocks.

B. Impact of clock insertion delay and timing

Table 3 illustrate that after applying our proposed algorithm, the clock insertion delay is reduced by ~36%. This kind of FFs placement also ensures the reduction of local skew within a specific functional module. Such placements also reduce the overall wirelength and loads from clock distribution networks. Results also indicate that the proposed algorithm has a very minimal impact on the timings of each sub-blocks.

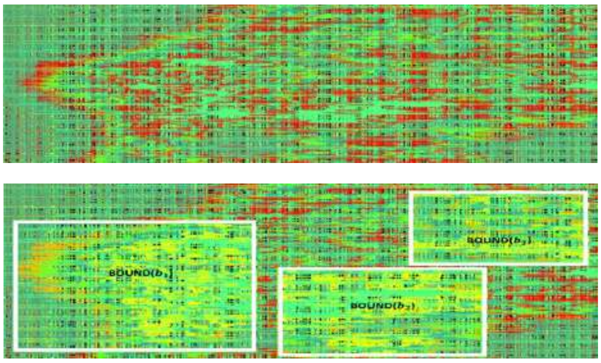
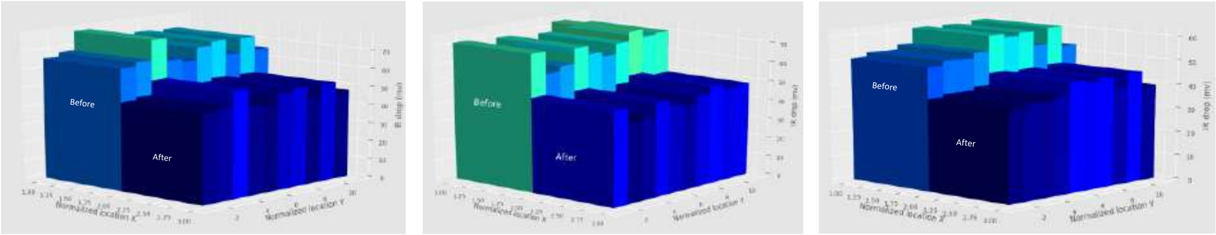


Fig. 5. IR hotspot maps before (Top) and after (Bottom) the application of the IR-mitigation technique.

VI. CONCLUSION

This paper presents a novel dynamic IR-drop optimization flow to eliminate IR hotspots along with the improvement of clock insertion delays. Our proposed approach is to convert the cell placement search into a deterministic linear search. We targeted to achieve the specified IR drop budget, rather than using greedy search to find the most optimal placement. Cell movement can be predicted accurately by the proposed model, and the proposed optimization scheme can mitigate IR-drop in quality and quantity without timing degradation. We acknowledge that the proposed algorithm has a limitation in reducing the hotspots located around macros, within the narrow channels, or around the chip boundary. These hotspots cannot be fixed by cell movement but can be mitigated by local PG augmentation, cell swapping, or decap cell insertion, etc. Our future work will be targeted to address these limitations.



(a) b_1 (b) b_2 (c) b_3

Fig. 6. IR hotspot regions before (Top) and after (Bottom) application of the IR-mitigation technique across different bounds.

TABLE 3. EXPERIMENTAL RESULTS USING PROPOSED ALGORITHM

Design	Bound	Reg-Count	Dynamic IR-Drop (mV)			Clock-Insertion Delay (ns)			Before				After			
									WNS		TNS		WNS		TNS	
			Before	After	% change	Before	After	% change	Setup (ns)	Hold (ns)	Setup (ns)	Hold (ns)	Setup (ns)	Hold (ns)	Setup (ns)	Hold (ns)
D1	b1	540	74	51	31.08	274	174	36.5	0.27	0.10	0	-0.57	0.29	0.11	0	-0.54
D2	b2	610	80	54	32.50	240	182	24.20	0.32	0.13	0	-0.63	0.35	0.15	0	-0.59
D3	b3	630	63	47	25.39	302	209	30.1	0.35	0.15	0	-0.66	0.36	0.17	0	-0.61

REFERENCE

[1] Z. Wang, et al. "Diagnosis of hold time defects" ICCD 2004

[2] S. Edirisooriya and G. Edirisooriya, "Diagnosis of scan failures", in Proc. VLSI Test Symposium 1995, pp. 250-255

[3] R. Guo and S. Venkataraman, "A technique for fault diagnosis of defects in scan chains", in Proc. Inti. Test Con! 2001. pp. 268-277

[4] S. Kundu, "Diagnosing scan Chain faults", IEEE Tran. On VLSI Systems, vol. 2,(4) 1994, pp.512-516

[5] D. Harris and S. Naffziger, "Statistical clock skew modeling with data delay variations", IEEE Trans. On VLSI Systems, vol. 9 (6), Dec. 2001, pp. 888-898

[6] I. Kim and H. B. Min, "Operation about multiple scan chains based on system-on-chip," 2008 International SoC Design Conference, Busan, Korea (South), 2008, pp. II-191-II-194, doi: 10.1109/SOCDC.2008.4815716.

[7] A.B.Kahng,BaoLiu,andQinkeWang.Supplyvoltagedegradationawar eanalytical placement. In Proceedings of International Conference on Computer Design (ICCD), pages 437–443, 2005.

[8] S. i. Heo, A. B. Kahng, M. Kim, L. Wang, and C. Yang. Detailed placement for IR drop mitigation by power staple insertion in sub-10nm VLSI. In Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE), pages 830–835, 2019.

[9] L. Bhamidipati, B. Gunna, H. Homayoun, and A. Sasan. A power delivery network and cell placement aware IR-drop mitigation technique: Harvesting unused timing slacks to schedule useful skews. In Proceedings of Computer Society Annual Symposium on VLSI (ISVLSI), pages 272–277, 2017.

[10] S. A. Swarnkar, K. Rahul, M. Anees and S. Yachareni, "Clock Tree Optimization of FPGA Semi-Custom Memory with SEU FlipFlops," *2020 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS)*, Vancouver, BC, Canada, 2020, pp. 1-4, doi: 10.1109/IEMTRONICS51293.2020.9216346.

[11] A. Tork, M. AbulMakarem and M. Dessouky, "Power Grid Automatic Metal Filling Algorithm Forming Maximum on-chip Decoupling Capacitance," *2007 2nd International Design and Test Workshop*, Cairo, Egypt, 2007, pp. 157-159, doi: 10.1109/IDT.2007.4437450.

[12] S. Chitwood and J. Zheng" IR Drop in High-Speed IC Packages and PCBs", Printed Circuit Design and Manufacture, April 2005

[13] "Synopsys ICC-II." [Online] Available: https://www.synopsys.com/implementation-and-signoff/physical-implementation/ic-compiler.html

[14] "Ansys Totem." [Online] Available: https://www.ansys.com/products/semiconductors/ansys-totem

[15] "Synopsys PrimeTime." Available: https://www.synopsys.com/implementation-and-signoff/signoff/primetime.html

Blockchain Technology to Manage the Energy Supply of Real Estate

Aleksandr Belov
Department of the Applied Mathematics
National Research University "Higher School of Economics
Moscow, Russia
0000-0001-7193-0633

Sergey Slastnikov
Department of the Applied Mathematics
National Research University "Higher School of Economics
Moscow, Russia
sslastnikov@hse.ru

Abstract— **In the framework of creating a digital ecosystem of commercial real estate objects, the main problem is the formation of a digital environment for managing all components of engineering systems that ensure the vital activity of the real estate object.**

The aim of this work is to develop a system for accounting for mutual settlements for electricity consumed on the basis of a distributed ledger using blockchain technology. To assess the effectiveness of the system a simulation model was built using AnyLogic system. Based on the model the system architecture was designed and a software application of the distributed ledger was developed.

Keywords— Blockchain technology, distributed ledger, simulation model, software application

I. Introduction

The electricity market in Russia, like in Europe, is changing dynamically. New energy sales companies are emerging that actively compete with each other for consumers. This competition is especially intensified in the retail electricity market. For organizations involved in the operation of real estate, the tasks of energy management and management of relationships with energy sales companies are extremely relevant. The use of smart contracts based on blockchain technology is becoming a serious competitive advantage for an energy company.

The blockchain is a technology that allows us to conduct transactions peer-based unified network (P2P network, peer-to-peer). Transactions of this type assume that each member of the network can carry out a transaction directly with any other member of the network without the involvement of a third-party intermediary [1]. The data chains are stored in a decentralized manner in a distributed ledger on the participants' devices. A distributed ledger is a distributed database that is stored on the device of each of its participants.

The advantage of this technology is that it has no governing body. It was precisely the absence of intermediaries in this system that aroused keen interest in it.

Lacking a controlling intermediary, the system must be distinguished by another idea: database security. With this in mind, each piece of information entered into this database is reviewed by independent contributors. The technology ensures the security of transactions, has confirmation of the authenticity of data and identity. If someone wants to transfer data, then he must be authenticated, and, as a result, a record of the input / transfer of data will be saved. The data participating in the blockchain is almost impossible to change if there is no access to the keys. The risk of a data breach is unlikely at this stage.

Due to its high degree of security and autonomy, the technology has become popular primarily in the financial industry [2]. The participants themselves check the authenticity of transactions; no third-party regulator is needed.

Energy industry is a promising area for using blockchain technology [3]. In particular this technology can find application at the stage of energy sales. For example, electricity can be sold either directly to consumers or through an independent utility company, and the technology under consideration could enable a utility company or a major supplier to bill consumers directly based on readings recorded on the blockchain.

Typically, electricity suppliers process a huge number of customer requests which undoubtedly affects the speed of decision-making, not to mention careful control over the supply process [4]. Blockchain technology can also be used in this process.

Along with billing for consumed energy blockchain technology can be seen as the backbone of energy logistics processes and metering of consumed electricity.

At the moment there are several well-known companies on the market that provide their services in the development and implementation of blockchain technology in energy companies.

These companies are:

- LO3 Energy and their TransActive Grid project, which is the decentralized open source platform for applications. Built-in tools allow you to measure the level of generation and consumption of electricity in real time, along with other indicators. Despite the fact that the project is under development, its first demo version is currently working in Brooklyn (USA) [5].
- Energy Blockchain Labs – was founded in 2016, the laboratory, together with other companies, is working on projects aimed at developing energy Internet technologies based on blockchain technology and solving problems in the field of generation, consumption, trade and energy management [6].
- Grid Singularity is a green blockchain company that is spearheading the development of an open, decentralized energy data exchange platform under the auspices of the Energy Grid Foundation (EWF) [7].

Now the Russian energy system consists of a small number of large participants between whom there are strong ties and direct communication with the state and vertical integration make the system inactive. Recently, the transition from a continental level to a smaller one (for example, urban or regional) is beginning to take on more and more economic meaning. Such a campaign can serve as

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

a solution to the problem of energy supply to remote territories (for example, the development of renewable energy sources in the Far North). Blockchain technology will ensure the most effective interaction between the buyer and the electricity seller and sell electricity not only without a trade margin and additional costs, but also change its cost depending on production volumes and needs [8].

As a rule, companies engaged in the operation of real estate process a huge volume of data from contracts and customer orders which undoubtedly affects the speed of decision-making. In addition parameter monitoring for energy engineering systems and processes is required. Blockchain technology can be effectively used for solving these tasks. Along with billing for energy consumption blockchain technology can be seen as the basis of energy logistics processes and measuring the amount of electricity consumed [9]. The aim of this work is to develop a system for accounting for mutual settlements for electricity consumed on the basis of a distributed ledger using blockchain technology.

II. FUNCTIONAL MODEL OF THE DISTRIBUTED LEDGER

To design the functional model of a distributed ledger of contracts for electricity consumers of real estate objects we used UML notation [10].

A distributed ledger consists of two types of nodes: a member and a manager. Each of them uses its own software. All actions in the network, such as registration of a new participant, execution of a contract for the supply of electricity to a real estate object, payment of consumed electricity are performed with the participation of the control unit. We will build functional models of operations in the network of the distributed ledger of contracts for the supply of electricity. The node acting as a buyer of electricity, generates a request with the contract data from its side and signs them with a digital signature for identification. The entire data request is signed to protect against changes.

Having received the application the control node checks the data for validity: the ownership of the digital signature and the terms of the electricity purchase agreement. If inaccuracies are found, it is sent for processing to the sender's node. After verification, the control node sends an application to the seller node when it appears on the network. The seller node checks the terms of the contract and, if agreed, carries out the formation of the contract on its part and signs it with a digital signature. Information protection is organized at the same level as at the buyer's site.

The second part of the contract from the seller node (electricity supplier) is sent to the managing node. The control node checks it and forms a block of two parts of the contract and information from the register about the previous block. And also signs it with a digital signature. For the signature, information such as the data of the contracts of both parties, the hash sum of the previous block and the digital signatures of the parties to the transaction are used, thereby the digital signature of the control node protects the entire block.

After that, the received block is sent to all network participants.

Main part is an algorithm for executing a transaction and checking its validity. This algorithm is presented on Fig.1.

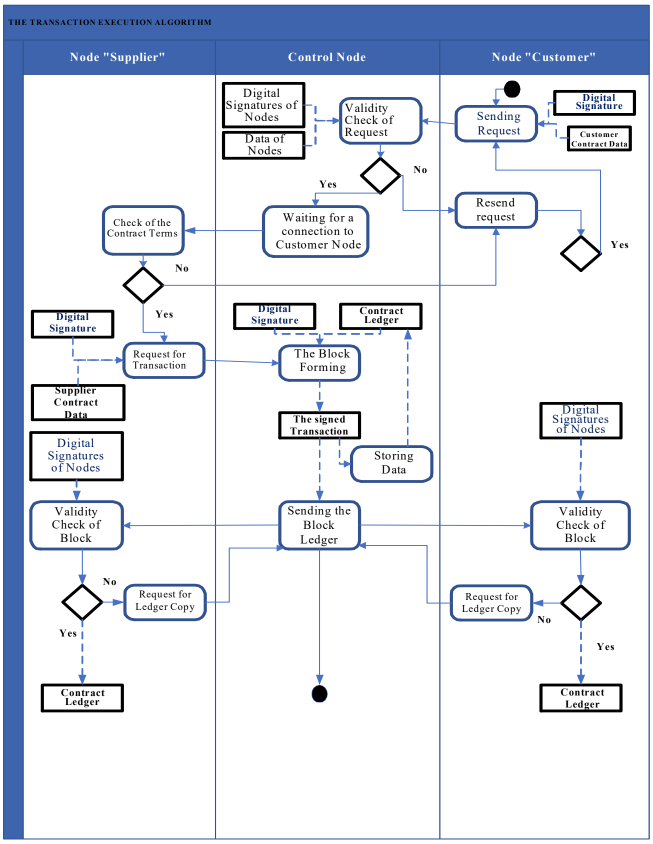


Fig. 1. FLOWCHART OF THE TRANSACTION EXECUTION ALGORITHM

III. MATHEMATICAL MODEL OF THE DISTRIBUTED LEDGER

The model of a distributed ledger is based on the client-server architecture. Additional tasks are performed on the server: identification, request verification, processing and generation and sending of a request error message. To achieve the reliability of our distributed ledger network we proposed additional application servers that will replace the failed main server. To develop a distributed ledger network we use model is based on Petri-Markov net [11].

Let’s consider Petri-Markov net which is presented on Figure 2.

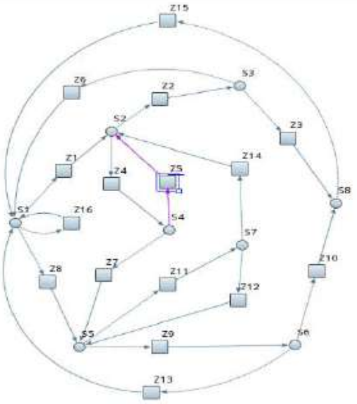


Fig.2 PETRI-MARKOV NET MODEL FOR THE DISTRIBUTED LEDGER

In this model $\{s_{1(s)}, ..., s_{j(s)}\}$ is the set of positions, $Z = \{z_{1(z)}, ..., z_{j(z)}\}$ is the set of transitions. Position S_1 - receipt of a request from the client, S_2 - operational state of the main server, S_3 - the ability to process the request by the main server, S_4 - inoperative state of the main server, S_5 - operational state of the backup server, S_6 - the ability to process the request by the backup server, S_7 - inoperative state of the backup server, S_8 - communication with the database server. Transition $\boldsymbol{Z_1}$ - client request to the main server, $\boldsymbol{Z_2}$ - user identification on the main server, $\boldsymbol{Z_3}$ - access to the database server from the server software, $\boldsymbol{Z_4}$ - failure of the main server, $\boldsymbol{Z_5}$ - operating state of the main server, $\boldsymbol{Z_6}$ - message from the main server about the impossibility of processing the request, $\boldsymbol{Z_7}$ - request transmission to another server, $\boldsymbol{Z_8}$ - client request to the backup server, $\boldsymbol{Z_9}$ - user identification on the backup server, $\boldsymbol{Z_{10}}$ - access to the database server from the backup server software, $\boldsymbol{Z_{11}}$ - failure of the backup server, $\boldsymbol{Z_{12}}$ - operating state of the backup server, $\boldsymbol{Z_{13}}$ - message from the backup server about the impossibility of processing a message, $\boldsymbol{Z_{14}}$- sending a request to another server, $\boldsymbol{Z_{15}}$ - sending processed data to clients, $\boldsymbol{Z_{16}}$ - preparing to receive new data.

According to the proposed Petri-Markov model one can only judge the attainability of states; for a complete assessment of the characteristics of the chosen algorithm, we need to build a simulation model.

IV. Computer Simulation Results

We used AnyLogic 8.3.3 as a simulation tool. It is a universal tool for simulation modeling [12]. The program has a modern graphical interface, is cross-platform and can run on Windows, MacOS and Linux.

The constructed simulation model of the interaction between the client and the control node in the AnyLogic program is shown in Figure 3.

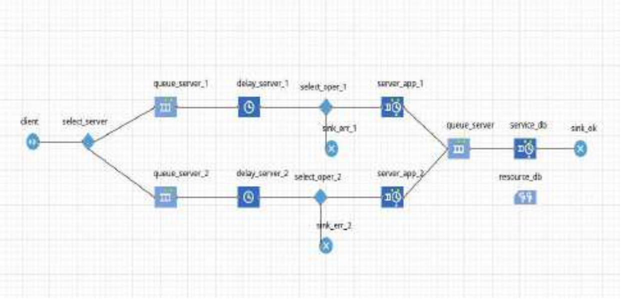


Fig.3 THE SIMULATION MODEL OF CLIENT INTERAGTION WITH THE MANAGING NODE

In this model are presented:

- client - a request from the client program.
- select_server - simulate the failure of the primary server and redirect the request to the backup.
- queue_server_1 and queue_server_2 - queue of requests to the primary and backup servers, respectively.
- delay_server_1 and delay_server_2 - delay for processing a request on the primary and secondary servers, respectively.
- select_oper_1 and select_oper_2 - simulation of operation processing, error content in the request to the primary and backup servers, respectively.
- sink_err_1 and sink_oper_2 - simulates the server response about an error when executing a request to the primary and backup servers, respectively.
- server_app_1 and server_app2 - simulation of transaction processing by the program on the primary or backup server, respectively.
- queue_server - database server queue.
- server_db - modeling the process of processing a transaction request on the database server.
- resource_db - placing transactions into the database.
- sink_ok - simulation of a response to the client about the successful completion of the operation.

We use this model to test the operability of the distributed register of contracts with electricity consumers of real estate objects. Model parameters: probability of failure of the main server - 0.20, latency on the server when processing a request from the client, and latency for applying a transaction on the database server - 75 msec. The rate of requests is measured by the number of requests per second.

The result of performing computer experiments is shown in Table 1.

Test	Request Rate (quantity/msec)	Number of Requests	Application Server Queue (quantity)	DataBase Server Queue (quantity)
1	1	10026	0	0
2	14	10836	1	1
3	20	10543	3	1
4	30	10344	47	1

TABLE I. COMPUTER SIMULATION RESULTS

According to the data obtained as a result of the computer simulation we can conclude that all requests are guaranteed to be served at a high intensity of their arrival in the network. You can also see that with a low intensity of requests the queue does not exceed one request, which means that requests with such an intensity will be processed without waiting in the queue.

When the request rate is more than 20, a sharp increase in the queue length begins, which affects performance.

V. Development of a Smart Contract System for Electricity Suppliers

To store the ledger on all nodes and user information on the control node, the SQLite DBMS was chosen. Each copy of the ledger, as well as user information, is available in single-user mode of the program installed on the node. The data in the SQLite database is stored in one file and when accessing them, direct access is used, this increases the speed of reading and writing.

The SQL language is used as the interface for interaction between the user applications and the DBMS.

To develop the software the C ++ programming language was chosen using the Qt cross-platform library. The Qt library which represents many dynamic link libraries is written in C ++. Its main advantage is cross-platform, that is, the written code can be used to create an application on any supported platform: Windows, Linux, Android, iOS, WindowsPhone, Embedded Linux and many others. Qt Creator was chosen as the development environment, it is the standard development environment for programs using the Qt library and comes in the same package with it.

Figure 4 shows a class diagram for a distributed ledger.

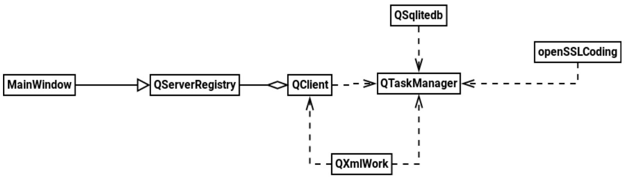


Fig.4 THE CLASS DIAGRAM FOR DISTRIBUTED LEDGER

A distributed ledger is made up of blocks combined into a chain (blockchain) that protects its integrity. The blocks themselves are provided with a digital signature which can be used to identify each of the parties who signed the Contract.

To access the distributed ledger the user uses a private key which is used to sign the block with a digital signature. This key is unique and its theft will not pose a threat to existing blocks in the distributed ledger but it can be used to sign new contracts without the knowledge of its owner, therefore, the storage of this key should be organized by each user of the distributed ledger network individually and convenient for him. The password itself is not stored on the computer only its hash is stored for verification.

In addition to the distributed ledger the control node stores confidential information about users registered on the network, as well as signature data [13]. Confidential user information is used by the control node as user verification; registration and storage is best organized according to the scheme shown in Figure 5. The same applies to a piece of information on a digital signature namely the private key.

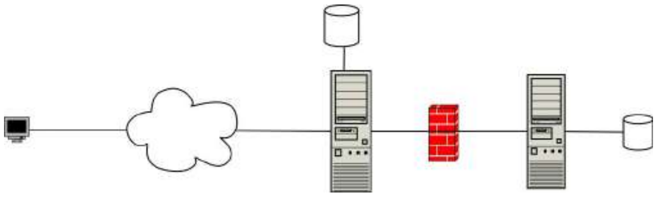


Fig.5 DATA PROTECTION SHEME ON THE CONTROL NODE

All external requests to the server with protected data are restricted by the firewall. Requests are accepted and sent only to the server of the control node which is located in the same local network. Requests for information are limited and can be in the form of true or false which makes it impossible to obtain information if you have access to the local network.

VI. Conclusion and Future works

Blockchain technology has already found wide application in the financial sector and is just beginning to develop in the energy industry, in particular in the electric power industry [14]. The following main advantages of using the technology in question in this industry were highlighted:

- Decentralized storage of transactional data will have a positive impact on the level of protection of this data and will allow for greater independence from the authority performing the centralized administration function.
- The use of blockchain technology will simplify the procedure for making payments using cryptocurrencies, digitalizing contracts, verifying transactions, managing digital content and executing trade operations.
- The technology under consideration will allow the exclusion of third-party intermediaries from the business model previously involved in it.
- Blockchain technology will provide an opportunity for consumers who are also electricity producers and have generating capacities (for example, solar panels and wind generators) at their disposal to sell the electricity they generate not only to their neighbors, but also to the operators of the electrical network. With this approach the purchase price of electricity by electricity grid operators will depend only on production volumes and will not include a trade margin.
- Through the use of blockchains all transactions will be executed in real time and the final settlement will be made only on the basis of the volume actually consumed. This will contribute to the fact that the operators of the electric network no longer need data for making mutual settlements.

The solution to design the distributed ledger of contracts between electricity consumers at real estate objects and electricity suppliers has been presented in this paper. The model of a distributed ledger is based on Petri-Markov networks. To assess the performance of the distributed ledger model at various queries intensities, an original computer simulation algorithm was proposed using the AnyLogic software. When the distributed register of contracts is transferred to the production phase real data will be compared with the simulation results.

The proposed model is based on a closed distributed ledger, in which there is a control node that performs the functions of a network administrator. In the future it is necessary to develop the resulting model and abandon the control node as an administrator in favor of a hybrid model, in which anyone can become a network participant without a lengthy and complex registration and data verification procedure. This will make it easier to use and increase the number of network participants.

Another area of research is the development of a solution that would integrate the transmission of data on consumed electricity using the Internet of Things technology into a smart contract system built using a distributed ledger.

References

[1] Blockchain-new opportunities for producers and consumers of electricity, review of world energy PwC., Available at http://www.pwc.com/utilities / (accessed 26/04/2020)

[2] Y. Guo, C. Liang. Blockchain application and outlook in the banking industry. J Financ Innov, 2016, 2 (1), p. 24

[3] C. Park, T. Yong. Comparative review and discussion on P2P electricity trading. Energy Procedia, 2017, Vol.128, pp. 3-9

[4] C. Eid, P. Codani, Y. Perez, J. Reneses, R. Hakvoort. Managing electric flexibility from Distributed Energy Resources: a review of incentives for market design. Renew Sust Energ Rev, 2016, Vol. 64, pp. 237-247

[5] LO3 Energy, Innovations empowering communities through localized energy solutions, Available at https://lo3energy.com/innovations/, (accessed 1 Feb 2020)

[6] E. Mengelkamp, J. Gärttner, K. Rock, S. Kessler, L. Orsini, C. Weinh ardt. Designing microgrid energy markets A case study: the Brooklyn Microgrid. Appl Energy, 2018, № 210, pp. 870-880

[7] Renewable Energy Markets. Available at https://www.energyweb.org/solutions/renewable-energy-markets/, (accessed 24 Feb 2020)

[8] Ahsan U, Bais A. Distributed big data management in smart grid. In: Wirel Opt Commun Conference (WOCC) 2017, IEEE, 2017, pp. 1–6

[9] M. Andoni, V. Robu, D. Flynn. Blockchains: crypto-control your own energy supply. Nature, 2017, 548 (7666) , p. 158

[10] Craig Larman. Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development, 3rd Edition, 2005, 736 pp.

[11] Murata T. Petri nets: Properties, analysis and applications. Proceedings of the IEEE, 1989, 77(4): pp. 541–580

[12] Artificial Intelligence and Simulation in Business. White Paper. https://www.anylogic.com/resources/white-papers/material-handling-simulation/, (accessed 03 Aug 2020)

[13] M. Mylrea, S.N.G. Gourisetti. Cybersecurity and optimization in smart ‘autonomous' buildings. Autonomy and Artificial Intelligence: A Threat or Savior?, Springer International Publishing, 2017, pp. 263-294

[14] Mihaylov M, Jurado S, Avellana N, Van Moffaert K, de Abril IM, Nowe A. Nrgcoin: Virtual currency for trading of renewable energy in smart grids. In: Proceedings of the 11th International Conference European Energy Market (EEM), IEEE, 2014, pp. 1–6

Evaluation and Mitigation of Crosstalk Effects in Silicon Bolometers Arrays

1st Mario Eduardo B.G.N. Silva
School of Electrical and Computer Engineering
University of Campinas
Campinas/SP, Brazil
mebgns@dsif.fee.unicamp.br

2nd Leandro Tiago Manera
School of Electrical and Computer Engineering
University of Campinas
Campinas/SP, Brazil
lemanera@unicamp.br

***Abstract*—Resistive infrared detectors arrays (bolometers) may show interferences between devices. Such interferences called crosstalk, or more specifically sneak path, reduce the signal quality, causing reading errors. This work proposes an interesting alternative solution to avoid sneak paths in bolometers arrays. The idea is to use simple diodes in series with the detectors, so all possible sneak paths are grounded and the detectors are only trigged when the diodes were conducting. Test were conducted in purely resistor arrays and diode-resistor arrays. The target resistance of one element was set to 105 kΩ. For both isolated devices and matrix devices elements, the target resistance was achieved, indicating that the sneak path effect has been avoided. So, this solution, in addition to improving the signal quality, is also a simpler construction when compared with transistors switches solution, requiring less metal levels, since no tracks are required to activate the transistor.**

***Index Terms*—bolometer; crosstalk; sneak path; resistor; diodes; infrared; detectors; array**

I. Introduction

The goal of this work is to analyze the signal interferences that occur in bolometric detectors arrays, and to propose an innovative solution to avoid the so called sneak path problem.

Bolometric detectors are made of resistors whose resistance varies with temperature when influenced by electromagnetic radiation, and one of their applications is the generation of thermal images.Thermal images are generated from a detector array whose thermal surface distribution contrast is observed as a function of time [1].

The term ”focal plane array” (FPA) refers to a set of individual image detector elements (”pixels”) located in the focal plane of an image system. Although the definition may include one-dimensional arrays, as well as two-dimensional (2D) arrays, it is often applied to the latter ones [2]. This work is focused on 2D arrays.

Fig. 1 shows a three-dimensional array connections scheme used in bolometer FPAs. The sneak path problem can be observed in this kind of arrangement.

A. The Sneak Path Problem

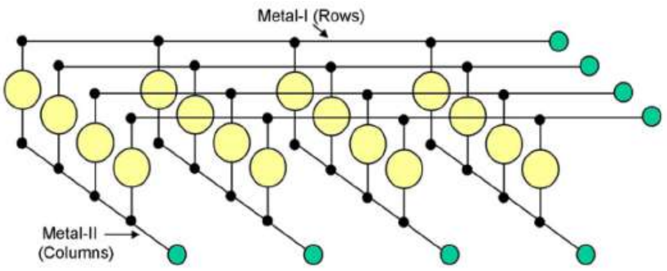


Fig. 1. Schematic diagram showing metals rows and columns connection scheme [3].

The sneak path problem can be described as a crosstalk interference between adjacent cells caused by the sneak path current. This is an undesired effect and may cause signal misinterpretations [4], [5], as shown in Fig. 2.

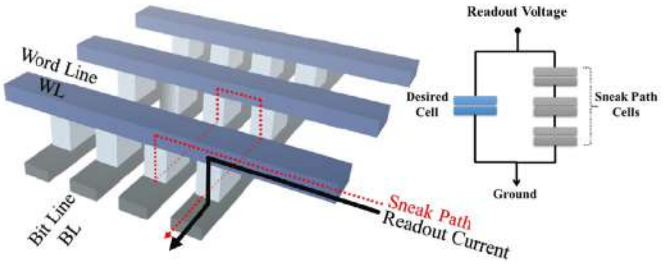


Fig. 2. Representation of sneak path cells and sneak path current in a array cell [6].

As can be seen in Fig. 2, the sneak path problem occurs when one element is addressed and other elements form a parasitic circuit in parallel. This parasitic circuit provides an undesired current pathway, since this current activate not only the addressed device, but also all other adjacent elements, leading to a signal misinterpretation.

Many solutions have been presented to solve the sneak path problem, including solutions based on transistors, diodes, volatile switches and non-linear devices [7].

The most common solution to solve sneak path problems in bolometric FPAs is the use of transistors in series with each detector. Memristors arrays technology devices have been using diodes as a popular way to solve the addressing problems, due to its simplest structure and easy manufacturing

This work was supported by Brazilian agencies CAPES (process nº BEX 1006/14-1) and CNPq (process nº 153630/2012-4).

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

processes [7].

As shown in Fig. 3, a diode is added to each memory cell producing a new diode and a memristor cell (1D1M). The sneak path problem is one of the main difficulties when building memories with three-dimensional design [6]

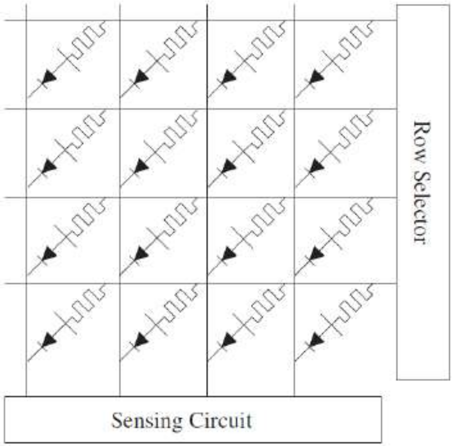


Fig. 3. Simple memory array with a diode in series with a memristor for each memory cell [5].

As can be seen in Fig. 4 the unaddressed diodes cells end up being reversed polarized, and do not conduct current, preventing the electrons from flowing in unwanted paths.

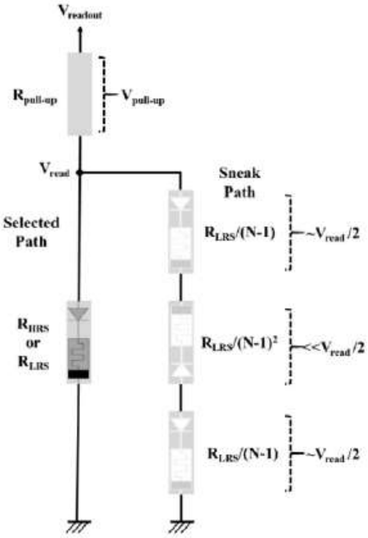


Fig. 4. Equivalent circuit of the 1D1R selection device [6].

The proposal in this work is to adapt the memristor diode solution for bolometric FPAs. Such solution is presented in [8]. Many works have been found in the literature having applications in memristors and pressure detectors [9], but none with any implementation in bolometric detectors.

II. Proposed Solution

The circuits are assembled in the configuration of NxN arrays, and the elements are read through row and column addressing. In order to test for the sneak problem effect, two different projects were fabricated and tested. Purely resistive elements and resistor diode elements. For purely resistive arrays with identical elements, it is possible to calculate the resistance measured between any pair of X (vertical) and Y (horizontal) wires by (1).

$$R_{eq} = R(\frac{2N-1}{N^2}) \tag{1}$$

Where:

N: Number of elements on the side of the square array.

A: Elements resistance.

Equation (1) was obtained by circuit analysis.

So for arrays composed of 100 kΩ resistors and using (1), one may find the value on any addressed element, as shown in the table I.

TABLE I
Calculated Resistances

Array	Resistance
Isolated Device	100 kΩ
2x2	75 kΩ
3x3	55,55 kΩ
4x4	43,75 kΩ
8x8	23,43 kΩ

As already mentioned, the solution proposed in this work have a diode added to each resistor producing a new cell of a diode and a resistor (1D1R). The advantages of this proposal are:

- Reduced manufacturing processes, since a diode is much simpler than a transistor.
- Reduction of metal levels, since no connections are required to power the transistor drives.
- Increased density sensors in the DIE.

An important observation is that in this work the detectors array and the sensing circuit are fabricated separately.

A. Implementation

Initially an electrical simulation was performed for both purely resistive elements and for 1D1R elements. Then the simulated arrays were implemented using a printed circuit board (PCB) and commercial devices (100 kΩ ± 5% and IN4148 diodes). In order to easily check the circuit configuration setup, electrical measurements were performed in both circuits by applying a DC voltage (HP3245A) and measuring the current using a standard multimeter (HP3458A).

After the measurements with the PCB, an integrated circuit (IC) design was started using the XFAB XC06 technology (Fig. 5). The manufactured IC was designed having an array 8x8 of resistors, an array 8x8 of 1D1R elements, and some isolated resistors and diode for testing purposes. Where used RPOLYH resistors (high resistive poly0 resistor) of 100 kΩ, and DP diodes.

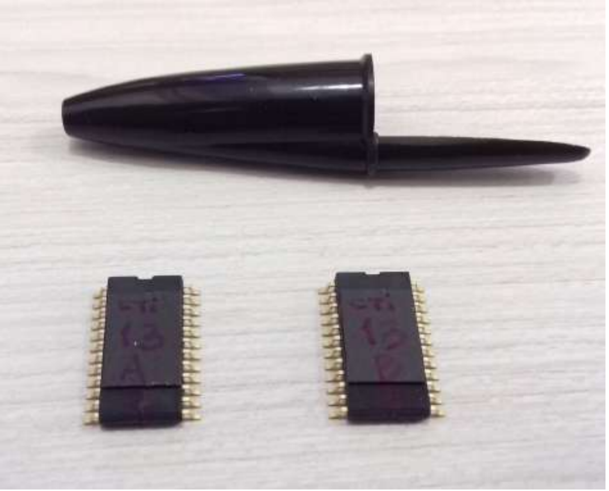


Fig. 5. fabricated ICs.

III. Results and Discussion

A. Electrical Simulations Results

Fig. 6 shows the IxV curves for an isolated resistor and arrays of 2x2, 3x3 and 4x4 elements. These values are i agreement with the results presented in table I.

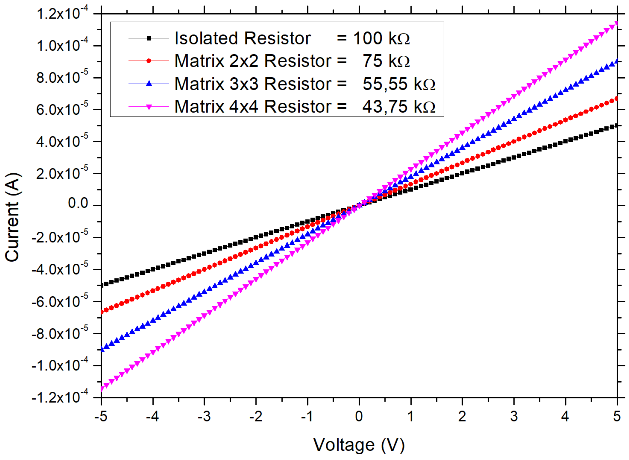


Fig. 6. Current vs Voltage of the simulated resistor arrays.

The simulated IxV curves for arrays with 1D1R elements are shown in Fig. 7. As expected, all values presented are practically identical, indicating that the “diode” solution have worked, avoiding sneak paths on devices that were not the measurement subject. The diode operation can be observed since all curves have a zero current value (diode leakage current in order of nA) in the negative voltage range and start to rise at about 0.3 V, reaching a linear operation at 0.7 V.

B. Electrical Measurements in PCB Arrays

The IxV curves for the PCB resistor arrays are shown in Fig. 8. These values are also in agreement with the results presented in the simulations and in table I.

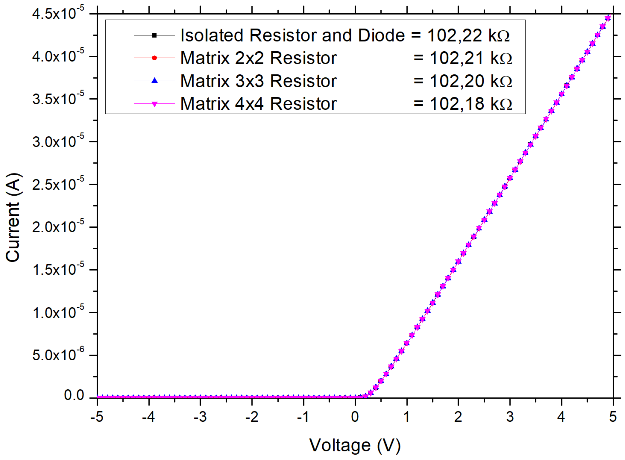


Fig. 7. Current vs Voltage of the simulated 1D1R arrays.

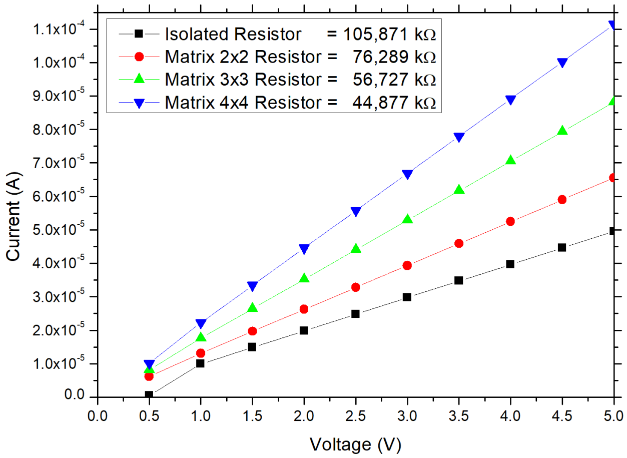


Fig. 8. Current vs Voltage of PCB resistor arrays.

The IxV curves for the PCB arrays of 1D1R elements are shown in Fig. 9. They are all overlapping indicating the effectiveness proposed by the diodes solution, resulting in no sneak paths problem.

C. IC simulations results

The IC design was performed using Cadence tools. The simulated IxV curves for the isolated devices are shown in Fig. 10. The isolated resistor have presented a 100 kΩ, as designed, for both negative and positive voltages. On the other hand, the 1D1R element has null value up to approximately 0.7 V and a resistance value close to 100 kΩ, as expected for a resistor diode configuration.

An interesting result can be observed in Fig. 11. For the IxV curve of 8x8 arrays, a device suffering from sneak path

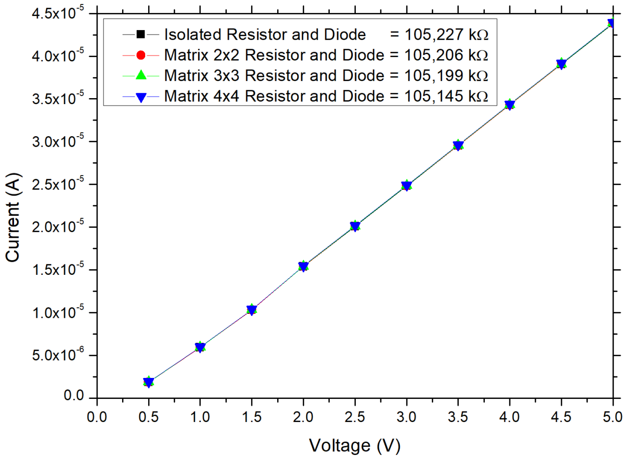


Fig. 9. Current vs Voltage of the PCB 1D1R arrays.

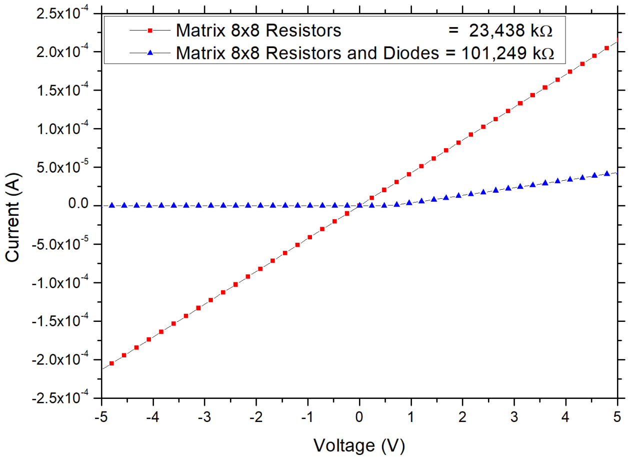


Fig. 11. Current vs Voltage of the 8x8 array simulated in Cadence

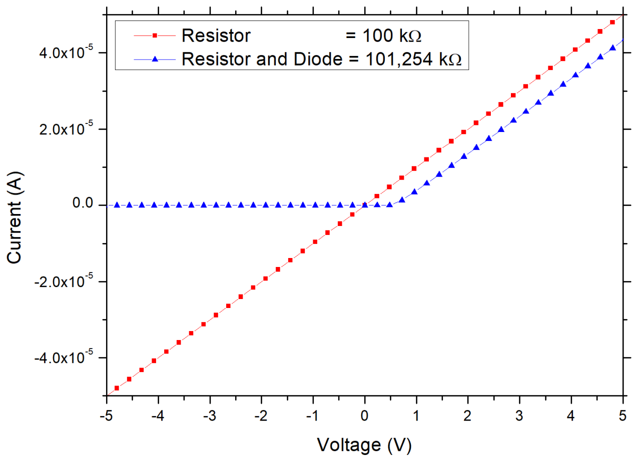


Fig. 10. Current vs Voltage of the isolated device simulation

problem, should present a resistance of about 23 kΩ, as pointed out in table I. As one can see, this value was reached for the resistor element array. Nonetheless, the cadence layout simulation for 1D1R elements have presented approximately 100 kΩ value and a resistor diode like curve, again, proving that the diodes were efficient in canceling the sneak paths.

D. IC Fabricated Measurements

Since all measurements and simulations have indicated so far that the solution proposed was indeed efficient, a IC measurement was performed. The chip was packaged and a test bench was designed for IxV measurements. The IxV curves measurements for the isolated devices are shown in Fig. 12. The curves follow almost in parallel, indicating a very close resistance value. The offset between the curves is due to the use of the diodes. Both numbers are in agreement with the designed target resistance value.

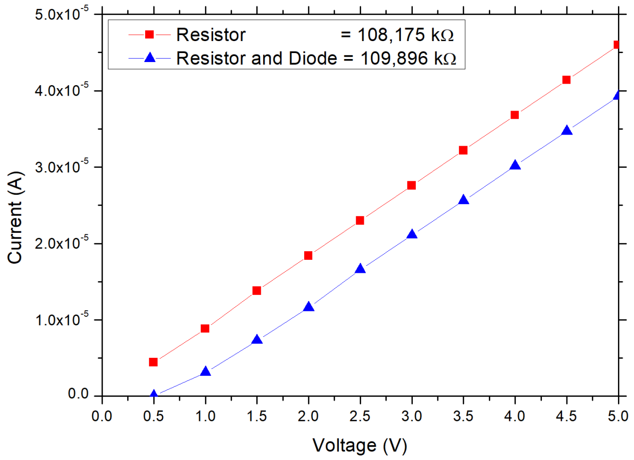


Fig. 12. Current vs Voltage of isolated devices fabricated inside the IC.

The IxV curve of 8x8 arrays, shown in Fig. 13, have presented misleading results for the 1D1R element. For the resistor array the expected 23 kΩ resistance value was obtained. In the other hand, due to lack of a guard ring for grounding the diodes, a non-expected behavior was achieved in the 1D1R elements.

As a summary, tables II, III, IV and V show all resistance values obtained through simulations and measurements. All implementations presented consistent results, except for the 1D1R elements of the IC 8x8 matrix. It was expected that the 8x8 1D1R IC array presented a similar value as the isolated resistor diode device, which did not happen, differing from all previous results. It was used the XFAB XC06 technology

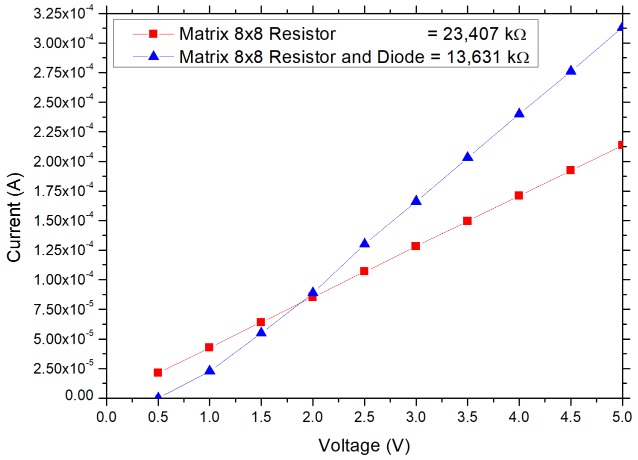


Fig. 13. Current vs Voltage of array 8x8 fabricated inside the IC.

that requires (mandatory) guard ring for grounding procedures. Since this requirement was not verified by the DRC, unfortunately the proper grounding has not been done. Some more testing were performed and it was possible to observe that, since some diodes are not properly grounded, they may conduct during the measurements. Therefore a future work the IC is being designed with guard rings.

TABLE II
RESISTANCES 1

Array	Calculated R	Simulation R	Simulation RD
Isolated Devices	100 kΩ	100 kΩ	102,22 kΩ
2x2	75 kΩ	75 kΩ	102,21 kΩ
3x3	55,55 kΩ	55,55 kΩ	102,20 kΩ
4x4	43,75 kΩ	43,75 kΩ	102,18 kΩ

TABLE III
RESISTANCES 2

Array	PCB R	PCB RD
Isolated Devices	105,871 kΩ	105,227 kΩ
2x2	76,289 kΩ	105,206 kΩ
3x3	56,727 kΩ	105,199 kΩ
4x4	44,877 kΩ	105,145 kΩ

TABLE IV
RESISTANCES 3

Array	Calculated R	Cadence R	Cadence RD
Isolated Devices	100 kΩ	100 kΩ	101,254 kΩ
8x8	23,43 kΩ	23,438 kΩ	101,249 kΩ

IV. CONCLUSION

In this work, sneak paths between resistive elements of bolometric FPA were analyzed. The goal of this work is to overcome the signal interferences that occur in bolometric detectors arrays, proposing an innovative solution to avoid this problem. For this purpose, isolated and arrays elements were designed, simulated, manufactured in PCB and fabricated trough IC design. Through the measurements of arrays made up of resistors, it was possible to clearly demonstrate the sneak paths effect. The proposed solution efficiency was demonstrated in the diodes resistors arrays. All electrical simulations and measurements performed in the PCB design have confirmed the effectiveness of the proposed solution. Tests were conducted in purely resistor arrays and diode-resistor arrays. The target resistance of one element was set to 105 kΩ. For both isolated devices and matrix devices elements, the target resistance was achieved, indicating that the sneak path effect has been avoided. However, for the ICs, it was expected that the 8x8 1D1R IC array presented a similar value as the isolated resistor diode device, and the same did not happen. This fact arises from the lack of a guard ring for grounding the diodes. Nonetheless, it was presented an effective solution that, in addition to improving the signal quality, is also a simpler construction when compared with transistors solution, requiring less metal levels, also avoiding sneak path problems at all.

TABLE V
RESISTANCES 4

Array	IC R	IC RD
Isolated Devices	108,175 kΩ	109,896 kΩ
8x8	23,407 kΩ	13,631 kΩ

ACKNOWLEDGMENT

The authors acknowledge to CEITEC (Porto Alegre/RS) for the samples fabrication, and also R. C. Teixeira (Electronic Packaging staff) from Renato Archer Information Technology Center (CTI-Renato Archer, Campinas/SP) for the samples packaging.

REFERENCES

[1] J. Caniou, *Passive infrared detection: Theory and Applications*. Springer, 2011.

[2] A. Rogalski and K. Chrzanowski, “Infrared devices and techniques (revision),” *Metrology and Measurement Systems*, vol. 21, no. 4, pp. 565–618, dec 2014.

[3] R. S. Saxena, R. K. Bhan, C. R. Jalwania, P. S. Rana, and S. K. Lomash, “Characterization of area arrays of microbolometer-based un-cooled IR detectors without using ROIC,” *Sensors and Actuators A: Physical*, vol. 141, no. 2, pp. 359–366, feb 2008.

[4] C. Xu, D. Niu, N. Muralimanohar, R. Balasubramonian, T. Zhang, S. Yu, and Y. Xie, “Overcoming the challenges of crossbar resistive memory architectures,” in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, feb 2015.

[5] M. A. Zidan, H. A. H. Fahmy, M. M. Hussain, and K. N. Salama, “Memristor-based memory: the sneak paths problem and solutions,” *Microelectronics Journal*, vol. 44, no. 2, pp. 176–183, feb 2013.

[6] F. Gül, “Addressing the sneak-path problem in crossbar RRAM devices using memristor-based one schottky diode-one resistor array,” *Results in Physics*, vol. 12, pp. 1091–1096, mar 2019.

[7] A. Chen, “Memory select devices,” in *Emerging Nanoelectronic Devices*. John Wiley & Sons Ltd, nov 2014, pp. 227–245.

[8] D. C. Osborn, “Sneak paths in X-Y matrix arrays,” *IEEE Journal of Solid-State Circuits*, vol. 4, no. 6, pp. 312–317, Dec. 1969.

[9] J.-F. Wu, “Scanning approaches of 2-d resistive sensor arrays: a review,” *IEEE Sensors Journal*, vol. 17, no. 4, pp. 914–925, feb 2017.

Statistical and Transfer Learning Model to Analyze Endurance Performance with Aging

He Huang
University of North Dakota
Grand Forks, ND, USA
he.huang@und.edu

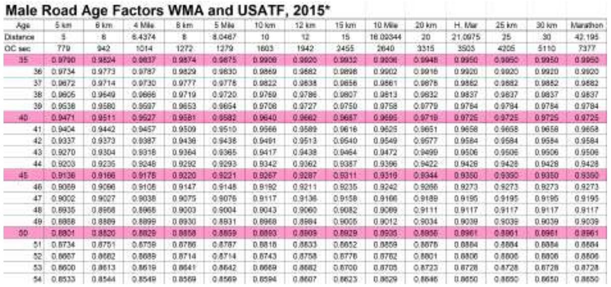
Ebrima N Ceesay
George Mason University
Fairfax, VA, USA
eceesay2@gmu.edu

***Abstract*—The decline of human physical ability with aging is an important factor to consider as to athletic performance evaluation. Physiological factors or real-world performance data have been used to build mathematical models. Marathon is a sports event attracting more and more participants. However, due to the limitation of field size, the race organizers usually must qualify each athlete for certain events based on age and gender. We use statistical and machine learning models for analyzing the Boston Marathon results to delineate the decrease in elder athletes' endurance performance. We also apply the transfer learning model on the mobile activity tracking social media platform Strava to classify beginners or professional athletes using performance data and applying sentiment classification to athletes' comments. Our results may provide valuable insights for athletes, coaches, and governing bodies for athletics.**

***Index Terms*—Machine learning, Statistics, Sport, Natural language processing**

I. Introduction

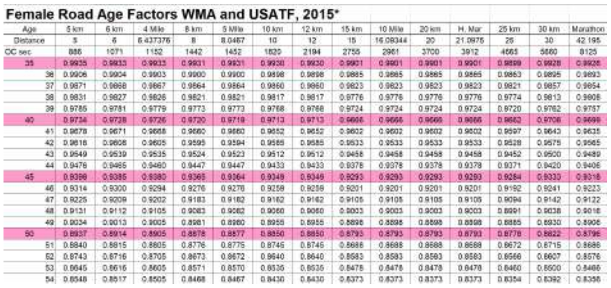
The change of endurance performance with aging is an interesting topic. It applies to the world-class athletes competing at the highest level, such as the Olympics and World Championship events. Also, it has drawn more and more attention from the general population with the popularity of road racing, cycling, and triathlon events [1]. For example, according to the "2018 U.S. Running Trends Report", near 18.3 million registrants entered road race events in 2017 [2]. We may expect a gigantic number of participants in all endurance events worldwide, which indicates a thriving industry of recreational endurance competition. Many of these athletes have passed their best physiological period for endurance performance. However, they still strive for excellence in their performance to the best of their ability. Understanding the change of human physical ability may provide important information to evaluate elder athletes' performance and guidance on how to manage expectations and design training programs for them. For instance, World Masters Athletics (WMA) adopts statistician Alan Jones' Age Grade Tables to make race performance of all ages comparable across the board. Fig. 1 and 2 show excerpts from World Masters Athletics (WMA) Running Age-Grade Tables by Alan Jones (https://github.com/AlanLyttonJones/Age-Grade-Tables). The table calculates the age-grading factor as the current world record of an event divided by the current world best time of each age in that event. The age-graded time of an athlete can be presented as the race time multiply by the age-grading factor of certain age and event [3]. The age grading system calculates athletes' race times as the percentage of the current world record times of corresponding age and gender, which gives a normalized scale to assess competitiveness.



Male Road Age Factors WMA and USATF, 2015*

Age	5 km	6 km	4 Mile	8 km	5 Mile	10 km	12 km	15 km	10 Mile	20 km	H. Mar	25 km	30 km	Marathon
Distance	5	6	6.4374	8	8.0467	10	12	15	16.09344	20	21.0975	25	30	42.195
OC sec	779	942	1014	1272	1279	1603	1942	2455	2640	3315	3503	4205	5110	7377
35	0.9790	0.9824	0.9837	0.9874	0.9875	0.9908	0.9920	0.9932	0.9936	0.9948	0.9950	0.9950	0.9950	0.9950
36	0.9734	0.9773	0.9787	0.9829	0.9830	0.9869	0.9882	0.9898	0.9902	0.9916	0.9920	0.9920	0.9920	0.9920
37	0.9672	0.9714	0.9730	0.9777	0.9778	0.9822	0.9838	0.9856	0.9861	0.9878	0.9882	0.9882	0.9882	0.9882
38	0.9605	0.9649	0.9666	0.9719	0.9720	0.9769	0.9786	0.9807	0.9813	0.9832	0.9837	0.9837	0.9837	0.9837
39	0.9538	0.9580	0.9597	0.9653	0.9654	0.9708	0.9727	0.9750	0.9758	0.9779	0.9784	0.9784	0.9784	0.9784
40	0.9471	0.9511	0.9527	0.9581	0.9582	0.9640	0.9662	0.9687	0.9695	0.9719	0.9725	0.9725	0.9725	0.9725
41	0.9404	0.9442	0.9457	0.9509	0.9510	0.9566	0.9589	0.9616	0.9625	0.9651	0.9658	0.9658	0.9658	0.9658
42	0.9337	0.9373	0.9387	0.9436	0.9436	0.9491	0.9513	0.9540	0.9549	0.9577	0.9584	0.9584	0.9584	0.9584
43	0.9270	0.9304	0.9318	0.9364	0.9365	0.9417	0.9438	0.9464	0.9472	0.9499	0.9506	0.9506	0.9506	0.9506
44	0.9203	0.9236	0.9248	0.9292	0.9293	0.9342	0.9362	0.9387	0.9396	0.9422	0.9428	0.9428	0.9428	0.9428
45	0.9136	0.9166	0.9178	0.9220	0.9221	0.9267	0.9287	0.9311	0.9319	0.9344	0.9350	0.9350	0.9350	0.9350
46	0.9069	0.9096	0.9108	0.9147	0.9148	0.9192	0.9211	0.9235	0.9242	0.9266	0.9273	0.9273	0.9273	0.9273
47	0.9002	0.9027	0.9038	0.9075	0.9076	0.9117	0.9136	0.9158	0.9166	0.9189	0.9195	0.9195	0.9195	0.9195
48	0.8935	0.8958	0.8968	0.9003	0.9004	0.9043	0.9060	0.9082	0.9089	0.9111	0.9117	0.9117	0.9117	0.9117
49	0.8868	0.8889	0.8899	0.8930	0.8931	0.8968	0.8984	0.9005	0.9012	0.9034	0.9039	0.9039	0.9039	0.9039
50	0.8801	0.8820	0.8829	0.8858	0.8859	0.8893	0.8909	0.8929	0.8935	0.8956	0.8961	0.8961	0.8961	0.8961
51	0.8734	0.8751	0.8759	0.8786	0.8787	0.8818	0.8833	0.8852	0.8859	0.8878	0.8884	0.8884	0.8884	0.8884
52	0.8667	0.8682	0.8689	0.8714	0.8714	0.8743	0.8758	0.8776	0.8782	0.8801	0.8806	0.8806	0.8806	0.8806
53	0.8600	0.8613	0.8619	0.8641	0.8642	0.8669	0.8682	0.8700	0.8705	0.8723	0.8728	0.8728	0.8728	0.8728
54	0.8533	0.8544	0.8549	0.8569	0.8569	0.8594	0.8607	0.8623	0.8629	0.8646	0.8650	0.8650	0.8650	0.8650

Fig. 1. Excerpts from World Masters Athletics (WMA) Male Running Age-Grade Tables by Alan Jones. The table calculates the age-grading factor as the current world record of an event divided by the current world best time of each age in that event. The age-graded time of an athlete can be presented as the race time times the age-grading factor of certain age and event.



Female Road Age Factors WMA and USATF, 2015*

Age	5 km	6 km	4 Mile	8 km	5 Mile	10 km	12 km	15 km	10 Mile	20 km	H. Mar	25 km	30 km	Marathon
Distance	5	6	6.437376	8	8.0467	10	12	15	16.09344	20	21.0975	25	30	42.195
OC sec	886	1071	1152	1442	1452	1820	2194	2755	2961	3700	3912	4665	5880	8125
35	0.9935	0.9933	0.9933	0.9931	0.9931	0.9930	0.9930	0.9901	0.9901	0.9901	0.9901	0.9899	0.9928	0.9926
36	0.9906	0.9904	0.9903	0.9900	0.9900	0.9898	0.9898	0.9865	0.9865	0.9865	0.9865	0.9863	0.9895	0.9893
37	0.9871	0.9868	0.9867	0.9864	0.9864	0.9860	0.9860	0.9823	0.9823	0.9823	0.9823	0.9821	0.9857	0.9854
38	0.9831	0.9827	0.9826	0.9821	0.9821	0.9817	0.9817	0.9776	0.9776	0.9776	0.9776	0.9774	0.9813	0.9808
39	0.9785	0.9781	0.9779	0.9773	0.9773	0.9768	0.9768	0.9724	0.9724	0.9724	0.9724	0.9720	0.9762	0.9757
40	0.9734	0.9728	0.9726	0.9720	0.9719	0.9713	0.9713	0.9666	0.9666	0.9666	0.9666	0.9662	0.9706	0.9699
41	0.9678	0.9671	0.9668	0.9660	0.9660	0.9652	0.9652	0.9602	0.9602	0.9602	0.9602	0.9597	0.9643	0.9635
42	0.9616	0.9608	0.9605	0.9595	0.9594	0.9585	0.9585	0.9533	0.9533	0.9533	0.9533	0.9528	0.9575	0.9565
43	0.9549	0.9539	0.9535	0.9524	0.9523	0.9512	0.9512	0.9458	0.9458	0.9458	0.9458	0.9452	0.9500	0.9489
44	0.9476	0.9465	0.9460	0.9447	0.9447	0.9433	0.9433	0.9378	0.9378	0.9378	0.9378	0.9371	0.9420	0.9406
45	0.9398	0.9385	0.9380	0.9365	0.9364	0.9349	0.9349	0.9293	0.9293	0.9293	0.9293	0.9284	0.9333	0.9318
46	0.9314	0.9300	0.9294	0.9276	0.9276	0.9258	0.9258	0.9201	0.9201	0.9201	0.9201	0.9192	0.9241	0.9223
47	0.9225	0.9209	0.9202	0.9183	0.9182	0.9162	0.9162	0.9105	0.9105	0.9105	0.9105	0.9094	0.9142	0.9122
48	0.9131	0.9112	0.9105	0.9083	0.9082	0.9060	0.9060	0.9003	0.9003	0.9003	0.9003	0.8991	0.9038	0.9016
49	0.9034	0.9013	0.9005	0.8981	0.8980	0.8955	0.8955	0.8898	0.8898	0.8898	0.8898	0.8885	0.8930	0.8906
50	0.8937	0.8914	0.8905	0.8878	0.8877	0.8850	0.8850	0.8793	0.8793	0.8793	0.8793	0.8778	0.8822	0.8796
51	0.8840	0.8815	0.8805	0.8776	0.8775	0.8746	0.8746	0.8688	0.8688	0.8688	0.8688	0.8672	0.8715	0.8686
52	0.8743	0.8716	0.8705	0.8673	0.8672	0.8640	0.8640	0.8583	0.8583	0.8583	0.8583	0.8566	0.8607	0.8576
53	0.8645	0.8616	0.8605	0.8571	0.8570	0.8535	0.8535	0.8478	0.8478	0.8478	0.8478	0.8460	0.8500	0.8466
54	0.8548	0.8517	0.8505	0.8468	0.8467	0.8430	0.8430	0.8373	0.8373	0.8373	0.8373	0.8354	0.8392	0.8358

Fig. 2. Excerpts from World Masters Athletics (WMA) Female Running Age-Grade Tables by Alan Jones. The table calculates the age-grading factor as the current world record of an event divided by the current world best time of each age in that event. The age-graded time of an athlete can be presented as the race time times the age-grading factor of certain age and event

Researchers have been exploring the underlying mechanisms of decline in endurance performance with aging from the perspective of human physiology [4] [5]. The main components include exercise economy, lactate threshold, maximal aerobic capacity, *et cetera* [6]. A comprehensive literature review systematically summarized the physiological factors responsible for the change of endurance performance with aging, for instance, age-related decrease in maximal oxygen

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

uptake (VO2max), maximum heart rate (HRmax), stroke volume, arteriovenous oxygen difference, active muscle mass, type II muscle fiber size, and blood volume [7]. Another recent review examined the contributions of vascular senescence to the degeneration of endurance and physical ability in the more senior population [8]. A pioneer 22-year longitudinal study revealed a significant decline of VO2max in a group of fifty-three former elite distance runners [9]. Although VO2max has been regarded as a poor indicator for endurance performance and one of the most misused physiological parameters [10], this work provided a valuable perspective of how aging may affect the human body. A more informative set of results would present measurements at multiple time points during the investigation period and backed up by relatively larger sample size. An investigation of oxygen uptake at the anaerobic threshold (VO2AT) in elite mountain runners showed that a significant decrease of VO2AT began after 49 years old [11].

Interestingly, a study about the effects of aging on indoor rowing performance revealed gender differences – while the decline with aging displayed a linear relationship in women, a biphasic model was more appropriate to fit the data collected from male subjects [12]. Later, an article showed that from age 25 to 85, women's performance in stationary rowing declined twice as faster as men's [13]. Another study investigating the change of swimming performance with aging demonstrated gender difference in swimming events of some distances [14]. These results suggested that it would be imperative to differentiate genders while evaluating the physiological effects caused by aging and build a robust data model. Moreover, an analysis of top Masters performance in athletics, swimming, rowing, cycling, triathlon, and weightlifting demonstrated that aging might affect various sport events differently [15]. Noticeably, an interview study explored the non-physiological factors that may account for the changes of performance with aging, such as psychological factors [16]. Their research suggested that analyzing physiological data may not provide a complete picture of aging's effects on sports performance.

As recreational endurance activities have been in blossom for many decades, a large amount of data is constantly being generated, which is characterized by the properties of big data – volume, variety, and velocity. In a study in 2012, they analyzed the top ten performances of each group based on age and gender in the New York City Marathon (1980-2009) [17]. The authors found out that despite the decline of aging performance, the race results within each age group had been improving. They attributed this trend to the overall increase of participation of master athletes in endurance sports, better training facilities, and a higher spirit of competitiveness. Although their study revealed a trend in a relatively long period, the top ten athletes' results could not well represent the broader population. Given this trend, analyzing current data is very helpful to disclose valuable patterns. Here, we propose to leverage the vast amount of data from recent race results to build a statistical model that can unveil the mathematical relationship between endurance performance and age. Our model can be applied to predict athletes' future performance and for the athletic governing bodies and race organizers to set and/or revise age-group categories and qualifying standards. For example, how do we properly define different athletics divisions: junior, open, masters based on statistics instead of just conventions. Additionally, as a reference, professional athletes and coaches must plan their careers accordingly since they want to reach the peak of their career before the decline of physical ability begins.

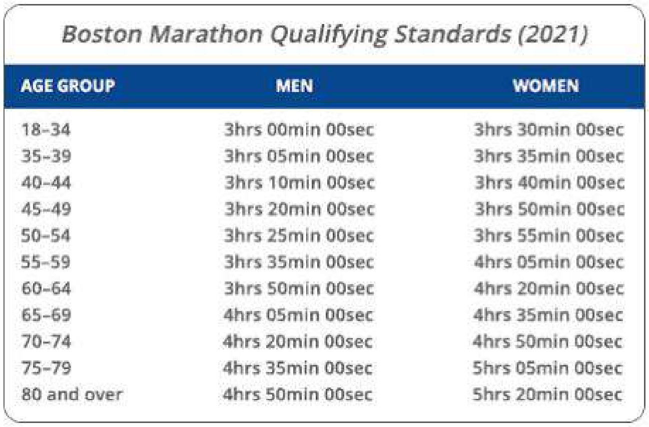
The user friendly applications that records activities during the exercise can be a valuable source of numeric and textual data [18] [19] [20], [21]. These applications create social media platforms where athletes can share their sport activities. It is important for these applications to provide reliable health information to prevent possible injuries for beginner athletes, especially while they are aging. Expert misinformation detection systems can enhance trust and reliability of health information which is provided by smart devices and applications [22] . Extraction of reliable textual information about the trail and body health in smart devices is a critical feature in creating safe environment for athletes [23].

II. Research design and data collection

Researchers can adopt two approaches, cohort study or cross-sectional study, to investigate the effects of aging on endurance performance. There are both advantages and limitations for these two approaches, which are discussed in detail below.

A. *Cohort study*

A cohort study design tracks the race performance of athletes for a relatively long time. The researchers evaluate the change of race performance and/or physiological parameters with aging. Those studies have been usually focusing on elite athletes. For example, a French group analyzed the data collected from individual athletes' careers in track and field and swimming events in three decades (from 1980 to 2009) and the ranking of chess players [24]. They discovered a relationship between performance and age. There are limitations in the design of cohort study when it comes to exploring the change of human physical abilities with the aging process. First of all, although performance data collected from elite athletes and world records may be considered an excellent representation of true human abilities at the highest level, the data are limited by sample size. Secondly, most professional athletes may only have a very short span of career during their peak phase. After they retire, they are no longer competing or competing at the level of efforts they used to be. Hence, we may not be able to collect data during a time window that is wide enough to demonstrate human ability change with aging. Moreover, even when they are at their peak phase, their performances may fluctuate because of other reasons instead of aging, such as prolonged injuries or illness, changes of training methods or coaches, adjustments to diet *et cetera*. Those factors may greatly affect the accuracy of the mathematical models; especially consider the relatively small sample size.



Boston Marathon Qualifying Standards (2021)

AGE GROUP	MEN	WOMEN
18-34	3hrs 00min 00sec	3hrs 30min 00sec
35-39	3hrs 05min 00sec	3hrs 35min 00sec
40-44	3hrs 10min 00sec	3hrs 40min 00sec
45-49	3hrs 20min 00sec	3hrs 50min 00sec
50-54	3hrs 25min 00sec	3hrs 55min 00sec
55-59	3hrs 35min 00sec	4hrs 05min 00sec
60-64	3hrs 50min 00sec	4hrs 20min 00sec
65-69	4hrs 05min 00sec	4hrs 35min 00sec
70-74	4hrs 20min 00sec	4hrs 50min 00sec
75-79	4hrs 35min 00sec	5hrs 05min 00sec
80 and over	4hrs 50min 00sec	5hrs 20min 00sec

Fig. 3. 2021 Boston Marathon Qualifying Standards [25]

B. *Cross-sectional study*

Researchers can investigate the race result data from different age groups in certain events for the approach of a cross-sectional study. The most prominent advantage of cross-sectional study design is leveraging the massive amount of data being generated. Millions of recent race results become available each year containing the performance and biographical data, from which we can build the performance-age model. Moreover, sampling from the general population may provide insights into the physical abilities at all levels, making the model more meaningful when applied to different groups. Apparently, on the other hand of the coin, data from recreational athletes may not reflect the true physical abilities because, unlike professional athletes who choose athletics as their careers, recreational athletes may lack the desire to compete and prepare for competing to their maximal potential. To overcome this problem, we need to select the relatively competitive events that host participants who are willing to train and compete to their true absolute abilities.

C. *Boston Marathon and Its Qualifying Standards*

Established in 1897, Boston Marathon is the oldest marathon and one of the world's most prestigious road racing events. Each year, around 30,000 runners all over the world compete in this race. The most notable feature of the Boston Marathon is enacting strict qualifying standards for their entrance procedure. Fig. 3 shows the Boston Marathon qualifying standards for men and women. The qualifying standards are set for each 5-year age and gender group. The times set in the qualifying standards have been adjusted many times during the last few decades to fit the race's capacity. In the recent years, the registration of the Boston Marathon opens at the beginning of September. Marathon performance on a certified course from the previous year can be used to qualify this race. Due to the limitation of field size, meeting the qualifying standard does not guarantee a race entry; it simply grants registration eligibility. The race organizer, Boston Athletic Association (BAA), ranks the athletes after the registration window is closed. Applicants with faster qualifying times are selected to enter the race first until the field has been filled up. This athlete selection procedure results in different cut-off qualifying times in practice each year. Implementing the entrance procedure with dynamic qualifying times guarantee the most competitive field of the race is assembled. For our purpose, race results from Boston Marathon may serve as excellent data set to build a performance-age relationship model. First of all, given near 25,000 data points produced per year, Boston Marathon race results have a large enough sample size to build reliable models. These data come from athletes of various, but relatively competitive, levels, not just a few world-class elite athletes. Models built upon these data are representative, and therefore may have broader applications. Secondly, the competitive field assembled at Boston Marathon ensures that race performance is evidence of athletes' real physical abilities. Thus, our model may demonstrate individuals' true physical limitations in different age and gender categories.

D. *Data Collection*

The 2017 Boston Marathon result data were used for our analysis, which contained 26,410 instances. The data set included the following attributes: bib number, name, age, gender, city, (US and Canada) state, country, citizenship, each 5-kilometer split time (from 5 to 40 kilometers), pace (time per mile), official time, overall place, gender place, division place. Also sport activities were colleted from Strava API [26]. Strava [19] is an application recording activity data during exercise.

III. Approach and Method

Regression models (median finish time versus age for males and females, respectively) were built to fit the Boston Marathon result data. The models might use different mathematical relationships to fit the data, such as linear, second, third-degree polynomial, and exponential regression. Then 10-fold cross-validations and statistical tests were performed to evaluate the models. The best fitting models were chosen and visualizations of the models were presented.

Clustering was used to reveal the reasonable age group separations. The official result data set contained the split time at each five-kilometer checkpoint from 5 to 40 kilometers. So, the latter split time contained the redundant information from previous ones. The data set was transformed into eight five-kilometer intervals to remove the redundancy. K-means clustering algorithm was applied to the instances containing the following attributes: age, overall place, eight five-kilometer split times. The data were scaled and centered. The clustering results were displayed with two dimensions, age and overall place.

IV. Results

A. *Linear and Polynomial Regression Model*

Fig. 4 and Fig. 5 showed the change of median finish times (male, and female athletes) with ages. From 20 to 70 years old, runners' finish times seemed to go exponentially slower. In the following steps, the 20 – 70 years old male

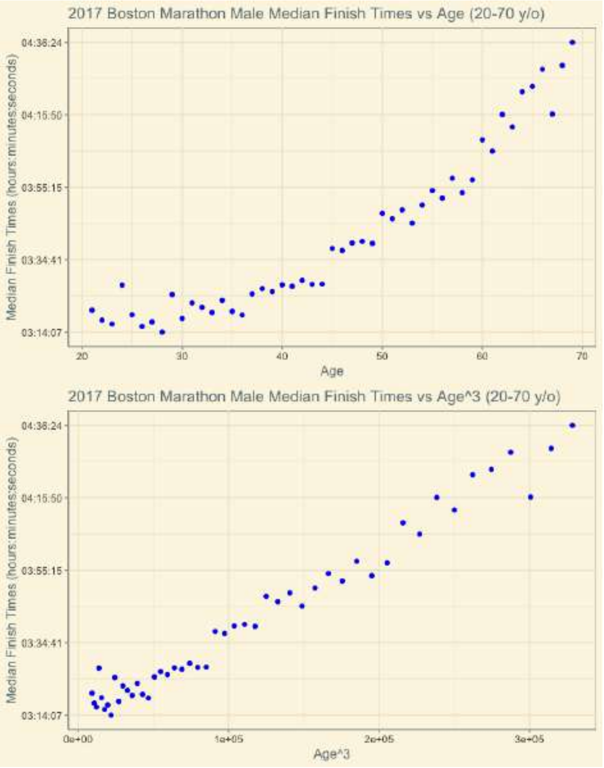


Fig. 4. Male 20 – 70 years old runners' finish times at 2017 Boston Marathon were plotted against ages and the cubic ages. The latter plot suggested a third degree polynomial regression model.

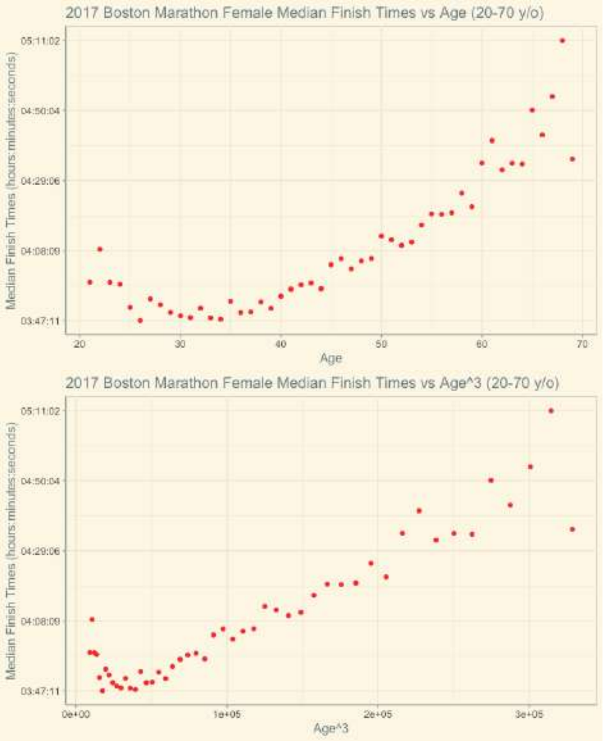


Fig. 5. Female 20 – 70 years old runners' finish times at 2017 Boston Marathon were plotted against ages and the cubic ages. The latter plot suggested a third degree polynomial regression model.

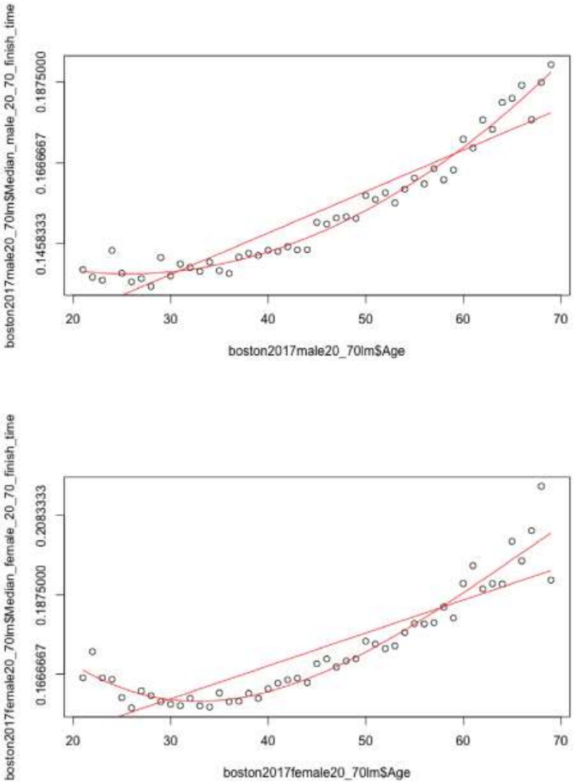


Fig. 6. The comparison of simple liner regression models and a third-degree polynomial regression models. The plots visually demonstrated that the latter ones fitted the data better for both male and female data. (Upper panel: Male; Lower panel: Female)

and female runners' median finish times were plotted against ages and the cubic ages, respectively. Apparently, the median finish times had a linear relationship with the cubic ages, which suggested a third-degree polynomial regression model (Fig. 6). Both simple linear regression models and third-degree polynomial regression models were built. ANOVA test was used to compare the two models. The results showed that a third-degree polynomial regression model fitted the data significantly better than a simple linear regression model. 10-fold cross-validation of the third-degree polynomial regression demonstrates a perfect fit (Fig. 7). Taken together, the analysis of 2017 Boston Marathon result data demonstrated that the athletes became slower as they aged, and there was an apparent third-degree polynomial relationship between runners' finish times and ages.

B. Clustering

The process of aging and the corresponding change of athletic ability may display different rates during different phases. Intuitively, we can sense the apparent physical changes caused by hormone levels, including the onset of puberty and menopause of women. However, the effects of hormone level alteration on athletic performance may take time to occur. Moreover, hormone level change is certainly not the only reason to explain everything in terms of aging. Mathematical

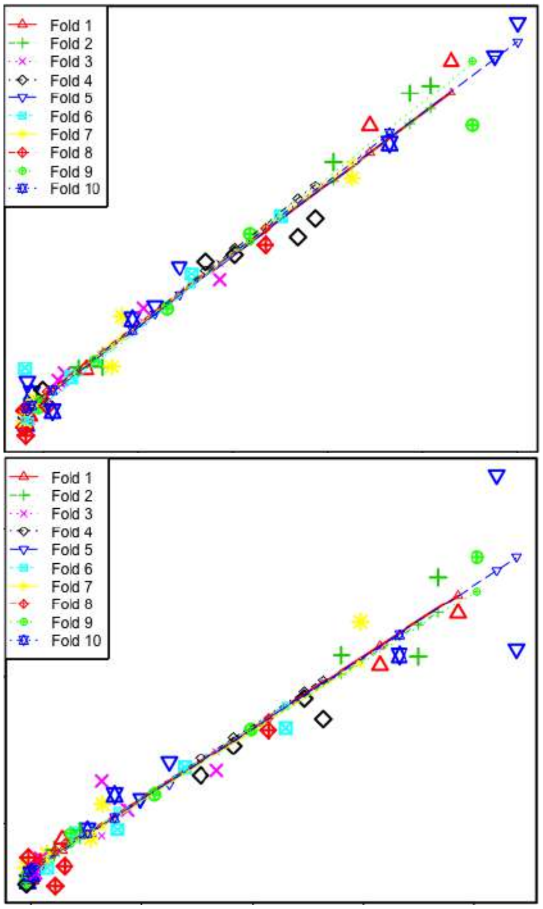


Fig. 7. 10-fold cross validation of the third-degree polynomial regression model demonstrates perfect fit and prediction (Upper panel: Male; Lower panel: Female).

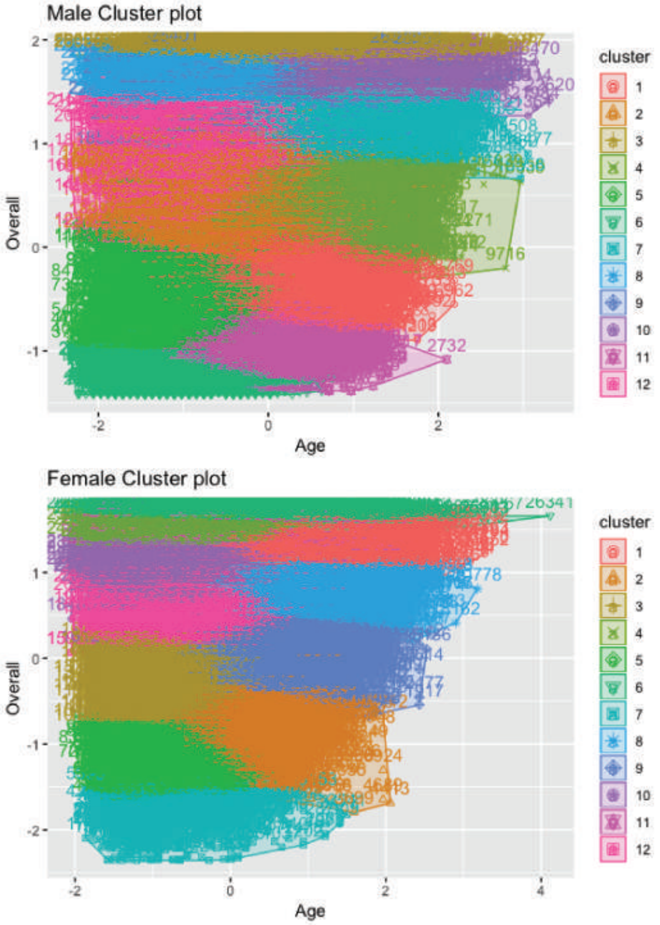


Fig. 8. The Cluster Plot of 2017 Boston Marathon result data. K-means clustering algorithm was applied to the instances containing the following attributes: age, overall place, eight five-kilometer interval times. The plots were projected on two dimensions, age and overall place. Data were scaled and centered: the mean was set at the 0 point in both axes.

models derived from real-world athletic performance data may be viewed as a function of the sum of various aging effects. In a model built by Fair and Kaplan [27], they used 40 and 70 years old as the differentiating points for different mathematical relationships. Another model suggested a single turning point around 30 years old [24]. Currently, the age boundary used to separate open and masters division in all track and field events is 40 years old, regardless of gender. Apparently, there would be a potential problem that males and females may experience different rates of aging, and athletes' performance may peak at different ages for different events. Here a clustering algorithm was used to discover if natural clusters in the performance-age two-dimensional space might be observed. Consider the gender differences of aging effects discovered by previous studies [12], [14]; male and female athletes were investigated separately. This step could provide the foundation for categorizing the populations into reasonable numbers of divisions and the boundaries of separation. Our clustering results suggested that for males and females, slightly different borders were discovered. As shown in Fig. 8, there was a line naturally formed by clusters around the mean age (44 years old for males), whereas for females, the line existed around the mean of 39 years old. Hence, our results suggested a different method of dividing open and masters division i.e., 45 and 40 years old for males and females, respectively.

C. Transfer learning model and Strava data analysis

In addition to the statistical approach, a machine learning model was applied for performance prediction in this section. Performance data from Strava API, a mobile tracking application were used. Strava collects time, speed, distance, calories, heart rate, and other sport activity data. Also, the athletes share comments on Strava. Two methods in studies by Heidari et al [28] and [22] as a natural language processing approach for sentiment classification of Strava comments was utilized in our research. Also, The transfer learning model BERT [29] [30] was used for the sentiment classification of comments. BERT(Bidirectional Encoder representation from transformer) is pre-trained model, which can be fine-tuned for specific task [31] [28]. 50,000 IMDB movie reviews was used for fine-tuning BERT for sentiment classification of Strava comments. A sentiment score assigned to each athlete's comments was considered as a feature for the neural network model's input space. The BERT had a self-attention mechanism, making it a powerful transformer to extract the context from the comments. Also, the multi-head attention of BERT helped to extract different contexts from each comment. A feed-forward neural network model with two hidden layers were chosen in this study. The data for training was labeled as professional or beginner athletes based on performance. The model classified the person as a beginner or professional. The accuracy of the FFNN model was demonstrated to be 89%. One potential problem was that the Strava API imposed a limitation on collecting the history of performance for each user, which could affect the model accuracy.

V. Conclusion and future work

Our analysis of Boston Marathon result data revealed a third-degree polynomial mathematical relationship between performance and age. Our results suggested that the age of 45 and 40, for males and females respectively, may be appropriately used to separate open and masters division in marathon running. Using the transfer learning model and Strava data was a new approach for the classification of professional or beginner athletes. In the future, more data from various events may be used to enrich the model. We also plan to combine statistical and transfer learning models to improve the performance of the algorithm.

References

[1] T. Malkinson, "Current and emerging technologies in endurance athletic training and race monitoring," in *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, pp. 581–586, 2009.

[2] "U.s. road race participation numbers hold steady for 2017." https://www.coloradorunnermag.com/2018/06/30/u-s-road-race-participation-numbers-hold-steady-for-2017/, 2018.

[3] A. Jones, "Age grading running races." http://www.howardgrubb.co.uk/athletics/wmaroad15.html, 24 Jan 2015.

[4] M. Oytun, C. Tinazci, B. Sekeroglu, C. Acikada, and H. U. Yavuz, "Performance prediction and evaluation in female handball players using machine learning models," *IEEE Access*, vol. 8, pp. 116321–116335, 2020.

[5] A. P. Welles, D. P. Looney, W. V. Rumpler, and M. J. Buller, "Estimating human metabolic energy expenditure using a bootstrap particle filter," in *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 103–106, 2017.

[6] H. Tanaka and D. Seals, "Endurance exercise performance in masters athletes: Age-associated changes and underlying physiological mechanisms," *The Journal of physiology*, vol. 586, pp. 55–63, 02 2008.

[7] P. Reaburn and B. Dascombe, "Endurance performance in masters athletes," *European Review of Aging and Physical Activity*, vol. 5, 04 2008.

[8] G. V. Mendonca, P. Pezarat-Correia, J. R. Vaz, L. Silva, and K. S. Heffernan, "Impact of aging on endurance and neuromuscular physical performance: The role of vascular senescence," *Sports medicine (Auckland, N.Z.)*, vol. 47, p. 583—598, April 2017.

[9] S. W. Trappe, D. L. Costill, M. D. Vukovich, J. Jones, and T. Melham, "Aging among elite distance runners: a 22-yr longitudinal study," *Journal of Applied Physiology*, vol. 80, no. 1, pp. 285–290, 1996. PMID: 8847316.

[10] S. Magness, *The science of running: how to find your limit and train to maximize your performance*. Origin Press, 2014.

[11] M. Burtscher, H. Förster, and J. Burtscher, "Superior endurance performance in aging mountain runners," *Gerontology*, vol. 54, 2008.

[12] K. Seiler, W. Spirduso, and J. Martin, "Gender differences in rowing performance and power with aging," *Medicine and science in sports and exercise*, vol. 30, p. 121—127, January 1998.

[13] M. T. Galloway, R. Kadoko, and P. Jokl, "Effect of aging on male and female master athletes' performance in strength versus endurance activities," *American journal of orthopedics (Belle Mead, N.J.)*, vol. 31, p. 93—98, February 2002.

[14] A. J. Donato, K. Tench, D. H. Glueck, D. R. Seals, I. Eskurza, and H. Tanaka, "Declines in physiological functional capacity with age: a longitudinal study in peak swimming performance.," *Journal of applied physiology (Bethesda, Md. : 1985)*, vol. 94, pp. 764–769, February 2003.

[15] A. B. Baker and Y. Q. a. Tang, "Aging performance for masters records in athletics, swimming, rowing, cycling, triathlon, and weightlifting.," *Experimental aging research*, vol. 36, pp. 453–477, September 2010.

[16] N. J. Ronkainen, T. V. Ryba, and M. S. Nesti, "'the engine just started coughing!' — limits of physical performance, aging and career continuity in elite endurance sports," *Journal of Aging Studies*, vol. 27, no. 4, pp. 387–397, 2013.

[17] R. Lepers and T. Cattagni, "Do older athletes reach limits in their performance during marathon running?," *GeroScience*, vol. 34, no. 3, pp. 773–781, 2011.

[18] A. Rezaei, M. Khoshnam, and C. Menon, "Towards user-friendly wearable platforms for monitoring unconstrained indoor and outdoor activities," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[19] "Strava application." https://www.strava.com/, 24 Jan 2019.

[20] M. R. Dey, U. Satapathy, P. Bhanse, B. K. Mohanta, and D. Jena, "Magtrack: Detecting road surface condition using smartphone sensors and machine learning," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pp. 2485–2489, 2019.

[21] B. K. Mohanta, D. Jena, N. Mohapatra, S. Ramasubbareddy, and B. S. Rawal, "Machine learning based accident prediction in secure iot enable transportation system," in *Journal of Intelligent Fuzzy Systems ,2021*, pp. 2485–2489, 2021.

[22] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0542–0547, 2020.

[23] M. Heidari, J. H. J. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[24] G. C. Berthelot, S. Len, P. N. Hellard, M. Tafflet, M. Guillaume, J.-C. Vollmer, B. Gajer, L. Quinquis, A. Marc, and J.-F. Toussaint, "Exponential growth combined with exponential decline explains lifetime performance evolution in individual and human species," *AGE*, vol. 34, no. 4, pp. 1001–1009, 2011.

[25] "Marathon qualification." https://chance.amstat.org/2014/09/boston-marathon/, 24 Jan 2020.

[26] "Strava application api." https://developers.strava.com/, 24 Jan 2019.

[27] R. C. Fair and E. H. Kaplan, "Estimating Aging Effects in Running Events," *The Review of Economics and Statistics*, vol. 100, pp. 704–711, October 2018.

[28] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[29] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[30] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[31] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

Parallel algorithm for modeling temperature fields using the splines method

1st Hakimjon Zaynidinov
Head of the Department of Information Technology, Doctor of Technical Sciences, Professor
Tashkent University of Information Technologies, Tashkent, 100200, Uzbekistan
Tashkent, Uzbekistan
tet2001@rambler.ru

2nd Oybek Mallayev
Department of Information Technology
PhD
Tashkent University of Information Technologies, Tashkent, 100200, Uzbekistan
Tashkent, Uzbekistan
info-oybek@rambler.ru

3rd Muslimjon Kuchkarov
Doctoral student of the department of information technologies
Tashkent University of Information Technologies, Tashkent, 100200, Uzbekistan
Tashkent, Uzbekistan
muslimjon1010@gmail.com

***Abstract*—The paper is devoted to the construction of an algorithm for parallelization of temperature field modeling processes based on bicubic spline and the problems that arise in it. It takes a lot of time and memory on a computer to model existing large-scale signals of temperature fields in three-dimensional graphical form. Solving these problems on the basis of the proposed parallel algorithm, the development of parallel algorithms and software tools to qualitatively determine the coordinates of high-temperature hotspots of temperature fields and allow them to be monitored continuously, and to calculate the digital processing of existing large-scale signals of temperature fields. Based on this, the article presents a parallel algorithm for digital processing of signals of temperature fields on the basis of bicubic spline and ways to parallel its computational processes using Open MP technology.**

Keywords— Vectorization, parallelism, bicubic spline, parallel algorithm, Open MP, interpolation, temperature field, coefficient, approximation, parallel processes

I. Introduction

The methods of spline functions are widely used in solving signal analysis and recovery problems. Spline functions are an evolving area of function approximation and numerical analysis theory. An algorithm for parallelization of temperature field modeling processes using the spline method was built. The values of temperature fields x = 1sm, y = 1sm in the grid are known. Based on these values, it is possible to analyze the heating processes. Beta Soft Board (DT) software from Dynamic Soft Analisis Inc. (USA) is used to analyze heating processes. Approaching the measured temperature values on the basis of bicubic splines allows to determine any point of the temperature field in three-dimensional graphical view. Parallel calculation of large-scale values of temperature fields in a short period of time is one of the most difficult issues. Because presenting large volumetric values in a three-dimensional graphical view causes memory and time issues spent on computational processes on the computer. These problems are solved by creating a parallel algorithm for digital processing of signals based on bicubic splines. The main issue discussed in the article is the development of a parallel algorithm and software tool for digital processing of large-scale signals of temperature fields using a bicubic spline. In the studies studied, the temperature fields were modeled using a spline, but its three-dimensional graph and a parallel algorithm of computational processes were not developed.

II. Results of bicubic spline modeling of temperature fields

Creating a parallel algorithm in memory systems requires first of all acquaintance with the processor characteristics and methods and requirements of computer systems. For example, when the process of loading large volumes of temperature field signals from the computer's RAM to the processor, the processor bus idle time and busy time distribution are also limited by the system, which is one of the reasons for slowing down the work to be done. Especially the processor and its aspects related to hardware technology are both important and complex process. The emergence of the concept of multi-core in processors has also led to the concept of “multi-threaded” being more widely used in software. These two concepts are inextricably linked in modern information technology.

There are two different ways to divide computational processes and the operations performed on them into streams. Each method uses OpenMP's streaming directives.

The method of approximation of the values of the temperature of the printing board through bicubic splines allows to eliminate the specified defect (defect) and to parallel their calculation processes.

The grinder leads to the task of minimizing polynomial splines in S(x) form and integral form functions to expressions for F(x) approximate function derivatives or their differentiable analogues for inputs. For example,

$$J(\alpha,S)=\sum_{i=0}^{n}\alpha_i\left(S_m(x_i)-f(x_i)\right)^2+\alpha\int_a^b\left|\frac{d^r}{dx^r}S_m(x)\right|^2dx, \qquad (1)$$

where a is the control parameter: m is the spline level: i is the ordinal number of the spline values (i = 0, 1 ..., n), ai is the weight multiplier (positive numbers), and r is the order of the product from the spline.

Cubic splines S3 (x) with extreme properties are the most common.

It consists of

$$\Psi(f)=\int_a^b\left|\frac{d^2f(x)}{dx^2}\right|^2dx \qquad (2)$$

Consisting of a functional minimum in appearance, C2[a, b], which interpolates a given function, is achieved by a cubic spline S3(x) on the basis of so-called natural boundary conditions in the form S”(a) = 0, S”(b) = 0 between all functions of the space. For such splines, the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

functional expression (1) can be changed to the following form:

$$J(S)=\sum_{i=0}^{n}\alpha_i(S_3(x_i)-f(x_i))^2+\alpha\int_a^b(S_3^{''}(x))^2dx. \qquad (3)$$

The process of minimizing the function leads to the problem of solving a system of non-separate matrix algebraic linear equations in a diagonal-preferred band grid. Level 3 splines typically correspond to five-diagonal symmetrically positively defined matrices [1, 2].

The problem of grinding a variable f (x) function through splines can be summarized as follows. The values of the function at points i.e. f (x), f (x)… .f (x) are given in the table and the following system of ratios can be constructed:

$$-\beta_i\left(\frac{m_{i+1}-m_i}{h_{i+1}}-\frac{m_i-m_{i-1}}{h_i}\right)=S_3(x_i)-f(x_i), \qquad (4)$$

Here it is

$$\beta_i>0, h=x-x_{n+1}(i=1,2,........n), m_1=S''(x_1),$$
$$m_{n+1}=m_1=0, h_0, h_{0+1}\neq 0$$

It is required to find the function S3(x) that minimizes the functions in the following view.

$$F(S)=\int_a^b(S_3^{''}(x))^2dx+\sum_{i=0}^{n}\frac{1}{\beta_i}(S_3^{''}(x_i)-f(x)_i)^2 \qquad (5)$$

a system of linear algebraic n-1 equations is constructed to determine unknown values of mᵢ.

βᵢ (i=0,1,…,n) is a five-diagonal symmetric matrix with coefficients ai,ci,di, depending on the multipliers.

$$Ax=f \qquad (6)$$

System 6 has a unique solution in relation to the mi vector at hi values that differ from each other in a sufficiently wide range. Once the magnitudes are determined after its solution, f (x) has different nodes. The values of the grinding function in these nodes can be found if 4 reciprocal ratios are used [3, 4].

$$S_3(x_i)=f(x_i)-\beta_i\frac{m_{i+1}-m_i}{h_{i+1}}-\frac{m_i-m_{i-1}}{h_i} \qquad (7)$$

It is known that when grinding signals f(xi)=fi there is usualy an error in determining the values, ie as follows:

$$S(x_i)-f_i|\le\varepsilon_i| \qquad (8)$$

Inequality can be given. In particular, ε_1= ε=const can be given

Formula 5 shows an algorithm for finding such multipliers. several iterations must be performed to determine the exact values of βi. The end of the execution of the iterations is determined by the requirement that the magnitude e does not exceed a given value. As an example, we give the problem of calculating the temperature distribution along the surface of the circuit board caused by the heating of the chips for the uneven position of the nodes of the function T (x, y)

TABLE 1 MEASURED VALUES OF TEMPERATURES AT GIVEN POINTS OF THE PRINTING PLATE.

x, мм	T^0C,x	ST^0,x	y, мм	T^0C,y	ST^0,y
48.0	37.60	37.60	90.9	32.88	32.86
110.0	39.17	39.18	94.8	34.45	34.48
141.3	40.75	40.75	99.5	36.03	36.00
158.5	42.32	42.32	104.3	37.60	37.46
181.4	43.90	43.90	111.0	39.18	39.30
204.4	45.47	45.48	118.2	40.75	40.82
246.5	47.05	47.06	123.4	42.33	42.32
262.7	48.63	48.61	128.7	43.90	42.74
298.6	48.63	48.59	135.9	45.48	45.50
309.11	47.05	47.08	162.9	47.05	47.12
317.2	45.47	45.49	174.1	48.63	48.72
328.7	43.90	43.91	191.8	48.63	48.68
356.4	42.32	42.32	206.6	47.05	46.96

As mentioned above, it is difficult to determine the temperature values at a given board point in Beta Soft DT because all calculations are output as temperature gradients in certain intervals i.e. in practice it is convenient to determine the exact values of temperatures only within temperature zone boundaries. Table 1 shows the measured values of the temperatures at the given points on the horizontal and vertical straight lines.

For the values of ST°C(x) and ST°C(y) grinding bicubic splines calculated on the basis of formulas 4 and 7, it is possible to see the dependence of the temperature on the x and y coordinates based on the location of the values given in.

TABLE 2 THE EFFECT OF THE COEFFICIENT BX ON THE ACCURACY OF THE APPROXIMATION OF THE TEMPERATURE FIELD FUNCTION.

β_x	**0.2**	**0.5**	**0.8**	**1**	**2**	**5**	**10**
T^{0}C,x	47.05	47.05	47.05	47.05	47.05	47.05	48.63
CT^{0}C,x	47.05	47.05	47.05	47.05	47.06	47.07	48.60
ΔT^{0}C,x	0	0	0	0	0.01	0.02	0.03

TABLE 3 THE EFFECT OF THE COEFFICIENT BY ON THE ACCURACY OF THE APPROXIMATION OF THE TEMPERATURE FIELD FUNCTION

B$_y$	**0.1**	**0.5**	**1**	**5**	**10**
T^{0}C,й	48.63	48.63	48.63	48.63	43.90
CT^{0}C,й	48.63	48.64	48.65	48.69	43.74
ΔT^{0}C,й	0.00	0.01	0.02	0.06	0.16

An analysis of the errors that occur when constructing signal grinding splines shows that the characters can be changed depending on the location of the dots on the board in the first place. Second, the values of the interpolation and grinding splines differ significantly in the increase in the values of the βx and βy parameters. It is possible to evaluate the effect of the change of the multipliers βx and βy on the accuracy of the approximation by means of smoothing splines of the temperature field function. The values of both functions are given at the points where the distance between the measured value of the temperature and the spline value is the largest in the modulus [5, 6, 7]. The simplest formula for the approximation of linear combinations to the function of two variables f (x, y) was studied by V.I. Arnold.

According to his testimony

$$F(x,y) = c[f(x) + \Psi(y)] \qquad (9)$$

most of the functions in the view are not dense in the C(D) space. In this case, c is a constant coefficient, D is a two-dimensional field. This means that the approximation in the sum view cannot be performed with infinitesimal error. In practice, when c = 0.5, the maximum error ΔT on the modulus not exceeding 0.02 C for a given temperature area T (x, y) is formed, and the conclusion about the multiplier increase βx and βy is confirmed.

A more universal method of approximation of functions f (x, y) that allows the average square deviation to be minimized was seriously first considered in formula 7 and

$$f(x,y) \cong \varphi_0(x) + \psi_0(y) + \sum_{k=1}^{n} \varphi_k(x)\psi_k(y). \qquad (10)$$

the approximation formula in Figure 10 was used. However, the algorithm for calculating the approximation parameters requires finding the eigenvalues of a particular form and the eigenfunctions of the intergral equations, and this is a rather complex mathematical problem. Approximation of this type of computational field with a bicubic spline is done more efficiently.

III. PARALLEL ALGORITHM FOR MODELING TEMPERATURE FIELDS USING BICUBIC SPLINE.

In the theory of basis splines (B-splines) developed methods for determining the parameters of multidimensional approximation structures that do not require the solution of a system of algebraic equations. For two-dimensional fields, T (x, y) takes the following form:

$$T(x,y) \cong S_{m,m}(x,y) = \sum_{i=-m}^{n_1+m} \sum_{k=-m}^{n_2+m} b_{ik} B_i(x) B_k(y), \qquad (11)$$

In this case, the m-spline level, the number of values on the n1-x axis, the number of values on the n2-u axis, the b$_{ik}$-spline coefficient [8, 9, 10].

Formula 11 is mainly used to create a parallel algorithm for calculating the bicubic spline. The algorithm for calculating it uses a drink cycle. That is, there are 3 cycles. The first loop is used for the columns of the bicubic spline, the second for the rows of the loop, and the third is used to calculate Bi(x) and Bi(y). The parallel flows of the system are used in the programming of parallel computational processes. The organization and management of these parallel flows is done depending on the complexity of the parallel algorithm. That is, in a parallel algorithm, it is necessary to sort the cyclic processes. For example, in a program, it is necessary to create separate streams for drinking cycles. But creating parallel streams in excess of the norm leads to data loss. As a result, the result of the program is incorrect. The same problem is encountered when programming a parallel algorithm for calculating a bicubic spline. To overcome this, the parallel currents must be switched off after the end of the second cycle. They can then be reused again as you move on to the second step of the first cycle.

The stage of implementation of the parallel algorithm of digital processing of temperature field signals is as follows:

1- Identify cyclic processes that can be calculated in parallel in the implementation of the algorithm.

2- Create dynamic arrays in the sequential phase of the algorithm and calculate them.

3- Organization of the procedure for calculating bicubic spline coefficients (bi, j), bases (Bi (x) and Bi (y)).

4- Create and implement a procedure for assembling parallel arrays in computer cores.

5- Create and display an array of results using serial and parallel computational procedures.

The block diagram of the parallel algorithm is shown in Figure 1. The acceleration coefficient of the parallel algorithm is calculated by the following formula (11)

$$S_p(n) = \frac{T(n)}{T_p(n)} \tag{12}$$

Here, T (n) is the time spent running a serial program, and Tp (n) is the time spent running a parallel program. Table 4 shows the acceleration coefficient, the time spent in series and parallel calculations. At 1024, the number of input signals was 0.045 seconds for serial computing, 0.025 seconds for parallel computing, and the acceleration coefficient was 1.8. The acceleration coefficients at different values of the incoming signal using the parallel algorithm of bicubic spline calculation are given in column 4 of the table. As the number of input samples increased, so did the acceleration coefficient of the parallel algorithm. The research results were obtained on an Intel (r) Core (TM) i5-10300H CPU @ 2.50Ghz processor.

FIGURE 1. BLOCK DIAGRAM OF THE BICUBIC SPLINE PARALLEL COMPUTING ALGORITHM.

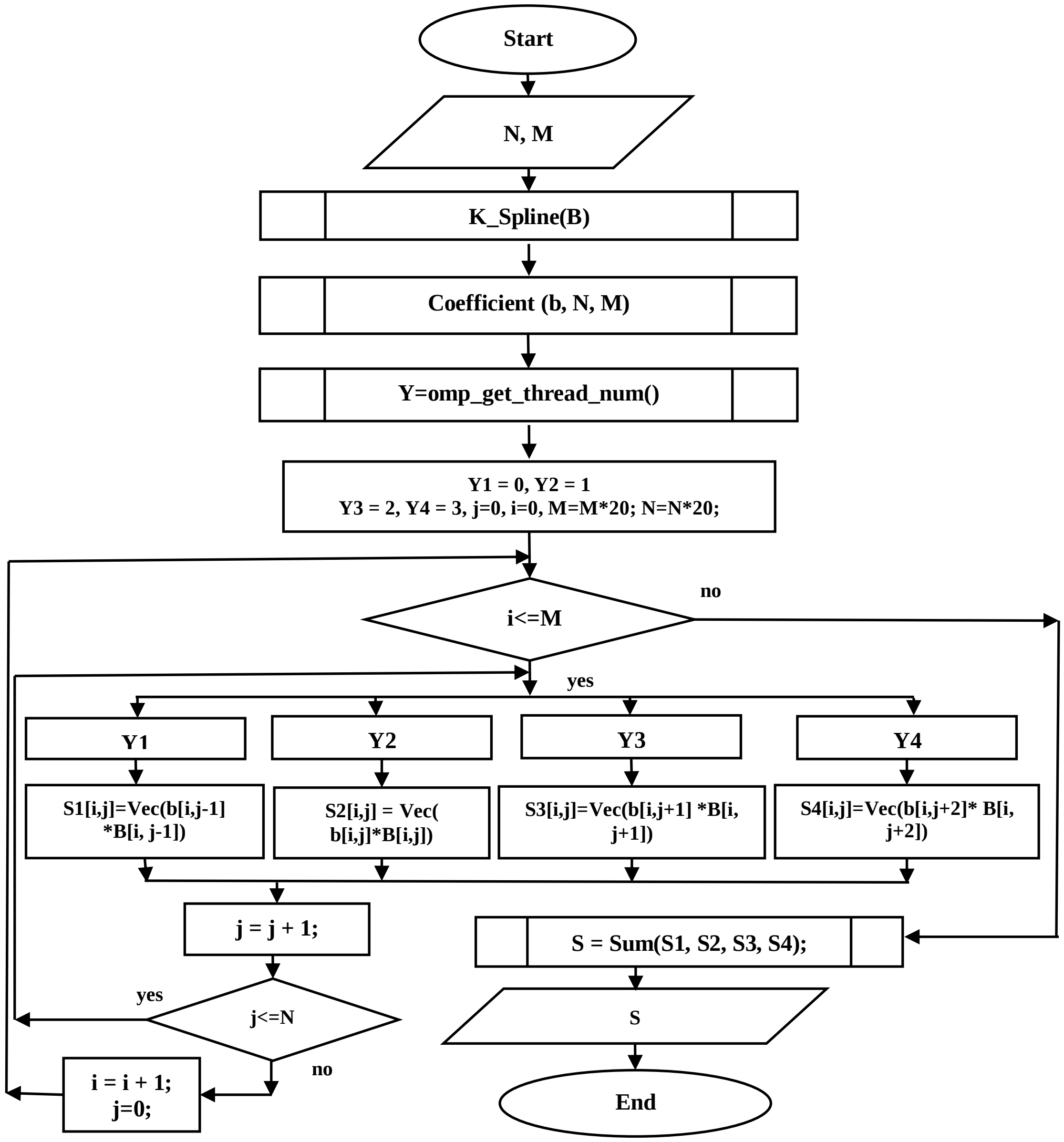

Start
N, M
K_Spline(B)
Coefficient (b, N, M)
Y=omp_get_thread_num()
Y1 = 0, Y2 = 1
Y3 = 2, Y4 = 3, j=0, i=0, M=M*20; N=N*20;
no
i<=M
yes
Y1
Y2
Y3
Y4
S1[i,j]=Vec(b[i,j-1] *B[i, j-1])
S2[i,j] = Vec(b[i,j]*B[i,j])
S3[i,j]=Vec(b[i,j+1] *B[i, j+1])
S4[i,j]=Vec(b[i,j+2]* B[i, j+2])
j = j + 1;
S = Sum(S1, S2, S3, S4);
yes
j<=N
S
no
i = i + 1; j=0;
End

Table 4 The results of a parallel algorithm on a quad-core processor

Number of input signal samples **N,M**	Sequentially (sec.)	Parallel (sec.)	Acceleration coefficient
1024	0,2 x10^{-4}	0, 1 x10^{-4}	2
4096	1,91 x10^{-4}	0,63 x10^{-4}	3,03
8192	3,48 x10^{-4}	1, 09 x10^{-4}	3,19

The results were obtained on an Intel (r) Core (TM) i5-10300H CPU @ 2.50Ghz processor. As the number of input samples increased, so did the parallel portion of digital processing and the acceleration coefficient. At N = 8192, it took 1.09x10-4 seconds to calculate the temperature field signals in parallel with the digital processing processes, and 3.48x10-4 seconds to calculate the series. The acceleration coefficient was 3.19. Since these results were performed on a quad-core processor, the maximum value of the acceleration coefficient was 3.19.

Parallel algorithm programming used the procedures and functions of OpenMP technology described below. The “#pragma omp parallel” directive to create parallel streams of the operating system, the “omp_get_thread_num()” function to select processor cores, the “omp_get_wtime()” procedure to determine the time taken to compute S1, S2, S3, S4 arrays distributed to cores, and each The directive “#pragma omp parallel for” was used to parallel the cyclic processes in the kernel. In addition, the method of vector calculation of the procedure "sum (S1, S2, S3, S4)" in the block diagram of the parallel algorithm shown in Figure 1 after a single step (clock) of parallel flows was developed [1,4]. This resulted in faster execution of multi-time looping processes and a further increase in the acceleration coefficient of the algorithm [11, 12, 13, 14].

A three-dimensional graphical representation of the available large-scale signals of temperature fields is organized as follows.

Figure 2. A three-dimensional graphical view of the initial state of the temperature field signals

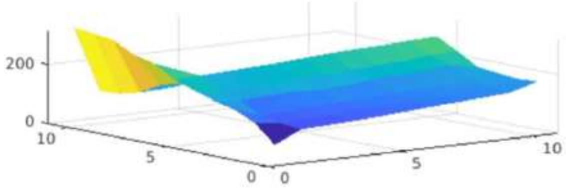
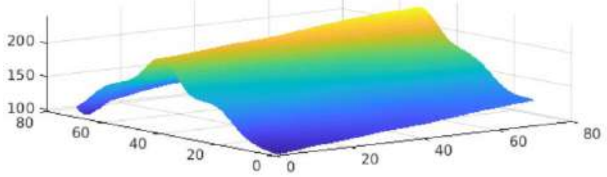


Figure 3. Three-dimensional graphical representation of temperature field signals modeled on the basis of bicubic spline.



In the graph, yellow is the heated part of the temperature fields. Created on the basis of a parallel algorithm, this graph allows tracking of heated areas in real time. The software allows you to view and analyze the exact coordinates of the heated areas and their values.

IV. Conclusion.

In conclusion, the results of two-dimensional thermal field modeling show that not only geophysical fields, but also seismic, acoustic and thermal and other complex multidimensional signals can be modeled using the proposed spline functions. The advantage of splines in computational problems of many variable functions is that they fully satisfy the conditions of parallelization of computational methods. This means that in the programming of parallel algorithms for memory systems, the correct organization of internal loop processes for parallel flows of the operating system and the adaptation of computational processes to the conditions of vectorization, the program eliminates errors and accelerates computational processes. The use of a one-dimensional array in the calculation of two-dimensional bases increases the program's ability to distribute parallel internal loop processes into parallel streams. This halves the calculation time.

References

[1] H. N. Zaynidinov, O. U. Mallaev, and B. B. Anvarjonov, “A parallel algorithm for finding the human face in the image,” in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 862, no. 5, doi: 10.1088/1757-899X/862/5/052004.

[2] H. N. Zaynidinov, O. U. Mallayev, and I. Yusupov, “Cubic basic splines and parallel algorithms,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, 2020, doi: 10.30534/ijatcse/2020/219932020.

[3] Z. Xakimjon and M. Oybek, “Definition of synchronization processes during parallel signal processing in multicore processors,” International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019, doi: 10.1109/ICISCT47635.2019.9012006.

[4] H. Zaynidinov, S. Bahromov, B. Azimov, and M. Kuchkarov, “Lacol Interpolation Bicubic Spline Method in Digital Processing of Geophysical Signals,” *Adv. Sci. Technol. Eng. Syst. J.*, vol. 6, no. 1, 2021, doi: 10.25046/aj060153.

[5] H. Zaynidinov, S. Bakhromov, B. Azimov, and S. Makhmudjanov, “Comparative Analysis Spline Methods in Digital Processing of Signals,” *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 6, pp. 1499–1510, Dec. 2020, doi: 10.25046/aj0506180.

[6] H. Yan, Z. Ma, and Y. G. Zhou, “Acoustic tomography system for online monitoring of temperature fields,” *IET Sci. Meas. Technol.*, vol. 11, no. 5, 2017, doi: 10.1049/iet-smt.2016.0303.

[7] Y. Kim, G. Lee, S. Park, B. Kim, J. O. Park, and J. H. Cho, “Pressure monitoring system in gastro-intestinal tract,” in *Proceedings - IEEE International Conference on Robotics and Automation*, 2005, vol. 2005, doi: 10.1109/ROBOT.2005.1570298.

[8] D. Singh, M. Singh, and Z. Hakimjon, “Evaluation methods of spline,” in *SpringerBriefs in Applied Sciences and Technology*, 2019.

[9] “Spline Functions,” in *The New Palgrave Dictionary of Economics, 2012 Version*,

Basingstoke: Palgrave Macmillan, 2013.

[10] K. Burak, V. Kovtun, R. Levytskyi, and M. Makoviichuk, “Determining density of regular grid for creating DTM using bicubic spline interpolation,” *Geomatics Environ. Eng.*, vol. 13, no. 2, 2019, doi: 10.7494/geom.2019.13.2.5.

[11] D. J. Poirier, “Spline Functions,” in *The New Palgrave Dictionary of Economics*, London: Palgrave Macmillan UK, 2018, pp. 12822–12825.

[12] J. S. Lim, *Two-dimensional signal and image processing*, vol. 710. 1990.

[13] A. Kroizer, Y. C. Eldar, and T. Routtenberg, “Modeling and recovery of graph signals and difference-based signals,” 2019, doi: 10.1109/GlobalSIP45357.2019.8969536.

[14] U. Khamdamov and H. Zaynidinov, “Parallel Algorithms for Bitmap Image Processing Based on Daubechies Wavelets,” 2018, doi: 10.1109/ICCSN.2018.8488270.

The Impact of Initial Swarm Formation for Tracking of a High Capability Malicious UAV

Jason Brown
School of Mechanical and Electrical Engineering
University of Southern Queensland
Springfield, Australia
https://orcid.org/0000-0002-0698-5758

Nawin Raj
School of Sciences
University of Southern Queensland
Springfield, Australia
Nawin.Raj@usq.edu.au

***Abstract*—The use of UAVs or drones for criminal or terrorist enterprises is an increasing problem. Many countermeasures have been proposed to prevent, deter, detect and/or mitigate the dangers posed by such malicious UAVs. One such countermeasure is to track or pursue a malicious UAV back to its point of origin using one or more surveillance UAVs in order to apprehend the UAV and possibly its owner. If the malicious UAV has a higher capability set than the surveillance UAVs, it will be able to outrun any one of them, and therefore the tracking responsibility must be distributed over a swarm of surveillance UAVs that are geographically dispersed across the tracking area of interest. One aspect of particular interest is how the initial formation of the swarm of surveillance UAVs impacts its ability to successfully track a malicious UAV. In this paper, we examine a specific circular initial swarm formation comprising uniformly spaced concentric rings of uniformly spaced UAVs. The total number of surveillance UAVs follows the sequence of centred hexagonal numbers as the number of rings increases. The tracking performance of this circular swarm of surveillance UAVs is compared to a reference swarm of the same size in which the initial locations of the UAVs are randomly chosen. Two tracking strategies are considered: 1) Reactive tracking, in which each surveillance UAV acts independently of the others and only pursues the malicious UAV when it itself detects it, and 2) Reactive tracking with predictive pre-positioning, in which once one surveillance UAV detects the malicious UAV, it communicates the estimated trajectory and speed of the malicious UAV to all swarm members so they can predictively move to a more optimum tracking position before the malicious UAV arrives. The results demonstrate that this particular circular swarm of surveillance UAVs has superior tracking performance relative to the reference randomly positioned swarm of the same size; this is true for both tracking strategies, but particularly when predictive pre-positioning is employed with a relatively small number of surveillance UAVs.**

Keywords — UAV, swarm, formation, communication, tracking

I. Introduction

There have been several high profile episodes of Unmanned Aerial Vehicles (UAVs) being used for nefarious purposes over the past few years. Perhaps the most well known example is the use of a UAV to invade the airspace of London Gatwick airport in 2018, causing massive disruption to civilian airport operations. More generally, such malicious UAVs can be deployed with a variety of motivations, including invasion of privacy (via aerial photography), trade of illegal substances (via delivery of small packages), causing physical damage or injury and causing critical service disruption [1-3].

Many technical techniques have been proposed to deter and/or mitigate the dangers posed by such malicious UAVs. These include physically catching UAVs, using high power laser beams to destroy UAVs, jamming of UAV control signals and GPS spoofing [3]. There are also possible administrative and regulatory countermeasures such as owner registration and licensing, and remote identification of UAVs [1-3].

From an academic perspective, one particularly interesting scenario involves a highly capable malicious UAV that may be able to evade current technical countermeasures. One option with such a UAV is to track or pursue it back to its point of origin using one or more surveillance UAVs in order to apprehend the malicious UAV and possibly its owner. Such a highly capable malicious UAV may be able to outrun individual surveillance UAVs, but this can be overcome by distributing tracking responsibility over a geographically dispersed swarm of such surveillance UAVs [4-6].

This leads to the question of how to design the initial formation of surveillance UAVs across the tracking area of interest so as to maximize tracking performance when a malicious UAV enters the tracking area. We assume that the tracking area of interest is circular with an important object of interest (e.g. an airport) at its centre. We further assume that the swarm of surveillance UAVs are static (i.e. they maintain their position by hovering) until one of the surveillance UAVs detects a malicious UAV in its vicinity, which may, for example, be via computer vision or on-board radar.

In this paper, we examine a specific circular initial swarm formation comprising uniformly spaced concentric rings of uniformly spaced surveillance UAVs. The number of surveillance UAVs in each ring increases linearly with radius so that the distance between an arbitrary surveillance UAV and its immediate neighbours is approximately constant. An example swarm is illustrated in Fig. 1 for $n = 4$ rings (not including the surveillance UAV at the centre) and a total of $N = 61$ surveillance UAVs. The rules for building similar swarms of different sizes are detailed in Section III. As will be seen, the available swarm sizes follow the sequence of centred hexagonal numbers.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

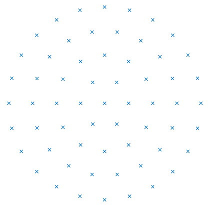


Fig. 1. Example regular circular formation with a total of $N = 61$ surveillance UAVs comprising $n = 4$ uniformly spaced concentric rings of uniformly spaced UAVs

We examine the tracking performance of these regular circular initial swarm formations via simulation in which the high capability malicious UAV moves along a diameter of the designated tracking area i.e. directly over the target object of interest at the centre of the tracking area. As a point of comparison, we also consider the performance of a reference swarm of the same size in which the initial locations of the surveillance UAVs are randomly chosen. Two tracking strategies are considered [5]: 1) Reactive tracking, in which each surveillance UAV acts independently of the others and only pursues the malicious UAV when it itself detects it, and 2) Reactive tracking with predictive pre-positioning, in which once one surveillance UAV detects the malicious UAV, it communicates the estimated trajectory and speed of the malicious UAV to all swarm members so they can predictively move to a more optimum tracking position before the malicious UAV arrives.

The contributions of this paper are as follows:

- A method of constructing a regular circular initial swarm formation of surveillance UAVs comprising uniformly spaced concentric rings of uniformly spaced surveillance UAVs, in order to detect and pursue a highly capable malicious UAV.
- Demonstration by simulation of the improvement in tracking performance afforded by using the regular circular initial swarm formation of surveillance UAVs compared to a randomly formed swarm. The performance improvement is particularly significant when a predictive pre-positioning tracking strategy is employed with a relatively small number of surveillance UAVs.

The paper organisation is as follows. In Section II, we review the literature on the topic of pursuit of malicious UAVs, and in particular discuss an optimal tracking guidance law for the surveillance UAVs to use in their pursuit. Section III discusses the building of regular circular formations of surveillance UAVs of different sizes. The available swarm sizes transpire to be the sequence of centred hexagonal numbers. Section IV discusses the simulation environment and specifies the simulation parameters. We consider a reference swarm in which the initial locations of the UAVs are randomly chosen in order to benchmark the tracking performance of the regular circular formations. The results of the simulation and analysis of the results are provided in Section V. Finally, conclusions and recommendations for future study are discussed in Section VI.

II. Previous Work

This section provides a review of the literature on the topic of pursuit of malicious UAVs. There has been very little research specifically on tracking of high capability malicious UAVs i.e. where the malicious UAV has a higher capability set than individual surveillance UAVs. Rather, most existing research focuses on tracking of a malicious UAV with similar or inferior capability to the individual surveillance UAVs, often by encircling it.

The work discussed in [7] involves the use of a swarm of *defense UAVs* (which is similar to the concept of a swarm of surveillance UAVs) to encircle a malicious UAV so as to restrict its movement. This is only possible because the malicious UAV is considered to have inferior capability compared to the defense UAVs. The paper discusses the various stages of encirclement including clustering, formation, chasing and escorting.

In [8], surveillance UAVs use the radio transmissions of a malicious UAV to pinpoint its location. A technique is described to increase the accuracy of the location fix by moving the surveillance UAVs into various positions relative to the malicious UAV. The paper does not discuss the relative capability of the malicious and surveillance UAVs.

In [9], research into the development of a testbed for a *Counter Unmanned Aerial System (CUAS)* is discussed. The surveillance UAVs in this system utilise computer vision in order to detect and pursue malicious UAVs. This testbed could in principle be used to investigate tracking of a high capability malicious UAV.

The work of [10] focuses on using on-board radar in a swarm of surveillance UAVs to facilitate detection and pursuit of a malicious UAV. This research makes an explicit assumption that the malicious UAV has inferior capabilities relative to the surveillance UAVs.

In [6], a guidance law was proposed to enable a surveillance UAV to set the optimum bearing to maximize tracking time of a highly capable malicious UAV once detected. Since we use this result in the simulation discussed in this paper, we provide an overview of it here. Fig. 2 shows a scenario where a surveillance UAV *S* has just detected a malicious UAV *M* as *M* arrives on the periphery of the detection zone of *S*. We define the following parameters:

- r is the maximum detection distance of *M* from *S*
- u is the maximum pursuit speed of *S*
- v is the speed of the malicious UAV (where $v > u$ since *M* is assumed to have higher capability than *S*)
- φ $(-\pi/2 \le \varphi \le +\pi/2)$ is the angle between the x axis and line joining *S* and *M*
- θ $(-\pi/2 \le \theta \le +\pi/2)$ is the angle relative to the x axis in which *S* moves to follow *M*

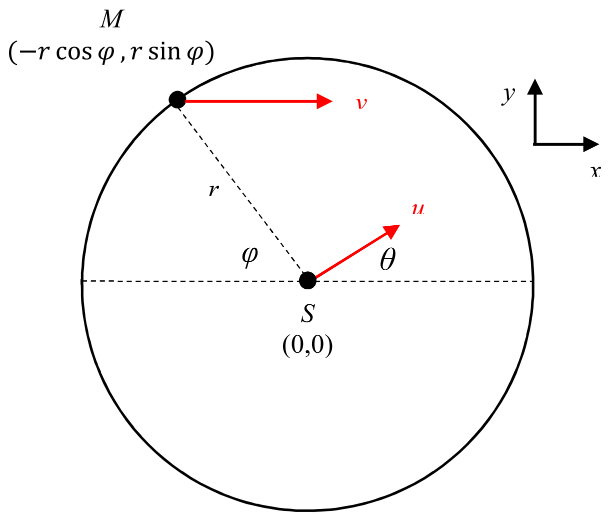


Fig. 2. Definition of parameters for guidance law

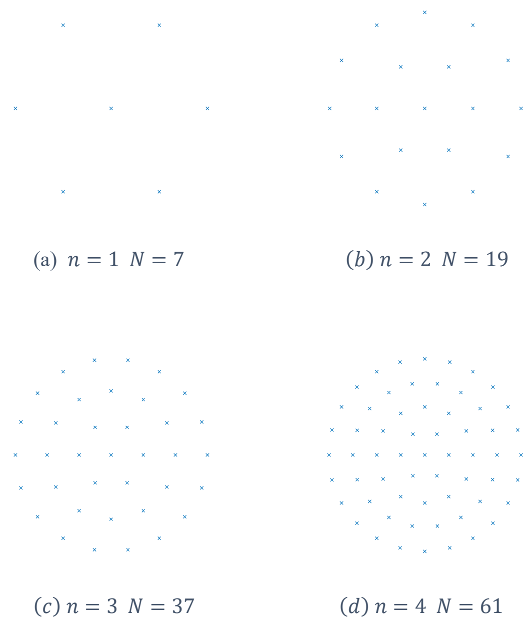
The optimal value of θ which maximizes the tracking time of M by S is given by [6]:

$$\theta_{optimal} = \cos^{-1}\left(\frac{2uv\sin\varphi}{\sqrt{u^4+v^4-2u^2v^2\cos 2\varphi}}\right) - \tan^{-1}\left(\frac{v^2-u^2}{[u^2+v^2]\tan\varphi}\right) \tag{1}$$

III. Regular Circular Swarm Formation

We first make an assumption that, in any regular swarm formation, there is a surveillance UAV at the centre of the circular tracking area i.e. hovering directly over the important object of interest to be guarded, such as an airport. When building a regular circular formation comprising uniformly spaced concentric rings of uniformly spaced surveillance UAVs, a question arises as to how to disperse the surveillance UAVs in the first ring out from the centre of the tracking area. If the distance between adjacent surveillance UAVs in this first ring is to be the same as the distance to the centre of the tracking area, there must be exactly six UAVs in the first ring which form the vertices of a hexagon, as illustrated in Fig. 3(a). In this formation, we have $n = 1$ ring (not including the surveillance UAV at the centre) and a total of $N = 1 + 6 = 7$ UAVs.

As the number of rings increases, the number of surveillance UAVs in each ring should increase linearly with radius so that the distance between an arbitrary surveillance UAV and its immediate neighbours is approximately constant. Therefore, a second ring of surveillance UAVs which has twice the radius of the first ring should have $2 * 6 = 12$ UAVs, as illustrated in Fig. 3(b). This implies for $n = 2$ rings, there will be a total of $N = 1 + 6 + 12 = 19$ UAVs. Similarly, this implies for $n = 3$ rings, there will be a total of $N = 1 + 6 + 12 + 18 = 37$ UAVs, and for $n = 4$ rings, there will be a total of $N = 1 + 6 + 12 + 18 + 24 = 61$ UAVs, as illustrated in Fig. 3(c) and Fig. 3(d) respectively.



(a) $n = 1$ $N = 7$

(b) $n = 2$ $N = 19$

(c) $n = 3$ $N = 37$

(d) $n = 4$ $N = 61$

Fig. 3. Regular circular formation comprising uniformly spaced concentric rings of uniformly spaced UAVs

Using the well known formula for the sum of terms in an arithmetic progression, it is simple to show that the general result is:

$$N = 1 + 3n(n+1) \;\; for\; n \geq 1 \tag{2}$$

This defines the sequence of integers sometimes referred to as centred hexagonal numbers.

IV. Simulation Environment

For the simulation, we compare the tracking performance of the circular swarm of surveillance UAVs discussed in Section III to a swarm of the same size (both in geographic terms and in terms of the number of UAVs in the swarm) in which the UAVs are initially randomly positioned. The meaning of “randomly positioned” in this context is open to interpretation. However, the regular circular swarm is built on the principle that the distance between adjacent UAVs is approximately the same, so for the reference randomly generated swarm, we should ensure that, on average, the same principle holds.

A naïve way to build the reference randomly generated swarm is to choose both the radial and angular coordinates of each swarm UAV from a continuous uniform probability distribution. That is, the radial coordinate ρ is chosen such that $\rho \sim U(0, R)$ where R is the radius of the swarm, and the angular coordinate ψ is chosen such that $\psi \sim (0, 2\pi)$. The problem with this approach is that, on average, the same number of UAVs are placed with radial coordinate ρ_1 as at radial coordinate ρ_2 even when $\rho_1 \neq \rho_2$. This results in a higher density of UAVs positioned close to the centre of the swarm than at the periphery as illustrated in Fig. 4(a).

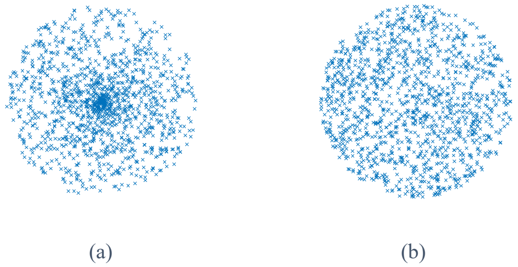


Fig. 4. Spatial distributions of reference randomly generated swarms comprising $N = 1387$ UAVs (a) radial coordinate uniformly distributed, and (b) square of radial coordinate uniformly distributed

Since circumference is a linear function of radius, the radial coordinate ρ should be chosen such that its probability density function $f(\rho)$ increases linearly as illustrated in Fig. 5(a), and therefore its cumulative distribution function $F(\rho)$ is as illustrated in Fig. 5(b). This ensures that, on average, the number of UAVs placed with radial coordinate ρ_1 is greater than the number placed with radial coordinate ρ_2 if $\rho_1 > \rho_2$ and furthermore, the principle that the distance between adjacent UAVs is approximately the same on average is satisfied.

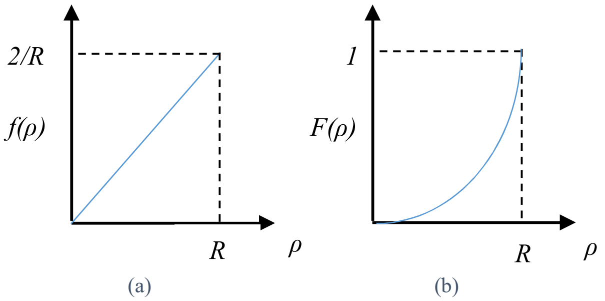


Fig. 5. Required characteristics of the radial coordinate ρ (a) probability density function $f(\rho)$ and, (b) cumulative distribution function $F(\rho)$

The required form of the cumulative distribution function $F(\rho)$ in Fig. 5(b) is as follows:

$$F(\rho) = \frac{\rho^2}{R^2} \qquad 0 \leq \rho \leq R \tag{3}$$

Now if $X = \rho^2/R^2$, X has a continuous uniform probability distribution i.e. $X \sim U(0,1)$ since its cumulative distribution function is linear. Given $\rho = R\sqrt{X}$, this implies that the radial coordinate ρ for each UAV should be chosen so as to be the product of R and the square root of a random variable with a continuous uniform probability distribution in the range (0,1). When this is implemented, the spatial distribution of UAVs is as illustrated in Fig. 4(b). Clearly the density of UAVs appears consistent across the region of interest.

In the simulation, the malicious UAV moves through a diameter of a circular tracking area. It is logical that the malicious UAV should pass through the centre of the tracking area, since this is likely to be the location of the target object of interest (e.g. an airport) that it is attempting to disrupt. The regular circular swarm of surveillance UAVs always has one UAV at its centre, therefore the malicious UAV will always be detected by at least this centred surveillance UAV. This is a key difference with respect to the reference randomly generated surveillance UAV swarm, for which it is possible that the malicious UAV will not be detected by any of the surveillance UAVs as it moves along a diameter.

For the regular circular swarm of surveillance UAVs, different diameters will pass through the detection zones of a different number of UAVs in the swarm. Given the malicious UAV follows a diameter of the tracking area, the angular orientation of the swarm must be randomised relative to the path of the malicious UAV in order to ensure a fair comparison with the reference randomly generated surveillance UAV swarm. This is achieved by rotating the circular swarm through an angle randomly drawn from a continuous probability distribution in the range $(0, 2\pi)$ prior to executing each iteration of the simulation, and then averaging the simulations results over all iterations. Different angular orientations of the same circular swarm are illustrated in Fig. 6.

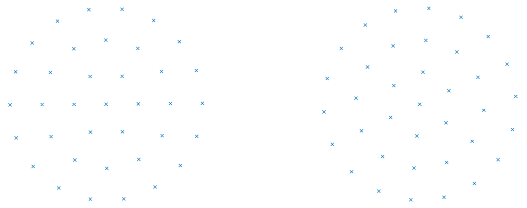


Fig. 6. Different angular orientations of the same regular circular swarm comprising $N = 37$ UAVs

The parameters of the MATLAB simulation are illustrated in Table I. As can be seen, two different tracking strategies are considered:

- reactive tracking only, in which a surveillance UAV only moves once it has itself detected the malicious UAV, and
- reactive tracking with predictive pre-positioning, in which the first surveillance UAV to detect the malicious UAV communicates the estimated speed and bearing of the malicious UAV to all other members of the swarm, allowing the other surveillance UAVs to predictively move into the optimum position to track the malicious UAV if and when it enters their vicinity.

The objective of the tracking is to maximize the time that the malicious UAV is within the detection range of at least one surveillance UAV from the swarm. Therefore, the simulation metric calculated is the proportion of time that the malicious UAV is being actively tracked by one or more surveillance UAVs while in the tracking area.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Simulation step	1m for the malicious UAV
Number of simulation iterations for each set of parameters	500
Tracking area shape	Circular
Tracking area size	15km radius
Number of UAVs in surveillance swarm (*N*)	Fixed for any one simulation run: the following subset of centred hexagonal numbers are considered $1 + 3n(n+1)\ for\ 1 \leq n \leq 21$ (7, 19, 37, 61, …, 1261, 1387)
Initial formation of surveillance UAVs in tracking area	Fixed for any one simulation run as either: Regular circular: see Section III Reference random: see Section IV
Tracking strategy	Fixed for any one simulation run as either: Reactive tracking only Reactive tracking and predictive pre-positioning
Surveillance UAV maximum speed (*u*)	Fixed for any one simulation run at 20m/s, 25m/s or 29m/s
Surveillance UAV direction on detecting malicious UAV (θ)	$\theta_{optimal}$ according to Eq. (1)
Surveillance UAV direction for predictive pre-positioning	Perpendicular to estimated malicious UAV path
Malicious UAV detection range (*r*)	100m
Malicious UAV maximum speed (*v*)	30m/s
Malicious UAV path	Diameter of tracking area
Surveillance UAV communication range (for predictive pre-positioning)	Not limited
Surveillance UAV communication method (for predictive pre-positioning)	Flood/Broadcast

V. RESULTS AND ANALYSIS

Fig. 7, Fig. 8 and Fig. 9 show the metric of interest, the proportion of time that the malicious UAV is directly pursued and tracked by at least one surveillance UAV while in the tracking area, versus the swarm size in terms of the number of surveillance UAVs. The figures represent maximum surveillance UAV speeds of *u*=20m/s, *u*=25m/s and *u*=29m/s respectively. Compared to the fixed malicious UAV speed of *v*=30m/s, they represent a large, medium and small capability mismatch between the surveillance and malicious UAVs respectively. All three figures illustrate plots for each combination of initial swarm formation (regular circular versus reference random) and tracking strategy (reactive tracking only versus reactive tracking with predictive pre-positioning).

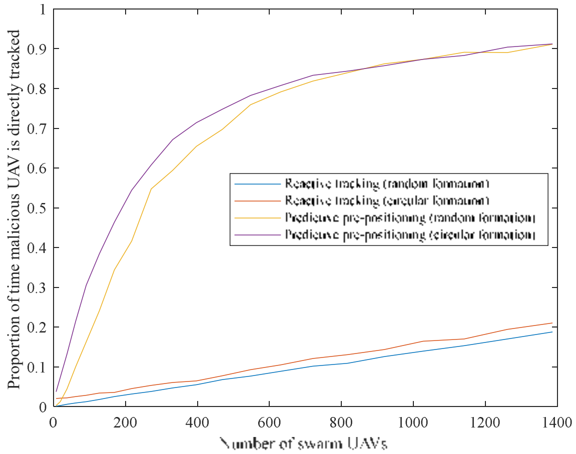


Fig. 7. Proportion of time malicious UAV is actively tracked against number of swarm UAVs for *u*=20m/s, *v*=30m/s and different swarm formations/tracking strategies

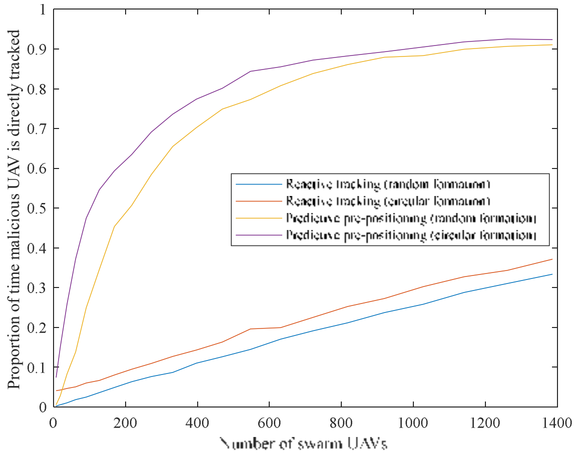


Fig. 8. Proportion of time malicious UAV is actively tracked against number of swarm UAVs for *u*=25m/s, *v*=30m/s and different swarm formations/tracking strategies

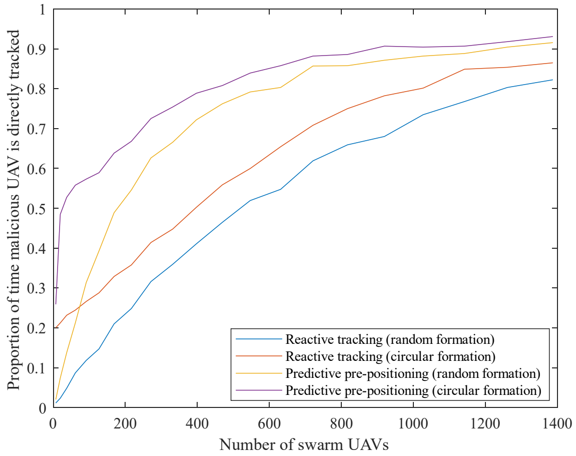


Fig. 9. Proportion of time malicious UAV is actively tracked against number of swarm UAVs for *u*=29m/s, *v*=30m/s and different swarm formations/tracking strategies

It is clear from all three figures that a regular circular swarm formation outperforms a reference random swarm formation for both tracking strategies i.e. the proportion of time that the malicious UAV is directly pursued and tracked by at least one surveillance UAV is higher for a regular circular swarm formation. The fundamental reason for this is that surveillance UAVs may be positioned in close proximity to each other by random chance in the reference random swarm formation, leaving relatively large voids where there are no surveillance UAVs present; hence there is a greater probability that the malicious UAV can fly between surveillance UAVs without being detected.

The difference in performance between the regular circular and reference random swarm formations is particularly notable when examining the plots for reactive tracking with predictive pre-positioning with a relatively small swarm size. For example, examining Fig. 9, for a small swarm size of $N = 37$ UAVs, the proportion of time the malicious UAV is actively tracked when the tracking strategy is predictive pre-positioning is 0.53 for a regular circular swarm formation versus 0.14 for the reference random formation. This large difference is in some respects due to the fact that, with the regular circular initial formation, the malicious UAV will always pass through the detection zone of at least one surveillance UAV (i.e. the surveillance UAV at the centre of the tracking area), whereas this is not guaranteed with the reference random initial formation. This difference is clearly most significant when the surveillance UAV swarm is relatively small – for larger swarms, there is a high probability that the malicious UAV will pass through the detection zone of at least one surveillance UAV even for the reference random initial formation.

As discussed in [5], reactive tracking with predictive pre-positioning outperforms reactive tracking only, particularly when the malicious UAV is significantly more capable in terms of maximum speed than the surveillance UAVs. This is the case even when the reference random initial formation of surveillance UAVs is employed.

VI. Conclusions and Future Work

This paper has demonstrated that a particular regular circular formation of surveillance UAVs has superior tracking performance (when tracking a highly capable malicious UAV) relative to a reference randomly positioned swarm of the same size. This is true for both tracking strategies (reactive tracking only and reactive tracking with predictive pre-positioning), but particularly when predictive pre-positioning is employed with a relatively small number of surveillance UAVs.

The next challenge is to devise regular initial formations of surveillance UAVs that outperform the regular circular formation examined in this paper. One proposal is to offset the angular orientation of adjacent rings of surveillance UAVs relative to each other in order to reduce the size of voids in the tracking area.

References

[1] J. O'Malley, "The no drone zone," in Engineering & Technology, vol. 14, no. 2, pp. 34-38, March 2019, doi: 10.1049/et.2019.0201.

[2] D. Schneider, "Regulators seek ways to down rogue drones: Growing antidrone industry offers radar, remote ID, and other tools - [News]," in IEEE Spectrum, vol. 56, no. 4, pp. 10-11, April 2019, doi: 10.1109/MSPEC.2019.8678424.

[3] J. P. Yaacoub, H. Noura, O. Salman and A. Chehab, "Security Analysis of Drones Systems: Attacks, Limitations, and Recommendations." Internet of Things, Volume 11, 2020, 100218, doi: 10.1016/j.iot.2020.100218

[4] C. Arnold and J. Brown, "Performance Evaluation for Tracking a Malicious UAV using an Autonomous UAV Swarm," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, 2020, pp. 0707-0712, doi: 10.1109/UEMCON51285.2020.9298062.

[5] J. Brown and N. Raj, "Predictive Tracking of a High Capability Malicious UAV," 2021 11th Annual Computing and Communication Workshop and Conference (CCWC), USA, 2021

[6] J. Brown and N. Raj, "Guidance Law for a Surveillance UAV Swarm Tracking a High Capability Malicious UAV", 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), Bandung, Indonesia, 2021.

[7] M. R. Brust, G. Danoy, P. Bouvry, D. Gashi, H. Pathak and M. P. Gonçalves, "Defending Against Intrusion of Malicious UAVs with Networked UAV Defense Swarms," 2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops), Singapore, 2017, pp. 103-111, doi: 10.1109/LCN.Workshops.2017.71.

[8] F. Koohifar, I. Guvenc and M. L. Sichitiu, "Autonomous Tracking of Intermittent RF Source Using a UAV Swarm," in IEEE Access, vol. 6, pp. 15884-15897, 2018, doi: 10.1109/ACCESS.2018.2810599.

[9] M. Pozniak and P. Ranganathan, "Counter UAS Solutions Through UAV Swarm Environments," 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 2019, pp. 351-356, doi: 10.1109/EIT.2019.8834140.

[10] A. Guerra, D. Dardari and P. M. Djuric, "Dynamic Radar Networks of UAVs: A Tutorial Overview and Tracking Performance Comparison With Terrestrial Radar Networks," in IEEE Vehicular Technology Magazine, vol. 15, no. 2, pp. 113-120, June 2020, doi: 10.1109/MVT.2020.2979698.

Detection and Prevention of Blackhole Attack in AODV of MANET

Asif Uddin Khan[1], Rajesh Puree[2], Bhabendu Kumar Mohanta[3], Sangay Chedup [4]
[1]Department of CSE, Sillicon Institute of Technology, Bhubaneswar-751024, India
[2]Department of Computer Science, Utkal University-751004, Odisha, India
[3]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram-500502, Andhra Pradesh, India
[4]Department of ECE, Jigme Namgyel Engineering College, Bhutan.
Email: asif.khan@silicon.ac.in[1], rajesh.puri132@gmail.com[2], bhabendukumar@kluniversity.in[3],
sangaychedup@jnec.edu.bt[4]

***Abstract*—One of the most dynamic network is the Mobile Adhoc (MANET) network. It is a list of numerous mobile nodes. Dynamic topology and lack of centralization are the basic characteristics of MANET. MANETs are prone to many attacks due to these characteristics. One of the attacks carried out on the network layer is the black-hole attack. In a black-hole attack, by sending false routing information, malicious nodes interrupt data transmission. There are two kinds of attacks involving a black-hole, single and co-operative. There is one malicious node in a single black-hole attack that can act as the node with the highest sequence number. The node source would follow the direction of the malicious node by taking the right direction. There is more than one malicious node in the collaborative black-hole attack. One node receives a packet and sends it to another malicious node in this attack. It is very difficult to detect and avoid black-hole attacks.Many researchers have invented black-hole attack detection and prevention systems. In this paper, We find a problem in the existing solution, in which validity bit is used.This paper also provides a comparative study of many scholars. The source node is used to detect and prevent black hole attacks by using a binary partition clustering based algorithm. We compared the performance of the proposed solution with existing solution and shown that our solution outperforms the existing one.**

***Index Terms*—MANET, AODV, Black hole attack, routing protocols, clustering.**

I. Introduction

MANET is a short term for Mobile Ad-hoc Network. It is also called wireless adhoc network, an unbroken network of mobile devices connected without using cables and without any infrastructure setup. Devices can travel in any direction in a MANET architecture independently and thus often change their ties with other devices. Wireless communications, from satellite transmission to home wireless personal area networks, have increased exponentially in recent years [1].

The wireless media communicates between a fixed node and a mobile node within its range. However, a permanent fixed infrastructure is required. The MANET system is a different model, however the broadcasting range of each node is restricted to the closeness of each other, and out-of-scope nodes are routed through intermediate nodes. However, the MANET system is designed to set up a system where necessary [2]. Each mobile node acts in such a network not only as a host, but also as the router, transmitting packets to other mobile nodes in the network that can not be directly wireless [3].

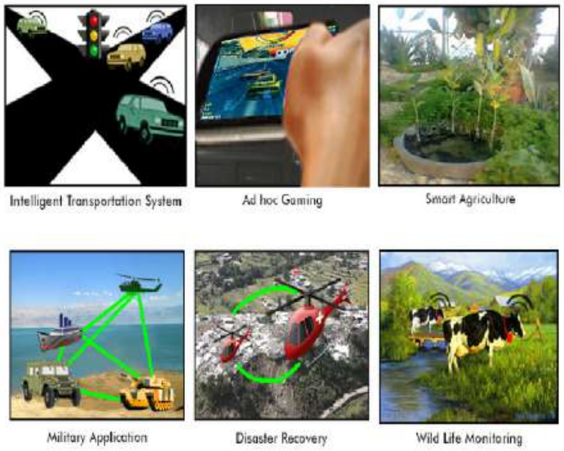


Fig. 1. Application of MANET in different areas.

One of MANET's recent applications is a sensor network consisting of several thousand small, low-powered sensing nodes [4]. Example of MANET implementations are shown in Figure 1. Security in these areas clearly is a critical problem. Secure communication is essential in

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

any wireless network [5] and [6] to make the system more trustworthy to the end users.

In MANET several attacks can be performed by the malicious node that prevents the operation of MANET. Blackhole attack and Grayhole attack are the most important attacks that needs to be investigated.In this paper we investigated the blackhole attack and proposed a solution to solve the issue.The organization of the paper is as follows.

In section-II literature survey is done. Section-III presents the motivation. In section-IV, we present the solution approach.Section-V discusses the performance analysis. Section-VI presents the simulation results. Finally in section-VII, we conclude and discuss the future work.

II. Literature Survey

In paper [7] Ashish Kimar Jain et al. proposed a method to detect black hole attack through first route reply mechanism in AODV protocol of MANET.

In paper [8] J. V. Vadavi et al. try to detect the black hole attack and avoiding them in participating in the network by enhancing the AODV protocol by making it delay aware.

Sushama Singh et al. try to enhance the performance of the AODV protocol by introducing a trusted AODV routing algorithm (tangent hyperbolic function is used to calculate the trust value) to detect the collaborative black hole attack. [9]

In paper [10] Heerendra Mahore et al. focused on agent based AODV protocol which means some sender nodes are assigned a task of the agent to check the RREP coming from the destination nodes to check and avoid the black hole attack.

In paper [11] Vidya Kumari Saurabh et al. proposed a clustering based AODV routing protocol in which each single node of the cluster in ping once to the cluster head to detect the difference between the data packets received and sent by the exact node to detect the black hole attack.

Sathish M. et al. in [12] proposed a method to detect the single as well as collaborative black hole attack by sending fake RREQ and Destination Sequence Number and when receives RREQs from nodes with higher DSN it will collect them in a list and inform all the other nodes that these are the malicious nodes.

III. Motivation

The method for the MANET black hole detection, proposed in [1] says that the validity of the message RREP is attached and stored on each node of the active path in an itinerary table. Whenever a route request is received for a node, the route response message is created by setting a value for a validity bit in RREP, whether that route is the target or if it has a legit route. This RREP will then return to the next hop it got RREQ from. The proposed route reply message varies from the basic AODV route reply message in terms of validity. The RREP message incorporates the validity mechanism. AODV RREP will have an additional header in a validity bit.

This new field is used to validate the route validity. Whenever a Route reply is received from a node, it is processed only if the RREP 's validity bit is specified. Only if the validity bit is set will an entry be made for this path. As an assailant, this mechanism is not known; it reacts without looking into its route table. This implies that the validity bit in the RREP sent by the attacker node would have a null value. A node with a validity value not identified will simply drop the RREP without entering the routing table [13].

The above technique is a great way to detect blackhole attacks [14], and a negligible overhead would be applied to the overhead network when a single bit is used in this strategy to detect attacks. However, if the attacker looks at the attackers' routing table and sets the validity value, and sends RREP to the source, then the source node does not detect a black-hole attack and sends data packets to that route assuming that the destination is reached by this route. The attacker would then drop data to the destination node without forwarding it.

The abbreviations used in this paper is described in TABLE IV.

A. *Flowchart of Existing Solution*

The above Fig.2 is the flowchart of existing solution from which we drew the motivation. The proposed work is the method to overcome the limitations of the above strategy.

IV. Solution Approach

In our proposed approach SOURCE node is used for the detection of black hole attack. After generating a route request, broadcast it and wait for RREPs. Source receive RREPs from k1+k2 no. of nodes. After receiving RREPs from k1+k2 no. of nodes store them in a list [15].

Make two groups of RREPs and store them in the list using a binary partition clustering algorithm such as k-means [16] based on destination sequence number. Then the average destination sequence number(ADS) is calculated for both the cluster that is (ADS1, ADS2). If ADS1 is greater than ADS2 and the difference is greater

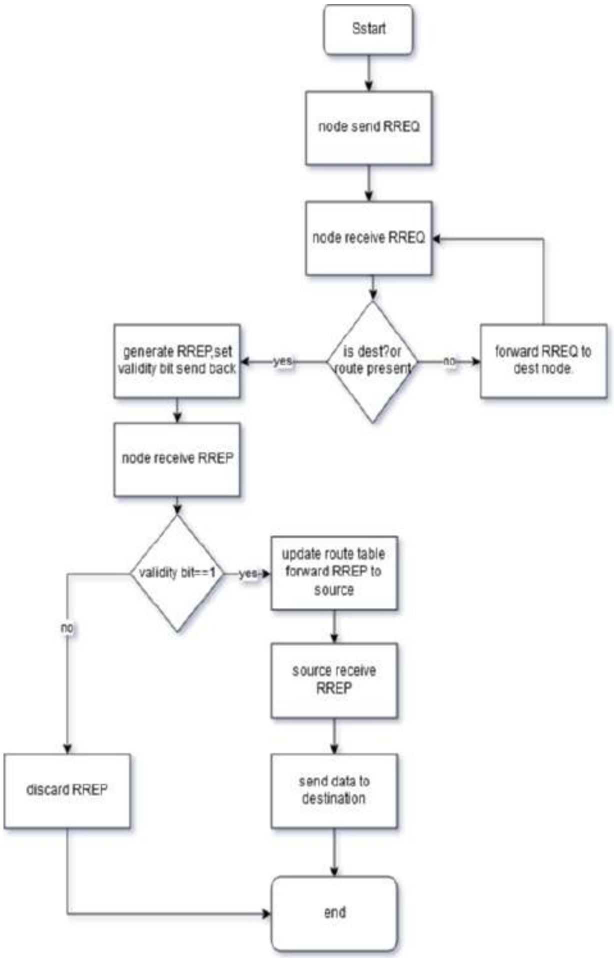


Fig. 2. Flowchart of existing solution

than a threshold value then cluster-1 may contain a set of black-hole nodes otherwise source node selects the best node from cluster-1, check validity bit, if set, send data [17].

If ADS2 is greater than ADS1 and the difference is greater than a threshold value then cluster-2 may contain a set of black-hole nodes otherwise source node select the best node from cluster-2 and check validity bit if set, send data through that node.

A. Proposed Method

The source generates RREQ for the establishment of a path to send data. The explanation of our solution approach is described as follows.

The source node generates RREQ and broadcast it through the network and waiting for the RREPs. Source node after receiving the RREPs from k1+k2 number of nodes stores it in a list.

The source node makes two groups using a binary partition clustering algorithm such as k-means [16] based on destination sequence number. Here cluster-1 and cluster-2 are having n11, n12... , n1k1 nodes and n21, n22, n23..., n2k2 nodes respectively we can see it in TABLE II and TABLE III.

In TABLE II it can be seen that cluster-1 contains the number of nodes n11 to n1k1 having destination sequence number ds11 to ds1k1. Similarly in TABLE III cluster-2 contains the number of nodes n21 to n2k2 having destination sequence number ds22 to ds2k2.

After that average destination sequence number is calculated for cluster-1 (ads1) and cluster-2 (ads2) by the equations given below.

$$ads_1 = \frac{\sum_{i=1}^{k_1} ds_1 k_i}{k_1} \tag{1}$$

$$ads_2 = \frac{\sum_{i=1}^{k_2} ds_2 k_j}{k_2} \tag{2}$$

After calculating the average destination sequence of cluster-1 and cluster-2, we check for cluster-1. If the average destination sequence of cluster-1 (ads1) is greater than the average destination sequence of cluster-2 (ads2) and their difference i.e. (ads1-ads2) is greater than a threshold value then cluster-1 is suspicious and may contain a set of black-hole nodes otherwise source node selects the best node from cluster-1, check the validity bit, if set, then send data through the node to the destination.

Similarly, the average destination sequence of cluster-1 and cluster-2 is calculated and check for cluster-2. If the average destination sequence of cluster-2 (ads2) is greater than the average destination sequence of cluster-1 (ads1) and their difference i.e. (ads2-ads1) is greater than a threshold value then cluster-2 is suspicious and may contain a set of black-hole nodes otherwise source node selects the best node from cluster-2, then check the validity bit is set or not, if set, then send data to destination through the node.

B. Proposed Algorithm

1. Start.
2. Broadcast RREQ (source).
3. Received RREP from k1+k2 nodes.
4. Store it in list.
5. Make 2 clusters using partition clustering algorithm(based on destination sequence number).

TABLE I
CLUSTER-1 CLUSTER-2 TABLE WITH DESTINATION SEQUENCE NUMBER

Cluster 1	Cluster 2
n_{11}	n_{21}
n_{12}	n_{22}
n_{13}	n_{23}
n_{14}	n_{34}

TABLE II
CLUSTER-1 TABLE WITH DESTINATION SEQUENCE NUMBER

number of nodes	destination sequence number
n_{11}	$ds_1 1$
n_{12}	$ds_1 2$
n_{13}	$ds_1 3$
n_{14}	$ds_1 4$
$n_1 k_1$	$ds_1 k_1$

TABLE III
CLUSTER-2 TABLE WITH DESTINATION SEQUENCE NUMBER

number of nodes	destination sequence number
n_{21}	$ds_2 1$
n_{22}	$ds_2 2$
n_{23}	$ds_2 3$
n_{24}	$ds_2 4$
$n_2 k_2$	$ds_2 k_2$

6. Calculate ads1 and ads2 by putting equations-1 and equations-2
7. If ads1 > ads2 and ((ads1-ads2)>ds-threshold) Cluster-1 is suspicious
8. Else source select best node, check validity bit and send data.
9. If ads2>ads1 and ((ads2-ads1)>ds-threshold) Cluster-2 is suspicious
10. Else goto 8
11. End.

where
ads1= average destination sequence number of cluster-1
ads2= average destination sequence number of cluster-2
ds-threshold= threshold difference of average destination sequence number.

TABLE IV
ACRONYMS AND THERE MEANING USED IN THIS PAPER

Acronyms	meaning
AODV	Ad hoc On Demand Distance Vector
MANET	Mobile Ad hoc Network
RREP	Route Reply
RREQ	Route Request
ADS	Average Destination Sequence Number
ADS1	Average Destination Sequence Number of cluster-1
ADS2	Average Destination Sequence Number of cluster-2

In step 1 start the algorithm, source broadcast route request, and in step 3 it received route reply from k1+k2 nodes then store it in a list in step 4.

In step 5 make two clusters using partition clustering algorithm(based on destination sequence number). then calculate ads1 and ad2 in step 6. In step 7 it checks If ads1 > ads2 and ((ads1-ads2)>ds-threshold) then Cluster-1 is suspicious.

In step-8 else source select best node, check validity bit and send data. In step 9 If ads2>ads1 and ((ads2-ads1)>ds-threshold) Cluster-2 is suspicious. In step 10 it goes to step 9 and in 12 it ends.

C. FLOWCHART

The below Fig. 3 is a flowchart that shows how a black-hole attack occurs with respect to time. It is further described with the timing diagram in Fig. 4.

D. TIMING DIAGRAM

The source node wants to send data to the destination node. In T1 time the source node generates RREQ and sends the RREQ in the network. Both the attacker and the destination node received the RREQ.

After receiving RREQ both attacker and destination send RREP to the source node or from where they received RREQ. In T2 time the source received RREP from an attacker and in T3 time receives RREP from the destination node. The RREP from the attacker having destination sequence number is higher than the RREP received from the destination node, source node thinks that this is the fresh enough route to the destination. So the source discards the RREP from the destination.

After discarding the RREP from the destination, the source node sends data to the attacker in T4 time then it drops all the data packet coming from the source without forwarding it to the destination.

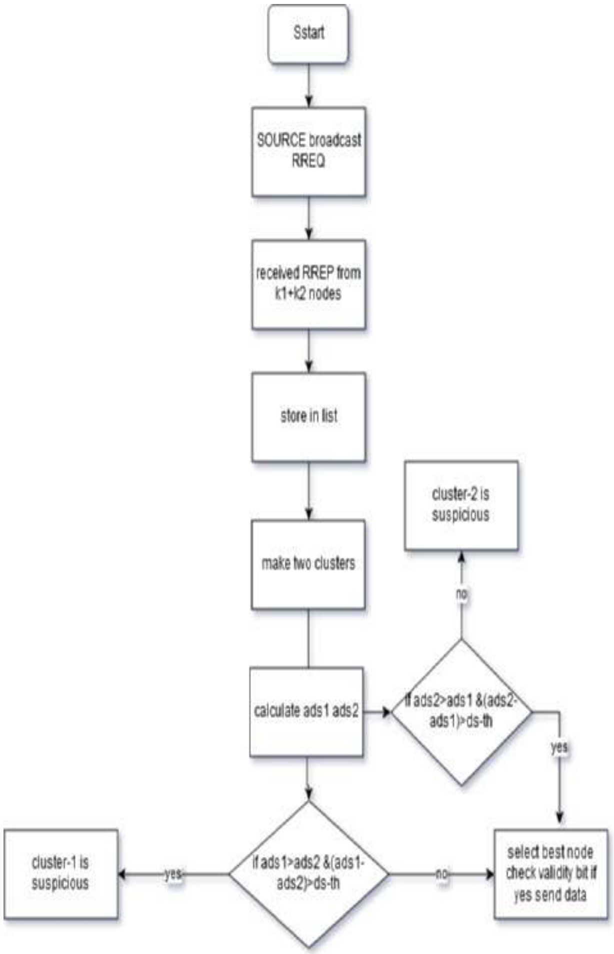

Sstart
SOURCE broadcast RREQ
received RREP from k1+k2 nodes
store in list
make two clusters
cluster-2 is suspicious
no
calculate ads1 ads2
if ads2>ads1 &(ads2-ads1)>ds-th
yes
cluster-1 is suspicious
yes
if ads1>ads2 &(ads1-ads2)>ds-th
no
select best node check validity bit if yes send data

Fig. 3. Flowchart of proposed solution

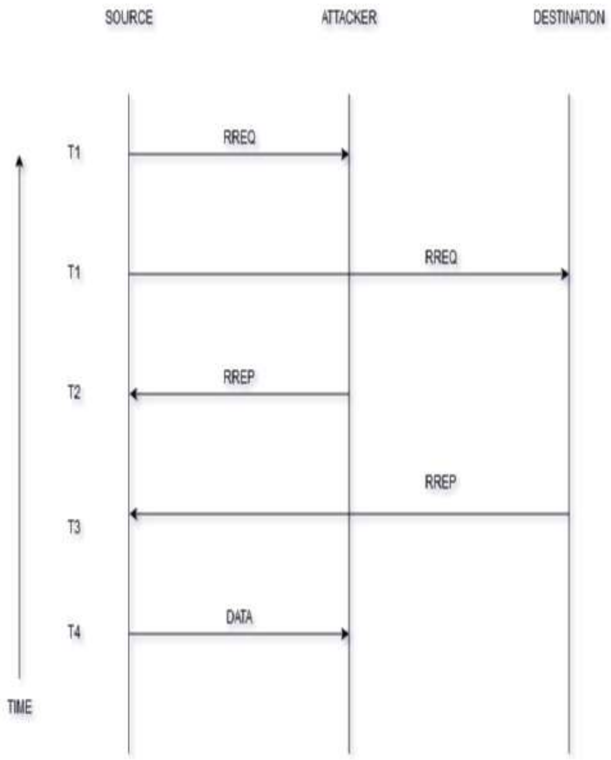


Fig. 4. timing diagram of black-hole attack

V. Discussion And Performance Analysis

In MANET, protection is a major problem. The entire network can be devastated by attacks. One of them is the Black-Hole attack [18]. Due to complex network topology, the location of the malicious node is very difficult to locate.

In this paper, we have proposed a method for detecting the black hole attacker using the proposed algorithm based on binary partition algorithm. In [1], the RREP message incorporates the validity bit mechanism.This new field is used to validate the route validity. Whenever a Route reply is received from a node, it is processed only if the RREPs validity bit is specified. Only if the validity bit is set will an entry be made for this path. As an assailant, this mechanism is not known; it reacts without looking into its route table. This implies that the validity bit in the RREP sent by the attacker node would have a null value. If a node receives a validity value not specified, the RREP can be simply dropped without entering the routing table.

We found that the above-proposed method is a good method that detects a black-hole attack. However, there is a problem, if the attacker will look to its routing table and set the validity value and send RREP to the source, then the source node will fail to detect the black-hole attack and send data packet in that route thinking that this is the destination path. Then attacker will drop data without forwarding it to the destination node. According to our proposed method,the source after generating a route request broadcast it and wait for RREPs. Source receive RREPs from k1+k2 no. of nodes. After receiving RREPs from k1+k2 no. of nodes store them in a list.

Make two groups of RREPs store in the list using binary partition clustering algorithm based on destination sequence number. Then we calculate the average destination sequence number (ADS) of both the clusters i.e ADS1 and ADS2. If ADS1 is greater than ADS2 and the difference is greater than a threshold value then cluster-1 may contain set of black hole nodes otherwise source node select the best node from cluster-1, check validity bit, if set, send data.

If ADS2 is greater than ADS1 and the difference is greater than a threshold value then cluster-2 may contain a set of black-hole nodes otherwise source node select the best node from cluster-2 and check validity bit if set, send data through that node.

An example is taken to discuss the proposed solution. The example is discussed in the following lines.

We take 11 RREPs nodes. Source receiving these 11 RREPs store it in a list. Then source makes two clusters namely cluster-1 and cluster-2. The cluster table is given in TABLE V and TABLE VI.

TABLE V
CLUSTER-1 TABLE WITH DESTINATION SEQUENCE NUMBER

Number of nodes	Destination sequence number
1	50
5	40
2	45
8	48
3	46

TABLE VI
CLUSTER-2 TABLE WITH DESTINATION SEQUENCE NUMBER

Number of nodes	Destination sequence number
4	10
6	15
7	14
9	11
10	12
11	13

The method is used to calculate the following
$ads1 = 50 + 40 + 45 + 48 + 46/5 = 45.8$
$ads2 = 10 + 15 + 14 + 11 + 12 + 13/6 = 12.5$
$ads1 - ads2 = 45.8 - 12.5 = 33.3$

let $ds - threshold$ value is 10.

From the proposed method, it is found that cluster-1 having the highest ads1 values as well as the difference between ads1 and ads2 is greater than the ds-threshold. From this, we can know that cluster-1 may contain a set of black-hole attacker nodes.

So source select the highest sequence number node from cluster-2 that is node-6 which have destination sequence number 15. The source node checks the validity bit, if it is set then send data through that node. In [1] the attacker can easily perform a black-hole attack by modifying the validity bit. But the proposed method solves the problem by segregating the attacker nodes from the normal nodes by using our proposed algorithm. Hence the sender can detect the attacker nodes easily and prevent the black-hole attack.

TABLE VII
SIMULATION PARAMETERS

Parameters	Values
Simulator	ns-2.35
Simulation Time	200 seconds
No of Nodes	100
No of attacker	1, 2, 3, 4, 5
Maximum Speed	30 m/s
Protocol	AODV
Packet Size	1024 Bytes
Traffic Model	TCP

VI. SIMULATION RESULTS

The Proposed solution is simulated using ns-2.35 and compared with Validity bit based approach [1].The simulation parameters are listed in table-VII. In figure-5, we plotted the graph of packet delivery ratio and in figure-6, we plotted the graph of routing overhead with the increasing no of attacker nodes from 1 to 5. From the figure-5, It is observed that packet delivery ratio is close to 1 and then decreases with the increasing no of attacker nodes but at all the points it is better than the validity based approach. Similarly in figure-6, it is observed that the proposed solution gives better result in terms of routing overhead with increasing no of attackers.

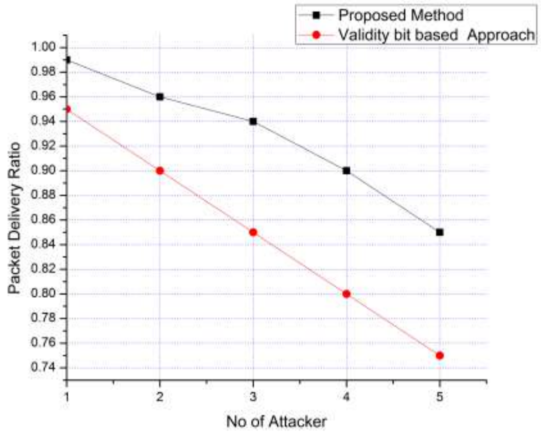


Fig. 5. Packet Delivery ratio Vs No of Attacker Nodes

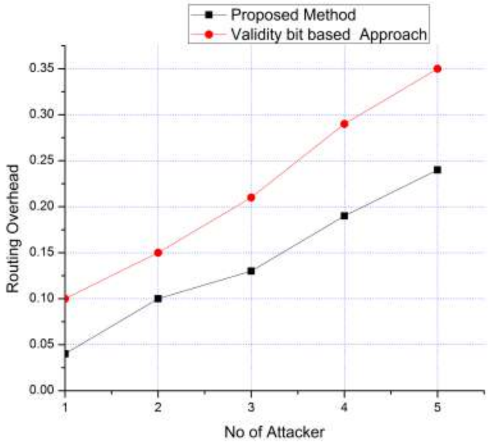


Fig. 6. Routing Overhead Vs No of Attacker Nodes

VII. Conclusion Future Direction

In this paper, we shown that, how the blackhole attack can be performed in the validity based approach [1] and proposed a solution for solving the issues. We compared our proposed method with the validity based approach and shown a better result using simulation result and analysis using example. In future we aim to extend this work for VANET in different scenario and more simulation and analysis results

References

[1] S. R. Deshmukh, P. Chatur, and N. B. Bhople, "Aodv-based secure routing against blackhole attack in manet," in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2016, pp. 1960–1964.

[2] A. Baadache and A. Belmehdi, "Avoiding black hole and cooperative black hole attacks in wireless ad hoc networks," *arXiv preprint arXiv:1002.1681*, 2010.

[3] R. Lakhwani, V. Jain, and A. Motwani, ""detection and prevention of black hole attack in mobile ad-hoc networks," *International Journal of Computer Applications*, pp. 0975–8887.

[4] P. N. Raj and P. B. Swadas, "Dpraodv: A dyanamic learning system against blackhole attack in aodv based manet," *arXiv preprint arXiv:0909.2371*, 2009.

[5] U. Satapathy, B. K. Mohanta, S. S. Panda, S. Sobhanayak, and D. Jena, "A secure framework for communication in internet of things application using hyperledger based blockchain," in *2019 10th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2019, pp. 1–7.

[6] S. S. Panda, D. Jena, and B. K. Mohanta, "A remote device authentication scheme for secure communication in cloud based iot," in *2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*. IEEE, 2019, pp. 165–171.

[7] A. K. Jain and V. Tokekar, "Mitigating the effects of black hole attacks on aodv routing protocol in mobile ad hoc networks," in *2015 International Conference on Pervasive Computing (ICPC)*, 2015, pp. 1–6.

[8] J. V. Vadavi and A. G. Sugavi, "Detection of black hole attack in enhanced aodv protocol," in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 2017, pp. 118–123.

[9] S. Singh, A. Mishra, and U. Singh, "Detecting and avoiding of collaborative black hole attack on manet using trusted aodv routing algorithm," in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016, pp. 1–6.

[10] H. Mahore, R. Agrawal, and R. Gupta, "Agent based black hole detection technique in aodv routing protocol," in *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, 2018, pp. 1–6.

[11] V. K. Saurabh, R. Sharma, R. Itare, and U. Singh, "Cluster-based technique for detection and prevention of black-hole attack in manets," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 2, 2017, pp. 489–494.

[12] Sathish M, Arumugam K, S. N. Pari, and Harikrishnan V S, "Detection of single and collaborative black hole attack in manet," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016, pp. 2040–2044.

[13] M. R. Dey, U. Satapathy, P. Bhanse, B. K. Mohanta, and D. Jena, "Magtrack: detecting road surface condition using smartphone sensors and machine learning," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 2485–2489.

[14] M. Al-Shurman, S.-M. Yoo, and S. Park, "Black hole attack in mobile ad hoc networks," in *Proceedings of the 42nd annual Southeast regional conference*, 2004, pp. 96–97.

[15] M.-Y. Su, K.-L. Chiang, and W.-C. Liao, "Mitigation of black-hole nodes in mobile ad hoc networks," in *International symposium on parallel and distributed processing with applications*. IEEE, 2010, pp. 162–167.

[16] K. Krishna and M. N. Murty, "Genetic k-means algorithm," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 3, pp. 433–439, 1999.

[17] R. H. Jhaveri, S. J. Patel, and D. C. Jinwala, "A novel approach for grayhole and blackhole attacks in mobile ad hoc networks," in *2012 Second International Conference on Advanced Computing & Communication Technologies*. IEEE, 2012, pp. 556–560.

[18] F.-H. Tseng, L.-D. Chou, and H.-C. Chao, "A survey of black hole attacks in wireless mobile ad hoc networks," *Human-centric Computing and Information Sciences*, vol. 1, no. 1, pp. 1–16, 2011.

Use of Vehicle Breathalyzers in the Reduction of DUI Deaths

1st Evan Yousif
dept. Computer Scinece and Informatics
Oakland University
MI, USA
eryousif@oakland.edu

2nd Dafer Alali
dept. Computer Scinece and Informatics
Oakland University
MI, USA
dalali@oakland.edu

3rd Surah Aldakhl
dept. Electrical and Computer Engineering
Oakland University
MI, USA
saldakhl@oakland.edu

4th Prof. Mohamed Zohdy
dept. Electrical and Computer Engineering
Oakland University
MI, USA
zohdyma@oakland.edu

Abstract*—For many years, there was a challenge of using detecting drivers who operated vehicles under the influence of alcohol. Getting tools that were accurate for measuring results was a major challenge. However, the innovation of Vehicle Breathalyzers assisted in boosting the reduction of Driving Under the influence [DUI] of alcohol or drug deaths. The existent errors in examining the DUI for drivers were resolved through the invention of the Breathalyzer whose capability to measure alcohol amounts has always helped to determine consuming alcohol amongst individuals. The Breathalyzer is one of the essential devices that identify whether a person being examined is drunk legal. This is *because* it provides the proportion of alcohol vapors that are exhaled into the air with a proportion that illustrates the alcohol content in the blood. The current Breathalyzer ignition systems usually require a driver to breathe into the device before starting the vehicle. If the driver alcohol content is above 0.08 percent blood concentration of alcohol, the driver can receive penalties according to the law. Studies have shown that the use of Breathalyzer has grown by a rate of 15 percent per year since their development. The Breathalyzer technology includes data systems which provide a recording of both driver and vehicle responses and rapid systems for monitoring the alcohol state in the driver. This paper provides a description of how these technologies can be applied by drivers to enhance safety in driving. This research will bring insight into how accidents are associated with drunk-driving crimes. It contributes also to the researchers who are looking for a way to reduce road accidents. Through the data collection, analysis, results, and findings, we will get the realization that using a breathalyzer to test drivers, especially at night will help police to identify drivers who have committed the crime easily.

Keywords— Breathalyzer, Monitoring devices Drinking Under Influence (DUI), Blood concentration of alcohol, Interlock programs.

Introduction

Driving under influence (DUI) was identified as the major cause of fatal traffic accidents by the Texas state government. However, a solution to detect drunk drivers was always a major challenge facing drunk drivers. Every hour, a person around the world dies because of alcohol-related driving. For many years, technology has played a major role in enhancing traffic safety. In many years' alcohol tests for drunk drivers had depended on the testing of urine and blood among the drivers. However, many state laws found these laws to be ineffective in providing sufficient outcomes, especially for drunk drivers. Current evolutions have led to the development of chemical testing methods that are ineffective among the drivers. However, the recent growth of these laws has led to the advancement of systems that can control drunk driving while at the same time providing sufficient measures that can control impaired drivers [1]. The present research examines the need for using alcohol Breathalyzers in curbing driving under the influence (DUI). Past studies have shown that drinking under the influence is the major cause of death on the roads. However, the installation of Breathalyzers of curbing DUI offenders has reduced accidents by 15 percent. This shows that the reduction of alcohol-related deaths can be achieved using these Breathalyzers whose focus is to stop the vehicle once it has detected that the driver has been under alcohol impairment. Most of the current vehicles have shown the existence of mandatory ignition interlock systems with the capacity to control alcohol-related crashes. The reduction of mandatory air bag laws and the 21-year minimum drinking age has been a major factor that has been identified for the reduction of alcohol consumption. [2]. The advancement in technology has a significant impact on the consolidation of rules versus DUI of alcohol and drugs. The evolution of Breathalyzer by Birkenstein is one of the most distinguished progression that have assisted to suppress DUI. This is because the Breathalyzer technology comprises a GPS that manages the driving and the position of the driver. Breathalyzers provide a real test of an intoxicated individual. The vehicles also connect to the GPS that would allow law enforcement officials to detect any suspected driver [3].

The major goal of this research work was to inspect Breathalyzer ignition devices and their efficiency to minimize alcohol intoxication between drivers. The selected studies in this paper investigated the performance of the Breathalyzer interlock programs on discontinuing drunk driving. The research studies utilized assistance to offer adequate results of the program through the selected individuals. The outcomes of the studies demonstrated that the use of Breathalyzer programs in stopping alcohol-linked deaths were considerable. However, the former testing of these technologies presented their efficiency in ensuring that alcohol-linked deaths have been reduced, it's evident that the quick advancement of these programs would help to improve

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

road safety in the future. This paper helps to outline that the developments in the interlock programs would play a main role in the special DUI courts in the monitoring of the vehicles with drunk drivers [4]. The remainder of the paper includes research significance, followed by a literature review which analyzed the approaches to the advancement of the road safety linked with drunk driving. It has been inclusively elucidated that the use of interlock systems, remarkably reduced the numbers of accidents. The methodology section concerned the estimation of data since the installation of Breathalyzer technologies as well as the sufficient analysis of the data to ensure that the control of death was efficient. The data and results section demonstrated that the use of Breathalyzers was effective in managing deaths related to DUI. The discussion section covered in-depth analysis of the gathered data associated with the research. The conclusions comprised a summary of the findings and recommendations.

I. RESEARCH SIGNIFICANCE

The significance of this research is that the results of data analysis of this research will be helpful to identify ways that drunk-driving crime can be controlled. This research also helps to bring an insight cover of the previous method that had been implemented but not effected. It also outlines the weakness of the previous methods, to create an idea of how engineers and medical practitioners could come up with a device that will assist the policemen in the day-to-day management of drunk trafficking. This research also will create a good foundation for future researchers who would like to make an invention on automatic machines or vehicles that raise a siren when the driver is drunk.

II. LITERATURE REVIEW

According to the United States Government Accountability officer in 2015, only 20 percent of drivers chose to use Breathalyzer. While the installation of the Breathalyzer is relatively cheap, it is evident that most individuals do not install it due to the interlock system that prevents individuals from using their vehicles, especially when drunk. Sufficient evidence conducted by the Boston State University in 2017 has shown that there is an increasing number of individuals who have installed the Breathalyzer in their vehicles. The focus is to comply with the laws [5]. Close monitoring of the effectiveness of the Breathalyzer interlock programs has shown that there is an increase of 65% in recidivism benefit during the period. This shows that monitoring of vehicles can be a major way that would help to control drunk driving in the United States and all over the world. While the installation of these devices has not been enforced by all states, it is evident that the increased installation of these devices has helped to curb the issue of drunk related driving in the country.

There is a need to identify measures that can be adopted to ensure that these systems are not used as punitive measures among individuals to control drunk related driving in the country. There is a need for the implementation of programs that would ensure that impaired driving has been reduced in the country. The installation of alcohol detection systems would pave way for improved management of road safety. Innovative ways to support the expansion of these programs would also help to control any alcohol-impaired driving in the country [6]. Figure 1 is a graphical representation of how Breathalyzers can be used in the detection of alcohol concentration in the blood. The importance of this graph is to show that the use of breathalyzers has helped to prevent any existence of alcohol-related crashes. This relationship also shows that the detection of alcohol consumption is greatly reduced beyond the 6 second blow time [7].

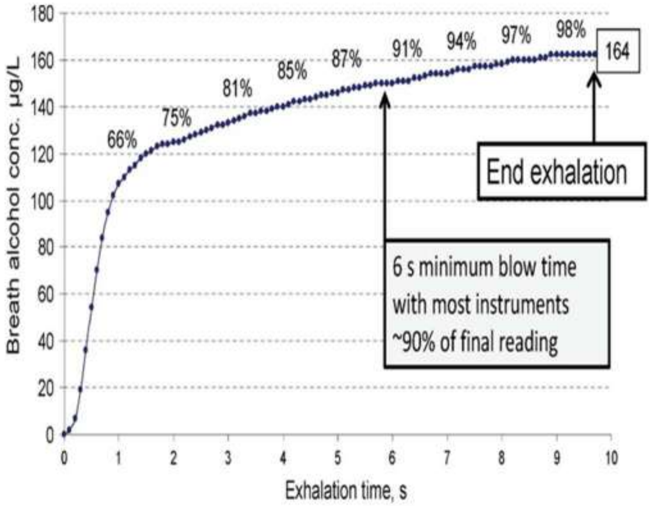


Figure. 1. Breath alcohol concentration as a function of exhalation time

The present study describes how these new electronic technologies can enhance safety on our roads. The first application of Breathalyzer technology by the Department of Transportation provided a solution to the prevalent concerns associated with license suspension, while at the same time protecting the public against the DUI offenders as representative to other drivers. The suspension and revocation of licenses through individuals who had repeated offenses help to curb the deaths associated with DUI [8]. According to Lebow, the main elements of the Breathalyzer technology are the interlock system, alcohol sensors, and an ignition sequence. The other functions of the program entail rolling retests, circumvention prevention, and the interlock categories. These systems help in the control of how this technology is used on the roads. The use of these technologies helps to monitor driving and controls the use of programs that control driving. The system contains technologies that ensure an adequate monitoring of the driver’s breath on a 24.7 basis [9].

The use of the GPS systems is to provide sensor advancements that would control the use of these technologies, hence preventing offenders from increased consumption of alcohol. Providing improved abstinence from alcohol is one of the major objectives associated with Breathalyzer technologies. The current use of continuous monitoring of vehicles using these systems has provided an alternative that would help to ensure that drunk driving has not been implemented in the country. The utilizations of these systems have provided a better alternative that would ensure that these systems have helped detect drinking habits in the population. Currently, there are more than 36 states that require the use of devices to help control DUI. Most of these laws are worse, especially for the offenders who have been found to have been culpable for any DUI offense. Repeat offenders are also found to have major cases related to drunk driving. Installation of ignition interlock has not been a mandatory practice in the country [10]. Numerous Studies have exhibited that the assessment of interlocking technologies such as Breathalyzers has become difficult over the years. Programs for assessing these technologies have not been established. However, data logger technology offers an effective program for managing the systems to guarantee that their performance has been identified. The effectiveness of these systems requires to be identified. Lockouts require to be specified for providing screening devices for the identification of drivers at high risk due to impaired driving [11]. State legislatures should identify means for managing any driving behaviors to determine a basis for the accomplishment of technologies that have been modified for the improvement of interlock experiences [12]. Statistics performed by the Federal Transport Department have demonstrated that car crashes that include the use of alcohol usually approaches 35% of all vehicle fatalities on the roads. This results in more than 10,000 individuals who die on the roads every year. Based on research carried out by the National Highway Traffic and Safety administration, more than one million driving convictions have taken place each year due to alcohol-related consumption. There is a need to undertake measures to ensure that these levels of alcohol consumption have been reduced on the roads. The present laws do not provide adequate prevention measures. However, the contrivance of alcohol breathalyzer offers a satisfactory measure as an interlock law that assists to guarantee that individual fatalities on the roads have been reduced [13].

In 2017, a study conducted by the National Safety Council indicated that the use of interlock systems was one of the most innovative ways to the reduction of road accidents. This was one of the measures that could be adopted to ensure that there were no road crashes that could affect individuals on the roads. The study indicated that the sufficient use of the roads with adequate ignition systems could play a major role that would ensure that road accident never occurred on the roads [14]. While the existing policies and interlock laws have helped to reduce drunk driving, few measures have been established to ensure that drunk driving has been controlled. There is a need for the establishment of policies that provide control for drunk driving to enhance road safety. These measures would be effective in ensuring that public health conversations can be enhanced without any fear of road crashes [15]. There is also a need for policy measures to ensure that public health engagements have involved the use of roads. Reforms of the interlock laws in all states should be carried out to ensure that public health conversations have been established. These policies would help to provide a comprehensive approach to driving. The alcohol culture in many states needs to be identified and reduced to ensure that the establishment of safety in driving has been enhanced. The establishment of these laws that help to curb drunk driving is one of the measures that can be established to improve road safety among the users [16].

There is a shred of strong evidence that shows the consistent use of alcohol breathalyzers for road users can enhance road safety. The drivers who have been advised to use interlocks are individuals who have been previously convicted of engaging in alcohol usage. There is a need to establish sufficient measures among road users that would help to curb and control the use of vehicles while under the influence of alcohol [17]. The adoption of these measures would provide a piece of strong evidence that indicates the need for road users to ensure that they have used sufficient interlock systems that would help them to control their alcohol use. These measures would provide adequate safety that would guide road users and ensure that there is immediate feedback when it comes to the consumption of alcohol among the road users. This would also help to advance road safety while at the same time improving appropriate alcohol consumption. The implementation of the ignition interlock system is one of the ways that can be used to ensure that the circumvention and data collection among road users has been enhanced. The durability and compliance among road users are one of the most important measures that can be used among road users to enhance road safety while at the same time increasing the ability to control alcohol usage among users on the roads [18].

The present research paper incorporates the approaches to the enhancements of the road safety associated with the drunk driving. There are many commitments that have been achieved in this work. It has been comprehensively demonstrated that the use of the interlock system significantly reduced the figures of road accidents. The rationale behind such an advancement on the road safety is the road transport department implementing the policy on the utilization of the Breathalyzer technologies. The legislation of the penalties for any driver who violate any of the road safety provisions improved the road safety across various states. Eventually, the results of the data analysis show a positive relation between the adoption of the breathalyzer interlock systems and the reduction in the number of road accidents across various states.

III. METHODS

This study identified both fatal and non-fatal injuries related to the DUI vehicle crashes from the National Automotive Sampling Systems General estimates data sets from 2015 to 2019. Similar reports were also obtained from the Fatality Analysis, Reporting systems in the same period. The estimation of data from the installation of the Breathalyzer technology was also identified. This estimate was also compared with the proportion of alcohol-related crashes. [19]. The number of deaths that were preventable because of the implementation of the system was also identified. The focus was to ensure that in the years during the installation the Breathalyzers helped to control the number of crashes that occurred as a result. The analysis also assumed that the economic savings and the lives that were saved during a similar period. The data were sufficiently analyzed to ensure that control of deaths was efficient. The device's effectiveness in the control of accidents was also identified to ensure that no deaths occurred as a result.

Probability sampling of 100 vehicles in which installed a Breathalyzer was used were compared with normal vehicles that did not employ the gadgets. The participants came from different states. Descriptive correlations were used to test and retest of all breathalyzes with 0.000 readings. The aim was to ensure that the correct readings were identified to measure the accuracy of the systems in measuring alcohol consumption. The participants in this research were approached to ensure that they had carried out a Breathalyzer test. Every participant had been requested to provide accuracy in their reading to ensure that correct feedback had been provided for the research purposes [20].

IV. RESULTS

From the results and findings, breathalyzer program is effective on installation on vehicles. Having done some analysis, the risk occur was 0.36 having 95% intervals of 0.2

and 0.63. Additionally, the program had been supported by controlled trials having the same test proofed that the installation of the breathalyzer on vehicles have helped to reduce the number of accidents caused by drunk drivers. The results also indicated that the use of Breathalyzers was effective in controlling deaths because of DUI. Having implemented this, with an alcohol limit of 0.08, it has significantly reduced the number of road accidents. [21]. From the results, a significant difference was identified. This was avoided by using correlations. Additionally, the large samples from different states were examined to give support to the results. Physical behavior of the alcoholic persons was also recorded to help me identify the real drunk driver. The participant reading using the One-way analysis of Variance (ANOVA) would be employed through this study to give samples. This contributed in coming up with a different in the collected samples. A 95% RCI difference would also be used as an efficient measure for the analysis of the research to accurately identify the samples. The identified absolute mean for the Breathalyzers was F (3, 115) = 0.35, p = 0.79.

The results of testing the Breathalyzer on its reduction of accidents related to DUI as follows. Figure II shows that the installations of Breathalyzer have a significant impact on the reduction of DUI offenders in the United States. The circled part shows the 5-month period on which the Breathalyzers interlock systems have been relicensed. The devices did not show any errors in the recording as there was more than a 15 percent indicated that drivers have been tested before they had driven their vehicles [22]. Table 1 showed the cumulative results of deaths associated with a DUI. Additionally, the number of records of captured drivers has been recorded which typically shows the trends on how the number of culprits is significantly reduced. Table 2 represented the record just from the last 12 months of the year 2020.

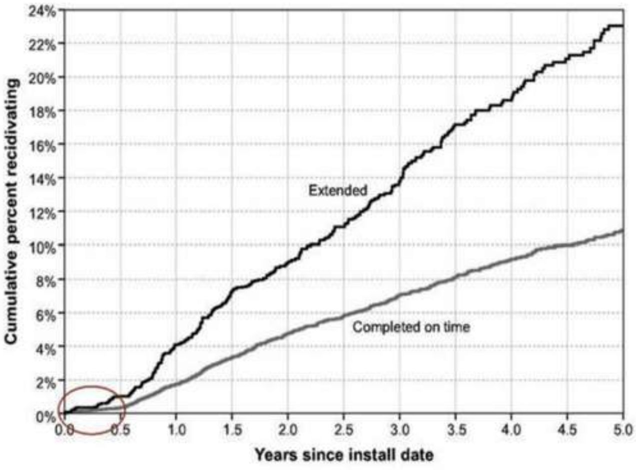
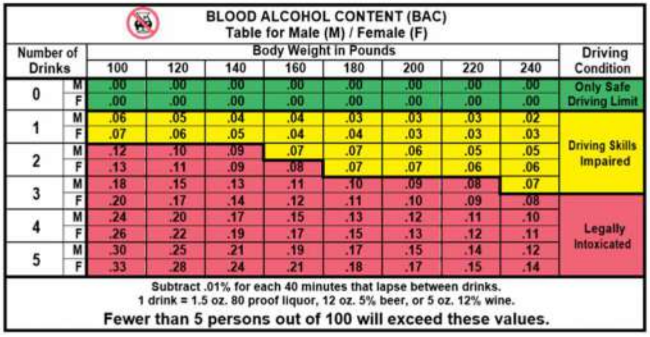


Figure 2. Cumulative percent recidivating as a function of years since install date.

TABLE 1. Percentage Rate of Accidents Caused by DUI for Last 3 Years

The percentage rate of Accidents Caused by DUI Per year			
Months	**2010**	**2015**	**2020**
January	88%	74%	34%
February	83%	62%	42%
March	85%	80%	33%
April	82%	71%	32%
May	71%	62%	27%
June	73%	59%	48%
July	87%	54%	32%
August	76%	73%	23%
September	89%	74%	21%
October	76%	63%	43%
November	80%	65%	12%
December	92%	77%	33%

TABLE 2. Blood Alcohol content (BAC)



BLOOD ALCOHOL CONTENT (BAC)
Table for Male (M) / Female (F)

Number of Drinks		Body Weight in Pounds								Driving Condition
		100	120	140	160	180	200	220	240	
0	M	.00	.00	.00	.00	.00	.00	.00	.00	Only Safe Driving Limit
	F	.00	.00	.00	.00	.00	.00	.00	.00	
1	M	.06	.05	.04	.04	.03	.03	.03	.02	Driving Skills Impaired
	F	.07	.06	.05	.04	.04	.03	.03	.03	
2	M	.12	.10	.09	.07	.07	.06	.05	.05	
	F	.13	.11	.09	.08	.07	.07	.06	.06	
3	M	.18	.15	.13	.11	.10	.09	.08	.07	
	F	.20	.17	.14	.12	.11	.10	.09	.08	Legally Intoxicated
4	M	.24	.20	.17	.15	.13	.12	.11	.10	
	F	.26	.22	.19	.17	.15	.13	.12	.11	
5	M	.30	.25	.21	.19	.17	.15	.14	.12	
	F	.33	.28	.24	.21	.18	.17	.15	.14	

Subtract .01% for each 40 minutes that lapse between drinks.
1 drink = 1.5 oz. 80 proof liquor, 12 oz. 5% beer, or 5 oz. 12% wine.
Fewer than 5 persons out of 100 will exceed these values.

TABLE 3. Percentage Number of Drunk Drivers per 100 Heads in the Year 2020

Months	**Percentage Number of Drunk drivers per 100 head in the year 2020**	
`	**Low**	**High**
January	20%	80%
February	22%	78%
March	48%	62%
April	50%	50%
May	52%	58%
June	40%	60%
July	66%	34%
August	72%	28%
September	78%	22%
October	82%	18%
November	90%	10%
December	91%	9%

V. DISCUSSION

The Breathalyzers have been designed to utilize the three technologies. These three technologies include semiconductor oxide sensor, an infrared spectrometer, and a fuel cell sensor. By using this technology, it facilitates the identification of alcohol in the blood. Normally called Blood Alcohol Content (BAC). The in-car Breathalyzer device can be described as ignition interlock systems whose focus is to measure the alcohol level in the breath hence preventing people from starting their vehicles. The alcohol content is measured from the breath through the breath alcohol concentration. The installation of the Breathalyzer is required to be used in law, especially by people who have been convicted of engaging in drunk driving. In many cases, the installation of these devices helps to ensure that driving privileges have been gained by individuals. The interlocks care of these devices helps to ensure that the driver has been efficiently controlled. This is due to the focus on the implementation of the systems on vehicles. It's clear that the interlock system data provide storage for the data and driving experience and locks the vehicle when the driver has been detected to have been drunk. The restriction on the driving experience of an individual is controlled through the reduced driving because of the use of the system for a long.

Table I indicated that the data collected show that in 2010, there was a greater number of a road accident caused by drunk drivers. We see the percentage generally is above 80%. As compared to 2015 where the number of accidents caused by DUI reduced. Comparing with 2020, the percentage rate had drastically reduced. This is because the introduction of Breathalyzers helps to control the number of drunk drivers on the road. For that year only we can see that the percentage rate is reduced from January to December where we had a slight high. This is for the fact that December has got holidays which, well, most of the people go out for fun. The gathered data presented in Table II clarify the percentage number of drunk drivers who linked to a low or high number of alcohols in their blood. Classified records were used for the year 2020. The registered data demonstrated that the driver's numbers with a high percentage of alcohol was high at the beginning. Subsequently, the numbers start to decrease with time. This happened because the drastic rules and regulations of the government may be established. Furthermore, moving towards the end of the year, the number of drunk drivers with a high amount of alcohol became lower. This is associated with the fact that most of the drivers had known that the Breathalyzer tool can give a clear evidence of the level of alcohol that the driver has taken. Any driving behavior that has been detected to have been unreasonable is always detected by the law enforcement agency. This is because the system also contains a GPS installed on it that can provide accurate location of the vehicle on a real-time basis. The improvements in safety behavior, especially in driving are also enhanced hence improving the driving experience for the drivers. DUI offenders have been forced by the government to use this system to control the driving experience on the roads. This helps to ensure that the driving among the road users has been controlled [23].

Over the years, the high cost of incarceration has contributed to the replacement of jail with monitoring programs. Breathalyzers have been identified as one of the most efficient systems that can be used to control the driving behaviors of major road users. This is because the technology contains an emphasis on the management of controlling electronic devices that are used by the drivers on the roads. There is a need for using interlocks that help to control any consumption habits among users on the roads. The interlocks provide efficient ways of managing these road users while at the same time providing consumption adjustments on the roads. These systems are efficient in providing control of the usage of the programs while at the same time providing sufficient measures for an advisory to the road users. To operate a car that has been installed with the Breathalyzer interlock system, the driver is required to provide a breath specimen. The alcohol concentration level is then monitored, which makes it easy to detect, especially if the road user has engaged in intoxicated levels of consumption. The use of random retests is also a requirement that should be identified and monitored early enough to ensure that road users have effectively been tested on their alcohol consumption levels. Studies have shown that the Breathalyzer often reacts to the potassium solution which turns green with the slightest detection of alcohol that is expelled in the air. The ability of this device to detect alcohol through chemical reaction is one that makes it an efficient system. Most of the available breathalyzes have a fuel cell sensor, semi conduct oxide sensor, and an infrared spectrometer. Most of the available fuel testers have a high degree of sensitivity and high accuracy for handheld and portable systems. Initially, drivers had tested for alcohol impairment through the roadside screening device. However, the development of the Breathalyzer has helped to provide real-time positive testing that helps to provide the identification of the motorists who have engaged in alcohol consumption over time. [24]

Breathalyzer devices are used for evidential breath testing to increase ethanol consumption. The accuracy of the device often depends on the breath being taken deep from the lungs. This means that the use of this device is not 100% guaranteed to identify a low concentration of ethanol consumption. However, increased testing using the device has proven that it is easy to identify alcohol consumption when an individual has taken too much to become impaired to drive. Figure 3 demonstrates that the breath alcohol concentration and how it is identified by a person who has consumed too much alcohol. According to the graph, individuals who have consumed large quantities of alcohol end up being detected earlier. However, less consumption is not identified using the Breathalyzer [25].

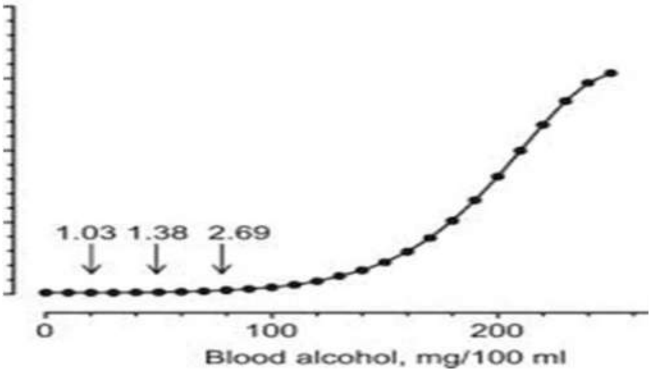


Figure 3. The breath alcohol concentration as a function of blood alcohol

The studies have indicated that road users are not fully informed about their transportation behaviors, especially how they should drive their vehicles. There is a need for educating the driver's on-road behavior to control any existing road accidents. Also, there is a need for established inspections that would be conducted periodically to ensure that vehicles do not endanger road users. There is also a need to have collision information to control the behavior of road users and ensure that transport information has been kept. This would increase the behavior of the road users and ensure that no accidents occur on the roads. Penalties for individuals who do not install the vehicle ignition systems should be established to ensure that road users have adequate systems that can be used to improve road safety. This would improve the transportation of vehicles while at the same time enhancing road control measures.

There is a need to have zero tolerance on drugs and alcohol especially when driving. The absolute adherence to road safety rules will ensure that a better driving experience has been established on the roads. This would also reduce the level of deaths and crashes that occur on the roads. The adoption of the present policies and laws on the roads will improve the statistics of road users. Since the statistics show that Breathalyzers can help, enforcement officials need to ensure that safety measures have been adopted hence improve road safety. This would be one of the measures that would help to ensure that road safety has been improved while that the same time ensuring that the culture of road measures and accident reduction strategies have been improved on the roads [26]. The advancement of the Breathalyzer needs some software code design with the communication. The software codes should be capable of giving a judge without human interference. The associated coded software should be integrated with the device. The software codes have a logic provided that, if the out-breath or the bloodstream of a drunk person comprises some alcohol, the software code can be designed to highlight a red alarm and indicate on the screen that the driver is drunk. Such software can make the Breathalyzer authoritative to be adopted as a control and judge tool for this purpose. Moreover, this software can be designed to give an alarm signal to the police department indicating that there is a drunk driver. Figure 4 is a diagrammatic representation of the Breathalyzer.

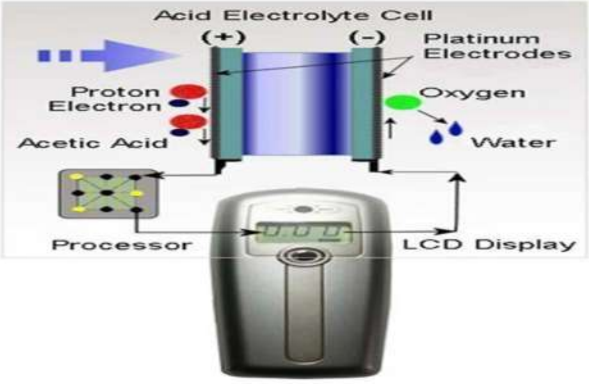


Figure 4. Diagrammatic representation of the Breathalyzer.

VI. CONCLUSIONS

To sum it up, while alcohol Breathalyzer work to reduce deaths associated with drunk driving, abstinence from drugs before driving is the only key to safe driving. The high costs of incarceration in the country have led to the implementation of DUI courts whose focus is to control drunk related offenses. The courts have established models and programs whose benefit is to enhance monitoring and treatment of drunk related offenses. When successful, these systems can help provide low cost to the government for the incarceration of offenders. The development of vehicle interlock programs can also provide sufficient management of alcohol consumption among drivers. The use of Breathalyzer technique will assist to minimize detention times, license suspension, and even reduction of fines among the users. The reduction of correction programs among the users is also another advancement that would be adopted, especially when the new systems have been implemented. Courts will also have more options for enhancing interlock programs while at the same time enhancing the confinement efforts for users. Since studies have shown that the Breathalyzer system has been effective in the control of drunk driving, measures should be established to ensure that there is improved durability, compliance, and also participation rates that improve the use of these systems to avoid the punishment of offenders.

VII. LIMITATIONS OF THE STUDY

This study has been limited due to the reduced nature of the samples selected. There was a need to conduct more data that could have helped to provide sufficient analysis of the study. This could have led to an increased examination of the study while at the same time enhancing the credibility of the findings. The selected sample size also contained limited data access, hence leading to insufficient sizes that could be used for efficient statistical measures. The use of a small sample size increased the margin of error in the research. For this reason, there was a need to increase the sample size to provide accurate representation of the statistical measures that were used in this research. This would have enhanced the outcomes of the research while at the same time providing more reliable information that could be used in the analysis. Moreover, this could have presented the researchers with better research outcomes that would have helped to provide sufficient data analysis [27].

Time constraints in also conducting this research were also minimized because of the existing data examination. The increased conflicts in personal bias and cultural issues were also major factors that limited the undertaking of the study, hence having negative outcomes on the consistency of the findings. Certain perspectives and phenomena in this research were required to provide a sufficient examination of the results, thus providing reliable information that would be used in the research process. One of the main limiting factors in this study was the lack of sufficient resources from the ministry of transportation to provide accurate and reliable results. The study could have increased its sample size and even used a more accurate methodology if the existing resources were sufficient. There was a need to provide transparency and honesty as a statistical measure to provide sufficient findings that could have been used by the ministry in the process [28].

VIII. RECOMMENDATIONS

The US Department of Transportation should adopt the Breathalyzer project and offer the required resources for its development to secure and enhance the road safety. The outcomes of the current research study have demonstrated that prohibiting every drunken driver from igniting his vehicle based on the interlock laws can considerably prevent car accidents and subsequent rescue peoples' lives. The establishment of the ignition policies on vehicles, can save greater than 15% of the lives, hence enhancing the road safety. Constructing of interlock programs and innovative measures to stop these vehicles can act as an adequate measure that can assist in managing road accidents while simultaneously providing innovative technologies for stopping road accidents [29]. State governments need to adopt sufficient measures for reducing car crashes on the roads through the adoption of policies that can help to curb road accidents. This can be achieved by providing a sensitivity and awareness of the breathalyzers and how they have helped to curb road accidents. This would be an efficient measure that can guide road users and give them an accurate understanding. One of the challenges that are faced on the roads. The adoption of sensible measures could provide an accurate understanding of the roads and how they pose danger to many users. Only interlock systems, vehicles should be allowed on the roads. This would make the vehicles more efficient for road users, as a result stop the road accidents that occur on a routine basis. Also, the limitation of driving while using cell phones should be identified as one of the measures that can be used to eliminate car crashes and enhance road safety [30].

The need for road safety needs to be established efficiently through the creation of a road safety culture. The focus is to ensure that organizations have adopted measures for road accident reduction. These measures would be efficient in providing employees with realistic measures and schedules that can help to provide road safety. A culture of road safety is one of the measures that need to be adopted to ensure that road users do not engage in activities that can threaten their safety controls. Apart from alcohol drinking, there is also a need for assessing driver's skills regularly with a need for checking and examining their driving behavior. This is an efficient measure that can help to improve the experience of a driver on the road, thus preventing them from engaging in behaviors that can contribute to road accidents [30].

REFRENCES

[1] R.F. Borkenstein, and H.W. Smith, “The Breathalyzer, and its Applications”. Medicine, Science, and the law, 2018. Retrieved from https://journals.sagepub.com/doi/10.1177/002580246200200103.

[2] J. M. Byrne, and D. J. Rebovich. “The new technology of crime, law, and social Control”. Monsey, N.Y: Criminal Justice Press, 2007.

[3] J. H. Coben, and G. L. Larkin, (1999). “Effectiveness of ignition interlock devices in reducing drunk driving recidivism”. American journal of preventive medicine, 16(1 Suppl),1999, pp. 81–87.

[4] S. Kyd, Driving offences: “Law, policy, and practice”. Burlington, VT: Ashgate, 2018.

[5] M. D. Laurence, and F. E. Zimring, “Social Control of the Drinking Driver”. Press. Washington, DC: National Academies Press, 2018.

[6] Geller, A., and Y. Negussie, “Getting to zero alcohol-impaired driving fatalities: A comprehensive approach to a persistent problem”. Washington, DC: National Academies Press, 2018.

[7] J. G. Mørland, and R. H. In Liu. “Alcohol, drugs, and impaired driving”: Forensic science and law enforcement issues. New York: Oxford Press, 2020, pp.720.

[8] R. B. Voas. “Enhancing the Use of Vehicle Alcohol Interlocks with Emerging Technology”. Alcohol research: current reviews, 36(1), 2014, pp.81–89.

[9] R. B. Leukefeld, Gullotta, T. P. and J. Gregrich. Handbook of evidence- “based substance abuse treatment in criminal justice settings.” New York: Springer, 2011.

[10] D. M. Malone, and, P. J. Zwier. “Effective expert testimony”. New York: Springer, 2014

[11] A. W. Jones. “Alcohol, drugs, and impaired driving”: Forensic science and law enforcement issues, 2020.

[12] P. C. Giannulli, and E. J. Imwinkelried. Scientific evidence. Newark, NJ: LexisNexis, 2007.

[13] R. J. Bonnie, and M. E. O’Connell. “Reducing underage drinking: A collective responsibility”. Washington, DC: National Academies Press, 2004.

[14] H. Lowry and M. E. Powell. “Defending drinking drivers”. Santa Ana, CA: James Publishers, 2016.

[15] K. In Aase, and L. In Schibevaag. “Researching patient safety and quality in healthcare: A Nordic perspective” 2017.

[16] R. Hughes, and United States. “Patient safety and quality: An evidence-based handbook for nurses”. Rockville, MD: Agency for Healthcare Research and Quality, U.S. Dept. of Health and Human Services, 2008.

[17] K. Decker. DUI bible: “Avoiding a drunk driving conviction”. Place of publication not identified: Xlibris Corp, 2008.

[18] Y. Negussie. “Getting to zero alcohol-impaired driving fatalities: A comprehensive approach to a persistent problem”. Washington, DC: National Academies Press, 2018.

[19] L. Taylor. “Drunk driving defense”. Boston: Little, Brown, 1986.

[20] K. M. Wolf. “California courts and judges”. Costa Mesa, CA: James Pub.,1996.

[21] J. B. Jacobs. “Drunk driving: An American dilemma”. Chicago: University of Chicago Press, 1989.

[22] G. Block. “Effective legal writing: For law students and lawyers”. Mineola (N.Y.: Foundation Press, 1991.

[23] M. Zwygart-Stauffacher. “Advanced practice nursing: Core concepts for professional role development”. New York, NY: Springer, 2017.

[24] B. H. Lerner. “One for the road: Drunk driving since 1900”. Baltimore: Johns Hopkins University Press, 2011.

[25] G. T. Savage, and E. W. Ford. “Patient safety and health care management”. Bingley, UK: Emerald JAI, 2018.

[26] J. G Trichter, and W. T. McKinney. “Texas drunk driving law”. Carlsbad, Calif:Michie, 1996.

[27] W. Wisconsin. “West's Wisconsin statutes annotated”. St. Paul: Thomson, 201.

[28] H. F. Ashdown, and R. J. Stevens. “Diagnostic accuracy study of three alcohol breathalyzers marketed for sale to the public”. BMJ open, 4(12), 2014, e005811. https://doi.org/10.1136/bmjopen-2014-005811

[29] K. Carroll. “Improving compliance with alcoholism treatment”. Bethesda, Maryland, 2017

[30] K. Toonika. “Biosensing Technologies for the Detection of Pathogens - A Prospective Way for Rapid Analysis”. Oxford: New York, 2018.

Improved Flower Pollination Algorithm-based Optimal Placement and Sizing of DG for Practical Indian 52 Bus System

S. Dhivya, *School of Electrical Engineering, Vellore Institute of Technology,* Chennai-600127, Tamil Nadu, India,
dhivya.s2019@vitstudent.ac.in

R. Arul, *School of Electrical Engineering, Vellore Institute of Technology ,* Chennai-600127, Tamil Nadu, India
arulphd@yahoo.co.in

***Abstract*— Nowadays, fossil fuel exhaustion is one of the significant environmental issues. Hence, the existing power system of Renewable Distributed Generations (RDG) is used to overcome this issue. The proper planning must be implemented to allocate the separate sources in the network of distribution areas. This paper proposed an optimal placement and sizing of Distributed Generation (DG) with an improved flower pollination algorithm (IFPA). This paper aims to find the optimal location and sizing of the Distributed Generations (DG). The objective of power loss minimization is considered with some constraints like power balance equations, voltage limits, DG sizing limits, and thermal limits are being done in this research. This proposed technique is evaluated on the practical Indian 52 bus system. The results of the optimal Distributed Generation location and sizing are obtained with the help of IFPA. Therefore, the percentage of loss reduction through IFPA is comparatively higher than the FPA and the chaotic FPA.**

***Keywords*— Distributed Generation (DG), Optimal placement and sizing, Improved Flower Pollination Algorithm, Power loss reduction, Practical Indian-52 bus system**

I. Introduction

Due to the fast growth of industries and population, there is a demand for electricity for the customers. Hence, 13% of the total generated power is wasted in distribution as line losses. The inductive loads are sustained mainly by the distribution network (DN). It sometimes gave poor power factor, voltage sag, and high system loss. Distributed Generation (DG) is installed to solve and overcome these problems, a possible solution [1]. The DN is the one portion of the entire system. It would join the transmission system of high voltage to consumers of low voltage. When the level of the target is at 7.5%, then the distribution loss is 15.5%.

Therefore, proper planning must be done for the DN of the primary and the second part. In the primary and the secondary distribution sectors, the total losses obtained were 70%, and the remaining 30% losses are obtained in the transmission and the sub-transmission lines. The Distribution Networks of the conventional arrangements are modified by the different sizes and formations established by the rapidly developed DG [2]. Its integrity result from the challenges provided from the technical and the economic perspectives. For ensuring the DN performance, optimal planning is pointing out the supreme importance. This DG encountered the stability of voltage, reliability, reduction of power loss, profitability, and the quality of power [3]. The losses can be minimized by installing the DG units at the proper positions. In remote areas, distributed generation plants such as wind turbines and photovoltaic (PV) energy are placed. The operating system required thorough integration into the distribution network and transmission network. This DG's motive is to reduce the cost, greenhouse gas emission, and losses by integrating all types of generation plants. To improve the stability, voltage, and power factor, DG is used. Thus, in DN, the DG sizing and its locations are playing a significant role. The DG means the electricity generation from the small-scale region to load centers. When the network had an expansion like the Distributed Generation, it took care of the areas of the new loads. Solar and wind are the most predominant renewable DG, which is eco-friendly in nature.

Depending upon the generation of the power, the DG is classified as,

- Type 1: the system injected only the actual power
- Type 2: the system had both the reactive and the real power
- Type 3: the system injected only the reactive power.
- Type 4: the system injected the real power but also absorbed the power, which is reactive.

The allocation of the optimal DG is an essential job in the whole system. If the DG is installed in a non-optimal location, it provides high system loss and voltage instability [4]. Finally, the results were provided the high operating cost. To increase efficiency and therefore get higher benefits, there would be an optimal allocation of DG to be obtained. Many research works are entitled to finding the optimal location and sizing the DGs. The Resistance/Reactance ratio (R/X) must be high in the Radial Distribution Network (RDN). Most published articles are found to work based on reduction in power loss during the location and sizing of Distributed Generations [5] for standard test systems of IEEE. This article proposed a new, improved Flower Pollination Algorithm for the allocation and sizing of DG. This research is evaluated with a real-time system such as the practical Indian 52 bus system.

II. LITERATURE REVIEW

The optimal operations of the Distributed Generations have been rendered by Ant Colony Search Algorithm (ACSA) method. The problem of placement of the optimal capacitor and the optimal feeder's reconfiguration is solved using the ACSA method [6]. The Cloud Mutation Flower Pollination algorithm is proposed for the problem of continuous optimization. In the optimization of the global zone, it added information in all the dimensions [7]. For the multi-objective findings like improvement in voltage profile, minimization in power loss, and operating cost, the Whale Optimization Algorithm (WOA) is proposed. It is evaluated in the 32-bus and 69-bus systems [8] [9] [10]. The Distributed Power Generation is used for the conversion of the power and the technology control. Here, the impacts are examined, and the connection strategies are enhanced [11]. The various DG performances are analyzed, and the advantages of those DG are increased [12]. The Combined Loss Sensitivity is proposed to find the location and size [13]—the Fuzzy-FPA (FFPA), a hybrid phase swapping algorithm. The deviation of the phase current is minimized here. Optimization of the fitness value is calculated [14]. The particle swarm optimization strategy is proposed with the hybrid genetic algorithm. For DG allocation, this algorithm is presented. It includes position, capacity, and number. Also, it identifies the required total power [15]. The main objective is to minimize the losses. To reduce the power losses, the DG addition is very effective [16]. To achieve less oscillation and maximum power, the conventional method is used [17]. The Flower Pollination Algorithm is proposed for the modeling and simulation of the MPPT SEPIC-BUCK converter series. For quickly finding the maximum power

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

point, this technique is proposed [18]. For the DG allocation, the Flower Pollination Algorithm is proposed, and the minimization of the total power loss was done. Hence the voltage of the system bus is improved. It is done with the help of MATLAB software [19]. The modified Flower Pollination Algorithm is used for the allocation and minimization of losses. By the DER unit allocation, the MFPA validation has been done [20]. The Chaos-FPA (CFPA) has been entitled to the DG placement problem. The equations of the Gauss map chaotic variable have been altered. The main aim is to increase the aggregator profit [21]. A non-linear nature of generation has been handled with the secured operation. The adaptability could be gained through chaotic variables [22]. The DG capacity is used for solving the problem of congestion management. The normalized weights have been used for converting the function of multi-objective into a single objective problem [23]. To control the speed of the BLDC motor, the FPA is proposed. The comparison has been made with the Ziegler-Nicholas method's help, which is conventional [24]. To reduce power loss and capacity of optimal DG, the method of FPA is being used. The photovoltaic DG is implemented [25]. For the dispatching of emission and the optimal load, the FPA and the hybrid Fuzzy were proposed. For optimum levels with the emission and economical combination, the FFPA method has been utilized [26]. In RDNs, the allocation of DG is solved with a chaotic symbiotic organism search algorithm. Also, Moth Flame Optimization is used for the allocation of PV-DG [27] [28].

The Improved single and Multi Harris Hawks Optimization Algorithms have been proposed to allocate and size DG. The main aim is to reduce the voltage deviation, total power loss and increase the voltage stability index [29] [30]. The comparison of Perturb & Observe, Modified Particle Swam Optimization, and FPA is made. For the reduction of the ripple current, the SEPIC converter is presented [31]. The allocation of the optimal capacitor is brought out with the help of FPA. The load flow analysis gives the power loss and flows with data structures' help [32]. Artificial Intelligence techniques have dominated the Distributed Energy Generation. To maximize the optimum operation schedule and reliability is the main objective, and to improve the operational efficiency, developments are being done [33].

III. PROBLEM FORMULATION

To allocate the proper Distributed Generations, more concern needs to be taken in the DN. The power loss minimization with optimal solar and wind DG allocation in RDN is the main objective in this paper. The various constraints in the distributed network need to satisfy the objective functions.

A. Objective Function

a) Minimization of Power Loss: When solar and wind DGs of the RDN are considered, the primary objective is its optimal location through power loss minimization. It is given in Eqn 1.

$$f_{obj}(X) = \sum_{k=1}^{nb} |I_K|^2 R_K \qquad (1)$$

Where, R_k, I_k are the resistance, k^{th} branch of the current magnitude, and the number of branches is denoted by *nb*.

b) Constraints: The various constraints in the DN need to satisfy the objective function given in Eqn 2– 5.

Power Balance

$$P_{slack} + \sum_{k=1}^{N_{SDG}} P_{SDG,k} = \sum_{k=1}^{N_L} P_{D,k} + \sum_{k=1}^{nb} P_{loss}(k) \qquad (2)$$

Voltage Limits

$$|V_{min}| \leq |V_k| \leq |V_{max}| \qquad (3)$$

DG sizing Limits

$$P_{SDG,min} \leq P_{SDG,k} \leq P_{SDG,max} \qquad (4)$$

Thermal Limits

$$I_{jk} \leq I_{jk}^{max} \qquad (5)$$

Where,

P_{slack} is used to indicate slack bus power
P_{SDG} is used to indicate DG power
P_D *is* used to indicate power demand
P_{loss} is used to indicate actual power loss
V_{min} and V_{max} Are used to indicate the minimum and maximum voltages, respectively.
$P_{SDG,min}$ and $P_{SDG,max}$ are used to indicate the minimum and maximum DG power
I_{jk}^{max} is used to indicate maximum current
j and *k* are used to denote branches
N_{SDG} *is* used to denote the number of DGs
N_L *is* used to denote the number of lines

IV. IMPROVED FLOWER POLLINATION ALGORITHM

Flower Pollination Algorithm is the concept from the flowering process of pollination [18]. The reproduction process of flowers does pollination. The pollinators are transferred from the pollens. This is classified as the abiotic and the biotic process. Based on the process of the execution, it is divided into cross-pollination and self-pollination. Some of the steps are involving in the Flower Pollination Algorithm. They are,

Step 1: Global pollination is called biotic or cross-pollination due to the pollinator's movement. It carries the pollen which following the Levy flight.

Step 2: The local pollination of the self or abiotic pollination is called an absence of pollination.

Step 3: There is a similarity in the reproduction probability of the associated two flowers.

Step 4: The switching probability's critical factor is p [0,1], which is an interaction between the local and global pollination.

In terms of an equation, steps 1 to 4 are updated. The pollinators are transferred from the gametes of flower pollen when the global pollination. Thus the pollen covers longer distances [32]. Hence the global pollination and the representation of constancy of the flower are given in Eqn 6.

$$x_i^{t+1} = x_i^t + L(x_i^t - g_*) \qquad (6)$$

Where the solution vector is denoted as x_i^t and the best current solution is denoted as g_* . L from Levy distribution represents step size, and the random number is denoted as GS.

$$\text{L}=\frac{\lambda\, \Gamma(\lambda) \sin(\frac{\pi\lambda}{2})}{\pi} \frac{1}{s^{\lambda+1}}\ , \quad (\text{s} > s_0 > 0) \qquad (7)$$

Hence, the standard gamma function is denoted as $\Gamma(\lambda)$ Where s>0.

From rule 2, the flower's constancy is expressed using the golden section ratio [34] in Eqn 8.

$$x_i^{t+1} = x_i^t + GS(x_j^t - x_k^t) \qquad (8)$$

Where the species of plants from the different flower pollens are denoted as x_j^t, x_j^t

Here, The Local pollination is rendered by golden section ratio. The technique of the IFPA has solved the problem of the power system [25]. In the identified allocation, the IFPA helps identify the placed solar and wind DGs. IFPA is easy for balancing because the parameters are significantly less controllable between the exploitation and exploration ability while optimization occurs.

A. Steps for solving the optimal location of solar and wind DGs with the help of Improved Flower Pollination Algorithm

1. To read the test system data
2. To initialize the IFPA parameters such as switch probability *p*=0.8, Iterations=150.
3. To identify the correct allocation of the DG placement.

4. To identify the best allocations, which is a Flower Pollination Algorithm input.
5. No DGs should be placed if rand>p in the installed locations
6. The concept of local pollination through golden search is adopted to identify the placed DGs.
7. To update the current best global solution.
8. To display the solution results. Stop the process if the values are correct; otherwise, repeat the process from step 4.

V. SOLUTION METHODOLOGY

The proposed system is evaluated in the practical Indian 52-bus system [35]. It is portrayed as RDN. There are three types of feeders used in the test system. This research aims to minimize the deviation of the total network loss [5]. There is a possibility of allocating two DG, which are considered here. In the simulation process, it is considered that Type-2 DGs are analyzed that would have both the reactive and the real power. The Indian practical 52-bus radial distribution system portrayed is considered for evaluation of IFPA. This test system consists of 3 main feeders, 51 branches, and 52 bus. It feeds a total network load of the absolute power of 4184 kW and reactive power of 2025 kVAr. The power factor of the connected load of this test system is 0.9 lagging. The base kV and MVA for this considered test system are considered as 11 and 1, respectively. The maximum and minimum bus voltage magnitude limits for the given test system are 1.05 p.u and 0.9 p.u., respectively. The vital objective of this work is the minimization of network loss and total load voltage deviation. The maximum number of DG units installed for this system is considered equal to the number of feeders.

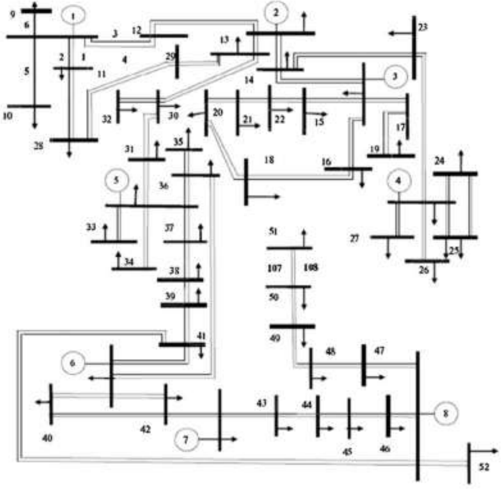


Fig. 1: Indian-52 Bus System

VI. SIMULATION RESULTS

There is a comparison obtained between the Flower Pollination Algorithm (FPA), Chaotic Flower Pollination Algorithm (CFPA), and the Improved Flower Pollination Algorithm (IFPA) are shown in Table 1. The Distributed Generation location and sizing are calculated for each algorithm [13]. The overall power factor and the loss reduction percentage are also calculated for each algorithm [14]. From Table 1, we can observe that the percentage of the loss reduction is higher in the Improved Flower Pollination Algorithm. But the power loss is very high while using the Flower Pollination Algorithm compared to the other two algorithms.

The Improved Flower Pollination Algorithm (IFPA) convergence characteristics are shown in Figure 2. This figure shows the best cost at various iterations. In this graph, we can clearly understand that while increasing the number of iterations, the best cost is attained. Here, the cost tends to our objective (i.e.,) Power loss.

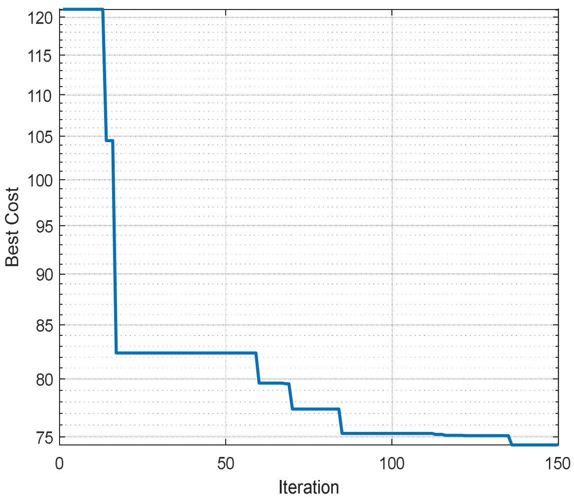


Fig. 2: IFPA Convergence Characteristics

In figure 3, the Power loss profile is obtained. This graph shows the power loss that occurred at various buses due to DG sizing and location. And it is calculated for 52 bus systems. At the point of 52 buses, the power loss is found to be decreased.

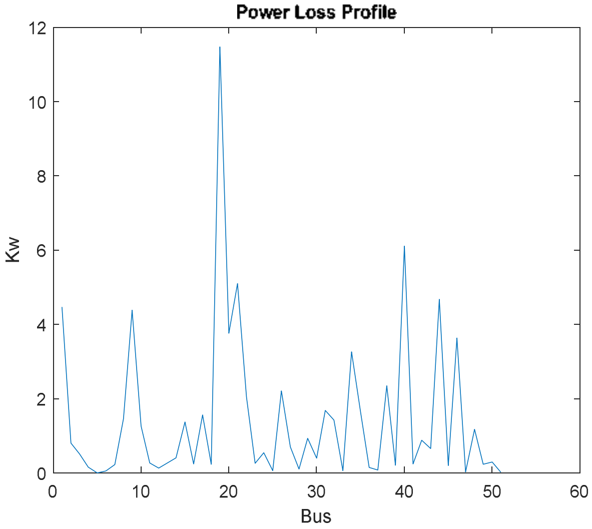


Fig. 3: Power Loss Profile at each bus in Indian-52 bus system

From figure 4, we can observe the voltage profile characteristics for the 52 bus system. And there, we can observe that a little much improvement in the voltage profile is obtained in the Indian - 52 bus system.

TABLE I. Analysis of Practical Indian 52 bus system

	Ploss (kW)	DG location	DG size (kW)	DG size (kVAr)	Loss reduction%
FPA[30]	89	10,47	922, 1102	478,647	79
CFPA[21]	84	11, 41	686, 1258	509 651	81
IFPA	79	41, 16	1261 ,968	683 362	82

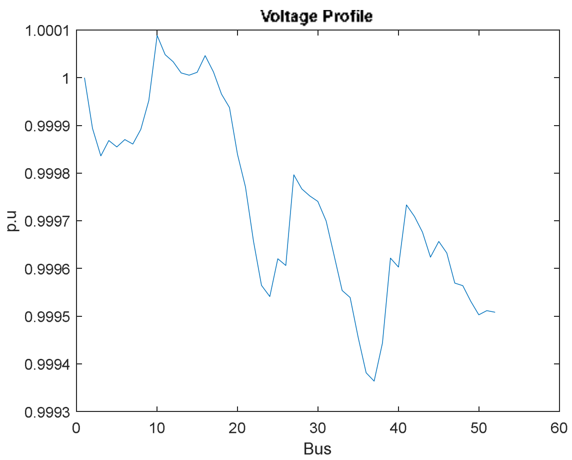


Fig. 4: Voltage Profile of Indian-52 bus system

VII. CONCLUSION

In this research, the practical Indian 52 bus, which belongs to the RDN problem, has been solved to allocate the optimal DG. Its location and sizing are determined using the Flower Pollination Algorithm (FPA) and chaotic FPA. Different type of heuristic algorithm like Improved Flower Pollination Algorithm (IFPA) is proposed in this work. From the above results, it is concluded that the power loss is minimized, and there is an improvement in the voltage profile factor. Hence, the efficient outcome is showcased with the practical Indian 52 bus system.

ACKNOWLEDGMENT

The authors are very grateful to the VIT University-Chennai campus authorities for carrying out this research work with extensive facilities.

REFERENCES

[1] Abdel-mawgoud, H., Kamel, S., Ebeed, M., & Aly, M. M., "*An efficient hybrid approach for optimal DG allocation in radial distribution networks,"* 2018 International Conference on Innovative Trends in Computer Engineering ITCE, (2018), doi:10.1109/itce.2018.8316643.

[2] Ali, A. H., Youssef, A.-R., George, T., & Kamel, S., "*Optimal DG allocation in distribution systems using Ant lion optimizer,"* International Conference on Innovative Trends in Computer Engineering *(ITCE) (2018),* doi:10.1109/itce.2018.8316645.

[3] Ankita, & Prakash, P., " Multi Sensitivity Factors and FPA for Loss Reduction and Voltage Profile Improvement by Optimal Placement and Sizing of Capacitor in Mesh Network," 2018 2nd International Conference on Power, Energy and Environment: Towards Smart Technology (ICEPE) (2018), doi:10.1109/epetsg.2018.8659003.

[4] Alam, A., Gupta, A., Bindal, P., Siddiqui, A., & Zaid, M., "*Power Loss Minimization in a Radial Distribution System with Distributed Generation,"* 2018 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) (2018), doi:10.1109/icpects.2018.8521619.

[5] Bhat, M. V., & Manjappa, N., "*Flower Pollination Algorithm Based Sizing and Placement of DG and D-STATCOM Simultaneously in Radial Distribution Systems*," 2018 20th National Power Systems Conference (NPSC) (2018), doi:10.1109/npsc.2018.8771803.

[6] C. F. Chang, "Reconfiguration and capacitor placement for loss reduction of distribution systems by ant colony search algorithm," IEEE Trans. Power Syst., (2018) vol. 23, no. 4, pp. 1747–1755.

[7] Chen, Y., & Pi, D., "*An innovative flower pollination algorithm for the continuous optimization problem,"* Applied Mathematical Modelling (2020) pp. 237-265, doi:10.1016/j.apm.2020.02.023.

[8] D. B. Prakash and C. Lakshminarayana, "Multiple DG placements in the radial distribution system for multi objectives using Whale Optimization Algorithm," Alexandria Eng. J., (2018), vol. 57, no. 4, pp. 2797–2806.

[9] Ezzat, M., "Reliability Enhancement in Distribution Systems Through Distributed Generator and Capacitor Allocation Using Flower Pollination Algorithm," 2019 21st International Middle East Power Systems Conference (MEPCON), (2019), doi:10.1109/mepcon47431.2019.9008225.

[10] Prakash, D. B., & Lakshminarayana, C., *"Multiple DG placements in the radial distribution system for multi objectives using Whale Optimization Algorithm"* Alexandria Engineering Journal (2018) doi:10.1016/j.aej.2017.11.003

[11] F. Blaabjerg, Y. Yang, D. Yang, and X. Wang, "*Distributed Power-Generation Systems and Protection*," Proc. IEEE, (2017), vol. 105, no. 7, pp. 1311–1331.

[12] G. Sabarinath, T.G, Manohar, "Optimal Sitting And Sizing Of Renewable Energy Resources For Power Loss Reduction In Radial Distribution Systems Using Whale Optimization Algorithm," 2018 Int. Conf. Emerg. Trends Innov. Eng. Technol. Res., (2018), pp. 1–5.

[13] G. Sabarinath, T.G, Manohar, "*Optimal Placement and Sizing of Distributed Generation using Flower Pollination Algorithm for Power Loss Reduction Maximization in Distribution Networks*," 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES) (2018), pp. 256-256, Doi: 10.1109/ICPEICES.2018.8897436.

[14] Mahendran, G., & Govindaraju, C., "*Flower Pollination Algorithm for Distribution System Phase Balancing Considering Variable Demand,"* Microprocessors and Microsystems, 103008, (2020), doi:10.1016/j.micpro.2020.103008.

[15] Mahmoud Pesaran, H. A., Nazari-Heris, M., Mohammadi-Ivatloo, B., & Seyedi, H., "*A hybrid genetic particle swarm optimization for distributed generation allocation in power distribution networks,"* Energy, 118218 (2020), doi:10.1016/j.energy.2020.118218.

[16] Manimegalai, R., Visalakshi, S., & Devi, S. L., "*Optimally locating microgrid for the minimization of losses,"* 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM) (2017), doi:10.1109/iconstem.2017.8261379.

[17] Murdianto, F. D., Nansur, A. R., & Hermawan, A. S. L., "Modeling and simulation of MPPT coupled inductor sepia converter using Flower Pollination Algorithm (FPA) Method in DC microgrid system" 2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA) (2017), doi:10.1109/elecsym.2017.8240364.

[18] Murdianto, F. D., Nansur, A. R., Hermawan, A. S. L., Purwanto, E., Jaya, A., & Rifadil, M. M., "Modeling and Simulation of MPPT SEPIC - BUCK Converter Series Using Flower Pollination Algorithm (FPA) - PI Controller in DC Microgrid Isolated System" 2018 International Electrical Engineering Congress (iEECON) (2018), doi:10.1109/ieecon.2018.8712290.

[19] Oda, E. S., Abdelsalam, A. A., Abdel-Wahab, M. N., & El-Saadawi, M. M., "*Distributed generations planning using flower pollination algorithm for enhancing distribution system voltage stability*," Ain Shams Engineering Journal, 8(4), 593–603 (2017), doi:10.1016/j.asej.2015.12.001.

[20] Oda, E. S., & Abdelsalam, A. A., "*Optimal DGs allocation in distribution networks using modified flower pollination algorithm*" 2017 Nineteenth International Middle East Power Systems Conference (MEPCON) (2017), doi:10.1109/mepcon.2017.8301370.

[21] Pandya, K. S., & Joshi, S. K., "*CHAOS enhanced Flower Pollination Algorithm for Optimal Scheduling of Distributed Energy Resources in Smart Grid*" 2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia) (2018), doi:10.1109/isgt-asia.2018.8467806.

[22] Rajagopalan A, Kasinathan P, Nagarajan K, Ramachandaramurthy VK, Sengoden V, Alavandar S. Chaotic self-adaptive interior search algorithm to solve combined economic emission dispatch problems with security constraints. Int Trans Electr Energy Syst. 2019;29: e12026. https://doi.org/10.1002/2050- 7038.12026

[23] Peesapati, R., Yadav, V. K., & Kumar, N., "Flower pollination algorithm-based multi-objective congestion management considering optimal capacities of distributed generations," Energy, 147, pp. 980–994 (2018), doi:10.1016/j.energy.2018.01.077.

[24] Potnuru, D., Alice Mary, K., & Sai Babu, C., "*Experimental implementation of Flower Pollination Algorithm for speed controller of a BLDC motor,"* Ain Shams Engineering Journal (2019), doi:10.1016/j.asej.2018.07.005

[25] Prasetyo, T., Sarjiya, S., & Putranto, L. M., "Optimal Sizing and Siting of PV-Based Distributed Generation for Losses Minimization of Distribution using Flower Pollination Algorithm" 2019 International Conference on Information and Communications Technology (ICOIACT) (2019), doi:10.1109/icoiact46704.2019.8938424.

[26] S, T., Kim, H.-J., & Ra, I.-H., "Hybrid Fuzzy and Flower Pollination Optimization Algorithm for Optimal Dispatch of Generating Units in the Existence of Electric Vehicles," 2020 6th International Conference

on Advanced Computing and Communication Systems (ICACCS) (2020), doi:10.1109/icaccs48705.2020.9074459

[27] S. Saha, "A novel multi-objective chaotic symbiotic organisms search algorithm to solve optimal DG allocation problem in the radial distribution system," no. January (2019), pp. 1–25.

[28] Seoul, S., Chenni, R., Hasan, H. A., Zellagui, M., & Kraimia, M. N., "MFO Algorithm for Optimal Location and Sizing of Multiple Photovoltaic Distributed Generations Units for Loss Reduction in Distribution Systems," 2019 7th International Renewable and Sustainable Energy Conference (IRSEC) (2019) doi:10.1109/irsec48032.2019.9078241.

[29] Selim, A., Kamel, S., Alghamdi, A. S., and Jurado, F., " Optimal Placement of DGs in Distribution System Using an Improved Harris Hawks Optimizer Based on Single- and Multi-Objective Approaches," IEEE Access, (2020), 8, pp. 52815–52829. doi:10.1109/access.2020.2980245.

[30] Sudabattula, S. K., Suresh, V., Subramaniam, U., Almakhles, D., Padmanaban, S., Leonowicz, Z., & Iqbal, A., "Optimal Allocation of Multiple Distributed Generators And Shunt Capacitors In Distribution System Using Flower Pollination Algorithm," 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe) (2019) doi:10.1109/eeeic.2019.8783417

[31] Suyanto, S., Mohammad, L., Setiadi, I. C., & Roekmono, R., "Analysis and Evaluation Performance of MPPT Algorithms: Perturb & Observe (P&O), Firefly, and Flower Pollination (FPA) in Smart Microgrid Solar Panel Systems" 2019 International Conference on Technologies and Policies in Electric Power & Energy 2019 doi:10.1109/ieeeconf48524.2019.9102532

[32] Tamilselvan, V., Jayabarathi, T., Raghunathan, T., & Yang, X.-S., "*Optimal capacitor placement in radial distribution systems using flower pollination algorithm"* Alexandria Engineering Journal. (2018) doi:10.1016/j.aej.2018.01.004

[33] Twaha, S., & Ramli, M. A. M., "A review of optimization approaches for hybrid distributed energy generation systems: Off-grid and grid-connected systems," Sustainable Cities and Society (2018) 41, pp.320–331. doi:10.1016/j.scs.2018.05.02

[34] J. Alikhani Koupaei , S.M.M. Hosseini and F.M. Maalek Ghaini. A new optimization algorithm based on chaotic maps and golden section search method, Engineering Applications of Artificial Intelligence, 50(2016)201–214, 10.1016/j.engappai.2016.01.034

[35] Sabarinath.G and T.Gowri Manohar. Application of Bird Swarm Algorithm for Allocation of Distributed Generation in an Indian Practical Distribution Network, *IJ Intelligent Systems, and Applications,* 7, 54-61. (2019), DOI: 10.5815/ijisa.2019.07.06

Data Storage in Blockchain Based Architectures for Internet of Things (IoT)

Munavwar Shaikh, Charles Shibu, Enrico Angeles, Deepa Pavithran
Information Security Engineering Dept, Abu Dhabi Polytechnic, *P.O. Box 111499*,
Abu Dhabi, United Arab Emirates.
{munavwar.shaikh, charles.gabriel, enrico.angeles,deepa.pavithran}@adpoly.ac.ae

***Abstract*— Data storage in IoT systems describe how data is pushed from the sensor and where it will be stored. In Bitcoin, there is a limit to the number of transactions that can be stored in the block, similarly in Ethereum it is limited by the gas limit in block. An IoT infrastructure have several applications that includes a sensor collecting data from physical environment. Storing a large amount data in a blockchain faces several challenges due to the distributed nature of the blockchain, high transaction processing time, and reduced scalability. Hence it is crucial to identify what data should be stored and how to store it in a secure way. In this paper, we provide how data is being stored in various blockchain based IoT applications and provides data storage compliance in treating IoT data in a blockchain environment.**

Keywords—blockchain, Internet of things, data storage, sensors, data compliance, cloud.

I. Introduction

Blockchain is a recent technology that can provide data security and privacy in IoT Systems. It uses a distributed ledger design that stores data in multiple nodes within the network. In a decentralized system, computing and storage tasks are redundant, which means every node in the decentralized network must store a complete copy of all data and perform the same operations of data. This is how blockchain maintains trust among nodes and provide security to the data. However, when the number of participating nodes increases or the amount of data to be stored is high, it will not be practical to replicate the entire data into all participating nodes and to perform concurrent computations on it.

The existing internet of things systems uses either offline or cloud-based information-sharing technology. It has several shortcomings including the high maintenance cost, low efficiency, and lack of an effective mechanism to ensure IoT equipment Identity, information validity, authenticity, consistency, and integrity of information in different systems [1].

Blockchain systems do not rely on trusted third parties. All nodes make decisions together to verify the legitimacy of the transaction. Even if some nodes are attacked or destroyed, it will not cause damage to the entire blockchain system.

II. Background

A. Blockchain

Blockchain is a distributed ledger where blocks are connected with cryptographic hashes and are distributed over all the nodes in the system. Generally, there are two types of blockchain: public and permissioned. In a public blockchain, anyone can join the network or participate in the consensus mechanism. An example of a public blockchain is Bitcoin. In a permissioned or private blockchain, only certain entities will have access to participate in the blockchain process. This is typically used in applications that involve multiple entities in a restricted way. It will have a certain control on who can access the data and who can participate in the consensus. The consensus is an agreement on how and what data are added to the blockchain.

B. Internet of Things

As defined in [2] Internet Of Things (IoT) is an emerging global Internet-based information architecture facilitating the exchange of goods and services in the global supply-chain network. The IoT data comes from a large number of devices generating billions of data objects. The IoT network has to collaborate with different devices to sample, process, and make this data useful for analytical and decision making. Realizing the full potential of IoT has not yet been achieved. This is due to the lack of standards, heterogeneous nature of devices, diversified communication protocol, and security [3].

III. Data Storage in Blockchain Based IoT Usecases

A majority of blockchain-based IoT Application is in the field of healthcare, Smart home, smart city, Industrial IoT and agricultural applications.

A. IoT Devices employing blockchain

The blockchain-based design for IoT data storage takes a distributed access control and data management approach. In this scenario, data ownership is transferred to users instead of following the traditional centralized access control system based on trust modeling. With the use of blockchain-based functions, secure and fail-safe data management with a verifiable and distributed access control layer on the storage layer is practiced. The storage of time-critical IoT data is facilitated at the edge of the network via a location-based decentralized storage system, which in turn is managed with blockchain technology [4]. The parallelism of large storage systems is suggested by Quanqing Xu[5] in order to shorten the execution time of many basic data analysis tasks, whereby blockchain can be used as intelligent contracts to facilitate the negotiation of a contract in the IoT and also to enforce it. OSD-based smart contract (OSC) approach is used in which IoT devices interact with such blockchains. For data analysis applications, the IoT device processors perform application-specific operations. That way, only the results are returned to the clients, rather than the data files they read [5].

B. Healthcare

Since the advent of smart IoT devices in the Healthcare industry, promising technologies such as cloud computing, ambient assisted living, big data, and wearables are being applied in various platforms within Healthcare [6]. E-health regulations related to data and policies worldwide to determine how they assist the sustainable development of IoT and cloud computing in the healthcare industry evolved as well with the thriving data in all these platforms [6]. The Medical data of a patient is treated with utmost care and secrecy. Hence the IoT devices in healthcare systems are devised to ensure the security of patients' healthcare data, realize access control for normal and emergency scenarios, and support smart deduplication to save the storage space in big data storage system. The medical files generated by the healthcare IoT network are usually encrypted and transferred to the storage system, which can be securely shared among the healthcare staff from different medical domains leveraging a cross-domain access control policy. A smart secure deduplication method could be followed as proposed by Tang et al. [7] to ensure the medical files or data can be accessed by all the data users after deduplication and authorized by the different original access policies.

C. Industrial IoT(IIoT)

The main goal of industrial IoT is to improve operational efficiency, increase the scale of production and better management of industrial components and processes through product customization, condition of machines getting monitored in an intelligent way, smart monitoring applications for production workshops as well as predictive and preventive maintenance of industrial equipment [8].

In industrial usage of IoT devices, storage and retrieval are done by integrating fog computing and cloud computing [9] and a flexible and economic framework for data processing was made by the authors. The data in these devices are preprocessed in the edge server with the addition of timestamps and stored locally. The remaining data gets sent to the external storage medium and away from the local storage for information retrieval and data mining [9]. This set of data is not time-sensitive.

D. Smartcity/Smart Home

Smartphone applications are used to control and monitor home functions using wireless communication technologies. Domb in [10] examines the concept of the smart home and the data associated with it, incorporating IoT services and cloud computing It also discusses embedding intelligence in sensors and actuators, networking intelligent devices using the appropriate technology, saving storage space and improving data exchange efficiency, and facilitating data interaction with intelligent things using cloud computing and storage for easy access to various locations. A precise, fast, open and shared information system is the basis for Smart City applications. In view of this massive, distributed, heterogeneous and complex state data, the storage and management of traditional data will encounter great difficulties. The traditional infrastructure uses a centralized approach, an expensive large server, hard disk storage hardware, and a relational database management system. This leads to poor scalability of the system, higher costs and, for the most part, difficult adaptation to the demand for higher reliability of real-time status data from smart city applications [11].

E. Agricultural Applications

Some IoT-based smart farming surveillance systems use a two-tier approach to data storage to store and secure large amounts of data from any IoT device. Tier-1 focuses on collecting data from various sensors and storing them locally using the SD card. Tier-2 uses a cloud server to store the large volume of IoT sensor data [12]. Gill et al. [13] explains the benefits of a cloud-based autonomous information system for the delivery of Agriculture-as-a-Service (AaaS) using cloud and big data technologies. Accordingly, the Aaas system collects information from various users via pre-configured devices and IoT sensors and processes it in the cloud with the help of big data analysis and provides the users with the necessary information automatically.

IV. DATA STORAGE ARCHITECTURES IN IOT BLOCKCHAIN

A. Storage in cloud

Permissioned blockchains use cloud servers to store encrypted data blocks. All transactions in a different block and create a combined hash of each block using Merkle Tree and transfer it to the distributed network. This way, any changes in cloud data can be easily detectable. Doing the storage in this manner also preserves decentralization to some extent [14]. Fig.1 shows blockchain based IoT architecture where a cloud storage is being used to store the IoT Data. Cloud storage uses users' data in an identical group. The Hash of the data stored in the cloud is sent to the overlay network. The hierarchical storage structure in storing the majority of the blockchain in the clouds helps faster upload of one day’s blockchain to the cloud, it maintains the most recently created blocks in a blockchain overlay network. The blockchain connector and the cloud connector are the two software interfaces that are defined to create the blocks and coordinate to the clouds but one disadvantage of this method is a need to be implemented in a more real IoT application to find out if this technology is reliable [15].

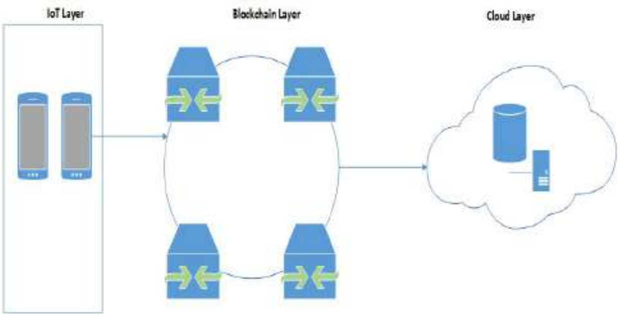


Fig:1 Storage in cloud

B. Storage in Decentralized way

This is a fully decentralized way of storing data. Liu et al. [16] integrates decentralized blockchain network and distributed storage network. Data produced by the IoT device is stored in the distributed storage nodes in the peer-to-peer network, whereas the reference to data that serve as the identifier is stored in the Blockchain. Blockchain only stores the digest of data but not data itself. Hence the amount of data on the blockchain is greatly reduced. The use of blockchain technology in storing data and with the use of management hub including the integration of IoT and decentralized access control to the blockchain solves the issues in managing several controlled IoT devices, it was able to cope up with different IoT scenarios but there are possibilities of threat and vulnerabilities in the IoT since it is no longer part of the blockchain technology [17]. Fig.2 shows a blockchain architecture that uses decentralized storage for storing IoT data. Decentralized cloud architecture was used in which small-scale data centers meet low latency and high bandwidth because these data centers are located closer to the users. The locality-aware data storage and the processing development provide its full potential with the decentralized access control layer. Data streams that are chucked at pre-defined lengths show reasonable results but this one is still in the initial stage and needs improvement[4]. Sapphire system which is a large-scale blockchain storage system used for data analytics in the IoT it uses an object-based storage interface that provides richer semantic information for the stored object to optimize its performance more effectively than other storage systems[5]. A secure structure for IoT data storage and protection which is based on blockchain technology incorporating edge computing helps manage data storage and assist small IoT devices to accomplish computations. The authentication system adapted the certificate-less cryptography, but it needs to improve the authentication scheme for the blockchain-based systems[18]. The integration of IoT, blockchain, and cloud technologies allows observation of vital signs of the patient in a protected approach. The storage and access of data used in the blockchain technology data sharing significantly increase overall throughput but need to be tested using various IoT frameworks[19]. Blockchain technology solves the problem of IoT information sharing security but performance needs to be enhanced when applied to a specific industry [1].

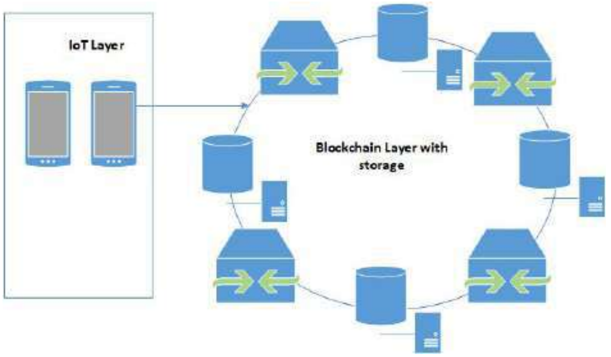


Fig:2 Storage in Decentralized way

C. Storage in Hybrid way

In this method data will be stored in a hybrid way, where processed/ or raw data from the sensor is stored in a cloud server and a reference to the data will be stored in the blockchain. Si et al. [1] propose a double-chain model combining data blockchain and transaction blockchain. IoT data is divided into lightweight data and multimedia data, where multimedia data is compressed and integrated to reduce data capacity and to improve data quality. The processed data is divided into account book data and outsourced storage data (multimedia data) stored in a fog node that can be easily downloadable. Fig.3 shows blockchain architecture that uses hybrid storage. A comparison of various storage options in IoT based blockchain architectures is provided in Table1.

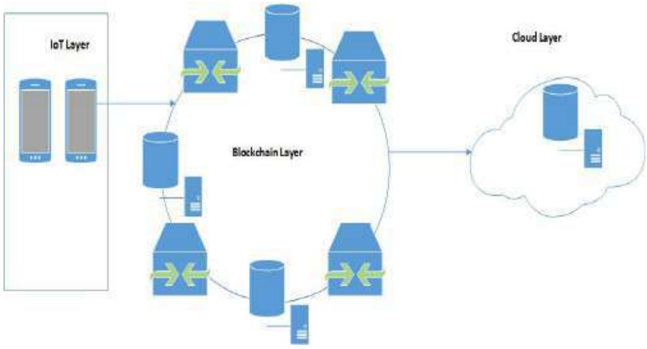


Fig:3 Storage in a Hybrid Way

Table 1: Comparison of various data storage options in IoT Blockchain

Type	Storage Design	Advantages	Disadvantages	Reference paper
Cloud	Hierarchical storage structure	• Faster upload of one-day's blockchain to the cloud	• Not yet implemented in a more real IoT application	[15]
Decentralized	Blockchain system with Management Hub	• It performs best with the use of management hub node	• It might face different threat and vulnerabilities since IoT is no longer part of the blockchain technology	[20]
Decentralized	Data streams are chucked at pre-defined lengths	• Initial results show reasonable overhead	• Still at the initial stage and needs for improvement	[4]
Decentralized	Object-based Storage interface	• Has richer semantic information for the stored object to optimize its performance more effectively than other		[5]

		storage systems		
Decentralized	Blockchain system combined with Edge Computing	• Secure scheme for IoT data storage and protection	• Need to improve authentication scheme for blockchain-based system	[18]
Decentralized	Data Sharing using Blockchain technology	• The overall throughput increased significantly	• Need to be tested using various IoT frameworks	[19]
Decentralized	Blockchain technology	• Solve the problem of IoT information sharing security	• Performance needs to be enhanced when applied to a specific industry.	[1]

V. Compliance on Data Storage

To start with knowing the objectives of the organization. This interprets to knowing which regulations will apply to the organization directly related to the regional legal and regulatory requirement and which types of data are expected to manage or not, how long to retain it and how to protect it. In the organization the compliance officers can be the ones to provide information.

Organizations need to focus on data classification and data mapping. Data classification and data mapping are crucial in discovering the types of information that are being held in storage systems and how they are being moved across the network. Not only are they essential factors in determining how regulated information is stored, but also a solid step in establishing compliant policies.

Continual monitoring: Storage compliance is not a set-and-forget affair. Continual monitoring is key in ensuring that regulated data is properly cared for during its lifecycle. Procedures must be in place to make sure this monitoring happens regularly.

More than security: Security and storage often go hand-in-hand, but storage compliance takes it to another level. Encryption will help, and many storage solutions feature support for the security-enhancing technology, giving both storage and security professionals one less thing to worry about.

Critical: Testing and audits will ensure compliance with policies and IT mechanisms are up to the task. It's best to work out the kinks now before having to explain to investigators why sought-after emails or transaction records have gone missing. There are many requirements based on the type of information and data the organization has. Some might involve using what is called DAR (Data Encryption at Rest), which encrypts the storage device so that if removed from the system, the data is nearly or totally impossible to access (the degree of difficulty depends on the encryption algorithm and the size, complexity and entropy of the key or keys for the device).

Understanding what is required from a governance point of view for the organizational data or the resulting information is based on things like best practices for the industry or regulations and agencies like the U.S. National Bureau of Standards (NIST), ISO, HIPAA, SEC, GDPR in Europe. And the resulting architectural or procedural changes are the types of things that will be needed to address as part of the architecture. Compliance is not easy, nor is it free. The cost depends on lots of factors but trying to force compliance after the architecture is planned and built is always far most costly than doing it beforehand.

Decentralization: The main benefit of using blockchain would be that no single authority would have control over the data generated by the IoT devices. That way, a distributed peer-to-peer network is born that permits the parties that don't know or trust each other to collaborate more smoothly. This type of network will also make it possible to unify IoT devices and streamline the distribution of updates throughout the network.

Security: The current security architecture of IoT has its shortcomings. When the data is managed by a central authority, the system is more susceptible to a single point of failure. Blockchain's unique security protocol normally described as transparent and immutable is a good solution to the largest issue of IoT development. Blockchains will store unalterable data history that can be consulted for each unique address. This lays the foundations of a platform that provides improved identification and authentication in IoT. The robust level of encryption that blockchain guarantees won't let the hackers overwrite data records.

Transparency: Anyone with authorization could track the transactions made on the network to follow up on what has happened in the past. This feature is useful to identify any leakage and take action.

Autonomy: Blockchain will reinforce the machine-to-machine economy that IoT is based on by offering a safe way to store information on different transactions. That way micropayments for services and data can be processed in a straightforward way. IoT devices that rely on blockchain can execute digital agreements automatically when the terms are met. Automating transactions between devices improves machine-to-machine communication.

Reduced costs: Integrating blockchain to the organizational processes would allow IoT companies to reduce costs. Eliminating the massive overhead costs related to IoT gateways will help them to reduce the strain on the company budget.

When defining compliance requirements, one should be looking to the future rather than the present because of the cost and challenge of shoehorning things in after the fact. That means that someone needs to be continuously studying compliance requirements in the given industry to which the organization belongs, along with best practices. Data will only become more important in the future, and we need to be up to the challenge.

VI. Conclusion

Due to the heterogeneity of applications, the amount of data generated by IoTs differs. Identifying what data to be added to blockchain is a primal task. Healthcare data should be treated in a confidential way and hence utmost care should be taken in sharing and storing of data. The IIoT data use external storage medium away from the local storage for information retrieval and data mining. This is achieved through Edge/Fog computing. Agricultural applications either store data locally or to cloud if the data generated is large. The decentralized nature of the blockchain requires the data to be stored in multiple locations. The most common method of storing data is in still in the cloud. However, decentralized ways integrates blockchain with distributed storage network. Hybrid storage is gaining more popular which divides the data into two, the one to be added to the blockchain network and the other to the cloud data storage. This is a more economical way that implements the concepts of blockchain while maintaining the network latency. The IoT will enable electronic objects to exchange huge amounts of data; storing it in a reliable way will be challenging. Organizational compliance group will know best how long data or information is required, but there are many other requirements that will have to address to ensure that the business objectives in the areas of performance, availability and data integrity, all of which need to be address for the life of the data and information.

REFERENCES

[1] H. Si, C. Sun, Y. Li, H. Qiao, and L. Shi, “IoT information sharing security mechanism based on blockchain technology,” *Futur. Gener. Comput. Syst.*, vol. 101, pp. 1028–1040, 2019, doi: 10.1016/j.future.2019.07.036.

[2] R. H. Weber, “Internet of Things - New security and privacy challenges,” *Comput. Law Secur. Rev.*, vol. 26, no. 1, pp. 23–30, 2010, doi: 10.1016/j.clsr.2009.11.008.

[3] S. A. Bragadeesh and A. Umamakeswari, “Role of blockchain in the Internet-of-Things (IoT),” *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 109–112, 2018, doi: 10.14419/ijet.v7i2.24.12011.

[4] H. Shafagh, L. Burkhalter, A. Hithnawi, and S. Duquennoy, “Towards blockchain-based auditable storage and sharing of iot data,” *CCSW 2017 - Proc. 2017 Cloud Comput. Secur. Work. co-located with CCS 2017*, pp. 45–50, 2017, doi: 10.1145/3140649.3140656.

[5] Q. Xu, K. Mi, M. Aung, Y. Zhu, and K. L. Yong, “A Blockchain-Based Storage System for Data Analytics in the Internet of Things Quanqing,” *New Adv. Internet Things*, vol. 715, pp. 119–138, 2018, doi: 10.1007/978-3-319-58190-3.

[6] L. Minh Dang, M. J. Piran, D. Han, K. Min, and H. Moon, “A survey on internet of things and cloud computing for healthcare,” *Electron.*, vol. 8, no. 7, pp. 1–49, 2019, doi: 10.3390/electronics8070768.

[7] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, “Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system,” *Inf. Sci. (Ny).*, vol. 479, pp. 567–592, 2019, doi: 10.1016/j.ins.2018.02.005.

[8] W. Z. Khan, M. H. Rehman, H. M. Zangoti, M. K. Afzal, N. Armi, and K. Salah, “Industrial internet of things: Recent advances, enabling technologies and open challenges,” *Comput. Electr. Eng.*, vol. 81, p. 106522, 2020, doi: 10.1016/j.compeleceng.2019.106522.

[9] J. S. Fu, Y. Liu, H. C. Chao, B. K. Bhargava, and Z. J. Zhang, “Secure Data Storage and Searching for Industrial IoT by Integrating Fog Computing and Cloud Computing,” *IEEE Trans. Ind. Informatics*, vol. 14, no. 10, pp. 4519–4528, 2018, doi: 10.1109/TII.2018.2793350.

[10] M. Domb, “Smart Home Systems Based on Internet of Things,” *Internet Things Autom. Smart Appl.*, vol. 11, no. 2, pp. 260–267, 2020.

[11] H. Y. Shwe, T. K. Jet, and P. H. J. Chong, “An IoT-oriented data storage framework in smart city applications,” *2016 Int. Conf. Inf. Commun. Technol. Converg. ICTC 2016*, pp. 106–108, 2016, doi: 10.1109/ICTC.2016.7763446.

[12] M. S. Ahmad and A. U. Zaman, “IoT-Based Smart Agriculture Monitoring System with Double-Tier Data Storage Facility,” pp. 99–109, 2020, doi: 10.1007/978-981-15-3607-6_8.

[13] S. S. Gill, R. Buyya, and I. Chana, “IoT based agriculture as a cloud and big data service: The beginning of digital India,” *J. Organ. End User Comput.*, vol. 29, no. 4, pp. 1–23, 2017, doi: 10.4018/JOEUC.2017100101.

[14] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, “A decentralized privacy-preserving healthcare blockchain for IoT,” *Sensors (Switzerland)*, vol. 19, no. 2, pp. 1–17, 2019, doi: 10.3390/s19020326.

[15] G. Wang, Z. Shi, M. Nixon, and S. Han, “ChainSplitter: Towards blockchain-based industrial IoT architecture for supporting hierarchical storage,” *Proc. - 2019 2nd IEEE Int. Conf. Blockchain, Blockchain 2019*, pp. 166–175, 2019, doi: 10.1109/Blockchain.2019.00030.

[16] S. Liu, J. Wu, and C. Long, “IoT Meets Blockchain: Parallel Distributed Architecture for Data Storage and Sharing,” *Proc. - IEEE 2018 Int. Congr. Cybermatics 2018 IEEE Conf. Internet Things, Green Comput. Commun. Cyber, Phys. Soc. Comput. Smart Data, Blockchain, Comput. Inf. Technol. iThings/Gree*, pp. 1355–1360, 2018, doi: 10.1109/Cybermatics_2018.2018.00233.

[17] O. Novo, “Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT,” *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1184–1195, 2018, doi: 10.1109/JIOT.2018.2812239.

[18] R. Li, T. Song, B. Mei, H. Li, X. Cheng, and L. Sun, “Blockchain for Large-Scale Internet of Things Data Storage and Protection,” *IEEE Trans. Serv. Comput.*, vol. 12, no. 5, pp. 762–771, 2019, doi: 10.1109/TSC.2018.2853167.

[19] D. H. Wang, “IoT based Clinical Sensor Data Management and Transfer using Blockchain Technology,” *J. ISMAC*, vol. 2, no. 3, pp. 154–159, 2020, doi: 10.36548/jismac.2020.3.003.

[20] O. Novo, “Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT,” *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1184–1195, 2018, doi: 10.1109/JIOT.2018.2812239.

Communication Chain in the Internet of Things with Spread-out Electronic Device System Abstraction.

1st Andrei Bragarenco

E

Chisinau, Republic of Moldova
andrei.bragarenco mib.utm.md

2nd Galina Marusic
C
E

Chisinau, Republic of Moldova
galina.marusic adm.utm.md

3rd Calin Ciufudean
C A
E

Suceava, Romania
calin eed.usv.ro

—Comm nication is the act of transferring information between entities. Existing architect ral comm nication approaches pro ide sol tions for data transfer from the sender to the recei er. hile those sol tions are targeting different domains, we can find a lot of similarities in their architect res. In this paper, we follow the interpersonal comm nication process principles with its comm nication components for defining a whole comm nication chain that combines the sensor act ator and data transfer items in one concept, pointing to the person to person and de ice to de ice comm nication similarities. e come with an approach of a methodology for the Internet of Things abstracted by the Spread o t Electronic e ice concept. e consider the whole IoT as a single electronic de ice with c t o t and bro en wires and replaced with comm nication chains.

—

I. Introduction

Nowadays, the new trend in technology, the Internet of Things (IoT), connects a massive number of smart objects, products, smart devices, and humans. In this trend, all mentioned things could be organized in groups or systems and cooperate in solving some specific tasks providing information and services to each other involving a typical communication process. Communication is the act of data exchange between things as an extensive system entity and plays an essential role in the system operation. There are a lot of technologies that implement the communication process for information transfer between entities. Curiously, almost all communication technologies are inspired by interpersonal communication, meaning communication between people.

One such example of communication is the online social networking tool Twitter, where users send and receive short posts called tweets. People who want to share information publish and those interested in someone s posts are subscribing to his tweets [1]. This social network platform s system is complex, but its communication relies mainly on the publish-subscribe message architecture.

There are many similarities between Twitter and one of the most popular protocols used for IoT applications - Message Queuing Telemetry Transport (MQTT). MQTT is a protocol specifically designed for machine to machine communication. MQTT protocol runs over TCP / IP stack and has a data packet size with a low overhead minimum (> 2 bytes) so that consumption of the power supply is also reduced [2]. MQTT is the perfect solution for Internet of Things applications. It is designed for devices with low bandwidth. Additionally, it is a simple messaging protocol that allows sending, reading, and publishing data from sensor nodes and much more. Comparing with Twitter, it is the same thing but for devices. MQTT protocol uses a publish/subscribe architecture. Its event-driven approach enables messages to be pushed to the client [3]. Devices are like users. Like Twitter, where users tweet a message to all the followers, a device publishes a message to all subscribers. And just like a Twitter user receives tweets from all people he follows, a device receives messages from all the devices it is subscribed to [4].

Robot Operating System (ROS) proposes another publish/subscribe architecture consisting of a collection of programs. That allows a user to control a mobile robot s operations efficiently. At its core, ROS is an anonymous publish/subscribe message-passing middleware with asynchronous communications. Some modules will issue a set of topics, while others subscribe to that topic. When new data is published, the subscribers can learn about the updates and can act on them. ROS provides the communication using a message-passing approach that forces developers to focus on pure interface logic. [5] The ROScore is a service that provides connection information to nodes so that they can transmit messages to one another. Every node connects to ROScore at startup to register details of the message streams it publishes and the streams it wants to subscribe to. When a new node appears, ROScore provides it with the information needed to form a direct peer-to-peer connection with other nodes publishing and subscribing to the same message topics. Every ROS system needs a running ROScore service so nodes can find other nodes [6].

Automotive is one of the most safety-critical domains. A typical automotive system consists of several Electronic Component Units (ECU) communicating on Controller Area Net (CAN) buses. Software stacks provide support for the computations and communication, including application tasks, the middleware, drivers, and bus peripherals. For communication purposes, tasks read input signals at their activation and write their outputs in shared variables at the end. Some application task reads the input data from a sensor, computes intermediate results sent over the network to other tasks, and, finally, another task, executing on a remote node, generates the outputs as the result of the computation [6]. The read/write operations with shared variables are similar to the publish/subscribe architecture for communication in the vehicle network.

All the aforementioned architectural approaches implement the communication processes that are most convenient in the domain those are operating, a worldwide social network for Twitter, a collection of devices for IoT with MQTT, a complex ECU with ROS, or an ECU interconnection for Automotive with CAN. Even those who

978-1-6654-4067-7/21/ 31.00 2021 IEEE

are targeting different technologies are using the same principles for data transfer – the subscribe-publish. This paper proposes a communication chain concept for the IoT systems abstracted to a spread-out electronic device, providing a methodology and an application example for validation.

II. Materials and Methods

A C C

For an Internet of Things system such as a spread-out electronic device, we will classify the communication process that is conducting the information through the entire system connecting its components. Here we present two primary classifiers – interaction and domain, describing a perspective and defining the classifier s aspects.

1

It refers to interactions in the entire system and will consist of the two types of interactions with devices and the environment.

Those interactions represent all the networking infrastructure that interconnect the parts of the spread-out electronic device together. In the classical references, it is referring to inter-device communication. Fig.1 present an abstraction for this classifier. This group includes all the connectivity techniques and methods, starting from the short-range wired link by an SPI or I2C protocol to the more complex like wireless via 4G or LoRa [7] as a physical link and TCP/IP stack for software abstraction. For the spread-out electronic device concept, the device-to-device communication mechanism abstracts the electrical interconnection between parts of the spread-out device, providing the application s higher abstraction layers to interact with the system as a whole.

E

Those interactions state for an abstraction to the whole system s interface to the physical environment, conceptually presented in Fig. 2. The spread-out device system concept assumes that all the network s sensor-actuator components are accessible to the application as being located on the same physical device [8]. The human-machine interface (HMI) is considered a particular case of sensor-actuator components designed for user interaction scope. Subsequently, those are also classified as device-environment interaction.

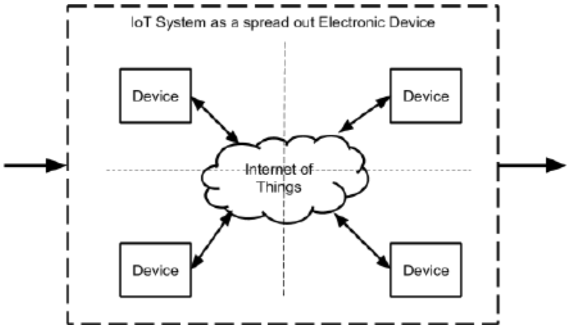


Fig.1. Spread-out electronic device with IoT interconnections [8].

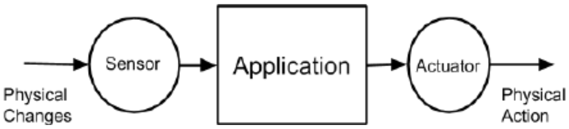


Fig. 2. Device to Environment interaction.

It classifies the communication interconnection by domain affiliation. This classifier ranks the communication interaction to the Hardware (HW) or Software (SW) domain.

The HW domain consists of interconnected electronic components like mechanical parts involved in the signal flow between the physical environment interaction, such as sensing or physical environment of the communication process, and the software domain. The sensor actuator stack will consist of all hardware domain features, such as transducing, power conversion, and hardware conditioning. The communication stacks will consist of the communication chips and the hardware protocol handling components. The physical communication channel itself, e.g., copper twisted pair or pulled-up I2C bus, refers to the hardware domain.

The SW domain represents the software resources that comprise all SW methods applied to the information channels that interconnect all the application components and conduct the signal through the application. The sensor, actuator, user interaction, and communication SW stacks, in this paper, are treated similarly, as information flow with transfer function chains for information transfer. In this domain, the inter-process communication mechanisms [9] and component interfaces [10] are involved in the information transfer between application components. Conceptually the interactions and domains are presented in Fig. 3.

C

1 C

Communication refers to the process of transferring information from source to destination via a communication channel. This statement is not limited to any specific domains.

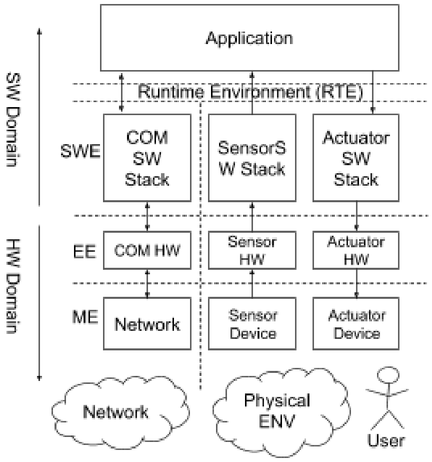


Fig. 3. Interactions and domains in the spread-out device concept.

There are a lot of similarities between interpersonal communication and IT communication. In both, there are similar components of communication, such as source and sender, channel, receiver and destination, message, feedback, encoding, decoding, and noise or barriers, as in Fig. 4.

Following the device to environment interaction concept from Fig. 2, the physical environment is the initial information source. The device collects the information by its sensing component and transfers it to the application components. The application evaluates the data and generates a reaction. Finally, the device sends the response back to the environment via its actuator stack. A wireless network is following the same concept of gathering information from the air. The device first senses the radio signal, conditioning it, and converts it into the information flow. A particular example that justifies the concept is the audio or lighting morse communication. Here, on receiving, the device detects the dot and dash first in the sound or light signal following the classic sensor conditioning signal flow before building a chars frame,

Considering theses from above, the communication chain for an IoT device will consist of the components included in the sensing-actuating chain merged with the components for information transfer or transportation. By this assumption, the entire component collection for the communication process will consist of the component chain presented in Fig. 5. Depending on the application needs, we could define various communication chains where some of the communication phases we could omit. For example, there is no need for conditioning when using a digital sensor and, similarly, no need for frame detection and decoding for an analog sensor. Still, for digital communication through an optical channel, we will need the whole communication chain.

The signal goes through several similar phases in the communication chain component set but complementary for receiving and transmitting processes. Those are following the diagram from Fig. 5, and detailed as follows:

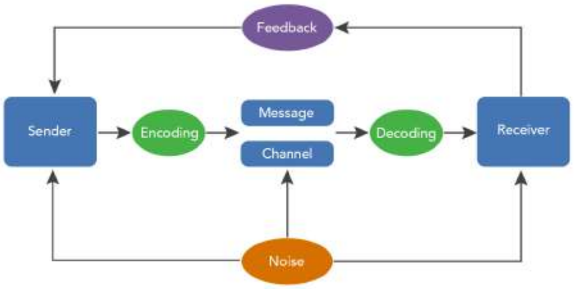


Fig. 4. Key components of communication [4]

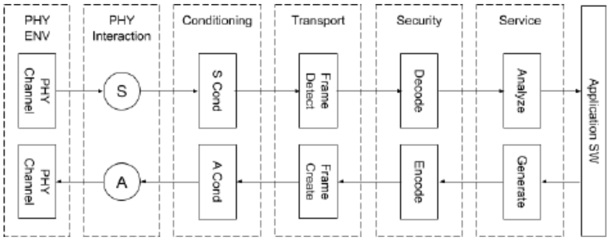


Fig. 5. Communication chain component set.

- PHY ENV phase stands for the physical communication channel that interconnects source and destination. The channel properties involve various chain components, such as conditioning or security, but not limited.
- PHY Interaction phase stands for converting the physical parameter values to the system internal information and vice versa. It consists of sensor and actuator components.
- The conditioning phase stands for information adaption and noise filtering gathered to form the channel or internal processes. We can omit this phase in cases like the wired digital communication on the same voltage level.
- The transport phase consists of detecting and creating frames according to a specific pattern or protocol. This phase is responsible for the safe transfer of the information through the channel.
- The security phase stands for securing the information and protect it from unauthorized access. The system encodes information on sending and decode on receiving.
- The service phase analyzes the received information, prepares it for the application, and generates information according to the application request.
- Application stands for the sender and receiver in the information flow.

C

Structurally, the system forms a collection of interconnected components that cooperate to produce a result [11]. According to the IoT concept as a spread-out electronic device mentioned above, physically, the system s components could be located on any of the IoT devices that are members of the system. Here an IoT system is considered a wholly electronic device that imaginary we cut in parts, and communication channels replace the broken connectivity lines. In this way, we will have two primary types of interconnections between components, as presented in Fig. 6.

The *C* *A* from Fig. 6. is an internal channel that connects *C* *A* and *C* from the application running on *A* . Internal channels are following the software interfaces and inter-process communication techniques [9].

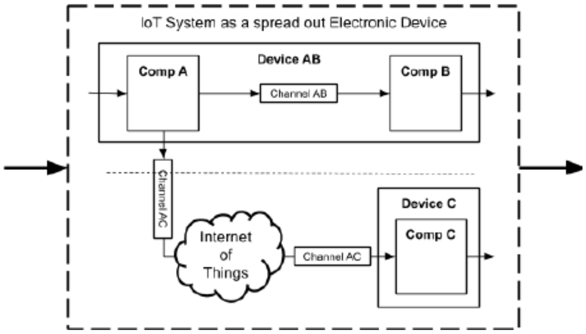


Fig.6. Communication channel types in a spread-out electronic device system.

The *C* *AC* connects *C* *A* and *C* *C* components that are part of applications running on different devices, *A* and *C*. Such connections are the target of the communication chain component set presented in the current work. The component interface is similar to the internal channel access to a component located on the same device.

Assuming that communication channels are abstracting the component s physical localization, at the flatten level, we could represent the whole system as components with internal transfer functions interconnected with channels, as illustrated in Fig. 7. Several similar signal channels that provide values for the same domain or parameter logically could be combined in a group and operated together. Following the subscriber-publisher concept, the publishing phase will update the component signal by a transfer function. The subscribing phase will connect to a specific component interface via a communication channel. In this work, we use the term to create a link or a channel between components for subscribing. Depending on which end of the communication channel has an active role, two types of channel linkages are possible – push and pull, as in Fig 8. Such connection involves a push or a pull method to operate the channel to apply or extract information. The same procedure we use for all channels included in the same group. Push means that the channels accept the change of its value by a specific method; Pull means that it is evaluated internally and accessible by an on-demand method. Reference [12] proposes a method for embedded systems development by configurations and code generation based on JavaScript Object Notation metamodels using the channel s concept defined in this paper and a tool proposal for operating the methodology.

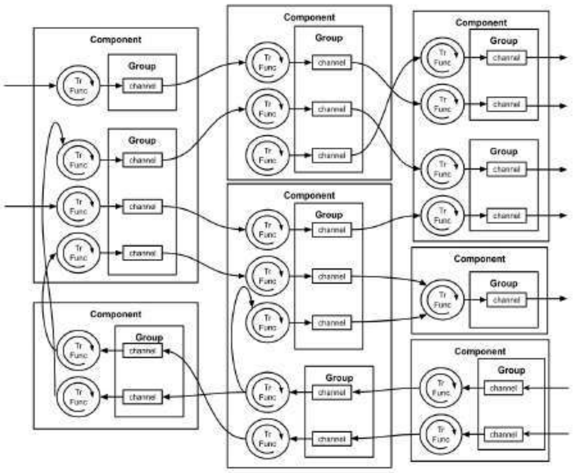


Fig.7. Interconnected compoenents of the system with channels at the flatten level [12]

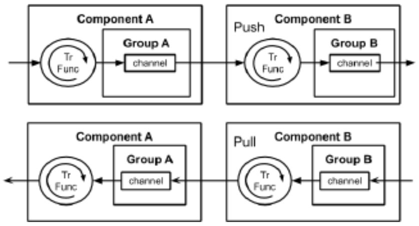


Fig. 8. Types of channel linkage [12]

III. Results

Following the concept of the spread-out electronic devices discussed in this paper, a remote mechatronic manipulator application is under test. The entire system consists of two robotic manipulators connected via a wireless network, here Wi-Fi. First is actuator-free used as a remote manipulator. It collects the arm joint s angles by sensors and transmits them via the network. The second, underactuated 3-DOF arm receives setpoints and uses them in its joints and head position. Fig. 9 presents the conceptual system architecture diagram. First tests are performed on a 3-DOF robotic arm, extending to a more complex mechatronic system with many DOF. Such an example could be an exoskeleton for helping people with locomotor diseases.

According to the layered architectural approach and the communication chain discussed here, the signal from the remote setpoint arm, which joints sensor data we use as the second arm s setpoint, passes through the whole communication chain and reaches the underactuated arm sensor service components. It can abstract the network and use it as the sensor is connected directly to the device. Fig. 10 presents the layered architecture of the system and the flows for the sensing and actuating signals. The resulting application allows handling a 3-DOF robotic arm via a copy of it, shadow, or an arm s software model. The control consists of sending joint setpoint angles to the underactuated arm through the wireless network. Cartesian coordinates of the arm head are sent instead of angle setpoints when inverse cinematics is involved.

IV. Discussions

The communication chain presented in this paper has a significant role in the spread-out electronic device concept. We follow it to interconnect the devices and threaten the entire IoT system as a single electronic device. By this abstraction, applications are developed considering that all resources are available through the platform s simple service via RTE. The paper [12] presents a method that allows the platform s configuration to define the communication chain through a metamodel. This way allows to specify only the chains from a single device, but the next steps are to extend the technique to define multidevice platforms within the same metamodel.

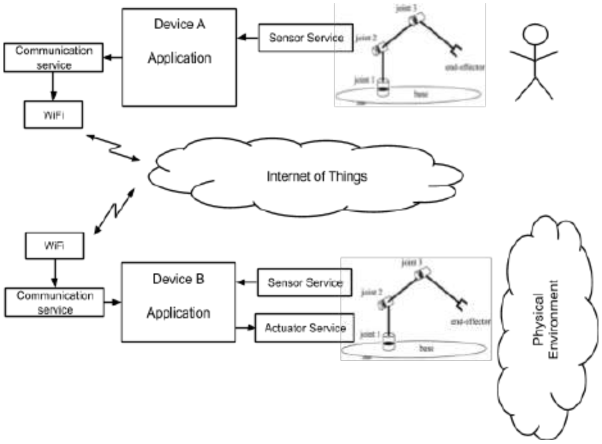


Fig. 9. Conceptual system architecture for remote manipulator control

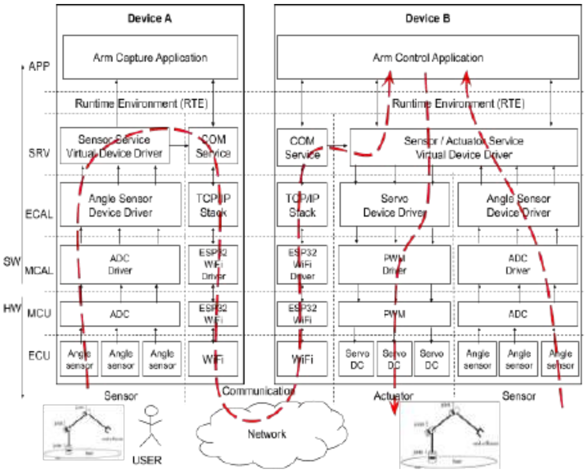


Fig.10. Layered architecture of the system and signal flow through communication chain.

The target of the research is to have a concept of systems with autonomous evolution of configurations. It will solve complex business chains considering that all the equipment and services in the world are available as services. This way, we can build complex systems, configurable templates, adaptable to the current needs or continuously changing environment. The bricks for such systems are component databases and configurable channels.

V. Conclusions

Communication is one of the most critical issues in the system component interconnection. Various projects implement many communication concepts. Still, almost all follow the general idea of the communication laws, which is far to be limited to the engineering domain. The method is inspired by interpersonal communication, containing all the communication components such as source and sender, channel, receiver and destination, message, feedback, encoding, decoding, and noise or barriers.

The primary information source is the physical environment. Considering that all the world devices are part of an enormous spread-out electronic device, the information should pass the whole communication chain to reach the application running on one huge device. A typical chain will involve interaction with the environment, signal conditioning, transport, and coding/encoding, providing the application s required service.

An application running on the spread-out electronic device system has been presented, consisting of a collection of components represented by its transfer functions, interconnected with communication channels. When interconnected components are not physically on the same device of the spread-out system, we insert a communication channel following the communication chain concept, including all required phases for information transfer.

The spread-out electronic device concept allows the IoT s abstract to a single device with access to Everything from Everywhere.

References

[1] F. Maclean, D. Jones, G. Carin-Levy, and H. Hunter, (2013). Understanding Twitter , in British Journal of Occupational Therapy. 76(6):295. DOI: 10.4276/030802213X13706169933021.

[2] R. Atmoko, R. Riantini, and M. Hasin, (2017). IoT real time data acquisition using MQTT protocol , in Journal of Physics: Conference Series. 853. 012003. 10.1088/1742-6596/853/1/012003.

[3] S. Shankar J, S. Palanivel, S. Venkateswarlu and M. Sowmya. (2019). MQTT in Internet of Thing , in International Research Journal of Engineering and Technology(IRJET) pp796-798

[4] MQTT broker – a Twitter for machines https://medium.com/ noobino/mqtt-broker-a-twitter-for-machines-b624ea2773b1 [accessed Dec 2, 2020].

[5] M. Anderson, Introduction to the Robot Operating System (ROS) Middleware , in Embedded Linux Conference OpenIOT Summit North America March 12-14, 2018 - Portland, OR

[6] H. Zeng, M.D. Natale, P. Giusto, and A. Sangiovanni-Vincentelli, Stochastic Analysis of CAN-Based Real-Time Automotive Systems in IEEE Transactions on Industrial Informatics, 5(4), 388–401. doi:10.1109/TII.2009.2032067

[7] A. Waret, M. Kaneko, A. Guitton and N. El Rachkidy, LoRa Throughput Analysis With Imperfect Spreading Factor Orthogonality, in IEEE Wireless Communications Letters, vol. 8, no. 2, pp. 408-411, April 2019, DOI: 10.1109/LWC.2018.2873705.

[8] A. Bragarenco, Sensor-Actuator Software Component Stack for Industrial Internet of Things Applications in 24th International Conference on System Theory, Control and Computing, pp. 540-545, October 8 - 10, 2020, Sinaia, Romania.

[9] L. Alawneh, H. Abdelwahab. Pattern Recognition Techniques Applied to the Abstraction of Traces of Inter-Process Communication , in 15th European Conference on Software Maintenance and Reengineering, 2011. doi:10.1109/CSMR.2011.27.

[10] A.R. Mamidala, Architecture of the Component Collective Messaging Interface , in International Journal of High Performance Computing Applications, 2010.

[11] A. Bragarenco, G. Marusic, and C. Ciufudean. Layered Architecture Approach of the Sensor Software Component Stack for the Internet of Things Applications , in WSEAS Transactions on Computer Research Volume 7 (2019): pp. 124–135.

[12] A. Bragarenco, Method for Embedded Systems Development by Configurations and Code Generation Based on JSON Metamodels , in Revista de tiin , Inovare, Cultur i Art Akademos 3(58) / 2020 pp 19-27, ISSN 1857-0461, doi:10.5281/zenodo.4269373

Economic Inclusion in the United States: Predictive Analysis of COVID-19 pandemic on County Rates of Unbanked Households

Timur Berdibekov
George Mason University
United States, Fairfax
tberdibe@gmu.edu

PruthviRaj Reddy Pasnoor
George Mason University
United States, Fairfax
ppasnoo@gmu.edu

Asmitha Rao Annamaneni
George Mason University
United States, Fairfax
aannaman@gmu.edu

Ebrima N Ceesay
George Mason University
United States, Fairfax
eceesay2@gmu.edu

***Abstract*—COVID-19 pandemic has a significant effect on the unemployment rate in the United States. However, the economic effect in different states is not the same for each household. In this work, Our goal is to capture and outline the relationships between pandemic incidence, economic inclusion, unemployment, and bank branch closures in order to understand the emerging relationship between the coronavirus pandemic, rates of economic inclusion, and economic well-being of localities. Furthermore, we machine learning algorithms to evaluate the predictive power of coronavirus incidence and fatality rates, county-level unemployment, and bank branch closure rates on rates of economic inclusion. Also, a natural language processing approach is used to analyze the unemployment COVID-19 textual data. We use BERT as a powerful transformer for sentiment classification on COVID-19 unemployment data.**

***Index Terms*—Machine learning, Natural Language processing, COVID-19, Transformer**

I. Introduction

The 2019 Novel Coronavirus (COVID-19) pandemic has presented fundamentally new disruptions and challenges in many areas. However, it continues to exacerbate existing long-term issues and underlying disparities in the United States with a significant impact on public health and the general economy [1] [2].

The economic and financial impact of the COVID-19 pandemic is also extensively monitored throughout the beginning of the pandemic through U.S. Census Bureau research, including the Household Pulse Survey launched in April 2020, as well as relevant data from the Current Population Survey and the Department of Labor [3] [4]. The coronavirus has a considerable influence on the employment sector in the United States, and researchers have noted an increase in unemployment rates, which have been increased to 14.7%, which is the highest ever since 1948. The link between the COVID-19 pandemic and unemployment rates has been demonstrated in recent research, with 46% of the adults whose salary is low having faced inconvenience paying their bills. There has been a loss of employment and reduced pay for most of the employees, eventually leading them to struggle to pay their house rent. Young workers aged 18 to 29, in comparison with those 50 to 64, are now employed at a different job because of pandemic [5] [6].

Emerging arguments in financial and public policy have focused on the potential of the COVID-19 pandemic to exacerbate existing issues in access to credit in the U.S. banking system, with a specific focus on the public benefits of enabling unbanked populations to quickly access federal relief funds in a low-cost and digital manner [7]. Furthermore, the physical banking branch network, or more specifically the presence of bank branches, in areas especially designated as Low and Moderate Income has been linked to greater credit and lending activity in a community, especially for mortgage originations and small-business lending [8] [9]. Retail bank branches also help collect hyper-localized soft information on bank customers that are not easily captured in automated or algorithmic underwriting, which can help identify more creditworthy personal and small-business borrowers [10] [11] [12] [13].

The Federal Deposit Insurance Corporation (FDIC), an independent agency tasked by Congress in maintaining stability and public confidence in the U.S. financial systems, as of 2009 actively seeks to expand financial inclusion as a means of expanding overall public confidence in the U.S. banking system and works both with the private and public sector in achieving that goal. Furthermore, the Federal Reserve Board of Governors and the twelve member Federal Reserve Banks actively promote financial inclusion as part of its dual mandate of ensuring maximum employment and stable-prices and moderate long-term interest rates. As such, both entities contribute to the body of research dealing with financial inclusion in the United States, and economists within the Federal Reserve System and FDIC have produced literature on removing the barriers to increasing financial inclusion [14] [15].

As far as industry, the American Bankers Association, a trade group representing U.S. banks, has partnered with the Cities for Financial Empowerment (CFE) Fund to promote the Bank On initiative, a national program for banks coordinating efforts in increasing access to bank accounts in the United States [16] [17] [18] furthermore, while themselves not significant contributors to existing research in financial

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

inclusion, Fintech startups that focus on deploying innovations in financial technology to compete with banks and traditional providers of financial services actively positioned themselves as potential solutions in increasing financial inclusion and contribute to trade groups and associations that produce industry-related research.

II. Methodology

A. Data Preprocessing

Data collected for this paper came from several sources. COVID-19 news and data comes from different online sources however Heidari et al [19] advanced bot detection tools helps us to avoid collecting data that contain misinformation and creates reliable data for our experiments. Also, we use official sources, such as The National Survey of Unbanked and Underbanked Households; a survey conducted every two years by the Federal Deposit Insurance Corporation to collected data on household participation in the banking system [20]. County-level coronavirus pandemic data was obtained from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [21]. The Local Area Unemployment Statistics (LAUS) program produces monthly and annual employment, unemployment, and labor force data and was used for county-level unemployment rates [22]. The Office of the Comptroller of the Currency's Corporate Applications Search was used to gather and aggregate the number of bank branch closures and openings in the United States, which was used to derive the number of banks closed or opened per population [23]. Furthermore, the United States Census Bureau's 2010-2019 county population totals dataset was used for 2019 U.S. county population estimates [24]. Lastly, Federal Information Processing Standards (FIPS) codes were used to map county names to unique codes for all datasets for consolidation, consisting of a unique two-digit code for state, including Washington, D.C., and Puerto Rico, and three-digit county or county-equivalent code unique to the state code.

A universal index was created, which mapped standardized county and county-equivalent labels to a unique FIPS code, a five-digit code consisting of two digits representing a particular county's state, and a three-digit representing the specific county or county-equivalent in that state. This representation enabled easy identification of individual counties across all datasets used, enabling more data aggregate on the economic inclusion dataset.

The initial FDIC supplemental dataset consists of hundreds of variables collected for over 70,000 U.S. households, requiring pruning of survey features not needed for this analysis. Since banked status was collected for only 47.9% of respondents, more than half of all respondents were removed from the dataset, resulting in approximately 32,904 observations of U.S. households where banked status was definitively known. At the geographical level, while the state and state-equivalent levels could be identified for every household response in the survey, accuracy to the county and the county-equivalent level was only possible for 12,990 households. These 12,990 households represented data from 280 uniquely identified counties across 41 states, in which the respondent household banked status was known.

We extract unemployment rates from U.S. Census Bureau contains unemployment rates and population estimates for more than 3,000 U.S. counties and county-equivalents. The unemployment dataset provided by the U.S. Census Bureau consisted of employment variables by month and by county for 13 months between August 2019 and September 2020.

The resulting data preprocessing yielded two datasets used for analysis. The first, more compact dataset consists of data specific to county and county-equivalents, including the rate of unbanked households based on 12,990 household survey responses, COVID-19 incidence and case fatality ratios, bank rate branch closures and establishments as reported to the OCC, and county-level unemployment rates. This dataset was used for predictive analysis using the Random Forest ensemble, Support Vector Regression, and Linear Regression algorithms to predict county-level unbanked rates using county-level data. The second, larger dataset excluded unbanked rates, allowing data inclusion for 3,219 U.S. county and county-level equivalent areas to visualize and understand overall county trends in unemployment, the COVID-19 pandemic, and many bank branch closures.

B. Data Merging

Firstly, the financial inclusion dataset is merged with the COVID dataset by merging the county code and unbanked rate from the financial inclusion dataset with the FIPS, Incidence rate, case fatality ratio, latitude, and longitude from the dataset. Then we merge the dataset of the financial inclusion and COVID with the unemployment dataset. Here the Civilian labor force, unemployed rate, and county code are merged into the dataset. To understand how the population has been changed in the year 2020 due to the pandemic, we merged the 2019 population estimate dataset using the county code followed by the bank closure data. As a result, the final dataset consists of the COVID-19 unemployment data, population, bank branch rate, geographical index data, bank closure, and opening rates for each county.

III. Exploratory data analysis

The 280 counties in which unbanked status was attributed to 12,290 nationally representative household respondents combined represent 149 million people or approximately 45% of the entire 2019 U.S. population estimate. Notably, the counties and county-equivalents with attributed economic inclusion status are more populated than a typical U.S. county and county-equivalent. Table I shows the unbanked rate in different states based on specific counties.

Relative to the overall weighted household, the national rate for unbanked rate is 5.3%. However, the unbanked rate mean for these counties is 4.5% while the median is only 2.2%, and a large share of counties do not have significant unbanked households. Figure 1 shows that the distribution of unbanked rates. The distribution skewed to the right which

TABLE I
FIVE LARGEST AND SMALLEST COUNTIES FOR U.S. COUNTIES WITH 2019 POPULATION ESTIMATES AND AVAILABLE UNBANKED RATES

County	State	Population (2019 Est.)	Percent of Unbanked Households
Los Angeles	*California*	*10,039,107*	*9.2%*
Maricopa	*Arizona*	*4,485,414*	*3.0%*
San Diego	*California*	*3,338,330*	*3.7%*
Orange	*California*	*3,175,692*	*1.9%*
Miami-Date	*Florida*	*2,716,940*	*5.3%*
Androscoggin	*Maine*	*108,277*	*0.0%*
Bartow	*Georgia*	*107,738*	*11.8%*
Miami	*Ohio*	*106,987*	*18.2%*
Franklin	*Missouri*	*103,967*	*1.2%*
Cecil	*Maryland*	*102,855*	*1.4%*
Average, All Counties		***101,969***	
Median, All Counties		***24,909***	
National Rate			***5.3%***

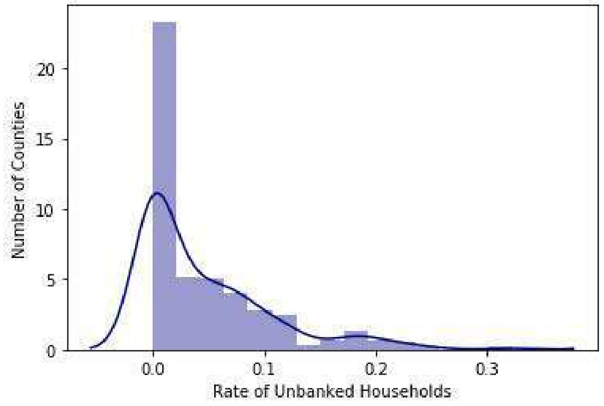


Fig. 1. The distribution of unbanked households rates for the 280 county

means that if one states has more counties, it does not mean it has higher unbanked household rate. Population-wise, counties with higher populations have higher unbanked rates compared with a typical county with less population. The small county with attributable banked rates has a 102,855 population, exceeding the average U.S. county population. So the cohort of 280 counties with attributable unbanked rates is more urbanized than a typical county due to higher population, and that the cohort consists of counties centered around either around 5% or 20% for the rate of unbanked households.

We use k-means clustering as an unsupervised machine learning algorithm and principal components analysis in 280 counties with known unbanked rates. We include county-level observations for COVID-19 incidence and case-fatality rates, unemployment level, rate of bank closures and openings, and unbanked rates. Figure 2 shows the results based on PCA and k-means clustering. Five clusters will yield the optimal results. Figure 3 shows a positive correlation between unemployment and unbanked households.

Figure 4 shows a positive correlation between branch rates in a county and the incidence rate of COVID-19. Also, The relationship between the unemployment rate and unbanked

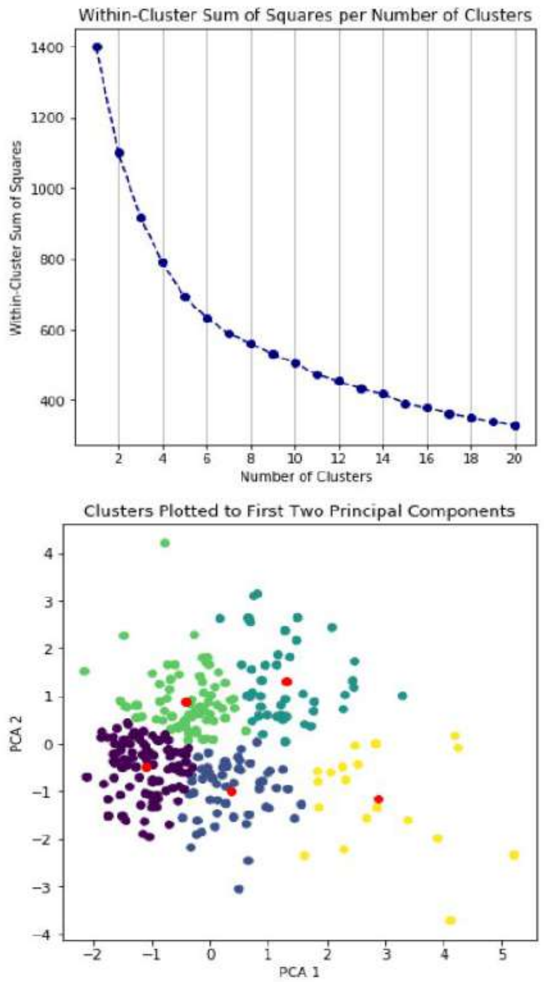


Fig. 2. Within-Cluster Sum of Squares for k-means clusters using one to 20 clusters, and a plot of clusters along principal components 1 and 2.

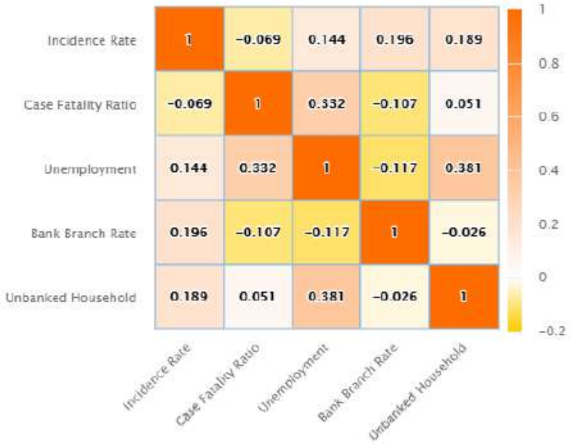


Fig. 3. Correlation Plot of Features for Select Counties

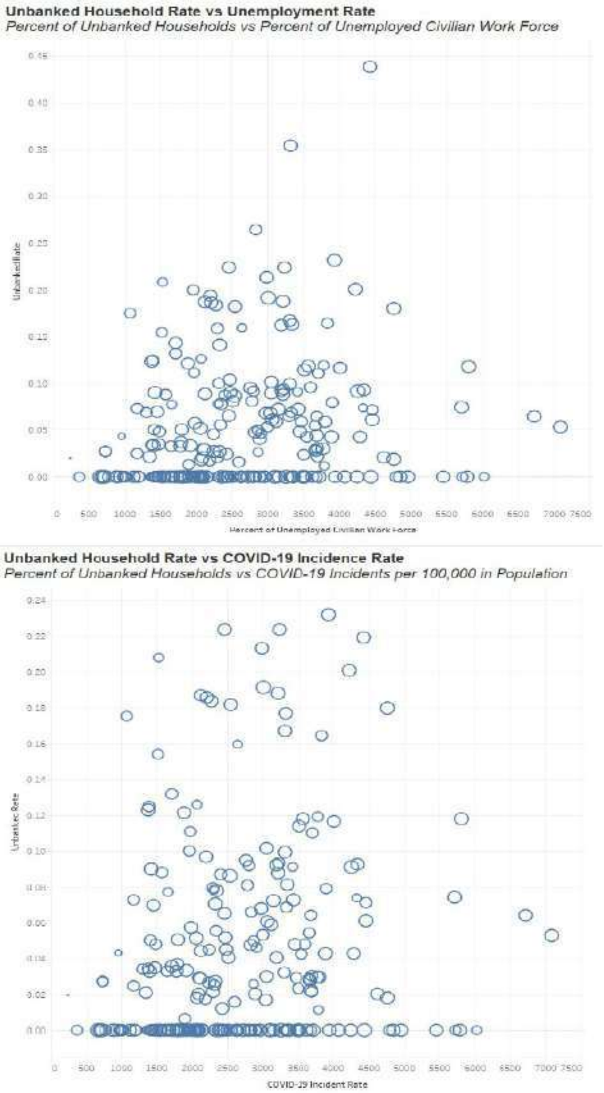


Fig. 4. The rate of unbanked households at the county level appears to increase slightly with an increase in unemployment and COVID-19 incidence rates.

households is positive. This analysis shows the colinearity between some features and the importance of the feature selection method in improving the algorithm performance. This research uses a combination of filter methods(PCA) and wrapper methods for feature selection.

Figure 5 shows Which state has the highest and the lowest number of confirmed cases in the united states. Texas has the highest, and Maine has the lowest number of confirmed cases in this research.

According to figure 6 New York has the highest and Vermont has the lowest number of deaths.

Figures 7 shows the highest and the lowest number of recovered cases in united states. New York has the highest 33,961 and Kansas has the lowest number of recovered cases.

Figure 8 shows the incidence rate in each county. In this treemap, we used the variables incidence rate, states, and county from the dataset grouped by shades to determine which county has the highest incidence rate and which county has the lowest rate. Above are the listed 50 states of Unites states according to the tree map's color shade. From the below map, Wheeler county in Oregon has the lowest incidence rate of 75, and Lincoln County in Arkansas has the highest rate of 18,082.

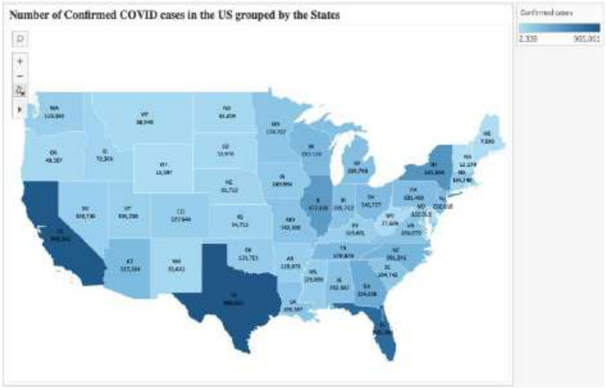

Number of Confirmed COVID cases in the US grouped by the States
Confirmed cases
2,330
905,001

Fig. 5. The number of confirmed cases in the US grouped by states.

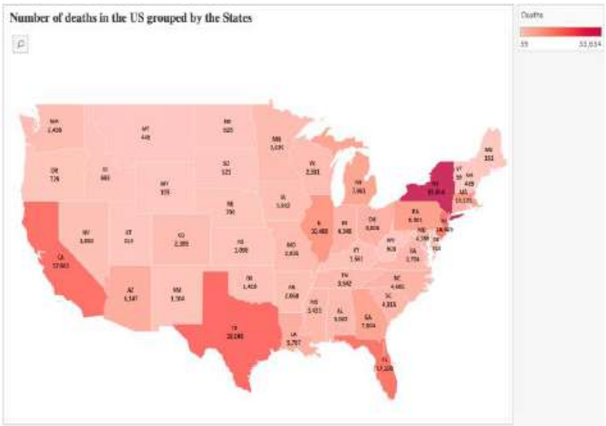


Fig. 6. The number of death cases in the US grouped by states.

After data analysis, we apply several machine learning models based on the Random Forest, Support Vector machine,

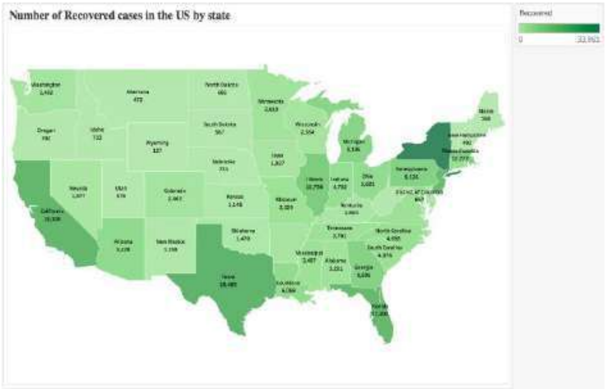


Fig. 7. The number of recovered cases in the US grouped by states.

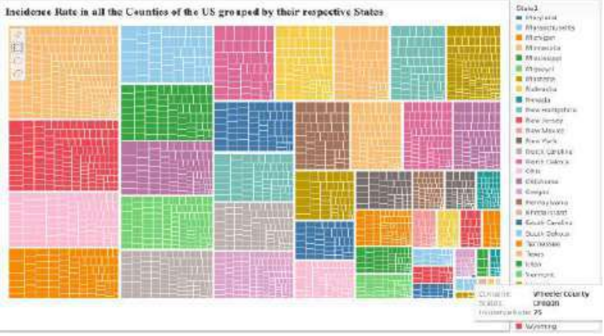


Fig. 8. The number of COVID-19 cases(incidence rate) in 50 states in United States

TABLE II
PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS

algorithm	f1_score	mcc
SVM	0.83	0.76
FFNN	0.90	0.83
RF	0.93	0.88
LR	0.86	0.84

Linear Regression, feed-forward neural network. Algorithms were used to accurately predict county-level unbanked rates based on COVID-19 pandemic rates, rate of bank closures, and unemployment rate. Table II shows the results. The best performing algorithm was Random Forest with an f1 score of 0.93, and the Neural network model ranked second with an f1 score of 0.90.

IV. USING TRANSFER LEARNING MODEL

In addition to applying machine learning models, we apply sentiment analysis on the COVID-19 unemployment data by using the natural language processing model in this section. For this task, We collect COVID-19 unemployment data by using api [25] of spatial AI and we remove bot comments by using method by Heidari et al. [26]. We compare three approaches for sentiment classification of COVID-19, which are TextBlob [27], VADER [28], and BERT [29] [30].

The TextBlob is based on the supervised learning method, the Naive Bayes classifier. TextBlob assigns a score of 1 for positive and -1 for negative, and 0 for neutral sentences.

VADER(Valence Aware sEntiment Reasoner) is a model used for sentiment analysis, which is sensitive to both polarity(positive and negative)and intensity of emotion.

Figure 9 shows a BERT model [29]. We use BERT [31]or the Sentiment classification task. BERT is a strong transformer model, and studies by [32] show how We can fine-tuning the BERT for sentiment classification in this work. Table III shows the result for sentiment classification model for COVID-19 data. Although the three models provide very similar accuracy for sentiment classification, we choose BERT as the best model since it is a transfer learning model, and we can fine-tune the model based on a specific task that we need on COVID-19 data.

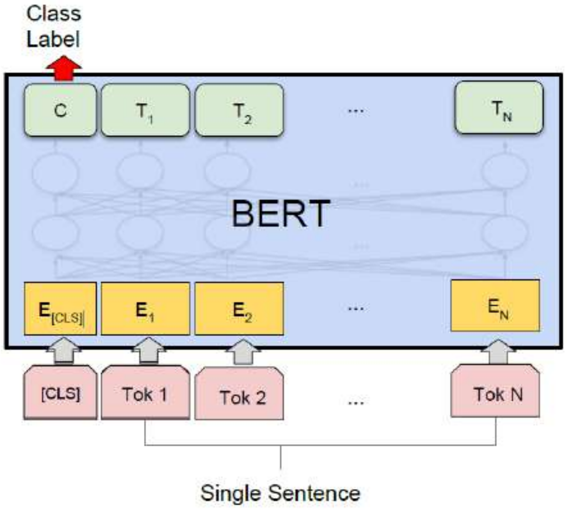


Fig. 9. BERT for sentiment classification

TABLE III
TEST CAPTION2

Model	Accuracy
TextBlob	0.78
VADER	0.80
BERT	0.83

V. CONCLUSION AND RESULTS

In this work, several factors could be contributing to the lack of predictive power. Primarily, features selected to predict economic inclusion rates could have limited relevance and usefulness to the models used in banked households. The relationships between economic inclusion, the impacts of the pandemic, unemployment rates, and bank branch closure rates are complex and not clearly delineated, and thus may share the same underlying dependencies that can contribute to model confusion, such as racial makeup, level of income, other demographic compositions of counties, and even the overall unbanked household rates at the state level.

As race and ethnicity have been linked to both the economic and public health impacts of the pandemic and access to the financial system, training the model on race and the selected features might yield a more informative model in terms of predictive power in future studies. One issue of this approach is that there is already a link between race and access to the financial system, and alleviating economic inclusion barriers caused by race is complex in policy implementation.

One novel point is that the rate of bank closings and establishments, used as a proxy to measure not only a community's access to credit but the general health of the economy, was not a significant feature for predicting unbanked household rates at the county-level. However, since the bank rate consisted of subtracting a specific county's bank closings from bank establishments and normalizing to a rate per 100,000 in population, it is possible that looking instead at the total number of banks present in a locale might be more beneficial for predicting rates of unbanked households, although this approach presents its challenges in terms of data availability. This paper represents a starting attempt to predict unbanked household rates with an appropriate degree of accuracy for use in informing broader policy decisions, especially in attempts at increasing economic inclusion through the adoption of novel and impactful methods. In this work, we apply transfer learning model BERT on the COVID-19 unemployment data to get more insight and context about people's ideas about the effect of a pandemic on unemployment. It is essential to consider textual information of different groups in society as semantic input for the machine learning models to provide more accurate results in future work.

References

[1] Centers for Disease Control and Prevention, "Coronavirus disease 2019 (covid-19) - community, work & school." https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html, 24 July 2020.

[2] P. Cihan, "Fuzzy rule-based system for predicting daily case in covid-19 outbreak," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–4, 2020.

[3] Center on Budget and Policy Priorities, "Tracking the covid-19 recession's effects on food, housing, and employment hardships." https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-recessions-effects-on-food-housing-and, 13 November 2020. Covid Hardship Watch.

[4] V. Z. Marmarelis, "Predictive modeling of covid-19 data in the us: Adaptive phase-space approach," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 207–213, 2020.

[5] R. Kochhar, "Unemployment rose higher in three months of covid-19 than it did in two years of the great recession." https://www.pewresearch.org/fact-tank/2020/06/11/unemployment-rose-higher-in-three-months-of-covid-19-than-it-did-in-two-years-of-the-great-recession/, 11 June 2020. Pew Research Center.

[6] C. STOLOJESCU-CRISAN, B. P. BUTUNOI, and C. CRISAN, "Iot based intelligent building applications in the context of covid-19 pandemic," in *2020 International Symposium on Electronics and Telecommunications (ISETC)*, pp. 1–4, 2020.

[7] A. Klein, "A big problem for the coronavirus economy: The internet doesn't take cash." https://www.brookings.edu/opinions/a-big-problem-for-the-coronavirus-economy-the-internet-doesnt-take-cash/, 25 March 2020. Pew Research Center.

[8] O. E. Ergungor, "Bank branch presence and access to credit in low-to moderate-income neighborhoods," *Journal of Money, Credit and Banking*, vol. 42, no. 7, pp. 1321–1349, 2010.

[9] A. D. Indriyanti, I. G. L. E. Putra, D. R. Prehanto, I. K. D. Nuryana, and A. Wiyono, "Development of mapping area software for dismissal people affected by covid-19," in *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)*, pp. 1–4, 2020.

[10] L. Ding and C. K. Reid, "The community reinvestment act (CRA) and bank branching patterns," *Housing Policy Debate*, vol. 30, pp. 27–45, Nov. 2019.

[11] A. Kyung and S. Whitney, "A study on the financial and entrepreneurial risks of small business owners amidst covid-19," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–4, 2020.

[12] N. Petrović, "Simulation environment for optimal resource planning during covid-19 crisis," in *2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, pp. 23–26, 2020.

[13] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733–2742, 2020.

[14] Y. L. Toh and T. Tran, "How the covid-19 pandemic may reshape the digital payments landscape." https://www.kansascityfed.org/en/publications/research/rwp/psrb/articles/2020/covid19-pandemic-may-reshape-digital-payments-landscape, 24 June 2020. Kansas City Fed.

[15] A. Khalid, M. H. Tahir, H. Muhammad Bilal Asghar, M. Munir, N. Arshed, and H. Rehman, "A meta-analysis of foreign direct investment and economic growth: An empirical evidence from pakistan during covid 19 policymaking," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1–6, 2020.

[16] BankOn, "Join the national bank on movement." https://joinbankon.org/certify/.

[17] I. McCulloh, K. Kiernan, and T. Kent, "Improved estimation of daily covid-19 rate from incomplete data," in *2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)*, pp. 153–158, 2020.

[18] F. Yakut and F. Ertam, "A digital forensics analysis for detection of the modified covid-19 mobile application," in *2020 5th International Conference on Computer Science and Engineering (UBMK)*, pp. 1–5, 2020.

[19] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0542–0547, 2020.

[20] Federal Deposit Insurance Corporation (FDIC), "Fdic survey of household use of banking and financial services." https://www.economicinclusion.gov/downloads/index.html#yearly, 2020.

[21] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, pp. 533–534, May 2020.

[22] U.S. Bureau of Labor Statistics, "Labor force data by county, annual averages: Local area unemployment statistics (laus)." https://www.bls.gov/lau/#cntyaa, 2020.

[23] Office of the Comptroller of the Currency, "Corporate applications search (cas) - branch closings and branch establishments - 1-1- 2019 to 12-6-2020." https://apps.occ.gov/CAAS_CATS/, 2020.

[24] U.S. Census Bureau, "Nation, states, and counties population: County population totals: 2010-2019." https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html, 2019.

[25] spatial.ai, "unemployment-sentiment-data-covid-19: 2010-2019." https://www.spatial.ai/post/unemployment-sentiment-data-covid-19, 2019.

[26] M. Heidari, J. H. J. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[27] https://textblob.readthedocs.io/en/dev/, 2018.

[28] https://pypi.org/project/vaderSentiment/, 2019.

[29] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[30] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[31] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[32] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

Implementation of a Low-Cost Automatic Baciloscopy System for the Diagnosis of Tuberculosis

Joaquin Gonzalez Villarreal
Department of Engineering
Pontifical Catholic University of Peru
Lima, Peru
joaquin.gonzalezv@pucp.edu.pe

Miguel Cataño Sanchez
Department of Engineering
Pontifical Catholic University of Peru
Lima, Peru
mcatano@pucp.edu.pe

Willy Carrera Soria
Department of Engineering
Pontifical Catholic University of Peru
Lima, Peru
wcarrer@pucp.pe

***Abstract*—According to the World Health Organization (WHO), tuberculosis continues to be a global problem of great importance. In Peru, around 31 120 cases of TB were registered annually, with an incidence of 116 new cases per 100 000 inhabitants. Peru still occupies first place in Latin America in the number of resistant tuberculosis cases. Overcrowding coupled with the lack of rapid diagnostic equipment accessible to the neediest populations has resulted in the disease remaining a problem. The diagnostic process is manual, leading to great variability in the results because it depends on the laboratory staff's experience and concentration. Although not the most sensitive, conventional baciloscopy techniques are the most widely used due to their speed and cost. These techniques involve a staining stage and a microscopy stage. Ziehl Neelsen (ZN) staining method is used in smear microscopy because it is effective and simple. This article shows a prototype that performs tuberculosis diagnosis for 4 samples, consisting of two automatical modules: the staining module and the microscope module. The staining module consists of a sample positioning sub-system, a reagent dispensing sub-system, and a heating sub-system. The microscope module captures fields' images of up to 4 samples stained using the Ziehl-Neelsen method and determines the diagnosis of Tuberculosis (TB) through image processing algorithms. The mechanical system of this module is made up of three subsystems: the main support that houses the electronic circuits, a mobile platform that moves the samples, and the microscope holding structure. Tests have been performed on 206 samples at the Dos de Mayo National Hospital in Peru. These tests involved several adjustments that allowed reducing the staining times and the quantities of reagents used.**

***Index Terms*—Tuberculosis diagnosis, microbiological diagnosis, staining methods, smear microscopy, microscopy, Ziehl Neelsen staining.**

I. Introduction

According to the World Health Organization (WHO), tuberculosis continues to be a global problem. Thus, according to the WHO World TB Report 2016, Peru still ranks first in Latin America in the number of cases of multidrug-resistant (MDR) and extensively drug-resistant (XDR) tuberculosis [1]. In 2017, only in Peru, there were about 31 120 new tuberculosis cases with an incidence of 116 new cases per 100 thousand inhabitants [2]. In 2019, there were 10 million cases worldwide, of which 1.4 million people were fatally affected [3].

TB diagnostic methods can be classified into smear microscopy, culture, and direct genetic detection in samples. The smear microscopy techniques are the most commonly used, involving a staining stage and a microscopy stage. Ziehl Neelsen (ZN) staining is the most-widely-used staining method in smear microscopy because it is effective, simple, economical, and fast; besides being the staining method recommended by the WHO for developing countries. This baciloscopy process consists of the next sub-processes: first, the sample is spread on a slide (smear), then the staining method to highlight Acid-Alcoholic Resistant Bacilli (AARB) is performed by the ZN method based on phenylated fuchsin, methylene blue, acid alcohol, and water. Finally, for diagnosis, the stained sample is examined under a microscope, where the tubercle bacilli present in the sample are counted. The diagnostic process involves the time and exposure of the personnel in charge, who perform an average of 70 diagnostic samples per day.

The Ziehl Neelsen method is performed manually by expert laboratory technicians who inevitably perform the process with wide variability, affecting post-diagnostic activities. Errors in AARB concentration can lead to the non-detection of TB patients (false negatives) who will continue the chain of transmission in the community or the futile treatment of non-tuberculous (false positives). [4]. This is why the guarantee or assurance of a standardized, rapid and reliable diagnosis is essential.

II. State of art

Currently, several scientific research centers have developed equipment for the automatic diagnosis of tuberculosis [5] [6] [7]. These medical kits correct the imperfections made by highly experienced laboratory technicians; therefore, a faster and more reliable medical diagnosis is obtained in comparison to the conventional method. The equipments developed in most

We are grateful to the National Innovation Program for Competitiveness and Productivity of the Peruvian Ministry of Production for financing this project, and the Dos de Mayo Hospital for their collaboration in the validation of the system.

cases are focused on genetic detection techniques. The present section is a review of existing diagnostic equipment initiatives.

The MODS method (microscopic observation drug susceptibility assay) is a technique based on culture methodology. It is based on a Middlebrook 7H9 liquid medium culture monitored through the observation of an inverted microscope. The formation of ropes, which are visible before colonies, is observed in an average of 8 days. Technologically, this is an inexpensive option and applicable to high incidence and low resource countries [8]. In Peru, the Universidad Peruana Cayetano Heredia (UPCH) and the Universidad Nacional de Ingeniería (UNI) have developed a MODS Plate Reader system that can diagnose multidrug-resistant tuberculosis (MDR-TB) in less than fifteen seconds, after the appearance of ropes. The system includes an automated plate reader for microscopic observation of MODS cultures. The reader automatically handles the MODS plates, and after running the autofocus algorithm, digital images of microscopic structures characteristic of Mycobacterium tuberculosis are obtained. Once the image is obtained, the diagnosis is made through MODS pattern recognition software. In this way, the automated reader reduces labor time and handling Mycobacterium tuberculosis (MT) cultures by laboratory personnel [6].

On the other hand, Cepheid, Inc. developed the Xpert MTB/RIF technique, an automated test for the diagnosis of tuberculosis based on the detection of specific Koch's bacillus nucleic acids with cartridges in an organic sample. Unlike conventional nucleic acid amplification tests (NAAT), the Xpert MTB/RIF simplifies the identification of mycobacterial DNA by automating the 3 processes required for PCR-based molecular testing: extraction, amplification, and detection [1]. The Xpert MTB/RIF technique can detect MT complex and rifampicin resistance within 2 hours of test initiation, with minimal technical intervention. This method has very high sensitivity and specificity. Additionally, the sample reagent, used to liquefy the sputum, has potent disinfectant properties (with the ability to kill TB bacteria) and largely eliminates biosafety issues during the testing procedure. These features allow the technology to be taken from a reference laboratory and used closer to the patient. The Xpert MTB/RIF equipment requires an uninterrupted and stable electrical power supply, temperature control, and the cartridge modules' annual calibration [9]. The Xpert has a high specificity for diagnosing tuberculosis, which in the series evaluated by the National Chest Institute of Chile was 95% in respiratory samples and 94% in non-respiratory samples (the standard for positivity is culture). Considering the cost of this test, it should not be indicated to all patients who are requested smear microscopy, but only in those cases with a well-founded suspicion of tuberculosis in which the smear microscopy is negative. As it is a susceptible method, it is not indicated for patients who have had tuberculosis because it is a PCR technique. Any mycobacterial DNA residue left over from the past will be amplified many times and may give a false positive result [1]. The most commonly used device is the GX4, which has 4 modules (allowing 16-20 tests per 8-hour shift) and its discounted cost for developing countries is approximately $ 17 000 [10].

An alternative to ZN-stained baciloscopy is fluorescence microscopy with Auramine-Rhodamine staining, which achieves better sensitivity and specificity results than the ZN-stained method. However, due to the high cost required to acquire a fluorescence microscope (a microscope with a mercury vapor lamp), ZN staining continues to be chosen in developing countries [7].

Finally, the Pontificia Universidad Católica del Perú has been developing automatic Ziehl Neelsen staining systems for the standardized diagnosis of TB. In a previous version [11] [12] [13] [14], the systems considered staining but not microscopy and also did not have a closed loop temperature control.

III. Description of Hardware and Software

The tuberculosis diagnosis prototype has a staining module and a microscopy module for 4 samples simultaneously. Each module is described in the following lines.

A. *Staining module*

The module for simultaneous staining of four samples is shown in Figure 1, and the Figure 2 shows the block diagram of the complete system.

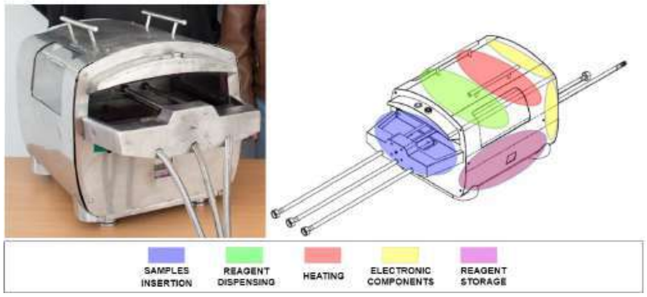


Fig. 1: Staining Module.

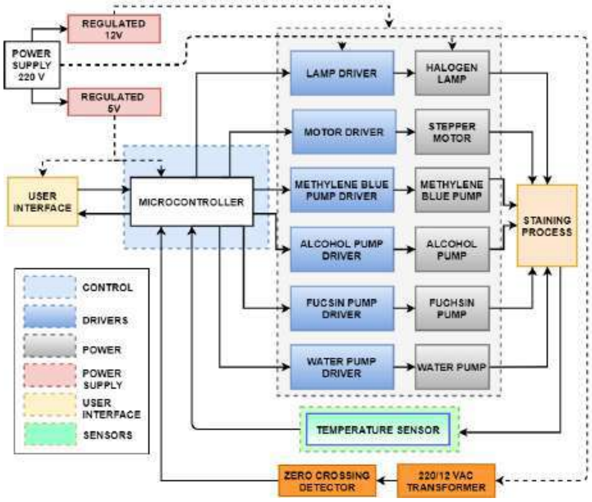


Fig. 2: Staining Module Block Diagram.

This staining module consists of a sample positioning sub-system, a reagent dispensing sub-system, and a heating sub-system. The stepper motor is part of the sample positioning sub-system. The pumps are part of the reagent dispensing sub-system. Finally, the halogen lamp and temperature sensor belong to the heating system. The microcontroller controls all sub-systems and the interaction between them.

The sample positioning sub-system receives the samples and moves them throughout staining process. The heating sub-system has a temperature control that maintains the temperature around 75 °C. The dispensing sub-system distributes the reagents to cover the samples and keep the surface tension on each sample.

*1) **Position sensor**:* The system has a limit switch sensor to define the zero position of the displacement. Figure 3a shows the schematic diagram of the conditioning circuit.

*2) **Temperature sensor, heating actuator, and actuator driver**:* The temperature sensor used is the LM35, working between room temperature and 100 °C. The sensor is located below the samples. Figure 3c shows the 10 mm distance between the reference sensor slide and the analysis slides. The sensor is in contact with the slide through the thermal paste.

The heating system has two OPALUX 220V-350V halogen lamps. Each lamp is inside a ceramic container and radiating heat on 2 samples. The system has an on-off control of the temperature with a margin of 5 °C error downwards to avoid exceeding the reference since the heating process is slow. In other words, when the reference is set at 80 °C, the lamps work at full power until the sensor reaches a temperature of 75 °C, where on-off control begins functioning. The control algorithm considers the error as the difference between the desired temperature and the measured temperature by LM35 sensor; the algorithm use a trigger angle of 1 ms (on) in case the error is greater than 5 °C; otherwise, there is no trigger angle (off).

The zero-crossing identification circuits is shown in Figure 3b and the firing angle control circuit for each lamp is as shown in Figure 4c. These circuits allow future implementations of more complex control algorithms.

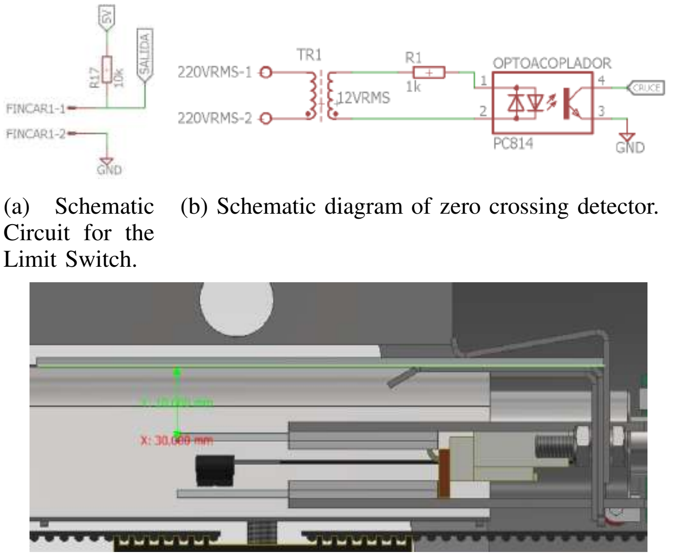
*3) **Pumps and pump driver**:* The four pumps used to dispense the three reagents and water are the RS-385, which are power by 12 V with a maximum current consumption of 0.5 A. The driver selected is a mosfet IRF530 with a 1N4004 diode and a pull-down resistor R1 of 100 KΩ. Figure 4a shows the driver circuit for each pump.

*4) **Stepper motor and stepper motor driver**:* A bipolar stepper motor, which is used in the sample displacement, can be controlled through micro-stepping and varying the rotation direction using the exciter A4988. Figure 4b shows the schematic diagram of the stepper motor driver for the movement on one axis. The stepper motors are energized with 12V.

The pulley that generates the linear displacement has 16 teeth separated by 2mm between teeth. One complete revolution of the motor shaft, 360° rotation, would be equivalent to 32mm. The motor is a NEMA 17 with 200 steps per revolution, equal to 1.8° each step, and as it is working at 1/8 step, this generates a linear movement of 0.02mm.

*5) **Power supply**:* The system is energized by a switching power supply RD65 of 5V at 6A and 12V at 3A.

*6) **Microcontroller and its software**:* The arduino platform was used but a custom board has been designed using the ATmega328P microcontroller [15]. The flow chart shown in the Figure 5 describes the whole staining process: inserting samples, fuchsin staining, heating samples, rinsing, dispensing acid alcohol, rinsing, methylene blue staining and the final rinse.

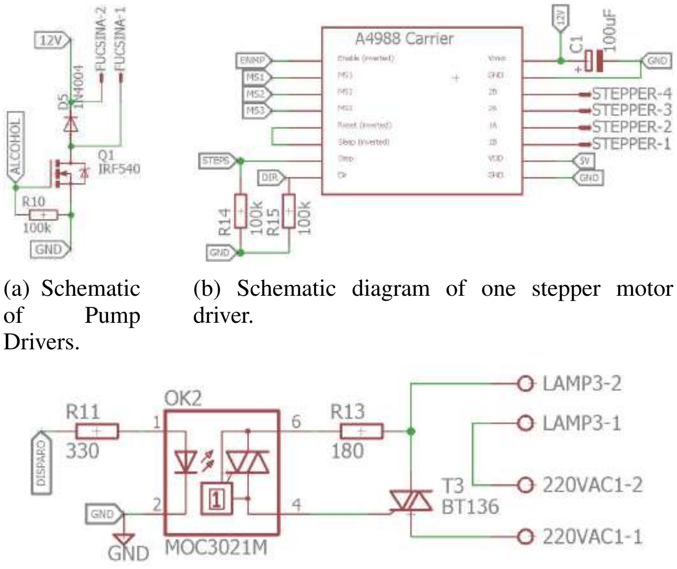


(a) Schematic Circuit for the Limit Switch.

(b) Schematic diagram of zero crossing detector.

(c) Temperature sensor position.

Fig. 3: Temperature sensor and conditioning circuits for the sensors.



(a) Schematic of Pump Drivers.

(b) Schematic diagram of one stepper motor driver.

(c) Trigger angle control circuit.

Fig. 4: Schematic diagrams of actuators’ drivers

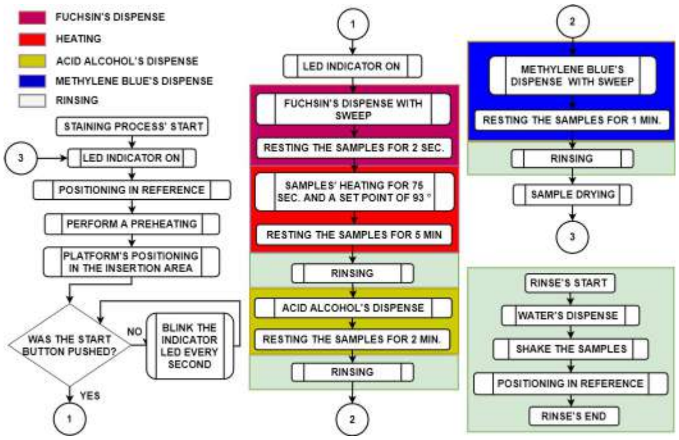

FUCHSIN'S DISPENSE
HEATING
ACID ALCOHOL'S DISPENSE
METHYLENE BLUE'S DISPENSE
RINSING
STAINING PROCESS' START
3
LED INDICATOR ON
POSITIONING IN REFERENCE
PERFORM A PREHEATING
PLATFORM'S POSITIONING IN THE INSERTION AREA
WAS THE START BUTTON PUSHED?
NO
BLINK THE INDICATOR LED EVERY SECOND
YES
1
1
LED INDICATOR ON
FUCHSIN'S DISPENSE WITH SWEEP
RESTING THE SAMPLES FOR 2 SEC.
SAMPLES' HEATING FOR 75 SEC. AND A SET POINT OF 93 °
RESTING THE SAMPLES FOR 5 MIN
RINSING
ACID ALCOHOL'S DISPENSE
RESTING THE SAMPLES FOR 2 MIN.
RINSING
2
2
METHYLENE BLUE'S DISPENSE WITH SWEEP
RESTING THE SAMPLES FOR 1 MIN.
RINSING
SAMPLE DRYING
3
RINSE'S START
WATER'S DISPENSE
SHAKE THE SAMPLES
POSITIONING IN REFERENCE
RINSE'S END

Fig. 5: General staining process flow diagram.

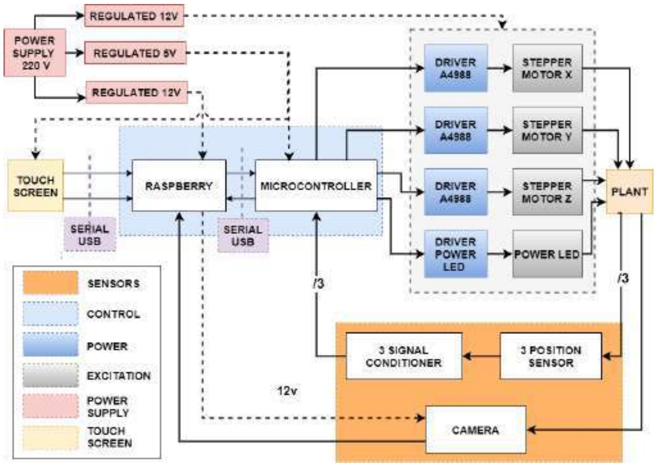


Fig. 7: Microscope Module Block Diagram.

B. Microscope module

This module's mechanical system is made up of three fundamental structures: the main support (housing); a mobile structure that contains and moves the samples, and the microscope optical structure. The touch screen, the electronic card and power supply, and other electronic components are in the main structure. The mobile structure carries out the movement of the platform through the X-Y–Z axes. The microscope structure is fixed, with the platform being the only one that moves on the Z-axis to perform the sample's focus.

Once illumination from the LED to the sample is activated, the camera and processing unit perform automatic focus while the platform changes its position relative to the microscope location. The module then acquires the focused image, processes it, and counts the bacilli.

Finally, the results are presented to the user through a graphical interface. Figure 6 shows the prototype's mechanical system and Figure 7 shows the block diagram.

*1) **Position sensor:*** The system has three limit switch sensors to identify the initial position in each axis, each of the sensors has a similar configuration and function as the sensor of the staining module.

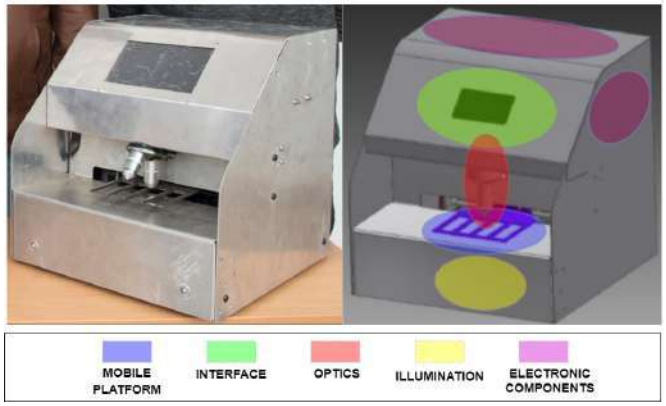


Fig. 6: Microscope Module.

*2) **Power LED and its driver:*** The system uses a power LED as the light source (12V, 0.9A, 6000-6500K). The driver considered is a mosfet IRF530 with a 1N4004 diode and a 100KΩ pull-down resistor R1. The ligth intensity of the LED is controlled by pulse width modulation (PWM).

*3) **Motors and their drivers:*** The microscope module uses three motors, each with the same hardware features as the staining module in order to generate movement in three axes. The X-Y plane's displacements do not need to be readjusted every time, so the microstep setting is fixed. In the case of Z axis motion, the micro-steps need to be readjusted according to the autofocusing algorithm's requirements.

The motor can rotate up to 1/16 of a 1.8° step. Ergo, it can rotate 0.1125°. Considering that the step between threads of a worm-screw is 2mm for the X and Y axes, 0.1125° corresponds to a linear displacement of 0.000625 mm. To move to the next field of 0.2 mm, the motor must rotate 36° or 20 steps. In the Z axis, there is a 1 mm between threads; therefore, a rotation of 0.1125° corresponds to 0.0003125mm.

*4) **User Interface-Touch Screen:*** The touch screen is a 7-inch Waveshare and it is controlled by the Raspberry platform. The development of the interface has been done with the Kivy Python framework. The Pony ORM object-relational mapper and the SQLite database manager have been used to store and search the records. The searches of the records can be carried out with information from any of the fields (Code, Diagnosis, Fields, Bacilli and Date). The user can easily initiate and terminate the process through this interface. Figure 8 shows the connections of the Raspberry to the Waveshare display. Figure 9 shows the main menu displayed on the screen.

*5) **Digital Camera:*** The system uses the C920E WEBCAM digital camera attached to the microscope tube through a structure created digitally in 3D printing. The camera sends the images to the Raspberry Pi 3 through the USB port. See Figure 8.

*6) **Processing Unit:*** It is composed of the Arduino Nano board based on an ATmega328 microcontroller and the Raspberry 3 board, a 64-bit ARM Cortex-A53 processor at 1.2

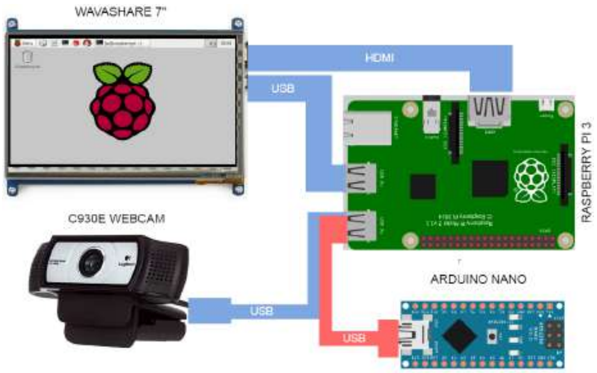


Fig. 8: Connection of the screen, the camera and the arduino nano to the raspberry.

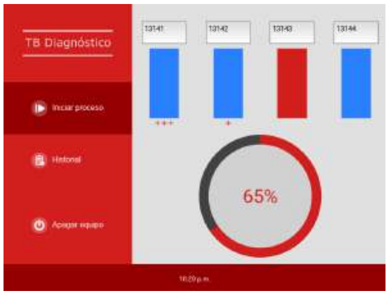


Fig. 9: User interface.

GHz. The first one is in charge of exciting the X, Y, and Z-axis motors and LEDs through its drivers. The second one is in charge of processing the acquired images; finding the best-focused image and making the samples' diagnosis, and controlling the user interface while communicating via USB so that the user can initiate, monitor, interrupt and terminate the session. Figure 8 shows the connections between them.

7) ***Electronic Software**:* Figure 10 shows the flow diagram of the automatic microscopy process.

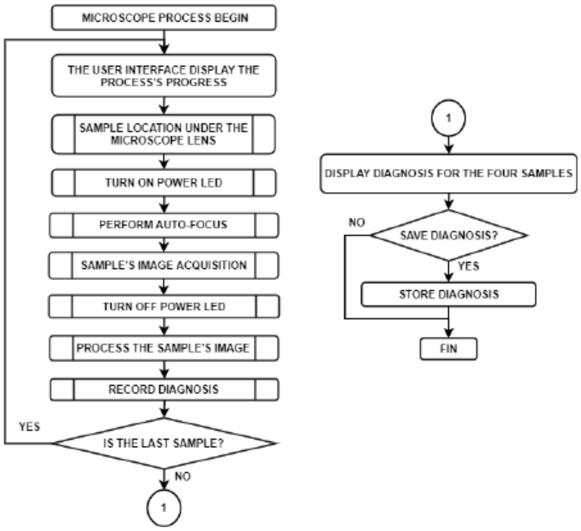

MICROSCOPE PROCESS BEGIN
THE USER INTERFACE DISPLAY THE PROCESS'S PROGRESS
SAMPLE LOCATION UNDER THE MICROSCOPE LENS
TURN ON POWER LED
PERFORM AUTO-FOCUS
SAMPLE'S IMAGE ACQUISITION
TURN OFF POWER LED
PROCESS THE SAMPLE'S IMAGE
RECORD DIAGNOSIS
YES
IS THE LAST SAMPLE?
NO
1
1
DISPLAY DIAGNOSIS FOR THE FOUR SAMPLES
NO
SAVE DIAGNOSIS?
YES
STORE DIAGNOSIS
FIN

Fig. 10: Microscope Module Flow Diagram.

IV. RESULTS AND DISCUSSION

The following lines present the factors that influence the performance of both modules.

A. *Sample inclination:*

The samples are clamped by tweezers, which break the reagents' surface tension when heated. The heating of the tweezers causes the fuchsin to flow and fall off the samples. A proper tilt allows the greatest amount of fuchsin to concentrate away from the clamp during heating. Hovever, this tilt pools the reagents to one side of their container. This improper storage prevents the actuator from pumping the reagents as the pumps are located on the opposite side, causing the reagents to remain in the containers.

B. *Fuchsin dispensing tubing:*

The pneumatic tubing used for fuchsin dispensing swells, clogs, and even releases particles. It is important to mention that the dispenser (not the tubing) of channel 1 became clogged after 40 sequences. Figure 11a shows the tubing after more than 15 sequences.

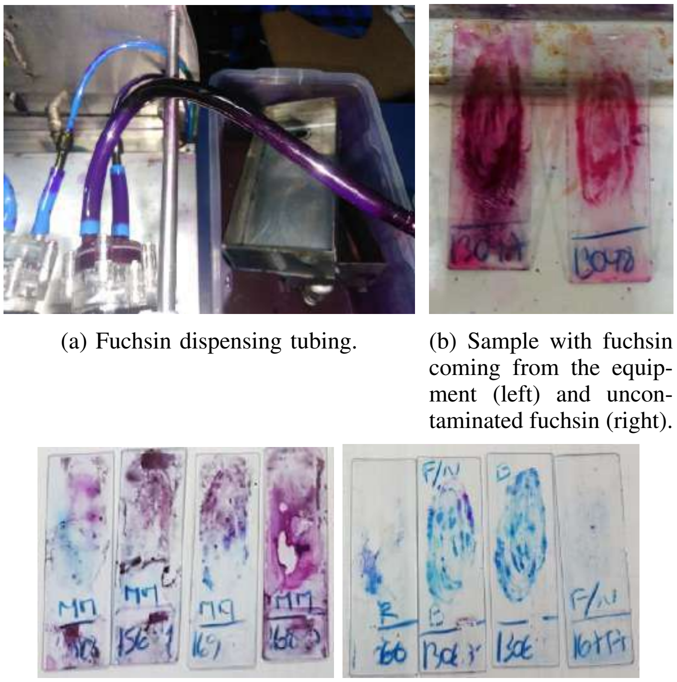
Figure 11b shows that the fuchsin coloration of the sample on the left is different from the sample's normal coloration on the right. This coloration is evidence of fuchsin contamination.

C. *Heating temperature*

The heating temperature can cause the samples to be dry and/or dark in color. By changing the equipment's inclination, the halogen lamps get closer to the samples, making it necessary to reduce the temperature. In the horizontal position, the reference temperature, based on the medical technologist's recommendation, was set at 100°C. In contrast, when tilted forward (due to the equipment structure's misalignment), the appropriate reference temperature was set at 93°C. Figure 11c (left) shows samples with an excess temperature, while in Figure 11c (right), the temperature is adequate.

With the temperature control active and the slides covered with water, the evolution of the reference sensor's temperatures on each of the 4 samples was validated considering several reference temperatures. The thermocouples of 4 Fluke 179 devices with a temperature sensor were used to measure each sample's temperature. The values displayed by the multimeters have been extracted in a synchronized way using image processing. The software used is a modification of the code presented in [16]. The figure 12a and 12b show the results for reference temperatures of 60 °C and 70 °C.

Subsequently, the ON / OFF control was validated with a sample covered with fuchsin; the result obtained is shown in Figure 13a. The Fluke 179 thermocouple multimeter was used to measure the temperature on the fuchsin-coated slide and the Fluke 175 multimeter was used to indirectly measure the temperature of the LM35 (via the sensor output voltage). It can be seen that the temperature difference is 10 ° C and the control is carried out properly. It was observed that the initial temperature of the LM35 does not significantly affect the final value of the sample's temperature; however, a preheating is

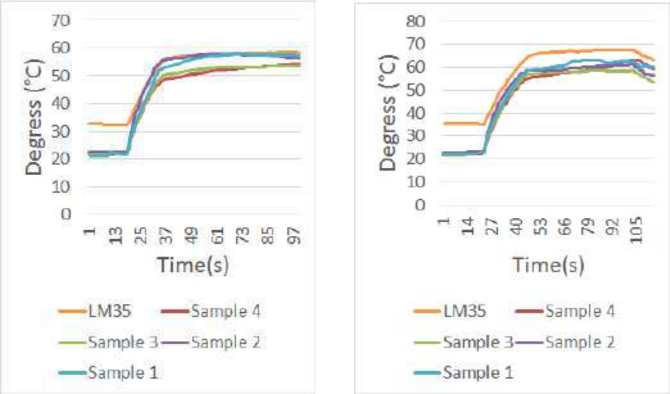


(a) Fuchsin dispensing tubing.

(b) Sample with fuchsin coming from the equipment (left) and uncontaminated fuchsin (right).

(c) Samples stained with excess heating (Left) and Samples stained with adequate heating (Right).

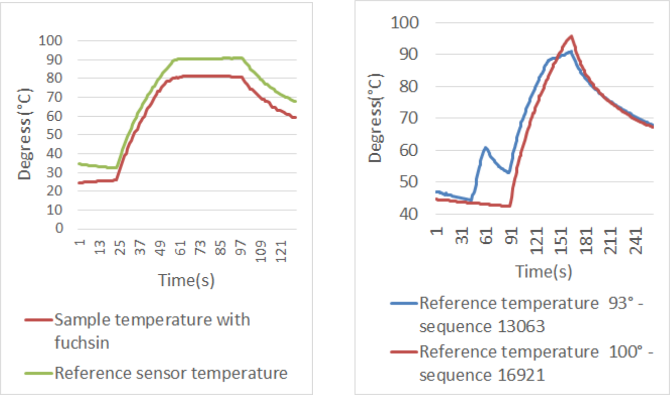
Fig. 11: Condition and influence of the materials and the mobile platform's inclination in the staining process.



(a) Response with 60 °C reference.

(b) Response with 70 °C reference.

Fig. 12: LM35 temperature sensor response and samples' temperature response with water.



(a) LM35 sensor temperature (set point of 93°) vs temperature in the sample with fuchsin.

(b) Reference sensor temperature during heating and preheating.

Fig. 13: Temperature control analysis.

established prior to the entry of samples to define the initial temperature condition of the reference sensor, this preheating does not change the traditional staining process.

Figure 13b shows the reference sensor values during heating and preheating for two samples in a single graph. The temperature control result is for the sample 13063 (set point of 93 °C) and for the sample 16921 (set point of 100 °C); the ambient temperatures were 21 °C and 21.6 °C respectively. For the 93 °C references, the heating position is given by setting #4 of Table I, while for the 100 °C references, it is given for setting #3. The heating time was seventy-five seconds. For the case of sequence 13063, there was a calibration preheating of the equipment, which had a reference of 60 °C.

D. Sample positions

Several adjustments have been made regarding the position of the limit switch sensor. The most relevant ones during calibration are shown in Tables I and II.

Regarding the changes between setting 1 and 2, it was observed that the dispensing was excessive and there was a waste of fuchsin; therefore, to avoid misuse of fuchsin, the sweep displacement is reduced from 3 cm to 2.6 cm. Besides, a small offset of 1 mm is added to the initial dispensing position to move it away from the clamps. The change between setting 3 and 4 is due to the fact that a two-millimeter adjustment was required for the halogen spotlight to align with the samples.

Table II shows the amount of fuchsin dispensed in ml for settings #1 and #4. Setting #5 is not included because fuchsin is lost during heating, causing the samples not to be completely coated. Figure 14 shows the coating of the slides after heating with setting #4.

TABLE I: Spindle positions in millimeters with respect to the location of the limit switch sensor.

Adj.	Heating position (mm)	Fuchsin dispensing end position (mm)	Initial fuchsin dispensing position (mm)	Initial position for sample entry (mm)
1	6	110	140	214
2	6	114	140	214
3	6	113	139	214
4	4	113	139	214

TABLE II: Variation of ml of fuchsin with respect to the dispensing interval.

Adj.	Dispensing interval (steps)	Dispensing interval (mm)	Sample 1 (ml)	Sample 2 (ml)	Sample 3 (ml)	Sample 4 (ml)	Total (ml)
1	1500	30	3.6	5	3.8	4.6	17
4	1250	25	3	3.55	3	4.5	14.1
5	950	19	2.1	2.75	2.25	3.1	10.2

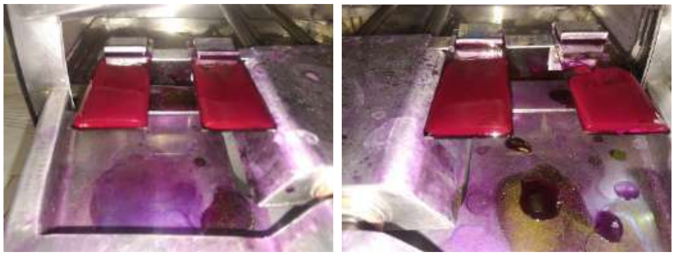


Fig. 14: Fuchsin coating of samples after heating.

E. Microscope module

The microscope module performs a proper displacement for the 4 samples; the X and Y axis have a minimum step displacement equivalent to 0.01 mm. The Z axis enables autofocus with a minimum displacement of 0.0003125mm or 16 microsteps. To implement the microscope module, a commercial one was purchased, and the part corresponding to the optics was cut. This reduced the robustness of the module. Therefore, the observed images have a small vibration that is corrected by the autofocus and the image capture software.

V. CONCLUSIONS AND FUTURE WORK

The staining module can stain 4 samples simultaneously in an automatic way. Regarding the reagent dispensing, it is concluded that the dispensing for all channels as a set can be controlled by varying the interval of dispensing positions. However, the amount dispensed is not uniform in all dispensers. Regarding temperature control, it is found that the temperature in the sample can be controlled from a reference temperature sensor. The mechanic is essential so that a proper clamping of the samples is crucial for the prototype's proper functioning; the inclination must be uniform so that it is the same for the four samples. Finally, the structures' material is also important; it has been proven that metal containers and pneumatic pipes are not appropriate.

The microscope module performs the desired displacements in all its axes, having a higher sensitivity in the Z-axis. However, this module can be improved; two of the most important things to improve are the cancellation of the vibrations that are observed in the samples caused by shocks in the base that supports the module, or when driving its motors; and the improvement of the coupling mechanism between the camera and the microscope.

For future work, it is proposed to evaluate the use of a temperature sensor that does not require contact and moves with the samples, in addition to considering additional temperature and humidity sensors for both the environment and the equipment. This information could be stored automatically in a local or remote database. To establish the initial temperature conditions in each sequence, we propose to evaluate the possibility of including a preheating of the samples before dispensing fuchsin or placing a fan.

Finally, it is possible to perform more detailed studies. That is to say, to carry out a study that allows determining the effectiveness of the equipment for certain types of samples: urine, feces, cerebrospinal fluid, among others.

Acknowledgment

Thanks to the National Innovation Program for Competitiveness and Productivity of the Peruvian Ministry of Production for funding this project, and the Dos de Mayo Hospital for their collaboration in the validation of the system.

References

[1] P. Vallejo V., J. C. Rodrigues D., A. Searle M., and V. Farga C., "Ensayo Xpert MTB/RIF en el diagnÃde tuberculosis," *Revista chilena de enfermedades respiratorias*, vol. 31, pp. 127 – 131, 06 2015. [Online]. Available: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0717-73482015000200010&nrm=iso

[2] W. H. Organization, *Global tuberculosis report 2018*. World Health Organization, 2018.

[3] ——, *Global tuberculosis report 2020*. World Health Organization, 2020.

[4] Q. N. V. L. Asencios, Luis, "Procedimientos para el conrtol de calidad externo de baciloscoía para el diagnóstico bacteriológico de la tuberculosis." *Ministerio de Salud del Peru*, 2014. [Online]. Available: https://bvs.ins.gob.pe/insprint/CINDOC/pub_ins/2014/procedimiento_control_calidad_externo_baciloscopia.pdf

[5] UNITAID, "Diagnostics Technology Landscape 5th Edition, May 2017," *WHO press*, no. May, pp. 1–90, 2017. [Online]. Available: https://unitaid.eu/assets/2017-Unitaid-TB-Diagnostics-Technology-Landscape.pdf

[6] G. Comina, D. Mendoza, A. Velazco, J. Coronel, P. Sheen, R. GILMAN, D. MOORE, and M. Zimic, "Development of an automated mods plate reader to detect early growth of mycobacterium tuberculosis," *Journal of microscopy*, vol. 242, pp. 325–30, 06 2011.

[7] B. N. Medina Leandro, "Diseno de un sistema automatico de tincion de cuatro muestras de esputo en simultaneo para el diagnostico de tuberculosis," 2018.

[8] J. Gonzàlez-Martin, "Microbiología de la tuberculosis," *Seminarios de la Fundación Española de Reumatología*, vol. 15, no. 1, pp. 25–33, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1577356614000025

[9] K. R. Steingart, I. Schiller, D. J. Horne, M. Pai, C. C. Boehme, and N. Dendukuri, "Xpert® mtb/rif assay for pulmonary tuberculosis and rifampicin resistance in adults," *Cochrane Database of Systematic Reviews*, 2014.

[10] C. M. Denkinger, S. G. Schumacher, C. C. Boehme, N. Dendukuri, M. Pai, and K. R. Steingart, "Xpert mtb/rif assay for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis," *European Respiratory Journal*, vol. 44, no. 2, pp. 435–446, 2014. [Online]. Available: https://erj.ersjournals.com/content/44/2/435

[11] L. Hipolo, W. Carrera, and M. Cataño, "Design and development of an automatic sequential sputum sample staining system to improve efficiency in the diagnosis of tuberculosis in developing countries," in *2019 7th International Engineering, Sciences and Technology Conference (IESTEC)*, 2019, pp. 526–531.

[12] Y. A. Pandzic Saba, "Diseño de un sistema automático de preparación de muestras de esputo para el diagnóstico de TBC," Jun. 2014. [Online]. Available: http://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/5357

[13] A. M. Rodríguez Chávez, "Validación del preparador automático de muestras de esputo para el diagnóstico de TBC," Jul. 2015. [Online]. Available: http://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/6144

[14] R. F. F. Avila De La Cruz and L. A. Ramírez Suárez, "Diseño y desarrollo de un prototipo preparador de siete muestras biológicas basado en la tinción de Ziehl-Neelsen para baciloscopía," May 2017. [Online]. Available: http://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/8552

[15] "From Arduino to a Microcontroller on a Breadboard." [Online]. Available: https://www.arduino.cc/en/Tutorial/BuiltInExamples/ArduinoToBreadboard

[16] A. Rosebrock, "Recognizing digits with OpenCV and Python," *PyImageSearch*, Feb. 2019. [Online]. Available: https://www.pyimagesearch.com/2017/02/13/recognizing-digits-with-opencv-and-python/

The Modified Proportional Integral Controller for the BLDC Motor and Electric Vehicle

Md. Rezanul Haque
Department of Electrical and Electronic Engineering
Independent University, Bangladesh
hrezanul@gmail.com

Shahriar Khan
Department of Electrical and Electronic Engineering
Independent University, Bangladesh
skhan@iub.edu.bd

Abstract–The recent dramatic growth in electric vehicles has brought renewed attention to their motors and power electronics. Of the motors used for electric vehicles, the DC motor and the three phase induction motor are well-studied. Developing countries have created their own solutions for electric vehicles, especially with the Brushless DC Motor (BLDC). The less investigated three phase BLDC motor presents challenges for control of the transient and steady-state response, overshoot, rise time, settling time, etc. A lapse in the control can cause the system to become unstable and reduce the life of components. In this paper, a BLDC motor for electric vehicles is studied with the Modified Proportional-Integral (MPI) controller and logic gates for the drives. The design allows reduction of overshoot to an estimated 5 % and the settling time to 2 seconds This study is expected to contribute to the applications of the BLDC, especially for electric vehicles.

Keywords–Electric vehicles, BLDC, Motor, Proportional Integral, PI controller, Logic controller, Electronics, Power electronics, Motor control, Overshoot, Transient.

I. Introduction

In the growing market of electric vehicles, the motors used include the DC motor, the synchronous motor, the induction motor, and the Brushless DC motor (BLDC). Of these, the DC motor, the synchronous motor and the induction motor are well-developed technologies. In comparison, the BLDC is a relatively new type of motor, made popular largely by advances in power electronics. To meet the growing market for electric vehicles, especially in developing countries, the recent trend has been to use the Brushless DC Motor. BLDC motors use an inverter to convert DC to AC which runs an AC motor. Compared to DC motors of the same power, the three phase BLDC motor is lighter with high torque to weight, requires less maintenance and can be better controlled [1-6].Unlike the DC motor, the BLDC motor does not produce any spark and noise [7] and is convenient to use in critical environments. BLDC motors are becoming popular in Electric vehicles, drones, residential and industrial applications [8, 9].

In electric vehicles, it is desirable to operate the BLDC motor with less overshoot and less settling time. For this purpose, Anti Wind PI (Proportional Integral) control, Artificial Neural Network and Fuzzy logic-based controllers have been proposed. All are mostly complex to implement, have large overshoots, and are otherwise costly [10,11].

This paper presents an inexpensive, robust, and easy-to-implement BLDC motor controller for Electric Vehicles using Modified Proportional Integral (MPI) and digital logic control.

II. Literature Review

Problems with PI control include high overshoot, and settling time. Kiron avoided the PI controller and proposed Auto tuned PID controller, with which he achieved 50% overshoot (at 1000rpm and 0.18N-m torque) [12]. Arulmozhiyal proposed Fuzzy PID controller and achieved 3.6% overshoot at 1500rpm with 5N-m load [13]. Neethu proposed a fuzzy controller and achieved 38% overshoot at 1000rpm [14]. As the overshoot is high, the BLDC motor takes high inrush current to initialize, which may reduce the life of the motor.

III. Theory of Operation

The proposed block diagram of the BLDC motor controller using Modified Proportional Integral (MPI) method is shown in Fig.1. The MPI block continuously checks the speed sensor, current/voltage sensor, and the desired speed and it tries to find the error using the MPI method. The logic operation selects the three-phase inverter gating sequence with the help of the hall sensor and the MPI output. Then the three-phase inverter controls the BLDC motor.

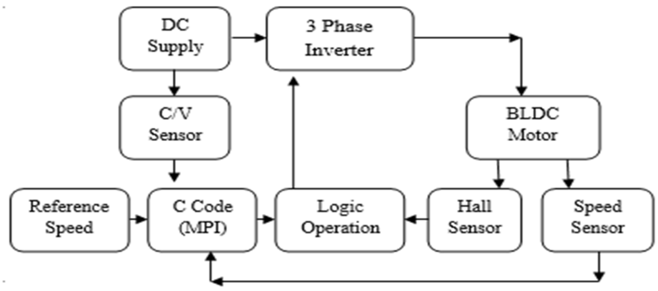


Fig.1: Block diagram of proposed system

IV. Controller Design

A. *MPI Controller*

The Modified Proportional Integral (MPI) controller is shown in Fig.2. To find the error between desired speed and actual speed, the MPI is required. The desired speed is sampled and further processed according to predefined tuned value. The desired speed which is instantaneous and coming from user is defined as

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

$$R'(t) = xu(t) \tag{1}$$

In sample and process block, the instantaneous desired speed is modified on the basis of equation 2.

$$R(t) = x\,r(t + 0) - x\,r(t - m) \tag{2}$$

Where x is the given desired speed and m is the desired settling time. The desired speed is then compared with actual speed and fed to PI block. The transfer function of a PI controller can be defined as.

$$\text{Output, } y(s) = e(s) \times (k_P + \frac{Ki}{S}) \tag{3}$$

The gain can be calculated as.

$$y(s)/e(s) = (k_P + \frac{Ki}{S}) \tag{4}$$

In time domain the transfer function of PI is.

$$y(t) = k_P\, e(t) + k_i \int e(t)dt \tag{5}$$

After solving trapezoidal rule, the overall expression becomes.

$$y(t) = K_P\, e(t) + K_i[\frac{Ts}{2}[e(t) + 2e(t-1)] + Past_{int}] \tag{6}$$

where $e(t)$ is the current error and Ts is sampling time. This equation was implemented in a proposed MPI controller in software using C code.

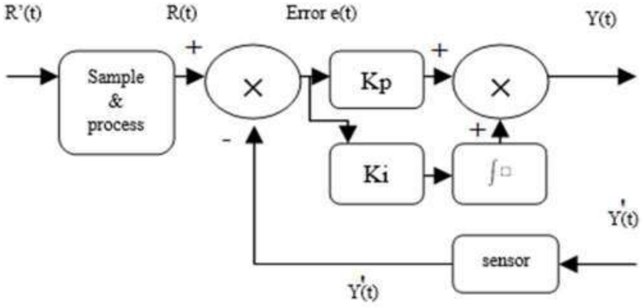


Fig.2: The algorithm of Proposed MPI block used in Software using C code.

B. *Logic Operation Controller*

For this project, three phase six switch inverters have been assigned. To control the switch gating sequence, the logic operation is required. Typically, the BLDC motor consists of three Hall effect sensors which give the information of rotor position. When a rotor magnet passes the Hall sensor, it provides a high signal for the North pole and low signal for South pole. The communication sequence would be $2^3 = 8$, which are 000, 100, 011, 110, 010, 101, 001 and 111. But as 000 and 111 are odd combinations, the logic operation will ignore them. Based on these values the emf can be calculated as.

$$EMF\ sa = Hall\ A\text{-}\ Hall\ B \tag{7}$$

$$EMF\ sb = Hall\ B\text{-}\ Hall\ C \tag{8}$$

$$EMF\ sc = Hall\ C\text{-}\ Hall\ A \tag{9}$$

Where $Hall\ A = 1$, $Hall\ B = 0$, $Hall\ C = 0$ for first sequence 100 and the EMF will be +1, 0 and -1 for sa, sb, and sc respectively. The software built in BLDC motor block gives the rotor position as bipolar communication (-1,0,+1) for each phase. Based on this, the inverter switch gating pulse must be decided to achieve the required terminal voltage which is similar to emf.

To achieve the three-phase terminal voltage, two switches must be on at a time for one phase, according to the emf sequence. When any upper switch conducts, the following phase becomes positive. When the lower switch conducts, the following phase becomes negative. When both switches open the following phase becomes zero. To avoid short circuit, no two switches in a line should conduct at a time. For instance, the first sequence (100) phase Y, B and G must receive positive voltage, zero voltage and negative voltage. To achieve this, switch 1 and switch 4 must be triggered. Based on this, table I and II are drawn below.

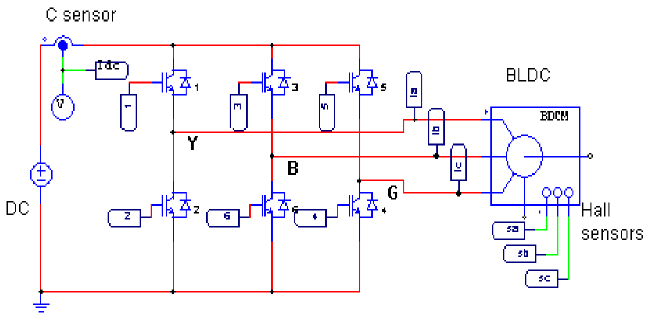


Fig.3: Three phase inverter with BLDC motor.

TABLE I. LOGIC LEVEL GENERATION FOR FORWARD ROTATION CALCULATED FROM ABOVE EQUATIONS

EMF			Forward Rotation					
sa	sb	sc	Switch					
Y	B	G	1	2	3	4	5	6
+1	0	-1	1	0	0	1	0	0
-1	0	+1	0	1	0	0	1	0
0	+1	-1	0	0	1	1	0	0
-1	+1	0	0	1	1	0	0	0
+1	-1	0	1	0	0	0	0	1
0	-1	+1	0	0	0	0	1	1

To run the motor in reverse direction the inverter switching pattern changes and Table II must follow in logic operation.

TABLE II. LOGIC LEVEL GENERATION FOR REVERSE ROTATION CALCULATED FROM ABOVE EQUATIONS.

EMF			Reverse Rotation					
sa	sb	sc	Switch					
Y	B	G	1	2	3	4	5	6
-1	0	+1	0	1	0	0	1	0
+1	0	-1	1	0	0	1	0	0
0	-1	+1	0	0	0	0	1	1
+1	-1	0	1	0	0	0	0	1
-1	+1	0	0	1	1	0	0	0
0	+1	-1	0	0	1	1	0	0

Based on these two tables, the logic operation block drives the BLDC motor using inverter.

V. SIMULATION RESULT AND DISCUSSION

After completing the logic operation table and the MPI tune value, a complete circuit was developed in the software to analyze the proposed system. Fig. 4. shows the complete circuit diagram of the BLDC motor controller with mechanical load of 3 N-m. The inverter block consists of six switches as shown in fig. 3. The logic operation maintains Table I and Table II, based on the reference input. The logic operation and MPI block consist of several C blocks written in C code. The MPI block consist of MPI equations explained in section IV.A and its monitor's instantaneous change in any current or voltage sensors involve speed, torque, direction bit and reference speed. The MPI block tunes the speed error, and PWM is generated at 10 kHz switching frequency, which is then fed to logic operation and generates the gating sequence. Fig.5 summarizes the overall operation of the proposed system.

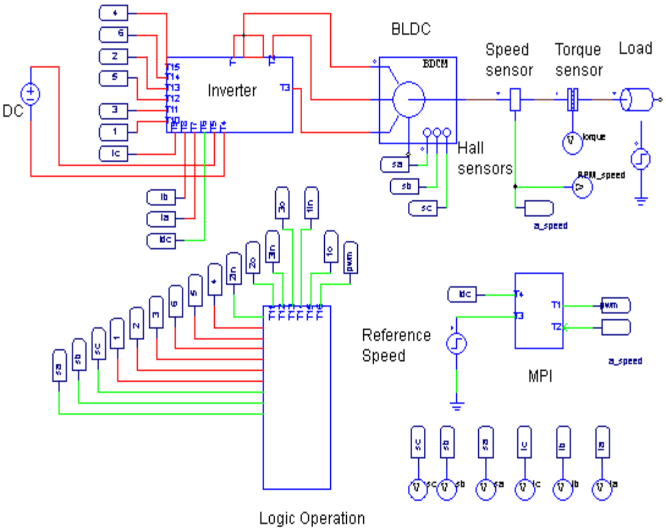


Fig.4: Circuit diagram of motor controller.

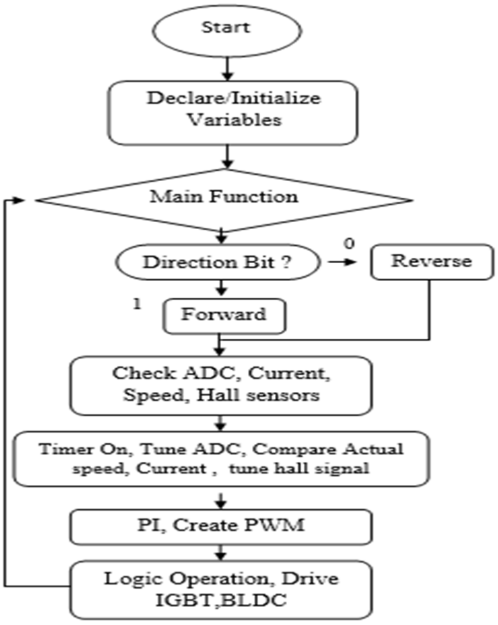

Start
Declare/Initialize Variables
Main Function
Direction Bit ?
0
Reverse
1
Forward
Check ADC, Current, Speed, Hall sensors
Timer On, Tune ADC, Compare Actual speed, Current , tune hall signal
PI, Create PWM
Logic Operation, Drive IGBT,BLDC

Fig.5: Main loop for proposed motor controller.

A. *Forward Operation*

Fig. 6 shows the forward rotation emf generated from BLDC motor. The emf is bipolar communication (-1,0,1) which contains the rotor position, based on which the logic operation follows Table I and generates six-switching pulse (Fig. 7).

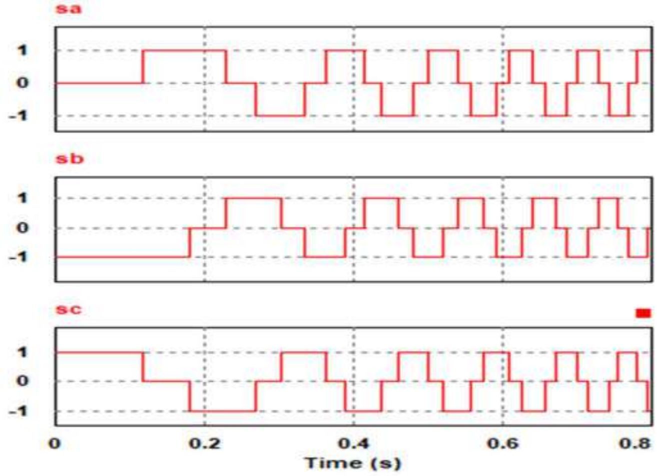


Fig.6: Simulated Forward rotation emf generation from BLDC motor which perfectly matches with calculated table I.

Fig. 6 and Fig.7 show that with the increase in rotor speed, the emf generated from the BLDC motor increases faster and the logic operation also generates six pulses faster which matches with calculated Table I.

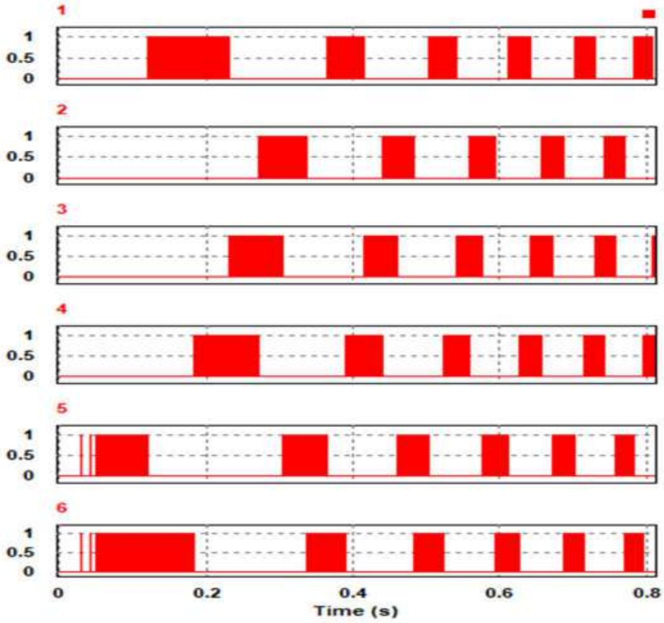


Fig.7: Simulated Gate signal generation from Logic Operation for forward rotation which perfectly matches with calculated Table I.

B. *Reverse Operation*

Fig. 8 shows the reverse rotation emf generated form the BLDC motor. This contains the rotor position and based on the logic operation, follows Table II, creating the six-switching pulse (Fig. 9). The switching sequence is visibly different from the forward rotation sequence.

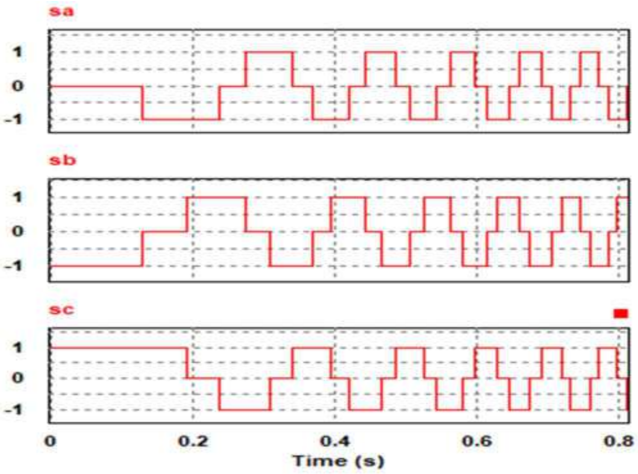


Fig.8: Simulated reverse rotation emf generation from BLDC motor which perfectly matches with calculated table II.

Fig.8 and Fig.9 show that the emf and the six pulses are generated faster as the rotor speed increases. This matches with the calculations in Table II.

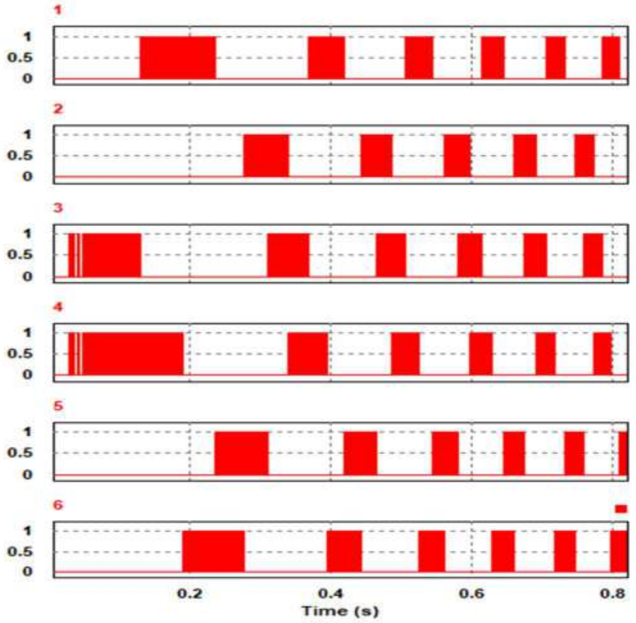


Fig.9:Simulated Gate signal generation from Logic Operation for reverse rotation which perfectly matches with calculated table II.

After successfully switching the three-phase inverter, the output current appears as shown in Fig. 10. The graph shows that there is no inrush current. The output current of the three phase inverter changes according to speed and torque.

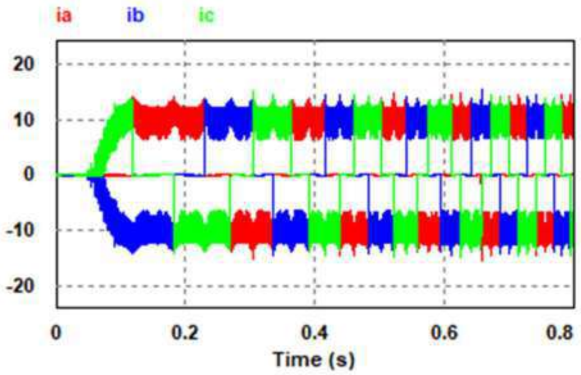


Fig.10: Simulated Inverter output Three phase current.

For forward rotation of 500 rpm and 1000 rpm with 3N-m load, the overshoot was 9.38% and 4.81% respectively, and the settling time was 1.01s and 2s respectively (Fig. 11 and Fig. 12).

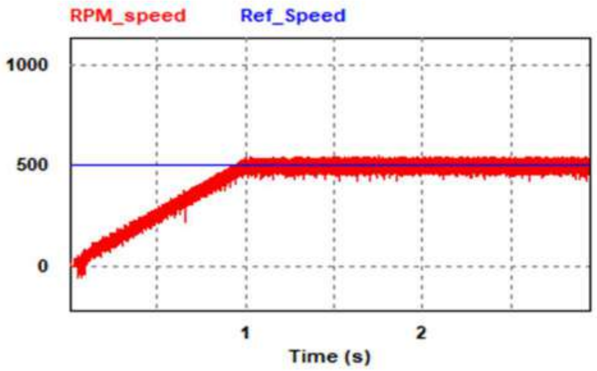


Fig.11: BLDC motor forward rotation at 500 RPM .

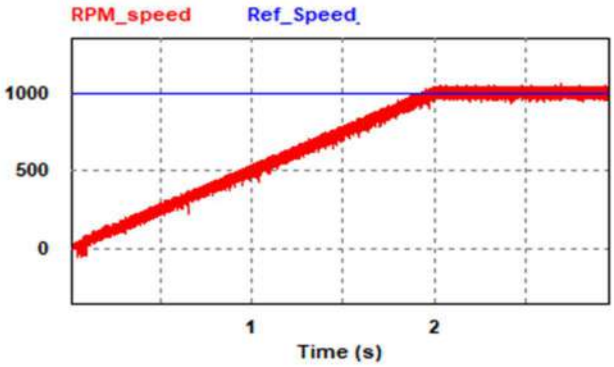


Fig.12: BLDC motor forward rotation at 1000 RPM .

For Reverse rotation of 500rpm and 1000rpm with 3N-m load, the overshoot was 8.12% and 5.18% respectively and the settling time was 1.18 s and 2.09 s respectively (Fig. 13 and Fig. 14).

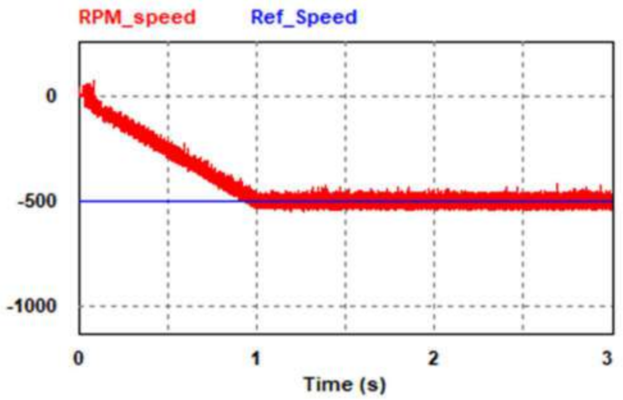


Fig.13: BLDC motor reverse rotation at 500 RPM .

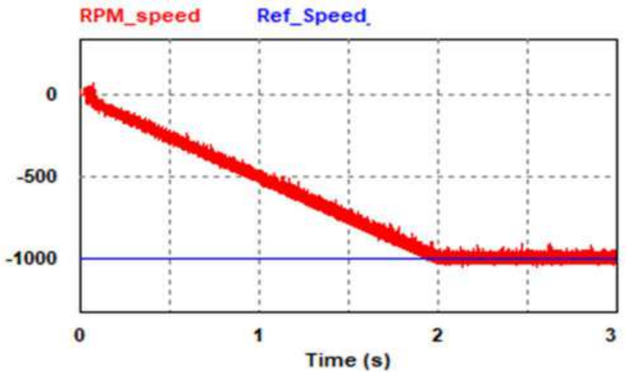


Fig.14: BLDC motor reverse rotation at 1000 RPM .

VI. Conclusion

This paper attempted to use the modified PI control for the electric vehicle's Brushless DC motor's speed and direction. The desired speed is further processed based on predefined tuned value. The PI controller calculates the error between the desired and actual speed and feeds to the logic operation. Accordingly, the proposed system is able to solve the high overshoot in transient response (1000rpm with 3N-m), which will be convenient for electric vehicles.

It is hoped that this study will contribute to applications of the BLDC motor in the electric vehicle industry.

References

[1] J. Karthikeyan, R. Dhana Sekaran, “Current Control of Brushless DC Motor Based on a Common DC Signal for Space Operated Vehicles,” in Electrical Power and Energy Systems, volume 33, pp. 1721-1727, August 2011.

[2] P. Pillay, R. Krishnan, “Modeling, Simulation and Analysis of Permanent-Magnet Motor Drives Part II: The Brushless DC Motor Drive”, *IEEE Transaction on Industry Applications*, September 2008, pp. 274-279.

[3] M. F. Bhuiyan, M. Rejwan Uddin, Z. Tasneem, M. Hasan and K. M. Salim, "Design, Code Generation and Simulation of a BLDC Motor Controller usuuing PIC Microcontroller," *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, Bhubaneswar, India, 2018, pp. 1427-1431

[4] S. Das, M. R. Haque and M. A. Razzak, "Development of One-kilowatt Capacity Single Phase Pure Sine Wave Off-grid PV Inverter," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 774-777,

[5] Shahriar Khan, Semiconductor Devices and Technology, Third Edition, ISBN: 978-094-33-5983-4, S. Khan, Dhaka, Bangladesh, June 3, 2018,.

[6] Shahriar Khan, Systems and Control, Third Edition, ISBN 978-984-33-3561, S. Khan, Dhaka, Bangladesh, May 2012.

[7] K. Tabarraee, J. Iyer, S. Chiniforoosh and J. Jatskevich, "Comparison of brushless DC motors with trapezoidal and sinusoidal back-EMF," 2011 24th Canadian Conference on Electrical and Computer Engineering(CCECE), Niagara Falls, ON, 2011, pp. 000803-000806,

[8] M. Bhuiya, N. Sakib, M. Uddin, K. M. Salim, "Experimental Results of a locally developed BLDC Motor Controller for electric tricycle.". 10.1109/ICASERT.2019, pp. 1-4.

[9] M. R. Haque, S. Das, M. R. Uddin, M. S. Islam Leon and M. A. Razzak, "Performance Evaluation of 1kW Asynchronous and Synchronous Buck Converter-based Solar-powered Battery Charging System for Electric Vehicles," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 770-773,

[10] M. Tariq, T. K. Bhattacharya,. N. Varshney, R. Dhilsha.. "Fast Response Antiwindup PI Speed Controller of Brushless DC Motor Drive: Modeling, Simulation and Implementation on DSP," Journal of Electrical Systems and Information Technology. 3. 10.1016/j.jesit.2015.11.008. (2016)

[11] C. Bohn and D. P. Atherton, "An analysis package comparing PID anti-windup strategies," in IEEE Control Systems Magazine, vol. 15, no. 2, pp. 34-40, April 1995.

[12] K. Gadekar, S. Joshi and H. Mehta, "Performance Improvement in BLDC Motor Drive Using Self-Tuning PID Controller," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 1162-1166,

[13] R. Arulmozhiyal and R. Kandiban, "An intelligent speed controller for Brushless DC motor," 2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA), Singapore, 2012, pp. 16-21,

[14] U. Neethu and V. R. Jisha, "Speed control of Brushless DC Motor: A comparative study," 2012 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Bengaluru, 2012, pp. 1-5,

Precision Gesture-based Inputs for Smartphone with Two Consecutive Simple Gestures

Tomoaki Amiya
Osaka Prefecture University
Osaka, Japan
sab01005@edu.osakafu-u.ac.jp

Ryo Katsuma
Osaka Prefecture University
Osaka, Japan
katsuma@cs.osakafu-u.ac.jp

***Abstract*—In recent years, smartphones and other mobile computers have become even smaller. Simultaneously, the number of buttons for user input has decreased, and their size has shrunk, which lowers the operability of user input. In this situation, gesture-based input is expected to be the new input method. gesture-based inputs have advantage in that it can work irrespective of device size Furthermore, there is no need to even look at the screen. Accelerometers equipped in smartphones are the primary tool for recognizing gesture-based inputs. However, such techniques are prone to misrecognitions. One factor contributing to this problem is the high number of gesture candidates. In response, we propose an input recognition method that uses a combination of two far simpler gestures types: trigger gesture (TG) and main gesture (MG). TG determine the start or content of the operation. MG provide the operational detail. In this paper, we test this concept on a music player app installed on a smartphone, which can be used while walking. The proposed method is found to improve recognition accuracy by 50%, assuming five types of input requests.**

***Index Terms*—Sensors and Systems, Signal Detection and Processing, Mobile Computing,**

I. Introduction

In recent years, smartphones and other mobile computers have become smaller, and the convenience of carrying them has been improved. However, the operability of user input has decreased with the number of buttons and screen size. Commonly, while walking or jogging, sound data are used for enjoying music, maintaining a pace, keeping rhythm, and absorbing news information or stories. So, when we want to operate something then, it has become necessary to touch the small screen while staring it, which can dangerously distract uses from their surroundings [1]. This has become a pertinent social problem. In Japan, according to the Tokyo Fire Department, from 2015 to 2019, 211 people have been transported by emergency because of related accidents [2]. This is expected to increase.

Therefore, owing to the ubiquity of smartphones, it has become necessary to operate the interface without looking at the screen. **Gesture-based Inputs** relies on accelerometers equipped in most smartphones. Accelerometers can detect hand motion and classify the types of movement. Gesture-based input is already well-implemented on game consoles, and the intuitive operation using mobile devices has been well-demonstrated. Gesture-based input is basically recognized by comparing acceleration values of input data with training data which are measured in advance [3] [4]. Gesture recognition consists of the following two steps:

1) Gesture detection: detects when a gesture is performed
2) Gesture classification: classifies which one in gesture candidates is performed

Fig. 1 shows the overview of gesture recognition.

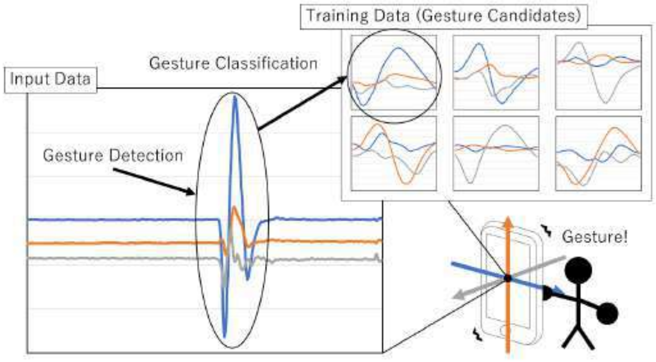


Fig. 1. Overview of gesture recognition

The problem with gesture-based input is that many misrecognitions occur. Noisy acceleration values are typically the problem, and a large number of gesture candidates thwart disambiguation. Misrecognition can be divided into the following three categories:

- False positive: outputs a detected gesture when no gesture is performed
- False negative: does not output results when a gesture is performed
- False classification: outputs the incorrect gesture

False positive and false negative occur with gesture detection. False classification occurs with gesture classification. Notably, there is a tradeoff between false positive and false negative accuracies. In this paper, we focus on improving false classification avoidance. Normally, the number of gestures candidates is the same as the number of input operations. If this number is large, it becomes necessary to prepare a huge number of training data in advance. This is seemed to lead

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

to the difficulty in classifying among them. We confirmed in preliminary experiments that increasing the number of gesture candidates leads to a lower recognition accuracy. Therefore, we propose a method to detect and classify gesture-based input with high accuracy by using two consecutive simple gestures and limiting the number of gesture candidates. The first gestures are trigger gesture (TG) which determine the start or content of the operation. The second are main gesture (MG) which is the actual input operation detail. Recognition accuracy is improved by limiting the number of MG candidates when TG is recognized. Applying a music-player app, we assume five types of input requests. We test these using an input pattern that does not limit MG candidates (baseline method) and an input pattern that limits MG candidates (proposed method). As a result of experimentation, we confirmed that the number of misrecognitions can be reduced by approximately 50% using proposed method.

II. Related work

The acceleration values acquired by accelerometers contain a lot of noise, even when applying a high sampling rate. When classifying gestures, both training data and input data must be matched in real time. Unfortunately, they both contain a lot of noise. Thus, misrecognitions commonly occur. Regarding the problem of noisy acceleration values, Murao et al. used nine accelerometers to acquire more accurate values [5]. It was shown that the recognition accuracy, when recognizing 27 types of gestures, was 62.7% on average. In their research, to reduce misrecognition caused by noise, they used a method of continuously calculating the average value of the acquired acceleration values.

Machine learning is often used in study to calculate the similarity of gestures and assign correct labels [6] [7] [8]. In this paper, however, we do not use feature-based machine learning, but instead classify gestures by calculating the similarity between time series data.

Even with the same gesture, the speed of hand motion changes depending on the time and the situation. In this case, the number of samples and the acceleration values of input data will be different, and it becomes difficult to calculate the similarity to training data. Dynamic Time Warping (DTW) method is an algorithm used for calculating the similarity between such data with respect to the problem that the gesture speed is not the same each time [9] [10] [11]. With DTW, by expanding and contracting the time in two time-continuous data, the difference in the number of samples and the acceleration values can be eliminated, and the similarity can be calculated with less overhead.

In this paper, we use a method of continuously calculating the average value of the acquired acceleration values for information extraction from training data. We apply DTW to calculate the similarity between training data and input data. Izuta et al. used a similarity calculation method via DTW to detect the start of a gesture [12]. It was shown that a gesture could be detected without waiting for its finish by continuously comparing its similarity with all training data.

Izuta et al. also mentioned that recognition accuracy would be degraded in case gesture candidates include similar ones [12]. So, it remains difficult to accurately classify gestures when the number of candidates increases. Therefore, in this study, using two consecutive simple gestures is applied to improve recognition accuracy.

III. Formal description of the problem

The purpose of this paper is to provide a method of recognizing gestures from the acceleration values acquired by accelerometers equipped in smartphones and use them as the input operation of a music player. In this section, we describe the assumptions, conditions, and target problem for gesture recognition.

We assumed that we acquire the acceleration values of gestures while holding a smartphone in one hand. The n gestures are prepared as gesture candidates, and are denoted by G_i $(i = 1, 2, 3, \cdots, n)$.

The variables of this gesture recognition problem include training data and input data. Training data are labeled with ones of the gesture candidates. The output, in turn, is the label with training data. The sequential acceleration values of G_i are training data represented by $B_i = (b_0, \cdots, b_{Ti})$. T_i represents the number of samples of training data B_i. Similarly, input data from the gesture start time t_0 to the finish time t_e, include the sequential acceleration values represented by $A = (a_{t_0}, \cdots, a_{t_e})$.

Almost all types of smartphones equip 3-axis accelerometers. These accelerometers can sample acceleration values for each of the x, y, and z axis. The acquired values include values of acceleration caused by the gestures and of gravitational acceleration. Thus, acceleration values b and a are acquired for the three axes of x, y, and z. Thus, they become vectors of three parameters. For example, in one sample, the parameters of training data, b, and input data, a, are (b_x, b_y, b_z) and (a_x, a_y, a_z), respectively.

With gesture recognition, the similarity between input data A and training data B_i is calculated. In this paper, the similarity is calculated using DTW and evaluated as a distance, $D(A, B_i)$. The smaller the distance of D, the more similar the two data become. Therefore, G_* which is $\min\{D(A, B_i), i = 1, 2, 3, \cdots, n\}$ is output.

IV. Baseline method

A. Gesture detection

It is necessary to detect the start and finish of gesture-based input because input data A is acquired continuously. We used the average of total acceleration as the condition for recognizing the start and finish of gesture-based input. The total acceleration $|a|$ at a certain time was set as following equation (1).

$$|a| = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{1}$$

From the start acquiring of input data, we continue to take the average value of the total acceleration. When the total acceleration, $|a|$, exceeds α by an amount greater than the average value before it for the first time, t_s, when the gesture starts is as following equation (2). α is a threshold for recognizing whether a gesture is performed.

$$t_s = \min\{\, t' \mid |a_{t'}| > \frac{1}{F} \sum_{t=t'-F}^{t'-1} |a_t| + \alpha \,\} \tag{2}$$

Next, it is recognized that the gesture finishes at the last time the total acceleration, $|a|$, exceeds the above threshold. The reason for doing this is that the total acceleration, $|a|$, may temporarily fall below the threshold during the gesture. Therefore, the time, t_f, when the gesture finishes, is as following equation (3).

$$t_f = \max\{\, t' \mid |a_{t'}| > \frac{1}{F} \sum_{t=t_s-F}^{t_s-1} |a_t| + \alpha \,\} \tag{3}$$

B. Gesture classification

After gesture detection, DTW was used to classify the type of gestures. We extract $A_t^* = (a_{t-F}, \cdots, a_t)$ and calculate the similarity between A_t^* and training data, B_i. calculate the similarity between A_t^* and training data, B_i. The similarity is evaluated based on the distance of DTW. In this paper, we use the Euclidean distance as the seed of DTW. The Euclidean distance, d, is defined as the distance between input data, a, and training data, b, at a certain time, as following equation (4).

$$d(a,b) = \sqrt{(b_x - a_x)^2 + (b_y - a_y)^2 + (b_z - a_z)^2} \tag{4}$$

Next, the algorithm of calculating the distance using DTW is shown. Considering the matrix, dtw, each element, $dtw[j][k]$, stores DTW distance defined as follows. We sequentially calculate the cost from $dtw[0][0]$ to $dtw[F][f_i - s_i]$ by dynamic programming to find a route that minimizes the cost of $dtw[F][f_i - s_i]$. The cost is the Euclidean distance, d, and the value of $dtw[F][f_i - s_i]$ is the distance, $D(A_t^*, B_i)$, which is used as the similarity. Fig. 2 shows the algorithm of calculating the distance using DTW.

We calculate the distance, $D(A_t^*, B_i)$, from extracted input data, A_t^*, for each training data, B_i $(i = 1, 2, 3, \cdots, n)$, by using DTW. Then output the gesture G_I having the smallest distances, $D(A_t^*, B_I)$, is performed.

$$D(A_t^*, B_I) = \min_{i=1,2,3,\cdots,n} D(A_t^*, B_i) \tag{5}$$

C. Preliminary experiment

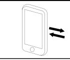
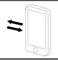
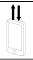
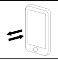
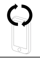
Basically, it is easy to find the correct answer with high accuracy to the alternative question. However, it is not easy to find just one correct answer from many choices. When this is applied to gesture recognition, it is expected that the recognition accuracy will decrease as the number of gesture candidates increases. To confirm the relationship between recognition accuracy and the number of gesture candidates, we conducted preliminary experiments.

```
Require: $A = (a_1, \cdots, a_F), B = (b_1, \cdots, b_{f_i - s_i})$
Ensure: $D(A, B)$
$dtw[0][0] \leftarrow 0$
for $j = 1, 2, 3, \cdots, F$ do
    $dtw[j][0] \leftarrow \infty$
end for
for $k = 1, 2, 3, \cdots, f_i - s_i$ do
    $dtw[0][k] \leftarrow \infty$
end for
for $j = 1, 2, 3, \cdots, F$ do
    for $k = 1, 2, 3, \cdots, f_i - s_i$ do
        $dtw[j][k] \leftarrow d(a_j, b_k) + \min \begin{cases} dtw[j-1][k] \\ dtw[j-1][k] \\ dtw[j-1][k-1] \end{cases}$
    end for
end for
return $dtw[F][f_i - s_i]$
```

Fig. 2. the algorithm of calculating the distance using DTW

Because we assumed that gestures are input consecutively, we adopted 13 gestures which can be performed while holding a smartphone in one hand, and return to the original positions after a performance(TABLE I, Fig. 3).

TABLE I
BASIC GESTURES

	Shake to the right
	Shake to the left
	Shake up
	Shake down
	Shake to the front
	Shake to the back
	Tilt to the right
	Tilt to the left
	Tilt to the front
	Tilt to the back
	Draw a circle
	Draw a triangle
	Knock the back twice

The content of the preliminary experiment was measuring the recognition accuracy by changing the number of gesture candidates. The recognition accuracy was defined as the per-

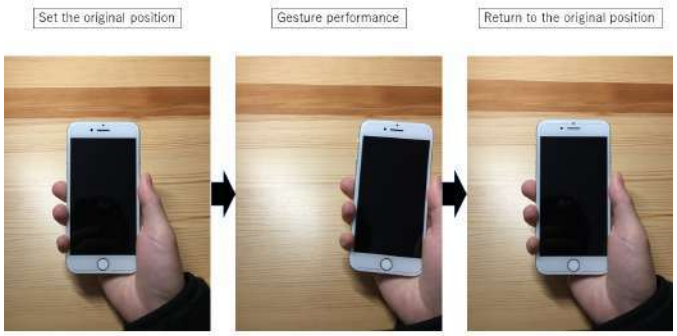


Fig. 3. Performing a gesture while holding a smartphone in one hand

centage that correct gesture labels were output, and the number of gesture candidates was changed from 2 to 13.

Gesture candidate lists were set to maximum of 78 combinations from TABLE I, and the recognition accuracy was calculated as the average value of trials. For the number of gesture candidates was 1, 2, 11, 12, and 13, we were able to examine all the combinations of gesture candidates. For the number of gesture candidates was 3 to 10, the number of possible combinations was too large, so we limited the number of gesture candidate lists to 78 patterns.

Fig. 4 shows the result of this. The accuracy for two choices was more than 99%, while it was 96% for correcting the gestures from 7 candidates, and 92% from 13 candidates. From the above result, the accuracy tends to decrease as the number of gesture candidates increases.

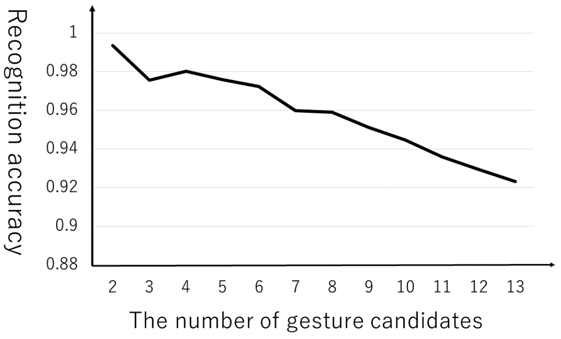


Fig. 4. Graph of recognition accuracy of changing the number of gesture candidates

V. Proposed method

A. System structure

In this paper, we implement the following system structure to recognize two consecutive simple gestures that are input operations. First, to remove the portion of the data that may cause an error in similarity calculation, training data preprocessing is performed. One gesture is recognized by continuously calculating the similarity between input data and training data. Then, two consecutive simple gestures are recognized from the set of candidates determined by which type of TG.

1) Training data preprocessing
2) One gesture recognition
3) Two consecutive simple gestures recognition

B. Training data preprocessing

When calculating the similarity via DTW, all data from the acquisition start time to the finish time are used as training data. Because they include the data acquired when gestures are not actually performed, which may cause an error in similarity calculation. Therefore, we extract the necessary information from training data. For that purpose, it becomes necessary to recognize when the gesture started and finished. We used the method same as baseline method to recognize them. As a numerical value required for recognition, the total acceleration, $|b|$, at a certain time is set as following equation (6).

$$|b| = \sqrt{b_x^2 + b_y^2 + b_z^2} \tag{6}$$

When the total acceleration, $|b|$, exceeded α greater than the average value before it for the first time, it recognized that the gesture started. Then, it was recognized that the gesture finished at the last time the total acceleration, $|b|$, exceeded the threshold. Therefore, the times, s and f, when the gesture started and finished were as following equations (7) and (8).

$$s = \min\{\, s' \mid |b_{s'}| > \frac{1}{s'-1}\sum_{t=0}^{s'-1} |b_t| + \alpha \,\} \tag{7}$$

$$f = \max\{\, f' \mid |b_{f'}| > \frac{1}{s-1}\sum_{t=0}^{s-1} |b_t| + \alpha \,\} \tag{8}$$

s_i and f_i are calculated for training data, B_i. In the following, training data, B, is $B = (b_s, \cdots, b_f)$. Next, the time required for one gesture, F, is determined. The time taken for each gesture, G_i, can be calculated by $f_i - s_i$. When recognizing a gesture of input data, it falls into a state in which the gesture is not known. Thus, the maximum time required for each gesture is F as following equation (9).

$$F = \max\{f_i - s_i, i = 1, 2, 3, \cdots, n\} \tag{9}$$

Fig. 5 shows the result of information extraction of training data when shaking a smartphone to the right.

C. One gesture recognition

Assuming that the current time is t and one gesture was performed in the past F samples, we extract $A_t^* = (a_{t-F}, \cdots, a_t)$ and calculate the similarity between A_t^* and training data, B_i. In the range of past F samples, $A_t^* = (a_{t-F}, \cdots, a_t)$, we continued to take the average value of total acceleration. At a certain time t, the smallest of the distances, $D(A_t^*, B_i)$, from each training data, B_I, is the output candidate. However, a gesture may not have been performed in the extracted input

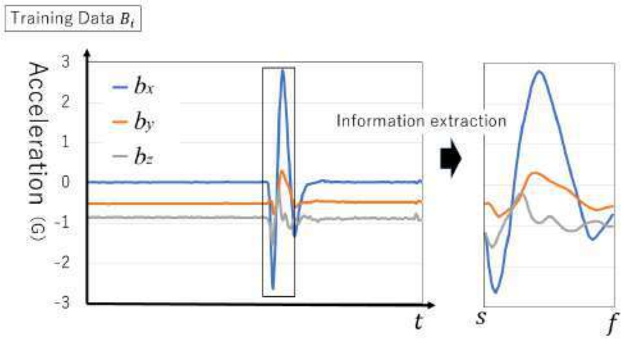


Fig. 5. Example of information extraction of training data (shake to the right)

data. If no gesture is performed, all distances will have similar values. Therefore, when the distance with training data, B_I, is clearly smaller than the distance with training data B_{II} having the second smallest distance, it will be recognized that the gesture, G_I, is performed. The condition at that time is as follows. $\beta(> 1)$ is the threshold for recognizing gestures.

$$\beta D(A_t^*, B_I) < D(A_t^*, B_{II}) \tag{10}$$

All distances have similar values when the gesture is not performed. This helps us recognize that the gesture starts when the above condition is first met and finishes when the condition is finally met. Therefore, the start time, t_s, and the finish time, t_f, of the gesture are set as following equations (11) and (12), respectively.

$$t_s = \min\{\, t' \mid \beta D(A_{t'}^*, B_I) < D(A_{t'}^*, B_{II}) \,\} \tag{11}$$

$$t_f = \max\{\, t' \mid \beta D(A_{t'}^*, B_I) < D(A_{t'}^*, B_{II}) \,\} \tag{12}$$

D. Two consecutive simple gestures recognition

In this paper, we focus on the operation of a music player. Table II shows the minimum 5 inputs required for operation.

TABLE II
INPUT REQUIRED TO OPERATE THE MUSIC PLAYER

Play/Pause
Play the next music
Play the previous music
Turn up the volume
Turn down the volume

Each gesture should be simpler because a complicated gesture puts a load on users. Therefore, we consider a simple gesture that is used for combination. We selected the gestures that require shaking smartphone once along each axis direction from the basic gestures. Table III lists gestures along each axis.

There are 6 types of simple gestures to be assumed, so we can select 6×6 of 36 combinations of two gestures. Of these, we assign them to the five inputs required to operate a music player. There are two possible methods for the role of TG.

TABLE III
GESTURES OF SHAKING THE SMARTPHONE ALONG EACH AXIS DIRECTION

axis	Positive (+) direction	Negative (-) direction
x	Shake to the right	Shake to the left
y	Shake up	Shake down
z	Shake to the front	Shake to the back

The first pattern is input-pattern A, which is an assignment method that selects only combinations that use the same TG for all MG. The second pattern is input-pattern B, which is an assignment method that selects combinations that use multiple TG. In input-pattern A, TG has the meaning of a signal to start input, and MG indicates the actual input contents. In input-pattern B, the input content is already determined when TG is input, and MG determines the degree. TABLE IV and TABLE V show the combinations of gestures for each input pattern. The combinations were decided so that the input content and the gesture could be connected intuitively. The parentheses in the table indicate the direction of the axis that primarily reacts.

TABLE IV
INPUT PATTERN A

Input operation	TG	MG
Play/Pause	Front(z+)	Back(z-)
Play the next music	Front(z+)	Right(x+)
Play the previous music	Front(z+)	Left(x-)
Turn up the volume	Front(z+)	Up(y+)
Turn down the volume	Front(z+)	Down(y-)

TABLE V
INPUT PATTERN B

Input operation	TG	MG
Play/Pause	Up(y+)	Back(z-)
Play the next music	Front(z+)	Right(x+)
Play the previous music	Front(z+)	Left(x-)
Turn up the volume	Right(x+)	Up(y+)
Turn down the volume	Right(x+)	Down(y-)

Both TG and MG are recognized by the same method, as shown in Section 4.3. In this paper, the interval between TG and MG is within 1 s. If MG is recognized within 1 s after the recognition of TG, the operation by the combination of TG and MG is output. If it exceeds 1 s after the recognition of TG, it means that no input has been made. Additionally, because the type of TG is classified at the time of MG recognition, MG is not output and is set as a new TG if MG does not match TG (Fig. 6). Therefore, in input-pattern A, it is easy to classify with only one type of TG. However, because the type of MG is not limited after TG is recognized, MG must be selected from five types. On the other hand, in input-pattern B, TG must be selected from three types. However, after TG recognition, the MG type is limited, and MG is selected from a maximum of two types.

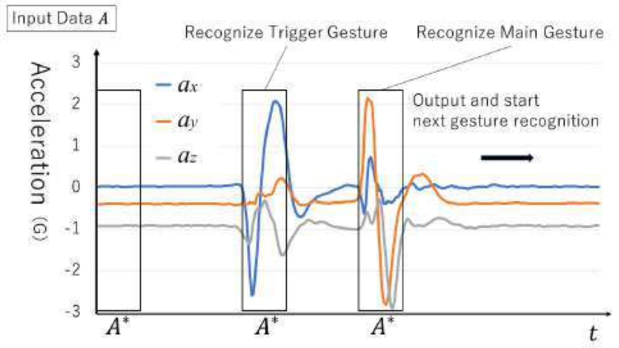


Fig. 6. Gesture recognition in input data

VI. Experiment

A. Experimental purpose

In this paper, we conducted experiments to evaluate the appropriate values of α and β, which are set as thresholds and to confirm whether the proposed method improves recognition accuracy. The environment was assumed to be a gesture-based input using a smartphone indoors. In this experiment, as a comparison method, we considered the input start time and finish time recognition using the total acceleration instead of similarity. This method is a basic recognition method used in vehicle airbags and has been improved to be applied to gesture-based input. As the threshold, the same α as the information extraction of training data was used, and the value having the highest recognition accuracy was found. The value of α obtained at this time was used to find the value of β having the highest recognition accuracy by using the proposed method. Finally, the recognition accuracy of each of the two input patterns in the proposed method was compared.

B. Environment

In this experiment, we assumed a gesture-based input using a smartphone indoors and all acceleration data was acquired indoors with the user stationary and holding a smartphone. The following items were used as experimental equipment.

- Smartphone iPhone8 (Apple A11 Bionic, Built-in 3-axis accelerometer, Sampling rate: 100Hz)

C. Experimental order

The experiments were conducted in the following order.

1) We measured the recognition accuracy by changing the value of α for the comparison method.
2) With the value at α which had the highest recognition accuracy by the comparison method, we measured the recognition accuracy by changing the value of β for the proposed method.
3) With the values of α and β measured in the above two experiments, we compared the recognition accuracy of the input pattern that limits the type of MG by the type of TG with the input pattern that does not limit the type of MG.

D. Evaluation method

As training data, we acquired each gesture shown in TABLE III once. We gave the same sequential acceleration values as input data to both the proposed method and the comparison method. In input data, combinations of TG and MG were performed seven times for each input pattern shown in the TABLE IV and V. While acquiring input data, the time when each gesture was performed was recorded and used as the correct answer label, indicating which gesture was performed. In each method, if there was a time with a correct label between the start time, t_s, and the finish time, t_f, of the gesture, it was assumed that gesture detection was correct. The gestures were recognized by each method, the outputs were classified, and a score was given to each, as shown in TABLE VI. The cumulative score was divided by the number of performed gestures and evaluated as the recognition accuracy.

TABLE VI
Classification and Score of Output

Output	Score
Correct recognition	+1
False positive	0
False negative	0
False classification	-1

VII. Results

A. Evaluate the value of α

The change in recognition accuracy of the comparison method when the value of α was changed by 0.1G is as shown in Fig. 7. When α=0.9G, the maximum recognition accuracy was 78.6%. This value was the best performance of the comparison method.

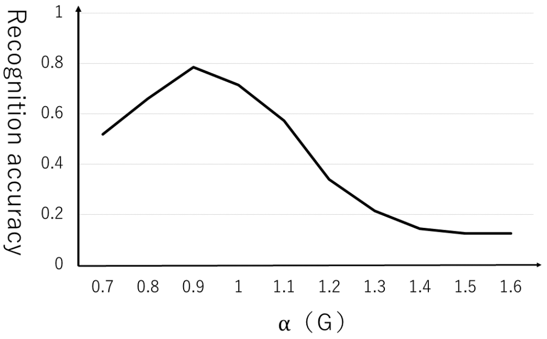


Fig. 7. Graph of recognition accuracy of comparison method

B. Evaluate the value of β

Fixing α=0.9G, the change in recognition accuracy of the proposed method was observed when the value of β was changed by 0.1, as shown in Fig. 8. When β=1.8, the maximum recognition accuracy was 96.4%. The best performance of the proposed method was clarified when the value of α was

0.9G. From the above, regarding the detection method of the time when the gesture was performed, it was confirmed that the proposed method in which the similarity was continuously compared obtained higher recognition accuracy than did the comparison method in which a threshold value was set for the acceleration value.

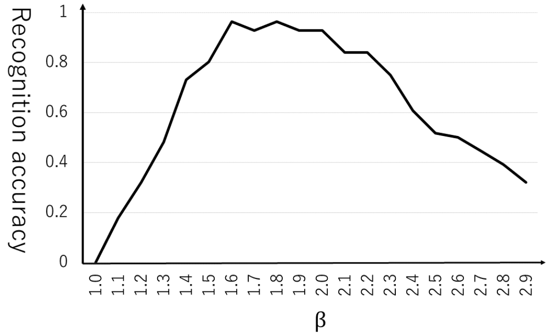


Fig. 8. Graph of recognition accuracy of proposed method

C. Comparison of recognition accuracy in each input pattern

In the proposed method, when α=0.9G and β=1.8, the recognition accuracy in the input pattern that did not limit the type of MG (Input pattern A) was 94.3%. On the other hand, the recognition accuracy of the input pattern that limited the type of MG by the type of TG (Input pattern B) was 97.1% (TABLE VII). The number of misrecognitions was reduced by 49.1%. From the above, it was confirmed that the recognition accuracy was improved by using two consecutive simple gestures.

TABLE VII
RECOGNITION ACCURACY IN EACH INPUT PATTERN

Input pattern	recognition accuracy
Input pattern A	94.3%
Input pattern B	97.1%

D. Recognition accuracy of single gesture

Under the same conditions as in Section VII-C, we calculated the recognition accuracy in the case of only one gesture for input. The average of the recognition accuracy was 95.2%. By comparing with TABLE VII, we confirmed that the recognition accuracy decreased when gesture-based input simply needs a combination of gestures, however, it was improved by limiting the types of gesture candidates even though a combination of gestures was needed.

VIII. Conclusion

In this paper, we proposed a gesture recognition method using two consecutive simple gestures to provide a music player's input operation, using a smartphone's accelerometer. From the comparison between the proposed method and the comparison method, it was confirmed that the recognition accuracy improved more when the gesture was recognized by comparing the similarities than when the gesture was recognized only by the acceleration values. Additionally, it was confirmed that the recognition accuracy was improved by limiting the types of MG corresponding to TG. Using two consecutive simple gestures resulted in complex gesture as a whole input operation, which reduced the recognition accuracy, however, we also confirmed that limiting the types of gesture candidates mitigated the reduction in recognition accuracy. In the future, we would like to calculate and compare the recognition accuracy by experiments on gestures while walking, which are expected to contain a lot of noise. We also expect that it will be necessary to discuss the optimal values of the thresholds, α and β, depending on the difference among users.

References

[1] S. Yoshiki, H. Tatsumi, K. Tsutsumi, T. Miyazaki, and T. Fujiki, "Effects of smartphone use on behavior while walking," *Urban and Regional Planning Review*, vol. 4, pp. 138–150, 2017.

[2] T. F. Department, "Watch out for accidents related to smartphones and other devices!" 2020. [Online]. Available: https://www.tfd.metro.tokyo.lg.jp/lfe/topics/201602/mobile.html

[3] T. Sakaguchi, T. Kanamori, H. Katayose, K. Sato, and S. Inokuchi, "Gesture recognition using gyroscopes and accelerometers," *Transactions of the Society of Instrument and Control Engineers*, vol. 33, no. 12, pp. 1171–1177, Dec 1997.

[4] J. BAEK, J. Ik-Jin, and Y. Byoung-Ju, "Recognizing and analyzing of user's continuous action in mobile systems," *IEICE transactions on information and systems*, vol. 89, no. 12, pp. 2957–2963, Dec 2006.

[5] K. Murao, T. Terada, A. Yano, R. Matsukura, and S. Nishio, "A study of gesture recognition with sensor-contained mobile devices," *The Special Interest Group Technical Reports of IPSJ. UBI*, vol. 25, pp. X1–X8, Mar 2010.

[6] L. JING, Y. ZHOU, Z. CHENG, and J. WANG, "A recognition method for one-stroke finger gestures using a mems 3d accelerometer," *IEICE Transactions on Information and Systems*, vol. E94.D, no. 5, pp. 1062–1072, 2011.

[7] Y. XUE, Y. HU, and L. JIN, "Activity recognition based on an accelerometer in a smartphone using an fft-based new feature and fusion methods," *IEICE Transactions on Information and Systems*, vol. 97, no. 8, pp. 2182–2186, 2014.

[8] X. HAN, J. YE, J. LUO, and H. ZHOU, "The effect of axis-wise triaxial acceleration data fusion in cnn-based human activity recognition," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 4, pp. 813–824, 2020.

[9] K. Murao and T. Terada, "A motion recognition method by constancy decision," *IPSJ Journal*, vol. 52, no. 6, pp. 1968–1979, Jun 2011.

[10] K. Murao, T. Terada, A. Yano, and R. Matsukura, "Evaluating sensor placement and gesture selection for mobile devices," *Information and Media Technologies*, vol. 8, no. 4, pp. 1154–1165, 2013.

[11] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2012, p. 262270.

[12] R. Izuta, K. Murao, T. Terada, and M. Tsukamoto, "Early gesture recognition method with an accelerometer," in *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia*, ser. MoMM ' 14. New York, NY, USA: Association for Computing Machinery, 2014, p. 4351.

Improving Accuracy of High-Speed Continuous Tap Recognition on Desk Using Acceleration and Sound Amplitude

1st Yuika Katayama
Osaka Prefecture University
Osaka, Japan
sbb01050@edu.osakafu-u.ac.jp

2nd Ryo Katsuma
Osaka Prefecture University
Osaka, Japan
katsuma@cs.osakafu-u.ac.jp

***Abstract*—The purpose of this study is to measure the time of tap action on a desk as accurately as possible using a mobile device, such as a smartphone and a tablet PC, placed on the desk. The existing method recognizes a tap based on whether the acceleration exceeds a threshold value. In addition, a waiting time is added after the tap judgment. However, this method often misjudges more taps when the user taps hard or taps continuously at high speed.**

Therefore, in this paper, we propose three new methods to solve this misjudgment problem by using the amplitude of sequential acceleration data and tap sound data: TIA (Threshold and Immediately preceding period for Amplitude), TIAA (Threshold, Immediately preceding period. and Attenuation for Amplitude), and TIAAS (Threshold, Immediately preceding period, and Attenuation for Acceleration and Sound wave). TIA uses the input wave amplitude threshold to find the time at which a tap may occur. TIA then checks the input wave for the last several milliseconds to determine if it was tapped at that time. TIAA uses the amplitude threshold of an input wave like TIA and the degree of attenuation for amplitude after the tap time. TIAAS uses the amplitude threshold of the sound for d seconds after the time the tap is recognized by TIAA for z-axis acceleration.

The experimental results show that TIA for sound amplitude and TIAAS recognize a tap more accurately than the existing method.

***Index Terms*—acceleration, sound amplitude, action recognition, tap**

I. Introduction

There is a rhythm game that scores the timing of the user's actions in response to objects (notes) that appear on the screen in sync with music and users aim for a higher score. The required user actions differ depending on the type of rhythm game, for example, tapping on specified positions the hands, pressing a button, or stepping on panels with feet. In some rhythm games, it is said that if the timing of the action is within about ±0.1, 0.066, and 0.033 seconds of the indicated timing, the score is higher in this order. One game is equivalent to one musical piece that contains several notes. A note indicates one action and its execution timing, and several notes appear in one music. Therefore, users need to perform the right action at the right time many times.

An array of notes that indicate the user's actions is called a chart, and there are various musical pieces and charts. A chart for beginners is low-density and simple, on the other hand, a chart for experts is high-density, complex, and more difficult. Rhythm games often require special large equipment, and many users play them at an amusement arcade. It takes a lot of practice to get a good score on a high difficulty chart. For this reason, players sometimes practice the timing of actions at home by tapping a desk as touch panels or buttons, while watching chart videos on Youtube. However, there is no way to judge if a user taps at the right time during practice.

In order to know if a user taps a desk at the correct time, it needs a method to identify the time when the desk is tapped and a method to determine whether the timing is correct. This paper focuses on the former method and aims to determine the time of tapping a desk as accurately as possible by placing a tablet on the desk in a quiet environment and measuring the shock wave of the tap.

There is already a method for recognizing taps by using the acceleration when a user taps the accelerometer directly. However, since the taps performed in rhythm games are powerful, a sensor may be damaged by tapping directly. Therefore, we placed a tablet on a desk so as not to tap a sensor directly, and measured the acceleration of the shock wave generated by tapping the desk. In this environment, acceleration measured by the tablet is noisier than tapping the sensor directly, which reduces the accuracy of the tap recognition by the existing method. In the existing method, after a tap is recognized, the next tap judgment is not performed until a waiting time. However, since the tap interval in rhythm games can be very short, and if the next tap occurs during the waiting time, the tap will not be recognized. As a result, there is a problem that the accuracy of tap recognition is reduced due to the increase of noise ratio and the inclusion of high-speed continuous taps.

In this paper, we propose the following three methods to determine the tapping time accurately. The first method is TIA (Threshold and Immediately preceding period for Amplitude), which uses amplitude threshold a of input wave to find the time at which a tap may occur. TIA then checks the input wave for the last b seconds to determine if it was tapped at that time. The second method is TIAA (Threshold, Immediately preceding period. and Attenuation for Amplitude), which uses

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

the amplitude threshold a of the input wave and the attenuation for the amplitude of the input wave from the time when the amplitude reaches its maximum. The third method is TIAAS (Threshold, Immediately preceding period, and Attenuation for Acceleration and Sound wave), which uses TIAA for the z-axis acceleration to determine the time of a possible tap, and only determines it a tap if the amplitude of the sound exceeds the threshold c in d seconds just after that time.

In our experiment, we measured z-axis acceleration (acceleration in the vertical direction relative to the desk surface), 3-axis acceleration, and sound with a tablet placed on the desk. For each proposed method and the existing method, we obtained appropriate values of parameters used in each method according to the value of T when the subject taps eight times at equal intervals T through experiments on actual devices and verified the best tap recognition accuracy. The results show that TIA for sound amplitude and TIAAS recognize a tap more accurately than the other methods.

II. Related Works

A. Acceleration-based action recognition

Norieda et al. has proposed a FingerKeypad that allows users to operate a device by tapping a finger around the back of the terminal [1]. FingerKeypad is equipped with nine bone conduction microphones on the backside of the terminal and identifies the tap position on the three parts of each finger and a tap is recognized when the extreme value after the fall of either sensor is below -1.5 V. Although the shape of the sensor is different, the method of identifying the tap time using threshold is the same as those described in this paper. Harrison et al. proposed Skinput, which uses the skin as an input surface by identifying the tap position on the arm, finger, and so on. Skinput uses an armband with 10 vibration sensors attached and is set two threshold values to recognize taps [2]. Murao et al. proposed a motion recognition method using three accelerometers attached to body [3].

In these studies, sensors are attached directly to the body to recognize the behavior. In contrast, in our study, we used a non-contact sensor to recognize the tap. Therefore, we must deal with the effects of strong noise influence and reverberation.

B. Sound-based action recognition

Nandakumar et al. proposed FingerIO, which tracks a finger using an ultrasonic speaker and two microphones [4]. Ishii proposed PingPongPlus, which projects images on the ping pong table according to the position of the ping pong ball hitting the table, using eight microphones on the ping pong table [5].

These methods use a large number of sensors, which make installation complicated and costly. In contrast, in this study, we used one accelerometer and one microphone mounted on a tablet placed on a desk to recognize the tap.

C. Action recognition using mobile devices

Methods of recognizing user behavior by attaching sensors to the body enable highly accurate recognition. However, wearing sensors on the body all the time can interfere with movement and the sensation of wearing the sensor can be stressful. Therefore, research is being conducted on methods for recognizing user action using accelerometers, gyroscopes, and microphones equipped in mobile devices such as smartphones and tablets.

For example, Iso et al. proposed a method for recognizing the following five actions while holding the cell phone: walking, running, going up/down stairs, and walking fast, which are recognized by using data obtained from the cell phone's built-in 3-axis acceleration sensor [6]. Kurasawa et al. proposed a method for estimating the storage location of a cell phone and the user's action by using the acceleration obtained from an acceleration sensor attached to a cell phone and the tilt of the phone [7]. Iketani et al. proposed a method for estimating the user's movement status using a mobile device equipped with an acceleration sensor [8].

Thus, smartphones and tablets can be used in this research because they enable measurements using multiple sensors simultaneously.

D. Action recognition by the combination of acceleration and sound

Ouchi et al. proposed a method to recognize the user's state in real-time using only the accelerometer and microphone of a cell phone [9]. Since the number of sound data increases by more than three orders of magnitude compared to acceleration data, this method avoids continuous use of sound data. It uses the acceleration sensor to roughly estimate the user's state, and then uses sound data according to the estimated state. In our study, considering the complexity of the calculation, acceleration is used to find the tap possibility, and then the sound amplitude is used to determine the tap.

E. Existing tap recognition method

This section describes the existing tap recognition method [10]. This is a method for recognizing two direct taps on an accelerometer. When the acceleration exceeds a threshold value (a_{THT}), there is a possibility of tapping. After that, when the time when the acceleration exceeds a_{THT} is less than or equal to a certain time (T_{DUR}), the tap is finally recognized. Just after the time of tap recognition, the next tap is not recognized during a waiting time (T_{LAT}), when the waiting time has passed, the next tap possibility can be judged.

Fig. 1 shows a graph of acceleration measured on the desk when tapping the desk once, where the horizontal and vertical axes represent time and z-axis acceleration, respectively. We will use this graph to explain how the existing method recognizes tap.

In Fig. 1, the time interval between points in the graph is 0.01 s. Suppose that T_{LAT} is 0.06 s. Since the acceleration exceeds the threshold a_{THT} at time t_1, there is a possibility of tapping. After that, the time when the acceleration exceeds

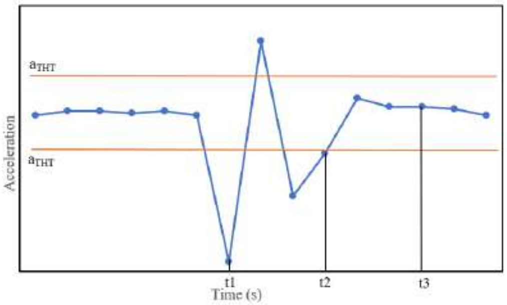


Fig. 1. How the existing method recognizes tap

the threshold is $t_2 - t_1$, so if this time is less than T_{DUR}, it is determined that the tap occurs at time t_1. Since the next tap is not recognized from time t_1 to T_{LAT} seconds, the next tap is not recognized until time t_3 in Fig. 1.

When T_{LAT} is long, the next tap within T_{LAT} will never be recognized. For a rhythm game at 60 fps, the interval between notes is 0.016 s when the notes come in consecutive frames. However, when a strong tap occurs, the amplitude of the input wave swings in the positive or negative direction, and then swings in the opposite direction due to the effect of vibration, so if T_{LAT} is shortened, the acceleration will exceed the threshold even at the time when the effect of vibration is still present, and it will be judged as a tap.

III. Problem Setting

In this section, we describe the assumed environment for recognize a user's tap and the purpose of this study.

A. Environments

A user taps the desk surface that is horizontal to the ground. A tablet is placed on the desk, with the widest side of the tablet parallel to the surface of the desk. When the user taps, a certain amount of vibration and impact generated from the tapping point on the desk are transmitted to the tablet. The tablet device is equipped with a microphone and a three-axis accelerometer. The z-axis of the accelerometer is always at a right angle to the desk surface, while the x-axis, y-axis, and microphone of the accelerometer are randomly oriented to the tapping point.

B. Problem input/output and objective function

Let P denote a set of tapping time $p_1, p_2, \cdots, p_n$. The input for the problem in this study is a time series of vibration and impact sound data detected by a tablet. The output is the set of estimated tapping time Q. The difference between $q \in Q$ and the actual tap time $p \in P$ must be small. If the actual tap time corresponding to q does not exist, it is a false positive, and the number of false positive must be reduced as much as possible. Therefore, the objective function f of the target problem is expressed by the following (1), and the purpose is to maximize f.

$$f = v - u \tag{1}$$

v is an evaluation value whose value increases when taps at a more correct time and u is an evaluation value whose value increases when the number of false positives increases. v is represented by (2). From the set of estimated tap times Q, the element that is closest to the actual tap time p is denoted by q^p.

$$v = \sum_{i=1}^{n} g(|p_i - q^{p_i}|) \tag{2}$$

$g(|p - q|)$ is a function that normalizes the difference between time p and q, and becomes 0 value when $|p - q|$ is greater than a constant value r. u is represented by (3).

$$u = \sum_{i=1}^{n} h(|p_i - q^{p_i}|) \tag{3}$$

$h(|p-q|)$ is a function that binarizes the difference between time p and q. It becomes 0 if the value of $|p - q|$ is less than or equal to a constant r, and w otherwise.

IV. Methods

In this section, we propose three kinds of tap recognition methods: TIA, TIAA, and TIAAS. As mentioned in II-E section, existing tap decision methods using thresholding have the problem of misjudges more taps due to tap vibration reverberation. When the amplitude of the input wave exceeds the threshold at time t, if the amplitude in the immediately preceding period also exceeds the threshold, time t may be the reverberation period. Therefore, TIA uses this mechanism to recognize tap more accurately. However, there are cases where TIA cannot cope with high-speed continuous tapping, because the next tap vibration may start before the reverberation of the previous tap disappears. Therefore, TIAA determines whether the amplitude exceeding the threshold is a new tap or a reverberation by comparing it with attenuation degree of the vibration over time. Since these TIA and TIAA can only be applied to time series data where the input wave is either acceleration or sound pressure value, we propose TIAAS, which recognition tap more accurately by using both of these time-series data. In TIAAS, TIAA is applied to the time series data of z-axis acceleration, and TIAA is also applied to the time series data of sound pressure for the time when there is a possibility of tapping, and when both are judged to be tapped, the tap is recognized. In the following sections, TIA, TIAA, and TIAAS are described in detail.

A. TIA

The amplitude of the input wave remains below the threshold for a certain time and then swings positively or negatively, and tap is recognized when it exceeds the threshold. Let A be the average value of z-axis acceleration when the desk is not tapped, which is measured in advance. Let a (> 0) be the acceleration threshold. Let $r_1, r_2, \cdots, r_{n-1}, r_n$ be the acceleration values detected in the b(b should be set appropriately according to the sampling rate of the device) seconds immediately before the current time t. When these values satisfy (4) and the acceleration value r at the current time t satisfies (5), the tap is recognized.

$$A - a \leq r_1, r_2, \cdots, r_{n-1}, r_n \leq A + a \tag{4}$$

$$a < |r - A| \tag{5}$$

Fig. 2 shows a graph of acceleration measured on the desk when tapping the desk once, where the horizontal and vertical axes represent time and z-axis acceleration, respectively. Using this graph, we will explain how TIA recognizes a tap.

In Fig. 2, the time interval between points in the graph is 0.01 s, the current time is time t_2, and suppose b is 0.05 s. The acceleration value between time t_1 and t_2 satisfies (4). Additionally, the acceleration value at the current time t_2 satisfies (5). So the tap is recognized at the time t_2. If the current time is t_3, the tap is not recognized because (4) is not satisfied at time t_2.

When judging using 3-axis acceleration or sound amplitude, A in (4) and (5) is set to 0.

B. TIAA

For the existing method and TIA, the amplitude of the input wave must not exceed the threshold before and after the time of judgment to recognize a tap. For this reason, both of false positives and false negatives are possible when high-speed continuous taps occur. Therefore, we define the degree of attenuation ratio overtime of the amplitude of the input wave to the maximum value as attenuation and use this attenuation degree (hereafter, we simply describe **attenuation**) to recognize a tap. We will use Fig. 3, which shows the acceleration during tapping, to explain attenuation.

In Fig. 3, the amplitude of the acceleration reaches its maximum at time t_1, and the value is $-1.048-(-1.332) = 0.284$. Let this amplitude be 100%. The amplitude at times t_2, t_3, and $t4$ are 0.132, 0.160, and 0.080, respectively, so the attenuation can be calculated as 46%, 56%, and 28%.

For sampling data where the measurement times of acceleration values are not equally spaced, resample them to make them equally spaced using the linear interpolation method. In this method, the threshold of the input wave amplitude is set to a (> 0), and when the input wave amplitude exceeds a, a tap decision is made by TIA. However, if a tap is recognized immediately before k seconds, the percentage of the amplitude of the current input wave $w_{current}$ to the amplitude of the input wave w_{last} at the previous tap time is calculated using (6).

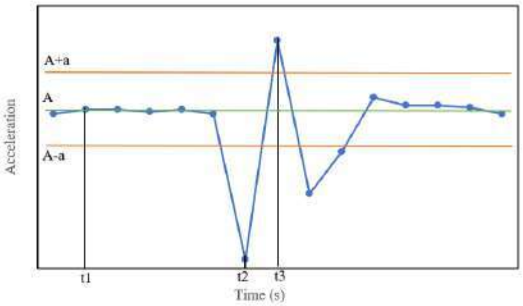


Fig. 2. How TIA recognizes tap

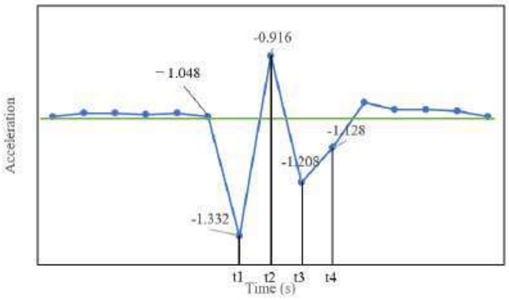


Fig. 3. Calculation of attenuation

$$\text{percentage} = \frac{w_{current}}{w_{last}} \times 100 \tag{6}$$

If the percentage calculated by the (6) is smaller than the predetermined attenuation of the amplitude of the input wave, the tap is not recognized. Fig. 4 shows a graph of acceleration measured on the desk when tapping the desk once, where the horizontal and vertical axes represent time and z-axis acceleration, respectively. Using this graph, we will explain how TIAA recognizes a tap.

In Fig. 4, the time interval between points in the graph is 0.01 s, the current time is time t_a, and suppose k is 0.04 s. At this time, the acceleration value in the k seconds immediately before the current time has not exceeded the threshold a, so the tap judgment is made using TIA. Since the acceleration value of t_1 exceeds a, the tap is recognized at time t_1. Next, when the current time is t_2, since t_1 has been judged as a tap, we move to the tap judgment using the attenuation for t_2. Next, when the current time is t_2, since tap has occurred at time t_1, TIAA recognizes the tap using attenuation. The red graph in the Fig. 4 is attenuation. In a tap recognition using attenuation, the percentage of the amplitude of the acceleration at time t_2 to the amplitude of the acceleration at t_1, the previous tap time, is calculated using (6) and checked to see if it is greater than attenuation. In the case of the Fig4, we check whether the distance from the central yellow-green line at time t_2 is greater in the red graph or the blue graph, and recognize a tap if the distance between red and yellow-green is greater than blue. At time t_2, the distance from the yellow-green line is greater in the red graph, so a tap is not recognized.

C. TIAAS

First, resample the z-axis acceleration data in the same way as TIAA and make the not equally spaced sampled data equally spaced. If TIAA is applied to the z-axis acceleration and recognizes a tap, TIAAS determines that it may be a tap. The same method as in Fig. 4 is used to make the tap possibility decision, and when the distance between the

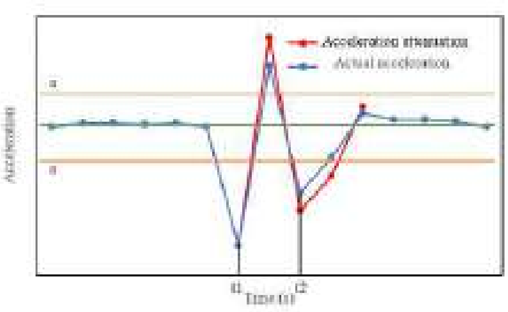


Fig. 4. How TIAA recognizes tap

acceleration value on the red line and the value on the yellow-green line is greater than the blue, the amplitude of the sound for d seconds from the current time is used to recognize the tap. During this time, if the amplitude of the sound exceeds the threshold c, the tap is recognized.

V. Experiment

We used two Nexus 9 tablets. One of them was used to obtain the correct answer data of the tapping time. It was placed lean against an object on the desk and its camera took videos of the tapping on the desk. The other one was placed randomly on the desk and measured 3-axis acceleration with Science Journal (Google app). Acceleration was sampled not equally spaced at about 16 Hz. In TIAA and TIAAS, acceleration was resampled at 16 Hz. Linear interpolation was used as the resampling algorithm. The sound was recorded at 44100 Hz extracted from the video and downsampled to 882 Hz. The time of the actual tapping was identified from the video image and was used as the correct answer data.

As an evaluation criterion for each tap recognition method, we explain the settings of v and u in the objective function described in section III. Let p be the time when the tap actually occurred. When the time when the tap is recognized to be $p \pm 0.033$ seconds (recognition range 1), the evaluation score is 1.01 points; when the time is $p \pm 0.066$ seconds (recognition range 2), the score is 1 point; and when the time is $p \pm 0.1$ seconds (recognition range 3), the score is 0.5 points. We calculate the total score v. However, if more than one score can be obtained, the one with the lower recognition range number is given priority, and no duplicate scores are obtained. If some taps are recognized around the time p, only the one with the earliest time is given a score, and the other tap recognitions are given 0 points. The number of false positives is u, and the number of the tap scored 0 points in the calculation of v is substituted. Each tap recognition method is evaluated by f in equation (1).

Fig. 5, shows how calculate score. Time t is actual tap time and time t_1, t_2 and t_3 are recognized tap time. The yellow, orange, and yellow-green range in Fig. 5 are recognition range 1, 2, and 3, respectively. Recognition range 1, 2, and 3 cover t_1, t_2, and t_3, respectively. In this case, the total score is 0.5 points for t_1, which means that t_2 and t_3 were false positive.

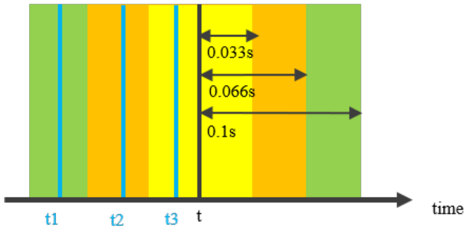


Fig. 5. Example of score calculation

A. Experimental

While increasing the BPM by 10 between 60 and 700, tap on the desk with one hand when the BPM was 60-440 and with both hands alternately when the BPM was 450-700 at equal intervals to the sound of the metronome application for each BPM. The total number of taps is nine at any BPM, one for time synchronization of the acceleration and sound, and eight for recognizing taps. For example, when the BPM is 60, after tapping once to set the time, the second tap is made after sufficient time has elapsed. The third tap is made just one second after the second tap, the fourth tap is made just two seconds after the second tap, and it takes seven seconds from the second tap to the last tap. We set a, b, c, d, T_{LAT}, a_{THT}, and T_{DUR} so that the evaluation value f is closest to the optimal solution of 8.08 for each method.

B. Preliminary experiment

We determined attenuation by measuring the z-axis acceleration, 3-axis acceleration, and sound for 100 taps. Fig. 6, Fig. 7, and Fig. 8 show the maximum and average attenuation values of z-axis acceleration, 3-axis acceleration, and sound amplitude, respectively. Fig. 9 shows a graph of the minimum attenuation of the z-axis acceleration. In TIAA, the percentage of the amplitude of the current input wave to the amplitude of the input wave at the previous tap time is calculated, and if the percentage is larger than the values in Fig. 6, Fig. 7, and Fig. 8, the tap is recognized. In TIAAS, the percentage of the amplitude of the current z-axis acceleration to the amplitude of the z-axis acceleration at the previous tap time is calculated,

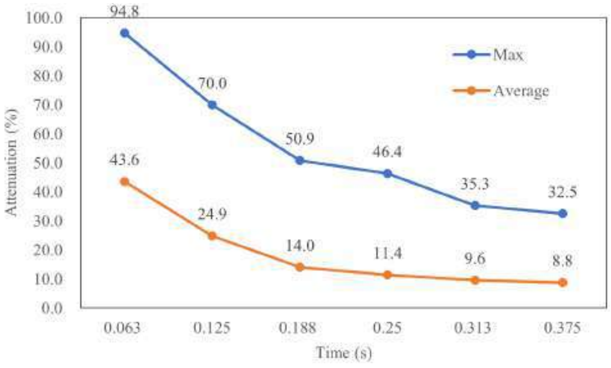


Fig. 6. Attenuation of z-axis acceleration

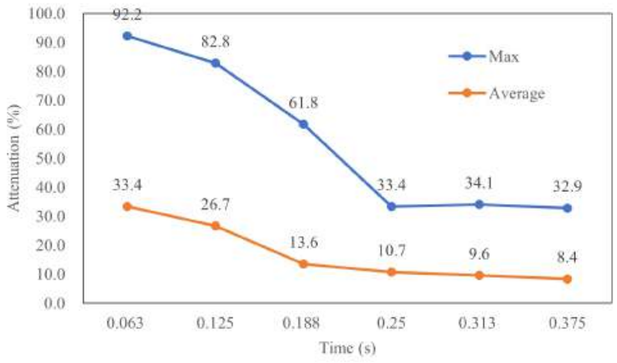


Fig. 7. Attenuation of 3-axis acceleration

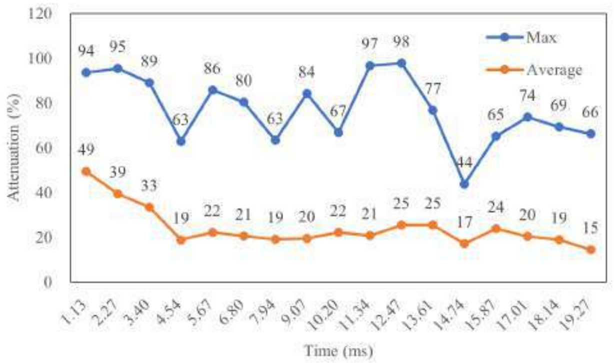


Fig. 8. attenuation of sound amplitude

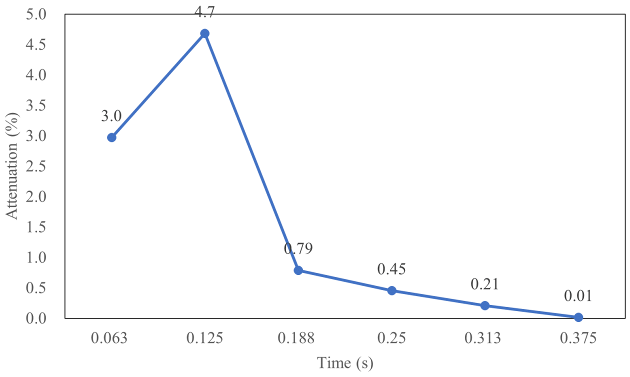


Fig. 9. Minimum attenuation of z-axis acceleration

and if the percentage is larger than the values in Fig. 9, there is the possibility of a tap.

VI. Results

A. Comparison of the results of each method

The result for each method is shown in Table I. The existing method used $T_{LAT} = 0.016s$. TIA used $b = 0.016s$. TIAA used the maximum value of the attenuation is used in TIAA, and when in TIAAS. The average number of false positives and evaluation value f are shown in Table I. Compared to the existing method, TIA was able to reduce the number of false positive for any sensor data and improved the problem of duplicate tap recognitions in a single tap. In particular, when using sound in TIA, f was high. We confirmed that the taps were almost completely recognized. However, TIA using only sound cannot work in a noisy environment. On the other hand, f of TIA using acceleration was lower than the existing method when the BPM was high. This is because recall value is greatly reduced due to the influence of reverberation. TIAA showed worse results than the existing method in terms of the number of false positives and f for recognitions based on z-axis acceleration and sound amplitude. The results using 3-axis acceleration in TIAA were better than the results using 3- acceleration using the existing method, but worse than TIA. TIAAS was able to achieve a higher f than the existing method, TIA using acceleration, and TIAA. Besides, TIAAS was able to reduce the number of false positives when using any sensor data and improved the problem of multiple taps recognitions in a single tap. The details of the results are described in the VI-B section and beyond.

TABLE I
THE NUMBER OF FALSE POSITIVE AND EVALUATION VALUE f IN EACH METHOD

Method		false positive	f
the existing method	z-axis	0.81	6.11
	3-axis	2.80	3.78
	sound	1.73	5.92
TIA	z-axis	0.10	5.82
	3-axis	0.16	5.45
	sound	0.12	7.95
TIAA	z-axis	0.92	3.88
	3-axis	1.46	3.51
	sound	22.55	-15.29
TIAAS		0.27	7.32

B. The existing method

We set T_{LAT} to 0.016 s and T_{DUR} to be long enough to make a tap recognition. The average of a_{THT} when f was the highest for each sensor data is shown in Table II.We set a_{THT} to maximize f at each BPM, and evaluation values when recognizing taps is shown in Fig. 10.In addition, T_{LAT}, T_{DUR}, and a_{THT} were set so that evaluation value would be the highest for each BPM. The evaluation values are shown in Fig. 11.

C. TIA

We set b to 0.016 s. The average of a when f was the highest for each sensor data is shown in Table III. We set a to maximize f at each BPM, and f when recognizing taps is shown in Fig. 12. In particular, when using sound had a high evaluation value. However, f using acceleration was lower than the existing method when the BPM was high. In addition, a and b were set so that the evaluation value would be the highest for each BPM. The evaluation values are shown in

TABLE II
a_{THT} WHEN f IS THE HIGHEST IN THE EXISTING METHOD

z-axis (m/s^2)	3-axis (m/s^2)	sound
0.36	0.52	11992

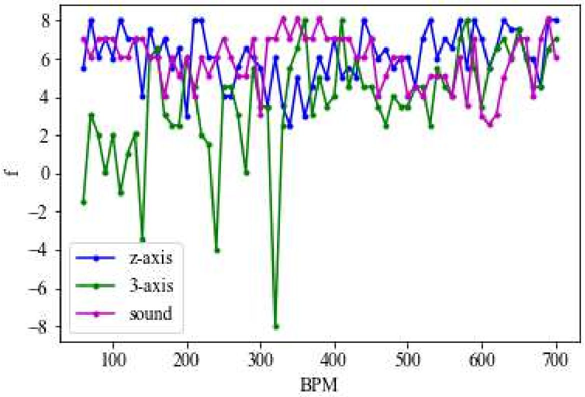


Fig. 10. Results of the existing method

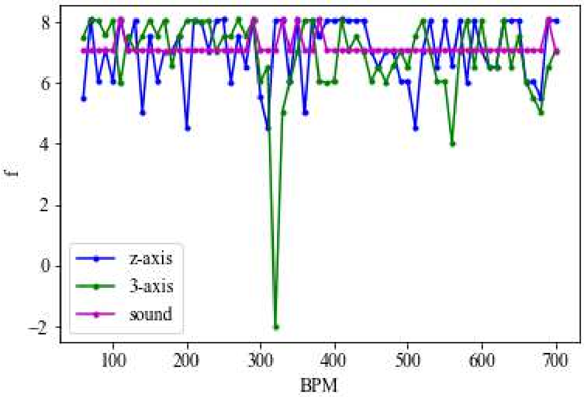


Fig. 11. Results when parameters are set to the optimal in the existing method

TABLE III
a WHEN f IS THE HIGHEST IN TIA

z-axis (m/s^2)	3-axis (m/s^2)	sound
0.37	0.62	7192

Fig. 13. The result is almost the same as when b is set to 0.016 s.

D. TIAA

For recognitions based on acceleration, $k = 0.375s$. For recognitions based on sound amplitude, $k = 19.27ms$. The average of a when f was the highest for each sensor data in TIAA using average attenuation is shown in Table IV. We set a to maximize f at each BPM. The evaluation values are shown in Fig. 14. Using the average attenuation drastically increased the number of times multiple taps recognitions in a single tap, resulting in a low f, especially for recognition using the sound amplitude. The average of a when f was the highest for each sensor data in TIAA using maximum attenuation is shown in Table V. We set a to maximize f at each BPM The evaluation values are shown in Fig. 15. Compared to using average attenuation, the number of false positives was reduced, resulting in a higher f. However, since the number of tap recognitions was reduced, f was lower than that of TIA.

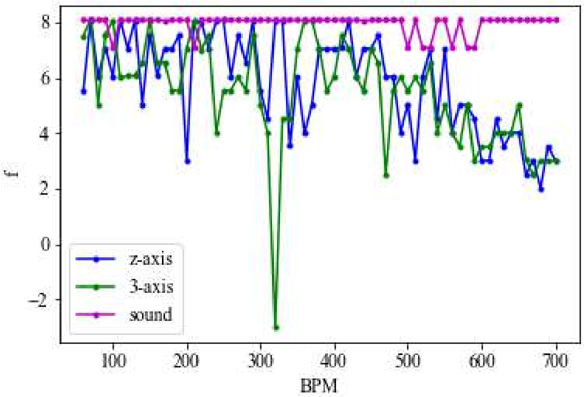


Fig. 12. results of TIA

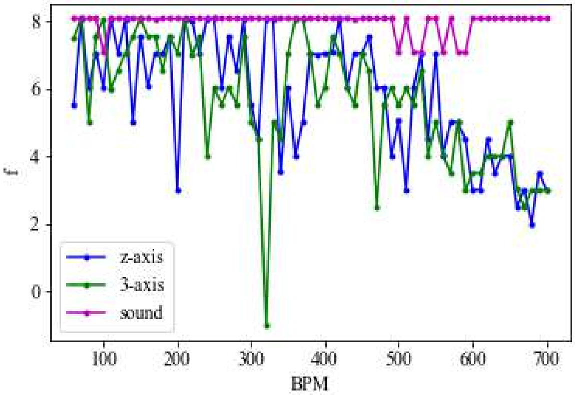


Fig. 13. Results when parameters are set to the optimal in TIA

TABLE IV
a WHEN f IS THE HIGHEST IN TIAA USING AVERAGE ATTENUATION

z-axis(m/s^2)	3-axis(m/s^2)	sound
0.40	0.44	10453

E. TIAAS

The average of a, c, and d when f was the highest for each sensor data is shown in Table VI. We set parameters to maximize f at each BPM and f are shown in Fig. 16.

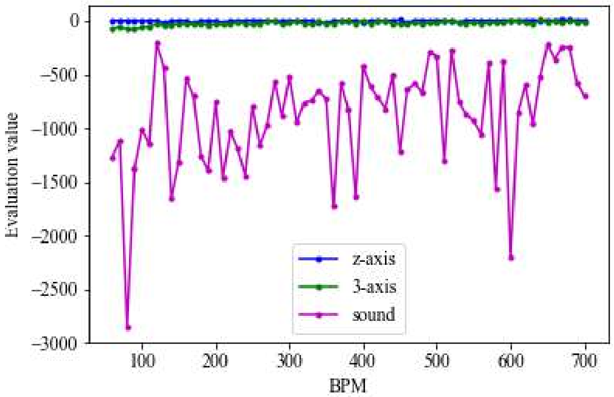


Fig. 14. Result of TIAA using average attenuation

TABLE V
a WHEN f IS THE HIGHEST IN TIAA USING MAXIMUM ATTENUATION

z-axis (m/s^2)	3-axis (m/s^2)	sound
0.30	0.41	8800

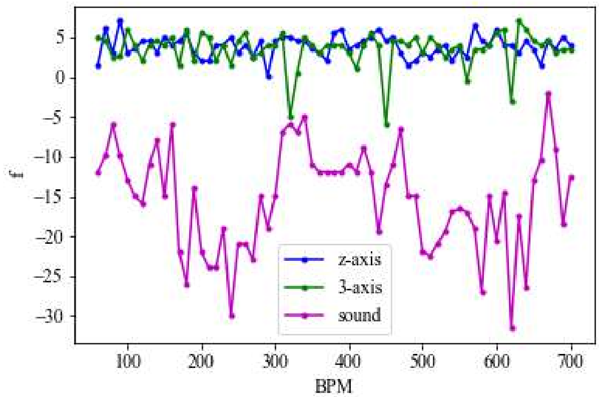


Fig. 15. Result of TIAA using maximum attenuation

TABLE VI
a, c, AND d WHEN f IS THE HIGHEST IN TIAAS

$a(m/s^2)$	c	d(s)
0.37	1953	0.021

VII. Conclusions

To solve the problem of duplicate tap recognition by the existing tap-recognition method, we proposed new methods; TIA, TIAA, and TIAAS. The results of the experiments showed that compared to the existing method, TIA was able to reduce the number of false positives and improved the problem of duplicate tap recognitions in a single tap. When using sound, TIA was able to recognize taps almost perfectly, but when using acceleration, the evaluation value was lower than the existing method due to the influence of reverberation. TIAA showed worse results than the existing method for recognitions based on z-axis acceleration and sound amplitude. TIAAS was able to achieve a higher f than the existing method, TIA using acceleration, and TIAA and improved the problem of duplicate tap recognitions in a single tap.

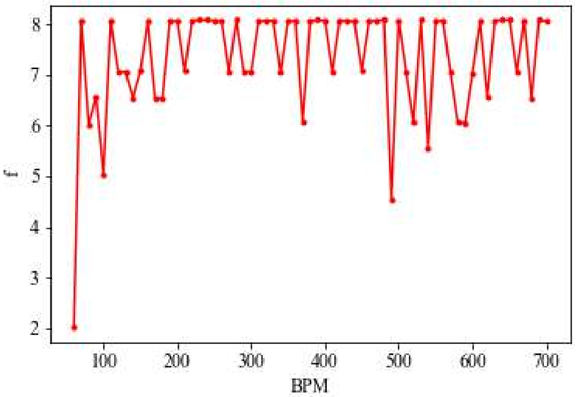


Fig. 16. Results of TIAAS

References

[1] Shin Norieda, Hideo Mitsuhashi and Makoto Sato, “FingerKeypad: Detection of Tap Location on Finger by Arrival Time Difference of Impact,” Human Interface Society, vol. 14, no. 4, pp. 393-402, 2012

[2] Chris Harrison, Desney Tan and Dan Morris, “Skinput: Appropriating the Body as an Input Surface, “ CHI’10, pp. 453-462, 2010

[3] Kazuya Murao and Tsutomu Terada, “A Motion Recognition Method by Constancy Decision, “Transactions of Information Processing Society of Japan, vol. 52, no.6, pp.1968-1979, 2011

[4] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, Shyam Gollakota, ”FingerIO: Using Active Sonar for Fine-Grained Finger Tracking,” CHI’16, pp.1515-1525, 2016

[5] Hiroshi Ishii, Craig Wisneski, Julian Orbanes, Ben Chun and Joe Paradiso ” PingPongPlus: Design of an Athletic-Tangible Interface for Computer-Supported Cooperative Play,” CHI’99, pp.394-401, 1999

[6] Toshiki Iso and Kenichi Yamazaki, ”Gait analyzer based on a cell phone with a single three-axis accelerometer,” MobileHCI’06, pp.141-144, 2006

[7] Hisashi Kurasawa, Yoshihiro Kawahara, Hiroyuki Morikawa and Tomonori Aoyama, “User Posture and Movement Estimation Based on 3-Axis Acceleration Sensor Position on the User’s Body,” UBI, vol. 2006-UBI-011, pp.15-22, 2006

[8] Naoki Iketani, Masaaki Kikuchi, Kenta Cho and Masanori Hattori, “Pedestrian Context Inference Using a 3-Axis Accelerometer;” UBI, vol. 2008-UBI-019, pp.75-80, 2008

[9] Kazushige Ouchi and Miwako Doi, “A Real-time Living Activity Recognition System by Using Sensors on a Mobile Phone,” Transactions of Information Processing Society of Japan , vol.53, no.7, pp.1675-1686, 2012

[10] Mono Wireless Inc. TWELITE 2525A Configuration Details. [Online]. Available: *https://mono-wireless.com/jp/products/TWE-Lite-2525A/manual_settings.html*

Pulse Rate and Blood Oxygen Monitor to Help Detect Covid-19: Implementation and Performance

Navid Bin Ahmed
Dept of EEE
Independent Univ, Banglaesh
Dhaka, Bangladesh
navidbinahmed@iub.edu.bd

Shahriar Khan
Dept of EEE
Independent Univ., Bangladesh
Dhaka, Bangladesh
skhan@iub.edu.bd

Nuzhat Arifa Haque
Dept of ECE
North South University
Dhaka, Bangladesh
nuzhat.haque@northsouth.edu

Md. Shazzad Hossain
Dept of ECE
North South University
Dhaka, Bangladesh
shazzad.hosain@northsouth.edu

***Abstract*—Covid has killed millions, and millions more are being infected every day. Severe infection causes Pneumonia associated with high pulse and breathing rates, and low blood oxygen. Pulse rate and blood oxygen are two parameters that doctors use to diagnose and measure Pneumonia and Bronchitis. Sensors for their measurement are widely available in developed countries, but their costs are prohibitively high for wide penetration in developing countries. Hardware prototypes incorporated with embedded software and IoT have been developed previously, but their costs have not been brought down sufficiently for wide penetration. We developed and tested the performance of an Atmel ATmega 328P MCU and MAX30100 sensor kit-based pulse rate and blood oxygen monitor. We applied the device on 12 subjects and compared the data with a commercially available, Rossmax SB150 pulse oximeter. A minimal deviation of 0.8175% for pulse rate and 0.425% for blood oxygen was obtained that endorses the accuracy of our algorithm and implementation. Cost analysis shows that implementation can be done at a cheap rate of about US$ 12 only.**

***Keywords*—** *Covid, pulse, blood oxygen, Pneumonia, Bronchitis, monitoring, cheap, oximeter, etc.*

I. Introduction

With coronavirus cases and deaths soaring across the world, the need for monitoring health parameters has increased greatly. Healthcare systems have been operating at their limits worldwide, coping with increasing numbers of patients. Infected patients have a higher pulse rate and lower blood oxygen. Pulse rate and blood oxygen may need to be monitored to prevent death. To make monitoring available to more people, sensor-based non-invasive devices have been developed that can be used at home. Thus researchers have great interest to develop measurement devices for mass production with high accuracy level and lower cost.

Due to the low penetration of health technology, especially in under-developed countries, low-cost oxygen measurement devices have been of interest in the literature [1]. IoT based low-cost remote pulse rate and blood oxygen monitoring has also received attention for telemedicine [2, 3, 4]. Though, most devices obtained good feedback, ensuring the validated performance of devices has great significance for quality assessment and diagnosis [5, 6, 7]. The goal of this study is twofold. One is to develop a low-cost device to measure pulse rate (BPM) and blood oxygen (%SpO2). The other is to compare the performance of the developed device with a standard device, Rossmax SB150.

Infrared (IR) and LEDs have sensor-based applications, known as Photoplethysmography for the measurement of heart rate and blood oxygen saturation. The MAX30100 has integrated Infrared (IR) and Red LED sensors at a low price. While many researchers have developed non-invasive blood oxygen saturation and pulse rate measurement devices [8, 9, 10, 11, 12, 13, 14], we developed a MAX30100-based portable system at the lowest possible cost and analyzed the performance on twelve subjects. A number of sensor-based monitoring systems have already been developed for affordable applications [15, 16, 17, 18], but keeping in mind the issues of sustainability and quality [19, 20], we have conducted a performance analysis after the implementation.

A recent study varied current through the Red LED of MAX30100 to measure blood oxygen, showing that acceptable performance is obtained at 14.2 mA [21]. Another study supported reducing the motion artefact of the sensor kit using an acceleration sensor [22]. A few subjects were assessed showing a high deviation of 5% for oxygen and 7% for pulse rate [23]. This was insufficient for correct estimation and diagnosis and the cost was high at US$ 20. To evaluate the performance of the sensor, a group of researchers analyzed data from 5 subjects and assessed the mean deviation [24]. We developed the device at a lower cost with higher accuracy. We analyzed a set of practical data from 12 subjects finding a deviation of only 0.8175% for pulse rate (BPM) and 0.425% for SpO2 (%). The cost analysis shows that our device costs US$ 12 only, which is cheaper than most previously implemented devices. We found minimal percentage error of the device and sensor data, ensuring a more effective performance than other systems to diagnose Covid.

II. Methodology

The objective was to study the current price of the available systems/devices and implement them at a low cost. Several

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

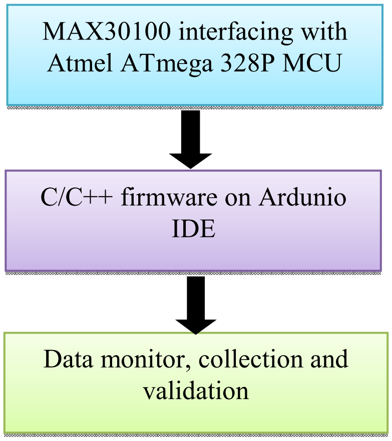

MAX30100 interfacing with Atmel ATmega 328P MCU
C/C++ firmware on Ardunio IDE
Data monitor, collection and validation

Fig. 1. Block diagram of pulse rate (BPM) and blood oxygen saturation (%SpO2) monitor

steps were taken for developing a prototype.

- We studied available devices on various e-commerce platforms. We focused on device materials, designs, prices, specifications, availability, and carried out a comparative study.
- We analyzed the cost, quality of the device and listed the components to purchase from online and physical shops.
- We studied previously published papers and focused on the most cited ones on ResearchGate and Google Scholar.
- We simulated the system before implementing the hardware prototype.

Available sensors and tools of acceptable quality and low cost were purchased from online and physical shops. C and C++ coding were applied on Atmel ATmega 328P MCU (Arduino UNO) using Arduino IDE. The same algorithm can be used for extending the prototype i.e. hardware implementation for other parameters. Our implementation steps are illustrated in figure 1.

III. Experimental Hardware Setup

Using the methodology in the previous section, we collected and verified the functioning of the sensors and microcontroller unit. We interfaced MAX30100 with Atmel ATmega 328P MCU. The analog data was observed on the serial monitor. We coded for the ADC to convert the analog data to digital. We calculated blood oxygen saturation (%SpO2) and then pulse rate (BPM) (equation 1).

$$\%SpO2 = \frac{\log(\frac{redACValueSqSum}{samplesRecorded})}{\log(\frac{irACValueSqSum}{samplesRecorded})} x100 \qquad \ldots\ (1)$$

Following a lookup table, we determined the oxygen saturation as (SpO2= SpO2LUT [index]). We observed the data on a serial monitor on the Arduino IDE. We then collected data from 12 subjects of different ages and validated comparing with a commercially available pulse oximeter, Rossmax SB150.

Finally, we calculated the error percentage to analyze the performance of our system. The sensor kit and its functional principle are shown in the following figures 2 and 3 respectively. Figure 2 upholds the sensor and 3 demonstrates how it functions (datasheet). Figure 4 shows the implemented device while being applied on the subject.

IV. Data Analysis: Collection and Validation

We experimented with the implemented device on several subjects of different ages. The collected data verified the accuracy of the sensor. We also validated the data comparing with a commercially available device, Rossmax SB150. Data set on the same parameters were also validated in [21] using NewTech PM100. We found a minimal deviation with an acceptable range of 0.8175% for pulse rate (BPM) and 0.425% for SpO2 (%) as depicted in the following Table I. The error was calculated as the following equation 2.

$$\text{deviation} = \frac{\text{experimental value} - \text{accepted value}}{\text{accepted value}} x100\% \qquad \ldots(2)$$

Figures 5 and 6 illustrate the validation of the collected data from various subjects that compare well with the pulse rate and oxygen saturation from a commercially available pulse oximeter. Figure 5 shows the number of subjects (X-axis) and corresponding pulse rates (Y-axis) collected from

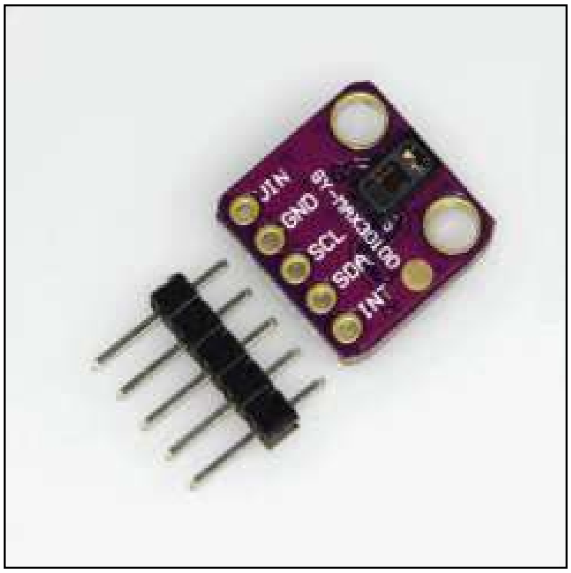


Fig. 2. MAX30100 sensor kit

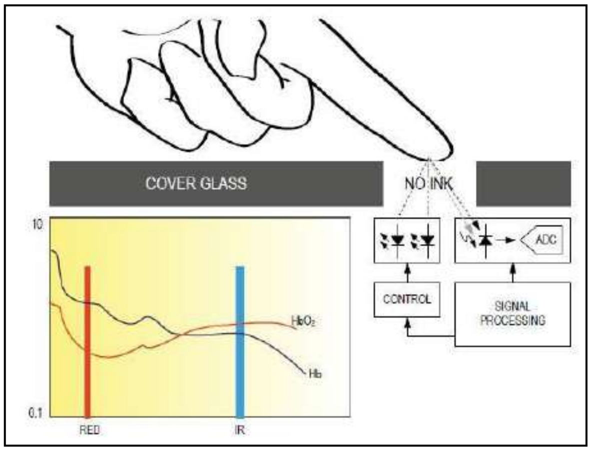


Fig. 3. Functional principle of the sensor (datasheet, 2019)

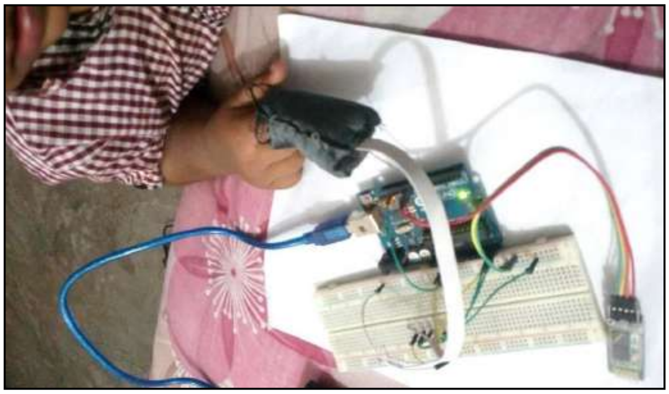


Fig. 4. Application of implemented Oximeter on children that measures SpO2 (%)

the implemented and reference devices, demonstrating a very marginal deviation in the logarithmic scale. Figure 6 shows the number of subjects (X-axis) and corresponding blood oxygen (Y-axis) from the implemented and reference devices, showing very close agreement in the logarithmic scale.

Table I. Collected data with minimal error %

Sub	Age	Heartrate (BPM)		Deviation %	SpO2 (%)		Deviation %
		Experimental	*Accepted*		*Experimental*		*Accepted*
1	29	67	66	1.5	98	98	0
2	24	85	85	0	97	98	-1.02
3	26	80	81	-1.23	96	96	0
4	28	83	83	0	98	97	1.03
5	32	74	75	-1.33	97	97	0
6	25	63	63	0	98	98	0
7	73	91	90	1.11	97	97	0
8	63	66	66	0	96	96	0
9	5	92	90	2.22	97	99	-2.02
10	28	74	73	1.37	96	97	-1.03
11	45	94	95	1.05	97	97	0
12	33	68	68	0	98	98	0

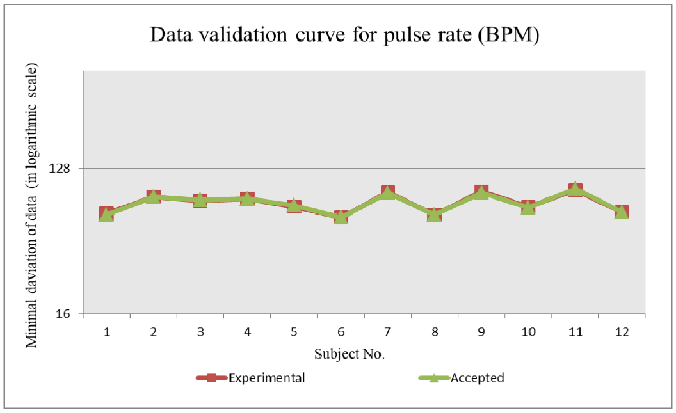


Fig. 5. Collected pulse rate showing minimal deviation as compared to a standard device (Rossmax SB150 pulse oximeter)

V. Scopes and Opportunities

While the number of dead and infected from the Covid-19 pandemic soaring across the world, the detection of Pneumonia is a global priority. Pneumonia kills more than 1.5 million children every year. Rural areas across low resource countries do not have enough qualified doctors and resources. People have to travel to cities for proper medical treatment, which they can barely afford. The possible applications may also be extended to the following categories.

A. Patient Monitoring

The developed device can be commercially deployed for monitoring patients. With the availability of telemedicine, wireless health monitoring systems have good potential in healthcare.

B. Telemedicine Application

Several organizations may come up with telemedicine programs. Such a tool has many possibilities to be incorporated as a remote data monitor in telemedicine.

C. Biomedical Innovation for Larger Application

This project can also be expanded as part of larger biomedical applications. Several sensors for other health parameters can be integrated.

VI. Cost Feasibility

As seen in the cost breakdown below, the implemented device is the cheapest among all other available devices. Implementation cost is around US$ 12 including the MCU, sensor and instrumentation.

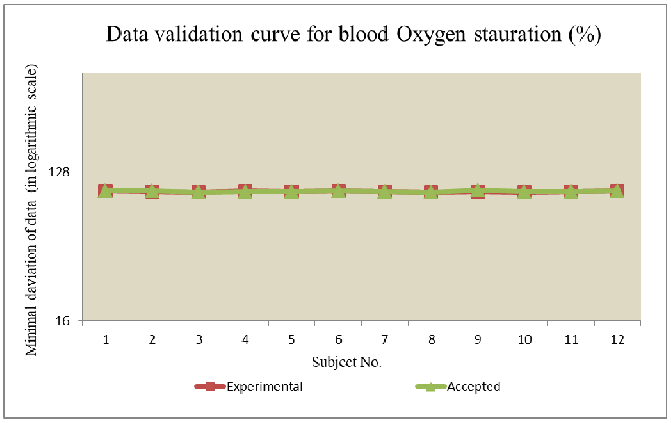


Fig. 6. Minimal deviation in collected blood oxygen saturation (%SpO2) with Rossmax SB150 pulse oximeter.

Table II illustrates the breakdown of cost:

TABLE II. Cost analysis

Apparatus	***Cost (US$)***
Atmel ATmega 328P MCU	4.00
MAX30100	7.00
Others	1.00
Total	12.00

Large scale implementation may greatly reduce the cost, owing to economies of scale. This pulse and blood oxygen data are prerequisites in disease detection and health condition analysis. There are widespread opportunities for further research and development. Some possible applications include remote Covid symptom detection, mobile app-based telemedicine, heart disease detection and analysis, wireless patient monitoring system, biomedical lab applications in universities and low-cost health checkup monitoring system for low resource countries. Research is being done to develop an Artificial Intelligence wearable health device, which can be a breakthrough in the field of healthcare.

VII. Conclusion

A device to measure heart rate and monitor blood oxygen (SpO2) has been developed with lower cost and higher credibility. The results and analysis validate the accuracy of the sensor. There is the possibility of integrating more parameters like breathing rate, chest-in-draw, body temperature etc. to develop even larger monitoring and detection system to fight Pneumonia and Covid. Cost analysis shows that the device is affordable. It is expected that this paper will contribute to the development of affordable technologies for measurement of health parameters.

References

[1] N. B. Ahmed, S. Khan, “Technological and medical progress in rural Bangladesh; A survey and study of the challenges”, IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, Canada, September 2020.

[2] N. Agarwal, S. Agarwal, A. Kumar, R. Kini, “Optimized low-power low-cost pulse oximeter for remote patient monitoring”, Texas Instruments India Educators’ Conference, 2013.

[3] K. M. Gamboa, A. G. Vargas, “Development of a low-cost pulse oximeter simulator for educational purposes”, IEEE Press, 2018.

[4] L. J. Frost, M. R. Reich, “Creating access to health technologies in poor countries”, Health Affairs, 28(4), pp 963-973, July 2009.

[5] W. Ahmed, S. Khan, K. A. Sidek, “Development of a photoplethysmography signal processing method for Oxygen measurement concentration”, 5th International Conference on Computing and Communication Engineering, 2014.

[6] G. W. J. Clarke, A. D. C. Chan, A. Adlar, “Effects of motion artifact on the blood oxygen saturation estimate in pulse oximetry”, IEEE International Symposium on Medical Measurements and Applications (MeMeA), Portugal, June 2014.

[7] Y. S. Yan, Y. T. Zhang, “A model-based artifact reduction method for the non-invasive estimation of blood oxygen saturation”, IEEE/ EMBS International Summar School of on Medical Devices and Biosensors (ISSS-MD), IEEE Press, 2004.

[8] L. Hakizimana, M. Wannous, “Distributed medical record system using portable devices, case study of Rwanda work in progress”, IEEE International Conference on Applied System Invention (ICASI), IEEE Press, New York, 2018.

[9] D. Mishara, S. Chandra, A. Chandra, S. Jain, M. Sarkar, “A portable system for real time non contact blood oxygen saturation measurements” 2017.

[10] G. Ates, K. Polat, “Measuring of oxygen saturation using pulse oximeter based on Fuzzy logic”, IEEE International Symposium on Medical Measurement and Applications, Budapest, Hungary, May 2012.

[11] F. M. Ferrera, F. J. Ferrero, C. Blanco, J. C. Vierra, M. G. Vega, J. R. Blanco, “Design of a low cost instrument for pulse oximetry”, IEEE Instrumentation and Measurement Technology Conference, Italy, 2006.

[12] N. Sutar, M. Parihar, R. Ijare, K. Gowari, “Design and development of SMD based wearable pulse oximeter”, IEEE International Conference on Communication and Signal Processing, India, 2016.

[13] F. Tang, X. Zhou, A. Bermek, “A low power and fast tracking light to frequency converter with adaptive power scaling for Blood SpO2 sensing”, IEEE Transactions on Biomedical Circuits and Systems, vol. 13, no. 1, pp. 26-37, February 2019.

[14] J. Wan, Y. Zou, Y. L, J. Wang, **“**Reflective type blood oxygen saturation detection system based on MAX30100”, IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), China, 2017.

[15] N. B. Ahmed, S. Khan, K. N. Tahsin, and I. Afrida, “Implementation of a low-cost heartbeat monitor as part of a larger health monitoring system”. 3rd International Conference on Inventive Computation Technologies, IEEE, pp. 813-818, 2018.

[16] N. B. Ahmed, K. N. Tahsin, S. Khan, M. Hoq and D. Alam, “Design and implementation of a microcontroller based flammable gas detector and automatic alarm system to ensure the industrial and domestic safety”, Asian Journal of Science and Technology, vol. 9, issue 1, pp. 7425-7429, January 2018.

[17] N. B. Ahmed, S. Khan, K. N. Tahsin, F. Hafiz, M. Hoq, M. A. S. Haque and F. Akter, “A low-cost microprocessor based line-follower robot for monitoring temperature and humidity under adverse conditions”, International Journal of Pure and Applied Mathematics, vol.119, pp. 12611-12619, Academic Publications Ltd, May 2018.

[18] N. B. Ahmed, S. Khan and F. Hafiz, “Design and implementation of an autonomous wireless information transceiver robot”, Proceedings of National Conference on Electronics and ICT, Bangladesh Electronics Society, Dhaka, April 2017.

[19] S. Khan, Telecommunication Engineering, S. Khan, June 2016, Dhaka, Bangladesh.

[20] T. S. Arulananth and B. Shilpa, “Fingertip based heartbeat monitoring system using Embedded Systems”, Proceedings of IEEE International Conference of Electronics, Communication and Aerospace Technology, Coimbatore, India, 2017, pp. 227-230.

[21] K. B. Saçan and G. Ertaş, "Performance assessment of MAX30100 SpO2/heartrate sensor," 2017 Medical Technologies National Congress (TIPTEKNO), pp. 1-4, Trabzon, 2017.
doi: 10.1109/TIPTEKNO.2017.8238126.

[22] L. Huang, M. Li, Y. Zhao, D. He, “Optimizing blood oxygen saturation measurement accuracy and consistency using fingertip acceleration and blood perfusion index”, IEEE International Conference on Mechatronics and Automation, China, 2018. **22**

[23] S. Chugh, J. Kaur, “Low cost calibration free pulse oximeter”, IEEE India Conference (INDICON), New Delhi, India, 2015.

[24] S. S. Randhawa, S. S. Ahluwalia, R. C. Gupta, “Evaluation of the performance of C.S.I.O. pulse oximeter”, Proceedings of RC IEEE-EMBS & 14th BMESI, IEEE Press, 1995.

Limitations of Loop Gain in Motion Models of Physical Systems

Tri Minh Tran, Yasunori Kobori, Anna Kuwana, Haruo Kobayashi

Division of Electronics and Informatics, Gunma University, Kiryu 376-8515, Japan

Email: trantri.ks@gmail.com; kobori@gunma-u.ac.jp; kuwana.anna@gunma-u.ac.jp; koba@gunma-u.ac.jp

***Abstract*— This paper presents some limitations of the classical concept of loop gain in many mathematical motion models of the real objects. It is well known that there are some advanced concepts in control theory and a lot of them are very heavy on the loop gain. Therefore, the loop gain is heard in many topics like frequency compensation, phase-gain margins, negative feedback systems. However, the loop gain cannot be used to analyze the overshoot phenomenon in many feedback amplifiers because it doesn't show the operating regions of high-order mechatronic systems. In addition, the theoretical calculation, the laboratory simulation and the practical measurement of loop gain are not unique. As a result, this is the biggest limitation in mathematical models of motion that a lot of engineers have trouble with loop gain. To overcome some disadvantages of the loop gain, the use of the self-loop function allows engineers to do the ringing test simpler because it is easily verified in MATLAB calculator, SPICE simulator, and Network Analyzer. In order to extend the application range of the Nichols chart of self-loop function in high-order mechatronic systems, the feasibility of complex functions and the explanation of overshoot phenomena are described in detail. It is shown that complex functions are useful tools for mathematical motion models of mechatronic systems.**

***Keywords*— Ringing Test, Nichols Chart, Phase Margin, Negative Feedback, Self-loop Function, Loop Gain**

I. Introduction

A. Motivation

Recent researches show that ringing occurs in many high-order mechatronic systems [1]. Ringing or overshoot makes these systems unstable. Hence, a general stability test for mechatronic systems should be introduced. The stability test using loop gain is a valuable tool for analyzing the stable region of high-order mechatronic systems. However, engineers often find the loop gain confusing and difficult to understand because the theoretical analysis and the practical measurement of the phase margin are not unique. In other words, the values of loop gain in these feedback systems are the assumption values. There are some limitations of the concept of loop gain because it is only used in classical control theory [2]. In addition, the operating regions of a high-order mechatronic system are not shown on the Nichols chart of loop gain.

In addition, there are some limitations of numerical methods. Differential equations are normally derived on the basis of a rate of change (velocity) or a second derivative (acceleration) that give a relation based on a physical law such as superposition principle. Expressing the force as a function of time allows one to solve for the acceleration, integrate that to get velocity, and integrate that to get position. Some differential equations are not as well-behaved, and show singularities due to a failure to model the problem correctly, or a limitation of the model that was not apparent. Some differential equations can be solved analytically in closed form, but most have to be approximated by numerical procedures, which can be unstable. Therefore, properties of complex functions are proposed to do the ringing test or overshoot test for high-order mechatronic systems [3].

The characteristics of complex functions give a whole new perspective to math-intensive mechatronic courses at most universities. A mathematical motion model captures the relationship between the signal source and the transmission network [4]. The principal motivation comes from the wide applications of complex function.

It is a main design methodology capable of providing theoretical examination of the operating region of a high-order mechatronic system. In this study, the results here not only provide a theoretical basis for analytically justifying some conventional concepts of existing complex functions but also open up the possibility of investigating the operating regions of high-order mechatronic networks by simply plotting the self-loop function.

B. Objectives and paper organization

Overshoot phenomena can be seen in many mathematical motion models of high-order mechatronic systems [5]. The impact of the overshoot on the mechatronic systems is usually addressed in the term “parasitic elements” [6]. Therefore, it is clearly understood that the occurrence, peak value, additional delay, and other damaging effects cannot be predicted or removed easily [7]. To do the ringing test for high-order mechatronic systems, a square wave is usually used to put in the input port [8]. Depending on the operating region of a high-order system, the output waveforms can be over-damping, or critical damping, or under-damping as shown in Fig. 1. The ringing test will be analysed in detail in section III.

The use of complex functions is an essential theory in many motion models of high-order mechatronic systems. In addition, transfer function is covered in many courses such as control theory, network analysis theory, but the complexity of network topologies requires to develop a specific approach in order to fully analyse a real circuit within the framework of complex function theory [9]. It must be stressed that complex function is present in all systems; therefore, it is important that a good understanding of complex function concepts in mechatronic systems should be developed [10].

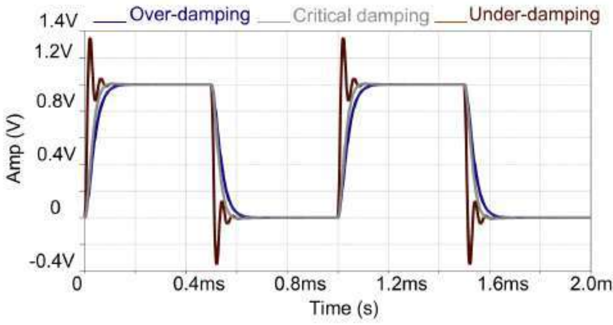


Fig. 1. Waveforms of output square signals in a high-order system.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

This research focuses on the various mechatronic networks, and their operating regions. This paper contains six sections. The research background for some motion models in mechatronic systems is presented in Section II. The limitations of loop gain are discussed in Section III. Several simulation results and some practical measurements of the ringing test are shown in Section IV. Future work and conclusion are presented in Sections V.

II. Research Background

A. Limitations of loop gain

This section points out some limitations of loop gain in some motion models in the mechatronic systems. Many challenges behind the ringing test for high-order mechatronic systems arise from the mathematical models of motion using the term "loop gain" [11]. A simplified transfer function is rewritten as in Eq. (1). Here, Aβ is called loop gain, A is the numerator function and β is the return-ratio. Fig. 2 shows the block diagram of a general control system. The input and output signals of the adaptive feedback systems can be either non-cyclic or periodic with the frequency variable [12].

$$H = \frac{A}{1 + A\beta} \approx 1 \qquad (1)$$

In an adaptive feedback control system, the reference signal and the feedback signal are subtracted as shown in Fig. 3. In general, a comparator is used to calculate the difference between the reference signal and the feedback signal [13]. In other words, the output signal is sampled and then fed back to the input to form an error signal that drives the plant or the actuating system. The actuating system is usually modeled by an open-loop function. The error signal can be used to eliminate or at least considerably reduce the effects of the variations of the actuating system [14].

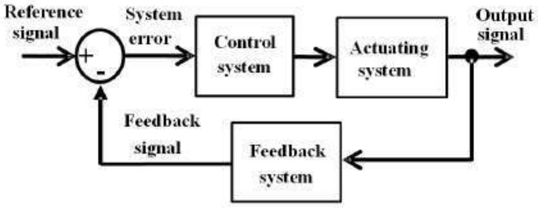


Fig. 2. Simplified model of a general adaptive feedback control system.

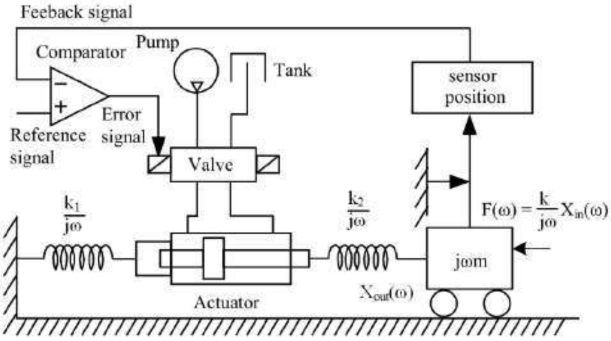


Fig. 3. Block diagram of a mechanical control system.

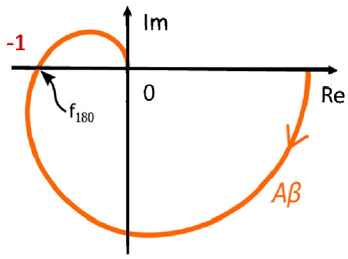


Fig. 4. Nyquist plot of a loop gain in an adaptive feedback system.

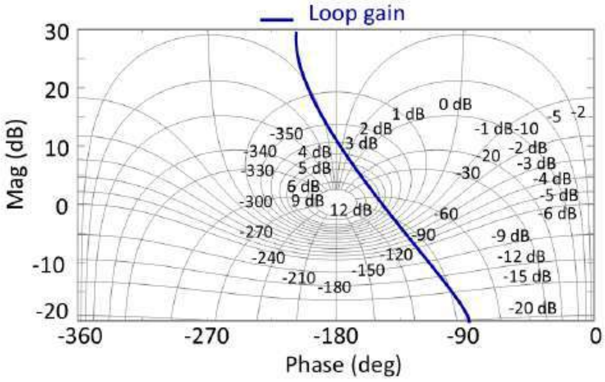


Fig. 5. Nichols plot of a loop gain in an adaptive feedback system.

Fig. 4 shows the conventional Nyquist chart of a loop gain in a control system. As Nyquist's stability condition does not show the operating regions of high-order mechatronic systems, the Nyquist diagram of loop gain cannot be applied for these systems. It is very difficult to derive and measure the exact value of the loop gain because it is an approximation model in MATLAB simulation [15]. In addition, the concept of loop gain is found in many negative feedback systems such as unity gain amplifiers, inverting amplifiers, negative feedback amplifiers while an exact value of loop gain is still not shown [16].

Fig. 5 shows the conventional Nichols plot of a loop gain in an adaptive feedback system. It is not widely used in both SPICE simulations and practical measurements because it is very difficult to predict loop gain when the properties of input and output signals are different. Hence, the characteristics of the control network and the linear network are significantly different [17].

B. Roles of complex functions

Complex functions give various meaningful insights in the motion models of mechatronic systems. They are applied in electrical, electronic, and mechanical systems [18]. Mechatronics in general would involve more physics with the math being somewhat trivial compared to complex functions. The transfer function H(ω) of a mechatronic network is the ratio of the output signal $V_{out}(\omega)$ to the input signal $V_{in}(\omega)$ as a function of the frequency [19]. A simplified transfer function is rewritten as in Eq. (2), where A(ω) is the numerator function, and L(ω) is the self-loop function. The phase margin at unity gain of the self-loop function shows the operating region of a high-order mechatronic system.

$$H(\omega) = \frac{V_{out}(\omega)}{V_{in}(\omega)} = \frac{A(\omega)}{1 + L(\omega)} \qquad (2)$$

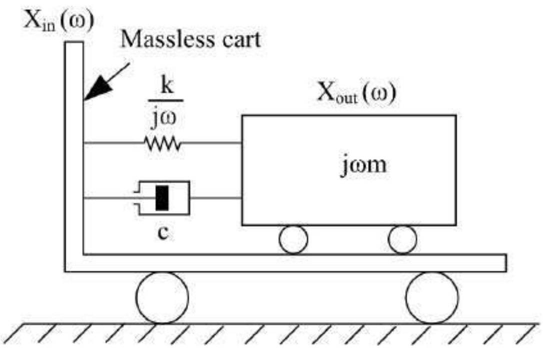


Fig. 6. Simplified model of a spring-mass dashpot system.

C. Behaviors of a second-order mechanical system

This section investigates the behaviors of a second-order mechanical system. A spring mass system is usually used to represent a complex mechanical system as shown in Fig. 6. Based on the superposition principle, Newton's first law points out the relationship between force and motion in a specific transmission space [20]. After applying the superposition principle at the output node for the translational system, the node equation is given in Eq. (3).

$$\left[j\omega m + c + \frac{k}{j\omega}\right] X_{out}(\omega) = \left[c + \frac{k}{j\omega}\right] X_{in}(\omega); \qquad (3)$$

Then, the transfer function and the self-loop function of this system are given in Eq. (4).

$$H(\omega) = \frac{X_{out}(\omega)}{X_{in}(\omega)} = \frac{b_0 j\omega + 1}{a_0 (j\omega)^2 + a_1 j\omega + 1}; \qquad (4)$$

$$L(\omega) = a_0 (j\omega)^2 + a_1 j\omega; \;\; b_0 = \frac{c}{k}; a_0 = \frac{m}{k}; a_1 = \frac{c}{k}$$

The operating regions of a second-order system are over-damping, critical damping, and under-damping as shown in TABLE I. In a 2nd-order mechatronic system, the operating regions are over-damping, critical damping, and under-damping. In case of under-damping, the overshoot will cause the ringing [21]. Three typical cases of a second-order system are given to determine the operating region of the system as shown in Eq. (5). Here, the coefficients of the transfer functions are under-damped (1:1:1), critically damped (1:2:1), and over-damped (1:3:1).

The Bode plot of these transfer functions, the Nichols plot of these self-loop functions are shown in Fig. 7, and Fig. 8, respectively. The phase margin is observed at unity gain of the self-loop function. The simulation results of the transfer functions and the self-loop functions are summarized in TABLE II.

$$H_1(\omega) = \frac{1}{(j\omega)^2 + j\omega + 1}; L_1(\omega) = (j\omega)^2 + j\omega;$$

$$H_2(\omega) = \frac{1}{(j\omega)^2 + 2j\omega + 1}; L_2(\omega) = (j\omega)^2 + 2j\omega; \qquad (5)$$

$$H_3(\omega) = \frac{1}{(j\omega)^2 + 3j\omega + 1}; L_3(\omega) = (j\omega)^2 + 3j\omega;$$

TABLE I. BEHAVIOURS OF 2ND-ORDER SELF-LOOP FUNCTION

Element	Value	
$\lvert L(\omega)\rvert$	$\omega\sqrt{(a_0\omega)^2 + a_1^2}$	
$\theta(\omega)$	$\frac{\pi}{2} + \arctan\frac{a_0\omega}{a_1}$	
Case 1	Over-damping ($\Delta = a_1^2 - 4a_0 > 0$)	
$\omega_1 = \frac{b}{2a}\sqrt{\sqrt{5}-2}$	$\lvert L(\omega_1)\rvert > 1$	$\pi - \theta(\omega_1) > 76.3^o$
Case 2	Critical damping ($\Delta = a_1^2 - 4a_0 = 0$)	
$\omega_1 = \frac{b}{2a}\sqrt{\sqrt{5}-2}$	$\lvert L(\omega_1)\rvert = 1$	$\pi - \theta(\omega_1) = 76.3^o$
Case 3	Under-damping ($\Delta = a_1^2 - 4a_0 < 0$)	
$\omega_1 = \frac{b}{2a}\sqrt{\sqrt{5}-2}$	$\lvert L(\omega_1)\rvert < 1$	$\pi - \theta(\omega_1) < 76.3^o$

TABLE II. SIMULATION RESULTS OF A 2ND-ORDER MECHANICAL SYSTEM

Case	Over-damping	Critical damping	Under-damping
Magnitude (transfer function)	**-12 dB**	**-6 dB**	**1 dB**
Phase margin (self-loop function)	**82° (observed at 98°)**	**76.3° (observed at 103.7°)**	**35° (observed at 145°)**

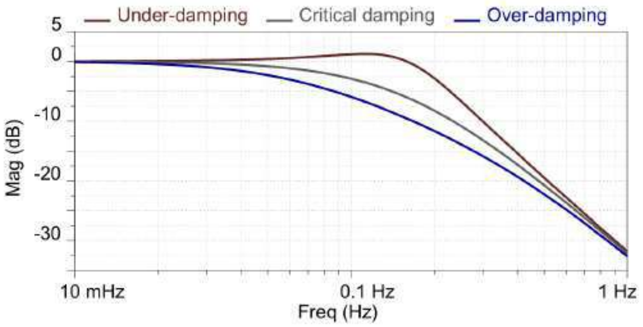


Fig. 7. Bode plot of the 2nd-order transfer function of a mechanical system.

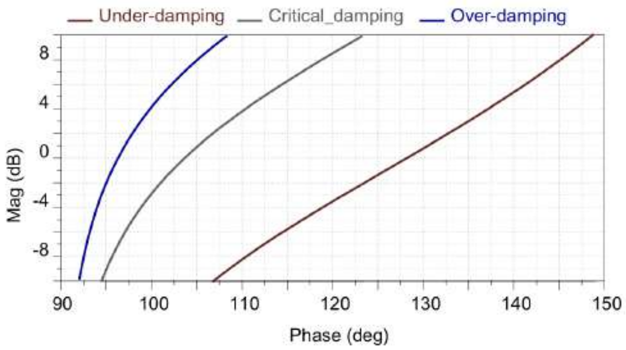


Fig. 8. Nichols chart of the 2nd-order self-loop function.

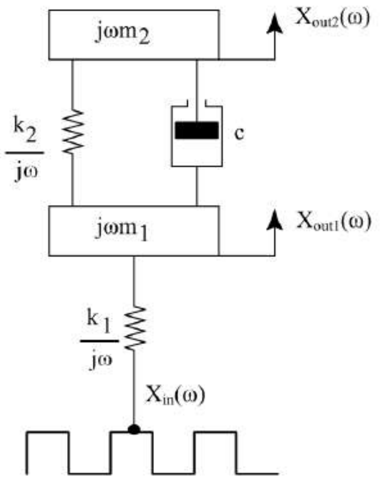


Fig. 9. Simplified model of a suspension system.

D. Operating regions of a fourth-order mechanical system

In this section, the behaviours of a 4th-order suspension system are investigated. In general, three basic types of suspension components are linkages, springs, and shock absorbers [22]. Fig. 9 shows a simplified model of a suspension system. Assume that displacements $X_{out1}(\omega)$ and $X_{out2}(\omega)$ are measured from respective mass positions in the absence of the input $X_{in}(\omega)$. Apply the superposition principle to the present system, the node equations are given in Eq. (6).

$$\begin{aligned}&\left(\frac{k_1}{j\omega}+m_1 j\omega+c+\frac{k_2}{j\omega}\right)X_1(\omega)=\left(c+\frac{k_2}{j\omega}\right)X_2(\omega)+\frac{k_1}{j\omega}X_{in}(\omega);\\&\left(m_2 j\omega+c+\frac{k_2}{j\omega}\right)X_2(\omega)=\left(c+\frac{k_2}{j\omega}\right)X_1(\omega);\end{aligned}\qquad(6)$$

Hence, the transfer functions and the self-loop function of this system are given in Eq. (7). The values of the constant variables of the suspension system are given in TABLE III. The use of Pascal's triangle is a simple way to calculate the binomial coefficients of a high-order mechatronic system as shown in TABLE IV. Three typical cases of a 4th-order system are given to determine the operating region of the system as shown in Eq. (8). Here, the coefficients of these transfer functions are under-damped (1:2:3:2:1), critically damped (1:4:6:4:1), and over-damped (1:9:10:9:1).

$$\begin{aligned}&H_1(\omega)=\frac{X_1(\omega)}{X_{in}(\omega)}=\frac{b_0(j\omega)^2+b_1 j\omega+1}{a_0(j\omega)^4+a_1(j\omega)^3+a_2(j\omega)^2+a_3 j\omega+1};\\&H_2(\omega)=\frac{X_2(\omega)}{X_{in}(\omega)}=\frac{b_1 j\omega+1}{a_0(j\omega)^4+a_1(j\omega)^3+a_2(j\omega)^2+a_3 j\omega+1};\\&L(\omega)=a_0(j\omega)^4+a_1(j\omega)^3+a_2(j\omega)^2+a_3 j\omega;\end{aligned}\qquad(7)$$

TABLE III. VALUES OF GIVEN VARIABLES OF A SUSPENSION SYSTEM.

Variable	Value	Variable	Value
b_0	$\frac{m_2}{k_2}$	a_1	$\frac{c(m_1+m_2)}{k_1k_2}$
b_1	$\frac{c}{k_2}$	a_2	$\frac{m_1k_2+m_2(k_1+k_2)}{k_1k_2}$
a_0	$\frac{m_1m_2}{k_1k_2}$	a_3	$\frac{c}{k_2}$

TABLE IV. PASCAL'S TABLE OF THE BINOMIAL COEFFICIENTS

	Coefficients of high-order mechatronic systems												
2nd					1		2		1				
3rd				1		3		3		1			
4th			1		4		6		4		1		
5th		1		5		10		10		5		1	
6th	1		6		15		20		15		6		1

$$\begin{aligned}&H_1(\omega)=\frac{j\omega+1}{(j\omega)^4+2(j\omega)^3+3(j\omega)^2+2j\omega+1};\\&L_1(\omega)=j\omega\left[(j\omega)^3+2(j\omega)^2+3j\omega+2\right];\\&H_2(\omega)=\frac{j\omega+1}{(j\omega)^4+4(j\omega)^3+6(j\omega)^2+4j\omega+1};\\&L_2(\omega)=j\omega\left[(j\omega)^3+4(j\omega)^2+6j\omega+4\right];\\&H_3(\omega)=\frac{j\omega+1}{(j\omega)^4+9(j\omega)^3+10(j\omega)^2+9j\omega+1};\\&L_3(\omega)=j\omega\left[(j\omega)^3+9(j\omega)^2+10j\omega+9\right];\end{aligned}\qquad(8)$$

Figs. 10 and 11 show the Bode plot of these transfer functions, the Nichols plot of these self-loop functions, respectively. The phase margins observed at unity gain of the self-loop function are summarized in TABLE V.

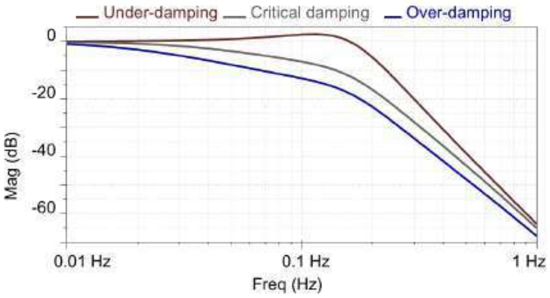


Fig. 10. Bode plot of the 4th-order transfer function.

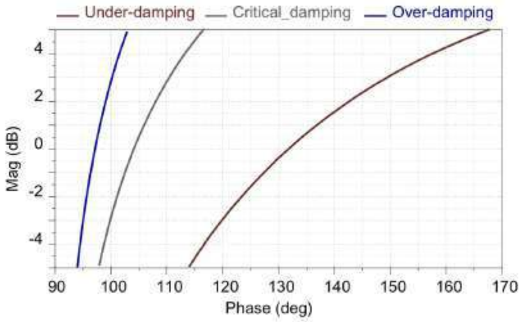


Fig. 11. Nichols chart of the 4th-order self-loop function.

TABLE V. SIMULATION RESULTS OF A 4ND-ORDER MECHANICAL SYSTEM

Case	Over-damping	Critical damping	Under-damping
Magnitude (transfer function)	-15 dB	-9 dB	1 dB
Phase margin (self-loop function)	82° (observed at 98°)	76.3° (observed at 103.7°)	48° (observed at 132°)

III. Behaviors of Feedback Amplifier Networks

A. Loop gain and self-loop function of inverting amplifiers

This section points out the differences between the loop gain and the self-loop function in the inverting amplifiers. Figs. 12 and 13 show the simplified models of two types of the inverting amplifier which are used in many electronic systems [23]. The open-loop function A(ω) of the op amp is described in Eq. (9).

$$A(\omega)=\frac{10^5}{1+j\dfrac{\omega}{2*10^2\pi}}; \qquad (9)$$

Eq. (10) gives the node equation at node V_A in the single-ended inverting amplifier which is derived based on the superposition principle.

$$V_A\left(\frac{1}{R_1}+\frac{1}{R_2}\right)=\frac{V_{in}}{R_1}+\frac{V_{out}}{R_2};\ -V_A A(\omega)=V_{out}\ ; \qquad (10)$$

Then, the transfer function and the self-loop function of this amplifier are given in Eq. (11).

$$H(\omega)=\frac{-\dfrac{R_2}{R_1}}{1+j\dfrac{\omega}{2*10^7\pi}\left(1+\dfrac{R_2}{R_1}\right)+\dfrac{1}{10^5}\left(1+\dfrac{R_2}{R_1}\right)}\approx-\frac{R_2}{R_1}; \qquad (11)$$

$$L(\omega)=j\frac{\omega}{2*10^7\pi}\left(1+\frac{R_2}{R_1}\right)+\frac{1}{10^5}\left(1+\frac{R_2}{R_1}\right);$$

Fig. 14 shows the loop gain in an inverting amplifier. In an inverting amplifier, the loop gain shows the gain reduction from the magnitude of the open-loop function to the magnitude of the inverting amplifier while the phase margin of the self-loop function indicates the operating region of this amplifier as shown in Fig. 15. The phase margin of this inverting amplifier is 90-degrees, this amplifier is absolutely stable.

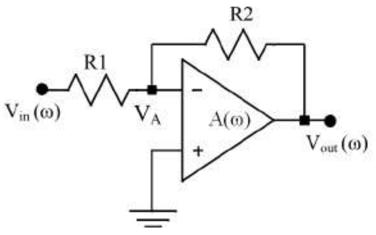


Fig. 12. Schematic of the single-ended inverting amplifier circuit.

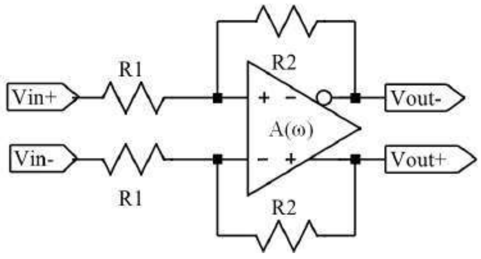


Fig. 13. Schematic of the fully differential inverting amplifier circuit.

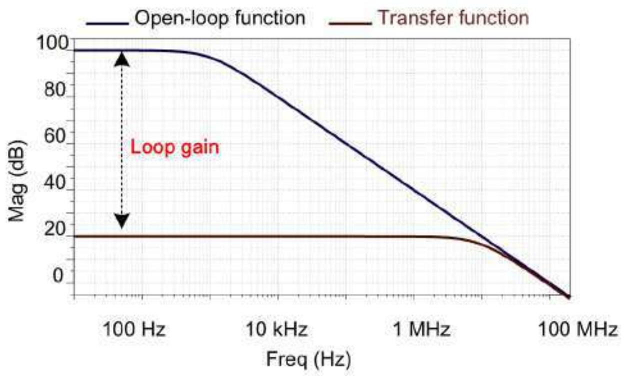


Fig. 14. Behaviour of the loop gain in an inverting amplifier.

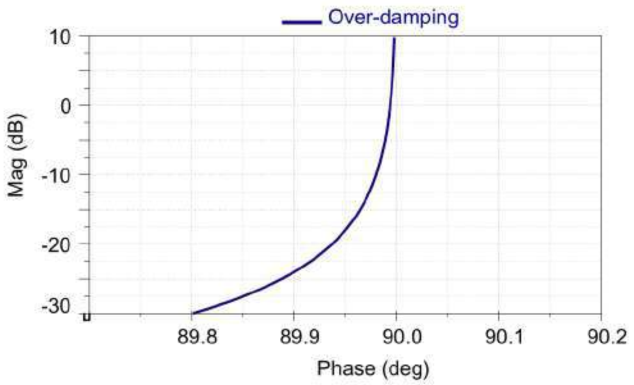


Fig. 15. Nichols plot of the self-loop function in an inverting amplifier.

B. Self-loop function of a 2nd-order inverting amplifier

In this section, the effects of the parasitic capacitor at the negative input of an inverting amplifier are investigated. Figs. 16 and 17 show the simplified models of the single-ended and the fully differential inverting amplifiers with the parasitic capacitors at the input of the operational amplifiers. After applying the superposition principle at node V_A in the single-ended inverting amplifier, the node equation is given in Eq. (12).

$$V_A\left(\frac{1}{R_1}+\frac{1}{R_2}+j\omega C_1\right)=\frac{V_{in}}{R_1}+\frac{V_{out}}{R_2};\ V_{out}=-A(\omega)V_A; \qquad (12)$$

Eq. (13) shows the transfer function and the self-loop function of this amplifier. They are second-order complex functions. TABLE VI presents the values of the constant variables in these complex functions.

$$H(\omega)=\frac{V_{out}(\omega)}{V_{in}(\omega)}=-\frac{b_0}{a_0(j\omega)^2+a_1 j\omega+a_2+1}; \qquad (13)$$

$$L(\omega)=a_0(j\omega)^2+a_1 j\omega+a_2;$$

TABLE VI. Values of Given Variables of Inverting Amplifier.

Variable	Value	Variable	Value
b₀	$\frac{R_2}{R_1}$	**a₀**	$\frac{R_2C_1}{2*10^7\pi}$
a₁	$\frac{1}{10^5}\left[R_2C_1+\frac{1}{2*10^2\pi}\left(1+\frac{R_2}{R_1}\right)\right]$	**a₂**	$\frac{1}{10^5}\left(1+\frac{R_2}{R_1}\right)$

The parasitic capacitor caused by the transmission line will change the operating region of the inverting amplifier. Assume that value of the parasitic capacitor $C_1 = 200$ pF, and values of resistors R_1 and R_2 are 1 kΩ and 10 kΩ. Figs. 18, 19 and 20 show the Bode plot of the transfer function, the Nichols plot of the self-loop function, and the transient response, respectively.

At around 1 MHz, the magnitude of the transfer function is greater than 20 dB and the phase margin at unity gain of the self-loop function is quite small 50-degrees (observed at 130-degrees). This amplifier works in the under-damping region; therefore, the ringing happens in the transient response.

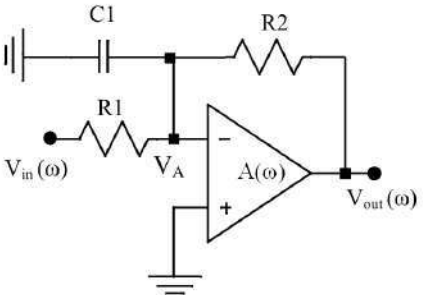


Fig. 16. Simplified model of the inverting amplifier circuit.

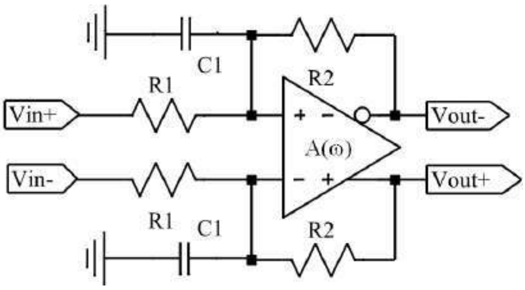


Fig. 17. Simplified model of the inverting amplifier circuit.

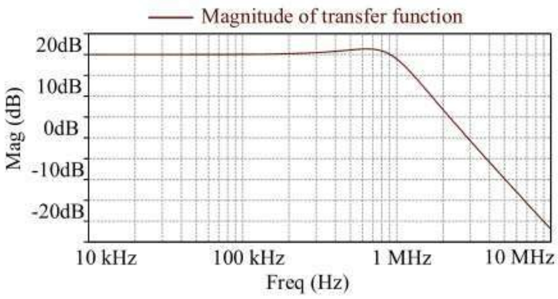


Fig. 18. Bode plot of the transfer function in the inverting amplifier.

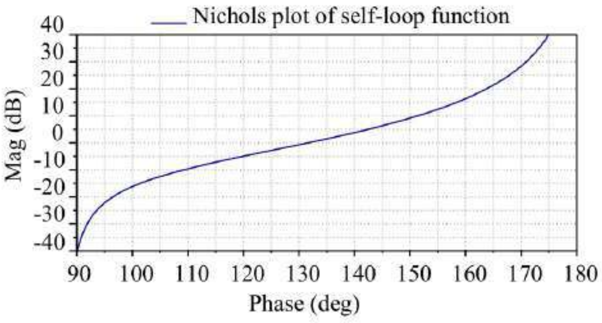


Fig. 19. Nichols plot of the self-loop function in the inverting amplifier.

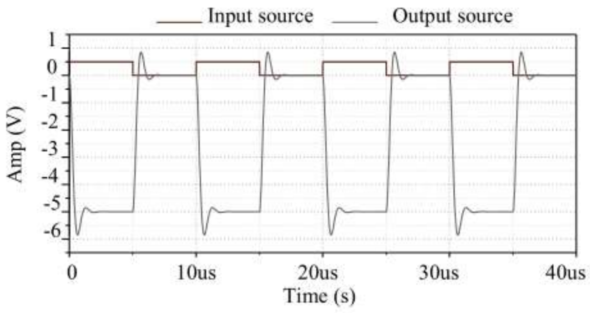


Fig. 20. Transient responses of the inverting amplifier.

C. Self-loop function of a 2nd-order non-inverting amplifier

In this section, the effects of the parasitic capacitor at the negative input of a non-inverting amplifier are investigated. From the view point of complex function, this non-inverting amplifier is also called negative feedback network. The term "positive feedback" is only used to indicate an oscillator or a signal generator network [24].

Fig. 21 shows the simplified model of the single-ended non-inverting amplifiers with a parasitic capacitor at the input of the operational amplifier. After applying the superposition principle at node V_A in the single-ended inverting amplifier, the node equation is given in Eq. (14).

$$V_A\left(\frac{1}{R_1}+\frac{1}{R_2}+j\omega C_1\right)=\frac{V_{out}}{R_2};\ V_{out}=A(\omega)\left(V_{in}-V_A\right); \quad (14)$$

Eq. (15) provides the transfer function and the self-loop function of this non-inverting amplifier. They are also second-order complex functions. The values of the constant variables in these complex functions are given in TABLE VII.

$$H(\omega)=\frac{V_{out}(\omega)}{V_{in}(\omega)}=\frac{b_0 j\omega+b_1}{a_0\left(j\omega\right)^2+a_1 j\omega+a_2+1}; \quad (15)$$
$$L(\omega)=a_0\left(j\omega\right)^2+a_1 j\omega+a_2;$$

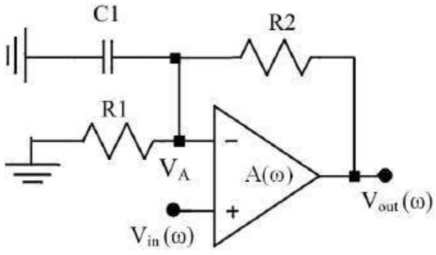


Fig. 21. Simplified model of the non-inverting amplifier circuit.

TABLE VII. VALUES OF GIVEN VARIABLES OF NON-INVERTING AMPLIFIER.

Variable	Value	Variable	Value
b_0	R_2C_1	**b_1**	$1+\frac{R_2}{R_1}$
a_0	$a_0=\frac{R_2C_1}{2*10^7\pi}$	**a_2**	$a_2=\frac{1}{10^5}\left(1+\frac{R_2}{R_1}\right)$
a_1	$\frac{1}{10^5}\left[R_2C_1+\frac{1}{2*10^2\pi}\left(1+\frac{R_2}{R_1}\right)\right]$		

Here, values of the passive components are capacitor $C_1 = 200$ pF, resistor $R_1 = 1$ kΩ, and resistor $R_2 = 1$ kΩ. Figs. 22, 23 and 24 show the simulation results of the non-inverting amplifier with a parasitic capacitor.

On the Bode plot of the transfer function, the magnitude is greater than 15 dB at around 3 MHz. And, the phase margin at unity gain of the self-loop function is very small 30-degrees (observed at 150-degrees). This amplifier also works in the under-damping region; so, the ringing occurs in the transient response.

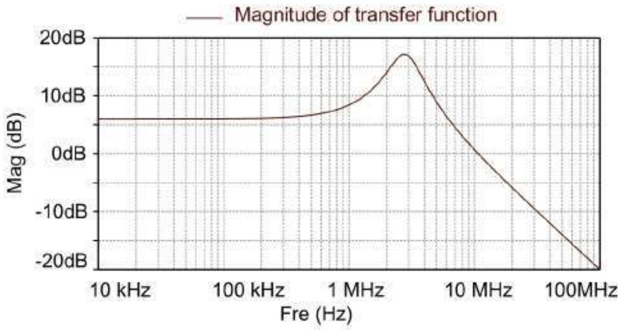


Fig. 22. Bode plot of the transfer function in the non-inverting amplifier.

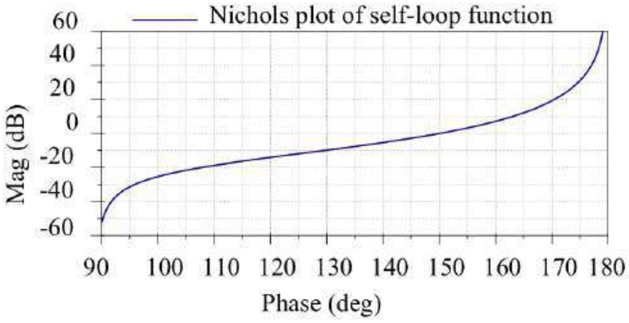


Fig. 23. Nichols plot of the self-loop function in the non-inverting amplifier.

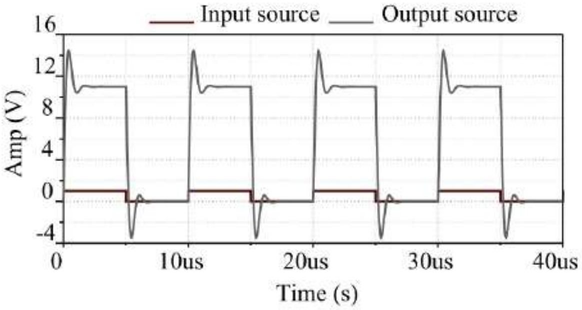


Fig. 24. Transient responses of the non-inverting amplifier.

D. *Self-loop function in shunt-shunt feedback amplifiers*

In this section, the properties of the shunt-shunt feedback amplifiers are investigated. Gain, gain stability, distortion, noise, input/output impedance and bandwidth and gain-bandwidth product will affect the behaviour of a negative feedback amplifier [25]. It is known that the conventional concept of loop gain is used to analyse four types of negative feedback amplifiers called shunt-shunt, shunt-series, series-shunt, and series-series. However, the approximated loop gain does not show the entire behaviour of these amplifiers.

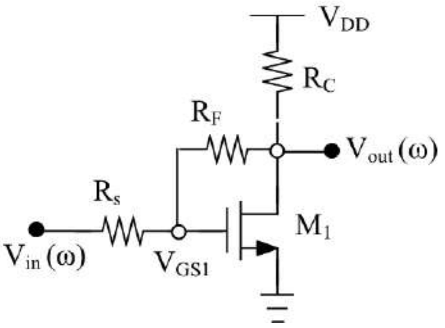


Fig. 25. Schematic of a CMOS feedback amplifier.

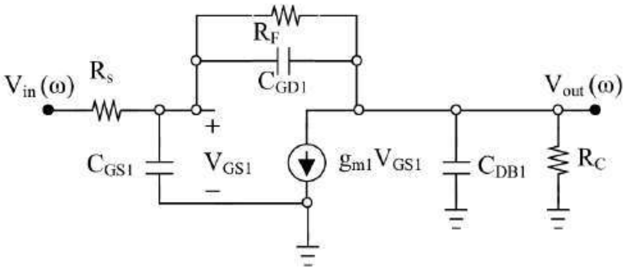


Fig. 26. Small signal model of a CMOS feedback amplifier.

Fig. 25 shows the schematic of a shunt-shunt CMOS feedback amplifier. The structure of this amplifier is also called self-bias common-gate CMOS feedback amplifier [26]. To derive the transfer function of this network, the small signal model is drawn in Fig. 26.

After applying the superposition principle at nodes V_{GS1} and V_{out}, the output voltage is given in Eq. (16).

$$V_{GS1}\left(\frac{1}{R_s}+\frac{1}{Z_{CGS1}}+\frac{1}{R_F}+\frac{1}{Z_{CGD1}}\right)=\frac{V_{in}}{R_s}+V_{out}\left(\frac{1}{R_F}+\frac{1}{Z_{CGD1}}\right); \tag{16}$$
$$V_{out}\left(\frac{1}{Z_{CGD1}}+\frac{1}{Z_{CDB1}}+\frac{1}{R_F}+\frac{1}{R_C}\right)=V_{GS1}\left(\frac{1}{R_F}+\frac{1}{Z_{CGD1}}-g_{m1}\right);$$

Eq. (17) shows the transfer function and self-loop function of the shunt-shunt CMOS feedback amplifier. The values of the constant variables of this circuit are given in TABLE VIII.

$$H_{CMOS}(\omega)=\frac{b_0 j\omega+b_1}{a_0(j\omega)^2+a_1 j\omega+1}; \tag{17}$$
$$L_{CMOS}(\omega)=a_0(j\omega)^2+a_1 j\omega$$

TABLE VIII. VALUES OF GIVEN VARIABLES OF A SHUNT-SHUNT CMOS FEEDBACK AMPLIFIER.

Parameter	Value
b_0	$R_F R_C C_{GD1}$
b_1	$R_C - R_F R_C g_{m1}$
a_0	$R_S R_F R_C\left(C_{GD1}C_{GS1}+C_{GD1}C_{DB1}+C_{DB1}C_{GS1}\right)$
a_1	$\left(R_C R_F^2+R_S R_C R_F g_{m1}+R_S R_F^2\right)C_{GD1}$ $+R_C R_F\left(R_S+R_F\right)C_{DB1}+R_S R_F\left(R_C+R_F\right)C_{GS1}$

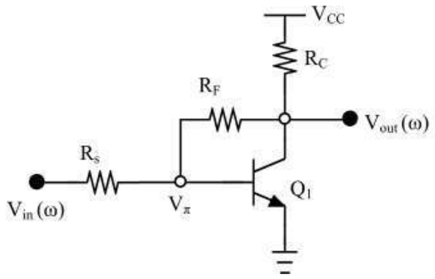


Fig. 27. Schematic of a BJT feedback amplifier.

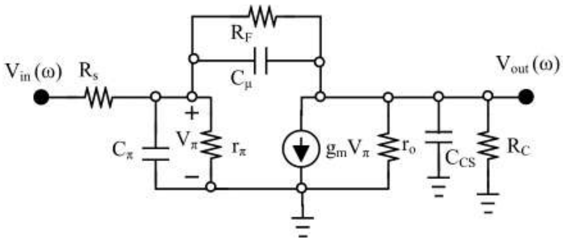


Fig. 28. Small signal model of the BJT feedback amplifier.

A self-bias common-emitter feedback amplifier is also called a shunt-shunt BJT feedback amplifier [27]. Figs. 27 and 28 show schematic and the small signal model of the shunt-shunt BJT feedback amplifier. The superposition theorem is also used to analyze this circuit. After applying the superposition principle at each node in the small signal model of this amplifier, the transfer function and the self-loop function of this amplifier are derived in Eq. (18).

$$H_{BJT}(\omega) = \frac{b_0 j\omega + b_1}{a_0 (j\omega)^2 + a_1 j\omega + 1}; \qquad (18)$$
$$L_{BJT}(\omega) = a_0 (j\omega)^2 + a_1 j\omega;$$

The values of the constant variables of this circuit are given in TABLE IX. A feedback resistor R_f = 1 kΩ is used to reduce the gain of the the shunt-shunt BJT feedback amplifier. Values of the chosen resistors R_C and R_S are 10 kΩ and 950 Ω, respectively. Fig. 29 shows the Bode plot of the transfer function of the common-emitter BJT feedback amplifier.

The simulated DC gain A_0 is 20 dB, and the simulated gain-bandwidth is 260 MHz at unity gain of the transfer function. The phase margin at unity gain of the self-loop function is 86 degrees at the unity gain of the self-loop function as shown in Fig. 30. The behaviour of this amplifier is over-damping.

TABLE IX. VALUES OF GIVEN VARIABLES OF A SHUNT-SHUNT BJT FEEDBACK AMPLIFIER.

Parameter	**Value**
b₀	$R_L C_{GD1}$
b₁	$-R_L g_{m1}$
a₀	$R_S R_L \left(C_{GD1} C_{GS1} + C_{GD1} C_{DB1} + C_{DB1} C_{GS1}\right)$
a₁	$R_L \left(C_{GD1} + C_{DB1}\right) + R_S \left(C_{GS1} + C_{GD1}\right) + R_S R_L g_{m1} C_{GD1}$

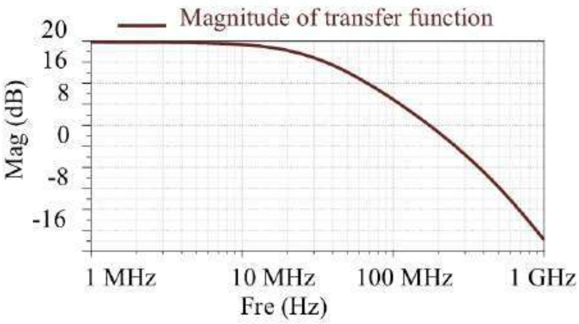


Fig. 29. Bode plot of the transfer function in the shunt-shunt BJT amplifier.

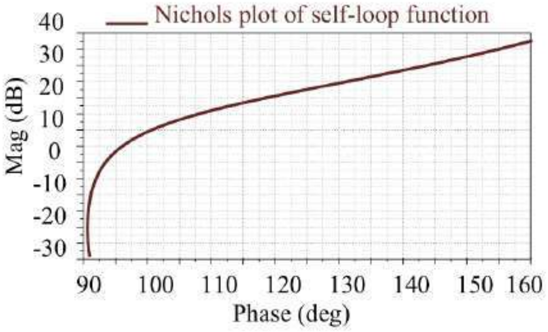


Fig. 30. Nichols plot of the self-loop function of the BJT amplifier.

IV. Ringing Test for Adaptive Feedback Networks

A. *Signal flow graph of DC-DC buck converter*

In this section, a switching DC-DC buck converter is considered as a control network. A step-down switching DC-DC buck converter is also called a dynamic load network. The propagated power at output will be changed regarding the actuating load. In other words, the variation of load will change the operating region of the power stage [28]. Therefore, a feedback loop is required for some switching power supplies because some parameters of the desired processing operation are changing as shown in Fig. 31. In this control system, a comparator is used to compare a referent voltage (DC voltage) with a feedback voltage, which is extracted from the output node [29]. The behaviours of the switching power supplies rely on negative feedback to maintain the output voltages at the specified values.

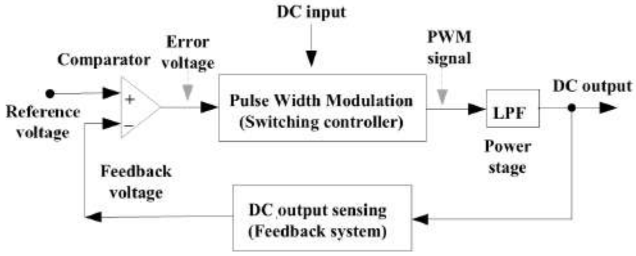


Fig. 31. Block diagram of a DC-DC buck converter.

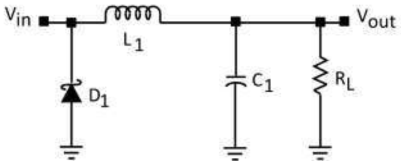


Fig. 32. Simplified model of the power-stage of the DC-DC buck converter.

B. *Behaviors of the power-stage of the buck converter*

In this section, the characteristics of a simplified model of the power-stage in a control system are analyzed. The power stage of the converter is an RLC network which is also called a 2nd-order low-pass filter in Fig. 32.

The function of this actuating system is to convert a multi-harmonic signal or a switching source into DC voltage [30]. Therefore, the transfer function and the self-loop function of the power stage are proposed to predict the maximal value of overshoot in this control system. The transfer function and the self-loop function of the power-stage are given in Eq. (19).

$$\begin{aligned} &H(\omega)=\frac{1}{a_0(j\omega)^2+a_1 j\omega+1};\\ &L(\omega)=a_0(j\omega)^2+a_1 j\omega;\\ &a_0=L_1C_1;a_1=\frac{L_1}{R_L}; \end{aligned} \tag{19}$$

To generate a maximal power, the balanced charge-discharge time condition for the power-stage of this control system is defined in Eq. (20).

$$\omega_{2RC}=\omega_{LC}\Leftrightarrow\omega=\frac{1}{\sqrt{LC}}=\frac{1}{2RC} \tag{20}$$

C. *Phase margin of the power-stage*

This section presents the measurement of the phase margin of the implemented circuit. Fig. 33 shows the schematic of the proposed design of a DC – DC buck converter circuit. In this implementation, the chosen passive elements are L_1 = 530 µH, C_1 = 220 µF, R_1 = 4 kΩ, R_2 = 2 kΩ, R_L = 5 Ω, C_2 = 470 pF, and L_2 = 50 µH.

The measurement setup for the implemented circuit is shown in Fig. 34. The performance of the proposed design of a 12 V – 5 V buck converter has been verified by the practical measurements.

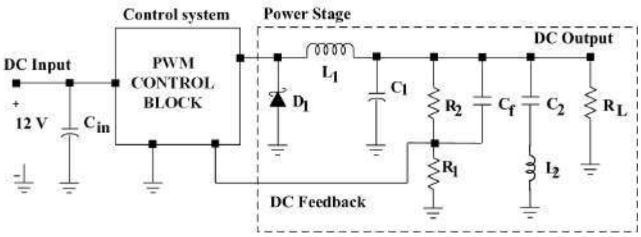


Fig. 33. Schematic diagram of a 12 V – 5 V buck converter circuit.

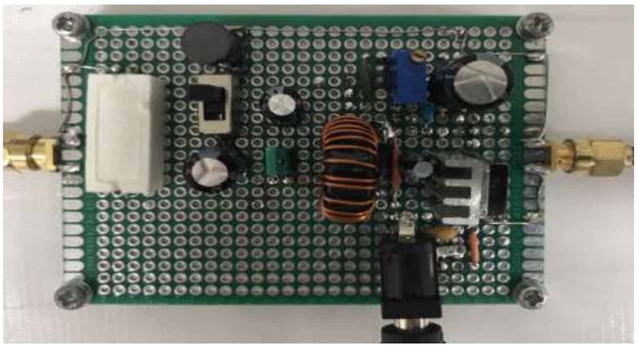


Fig. 34. Implemented circuit of a DC-DC buck converter.

Fig. 35 shows the measurement results of the Bode plot of the power-stage. The measured magnitude of the transfer function of the power stage at around the cut-off frequency is 2 dB. The chosen components in the implemented circuit make the power-stage of this control system always work in the under-damping region because there is a trade-off between the cost and size of the printed circuit board of the DC-DC buck converter circuit [31].

The measured phase margin at the unity gain of the self-loop function is 49 degrees (observed at 131 degrees) as shown in Fig. 36. The output ripple of the implemented DC-DC buck converter is kept very small 5 mVpp which is compared to the desired output voltage of 5 V as shown in Fig. 37.

This work overcomes some limitations of loop and the conventional Nyquist stability criterion and Nichols plot of loop gain in the control systems. The measurement of the phase margin at unity gain of the self-loop function at the actuating system is proposed to evaluate the quality of an adaptive feedback system [32].

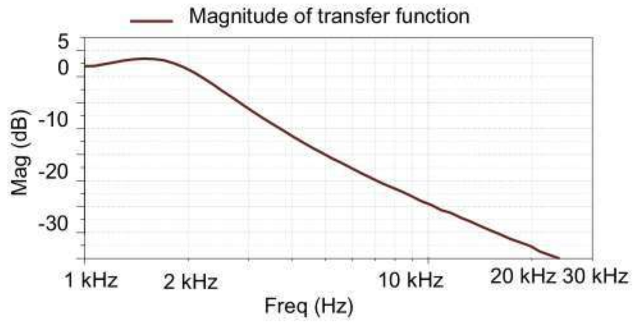


Fig. 35. Bode plot of the transfer function of the power stage.

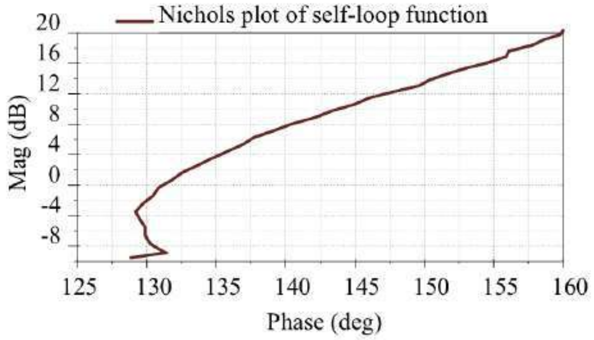


Fig. 36. Nichols plot of the self-loop function of the power stage.

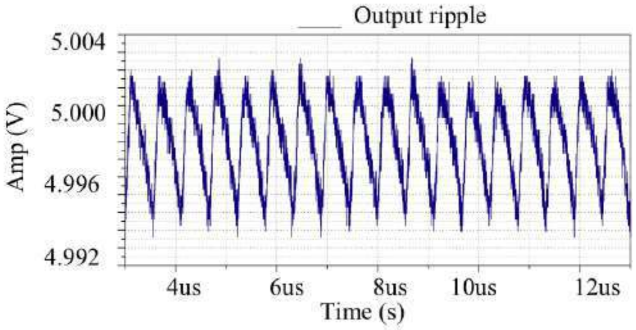


Fig. 37. Waveform of the output ripple of the DC-DC buck converter.

V. Conclusions

This paper has shown some limitations of the classical concept of loop gain in the negative feedback networks and investigated some motion models of high-order mechatronic networks. Loop gain cannot be used to analyze the overshoot phenomenon in high-order mechatronic systems because it doesn't show the operating regions of many high-order mechatronic systems. In addition, it is very difficult to derive the loop gain in many multi-loop feedback amplifiers such as unity gain amplifiers, inverting amplifiers, active filters, negative feedback amplifiers because loop gain is just a predicted value. In an inverting amplifier, loop gain only shows the gain reduction from the DC gain of the open loop function to the defined gain which is calculated by the ratio of two resistors.

The use of phase margin at unity gain of the self-loop allows engineers to do the stability test in high-order mechatronic systems. The Nichols plot of self-loop function is simpler than the Nichols plot of loop gain and the measurement of phase margin is easily verified by laboratory simulation and practical measurements. In case of under-damping, the overshoot or ringing makes the network unstable. This work also describes the approach to do the phase margin test for power-stage of a DC-DC buck converter. The phase margin of the power-stage is 49 degrees, and the output ripple is small 5 mVpp. Hence, the phase margin of the actuating system can be used to evaluate the quality of an adaptive feedback control system. In future works, ringing caused by the parasitic elements in other mechatronic systems will be investigated.

Acknowledgements

The authors would like to express their sincere gratitude to Prof. Hiroshi Tanimoto at Kitami Institute of Technology in Japan for his valuable comments.

References

[1] G. Ni, H. Jin, W. Lan, "Analysis of Linear Active Disturbance Rejection Control for Time-Delay Plants with Nyquist Plot," Chi. Cont. Conf., China, Jul. 2019.

[2] P. Wang, S. Feng, P. Liu, N. Jiang, X. Zhang, "Nyquist Stability Analysis and Capacitance Selection Method of DC Current Flow Controllers for Meshed Multi-Terminal HVDC Grids", J. Pow. & En. Sys., vol. 7, no. 1, pp. 114-127, July 2020.

[3] J. Ardila, H. Morales, E. Roa, "On the Cross-Correlation Based Loop Gain Adaptation for Bang-Bang CDRs," IEEE Trans. on CASI, vol. 67, no. 4, pp. 1169 – 1180, Apr. 2020.

[4] D. Tokmakov, S. Asenov, "Operational Amplifier Open Loop Gain Simulation in Electronics Engineering Education," 11th Nat. Conf. with Int. Parti. (ELECTRONICA), Bulgaria, July 2020.

[5] K. Ogata, Modern Control Engineering, 5th Ed., Prentice-Hall, Boston, 2010.

[6] R. Aisuwarya, Yulita Hidayati, "Implementation of Ziegler-Nichols PID Tuning Method on Stabilizing Temperature of Hot-water Dispenser," 16th Int. Symp. on Elec. & Comp. Eng., Ind., Jul. 2019.

[7] M. Liu, I. Dassios, G. Tzounas, F. Milano, "Stability Analysis of Power Systems with Inclusion of Realistic-Modeling of WAMS Delays", IEEE Trans. Pow. Sys., vol. 34, no. 1, pp. 627-636, 2019.

[8] L. Wang, M. Mirjafari, "Design of loop gain for load converters in a Distributed System," IEEE App. Pow. Elec. Conf. & Exp. (APEC), USA, March 2020.

[9] N. Kumar, V. Mummadi, "Stability Region Based Robust Controller Design for High-gain Boost DC-DC Converter", IEEE Trans. Ind. Elec., Feb. 2020.

[10] J. Lin, M. Su, Y. Sun, X. Li, S. Xie, G. Zhang, F. Blaabjerg, "Accurate Loop Gain Modeling of Digitally Controlled Buck Converters," IEEE Trans. on Indus. Elec., Ear. Acc., pp. 1-1, Jan. 2021.

[11] N. Mazlan, N. Thamrin, N. Razak, "Comparison Between Ziegler-Nichols and AMIGO Tuning Techniques in Automated Steering Control System for Autonomous Vehicle," IEEE Int. Conf. on Auto. Cont. & Int. Sys., Malay., Jun. 2020.

[12] Y. Yao, C. Huang, S. Liu, "A Jitter-Tolerance-Enhanced Digital CDR Circuit Using Background Loop Gain Controller," IEEE Trans. on CAS II: Expr. Brif. Op. Acc., pp. 1-1, Dec. 2020.

[13] P. Saini, P. Thakur, A. Dixit, I. Uniyal, "Controller Design based on Desired Phase Margin for Headbox Process," 2nd Int. Conf. on Adv. in Comp., Com. Con. & Net. (ICACCCN), India, Dec. 2020.

[14] S. Zhong, Y. Huang, "Comparison of the Phase Margins of Different ADRC Designs," IEEE Chinese Con. Conf. (CCC), China, July 2019.

[15] N. Sayyaf, M. Tavazoei, "Frequency Data-Based Procedure to Adjust Gain and Phase Margins and Guarantee the Uniqueness of Crossover Frequencies," IEEE Trans. on Ind. Elec. vol. 67, no. 3, pp. 2176-2185, Mar. 2020.

[16] L. Fan, Z. Miao, "Admittance-Based Stability Analysis: Bode Plots, Nyquist Diagrams or Eigenvalue Analysis", IEEE Trans. Pow. Sys., vol. 35, no. 4, Jul. 2020.

[17] R. Matušů, B. Şenol, L. Pekař, "Robust PI Control of Interval Plants With Gain and Phase Margin Specifications: Application to a Continuous Stirred Tank Reactor," IEEE Acc., vol. 8, pp. 145372-145380, Aug. 2020.

[18] R. Magossi, V. Oliveira, R. Machado, S. Bhattacharyya, "Stabilizing Set and Phase Margin Computation for Resonant Controllers," IEEE 58th Conf. on Dec. & Cont. (CDC), France, Dec. 2019.

[19] P. Pejovic, "Replotting the Nyquist Plot - A New Visualization Proposal," 20th Int. Symp. on Pow. Elec., Serbia, Oct. 2019.

[20] N. Nise, Control Systems Engineering, 7th Ed., John Willy & Sons. Inc., River St. Hoboken, NJ, 2015.

[21] R. Bhusal, B. Taner, K. Subbarao, "On the Phase Margin of Networked Dynamical Systems and Fabricated Attacks of an Intruder," 2020 Am. Con. Conf. (ACC), USA, Jul. 2020.

[22] F. Golnaraghi, B. Kou, Automatic Control Systems, 9th Ed., John Willy & Sons. Inc., River St. Hoboken, NJ, 2010.

[23] L. Ferreira, F. Guaracy, "Improved Independent and Uniform Multivariable Gain and Phase Margins in H-infinity/LTR Control," IEEE Trans. on Aut. Cont. Ear. Acc., pp. 1-1, Nov. 2020.

[24] A. Brito, "On The Misunderstanding of The Ziegler-Nichols's Formulae Usage," IEEE CAA J. Auto. Sini., vol. 6, no. 1, pp. 142-147, Jan. 2019.

[25] W. Jaikla, F. Khateb, M. Kumngern, T. Kulej, R. Ranj, "0.5 V Fully Differential Universal Filter Based on Multiple Input OTAs," IEEE Acc., vol. 8, pp. 187832 - 187839, Oct. 2020.

[26] Z. Guo, Y. Li, J. Xiang, "A Feedback Cone Design Synthesis and Its Corresponding Gain/Phase Margin," Chin. Auto. Cong. (CAC), China, Nov. 2019.

[27] P. Arya, S. Chakrabarty, "Reduced order controller for FO-IMC with desired phase margin and gain cross-over frequency," IEEE Bom. Sect. Sig. Conf. (IBSSC), India, Jul. 2019.

[28] T. Choogorn, "Estimation of HD3 and IM3 Distortion for Fully-Differential Gm-C Filter," 16th Int. Conf. on Elect. Eng. /Elect., Comp., Tel.. & Inf. Tech. (ECTI-CON), Thailand, Jul. 2019.

[29] T. Tran, Complex Functions Analysis, LAP LAM. Aca. Pub., ISBN: 978-620-3-46189-3, Mau., Feb. 2021.

[30] D. Seiferth, R. Afonso, F. Holzapfel, M. Heller, "Reduced Conservatism Proof of the Balanced MIMO Gain and Phase Margins," 59th Conf. on Deci. & Cont. (CDC), Korea, Dec. 2020.

[31] Y. Yan, A. Shehada, A. Beig, I. Boiko, "Auto-Tuning of PID Controller with Phase Margin Specification for Digital Voltage-Mode Buck Converter," IEEE Conf. Con. Tech. & App., Canada, Aug. 2020.

[32] T. Tran, A. Kuwana, H. Kobayashi, "Ringing Test for 3rd-Order Ladder Low-Pass Filters", 11th IEEE Ann. Ubiq. Com., Elect. & Mob. Com. Conf. (UEMCON 2020), USA, Oct. 2020.

Biomechatronic Embedded System Design of Sensorized Glove with Soft Robotic Hand Exoskeleton Used for Rover Rescue Missions on Mars

Paul Palacios
Space Physics & Engineering Division, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
paul.eng.94@ieee.org

José Cornejo
Space Physics & Engineering Division, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
jose.cornejo@ieee.org

Milton V. Rivera
Space Human Medicine Division, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
e15143035@uancv.edu.pe

José Luis Napán
Space Physics & Engineering Division,, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
jose.napan@pucp.pe

Walter Castillo
Space Physics & Engineering Division, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
walter.castillo@ieee.org

Victor Ticllacuri
Space Physics & Engineering Division, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
victor.ticllacuri@pucp.edu.pe

Andrés D. Reina
Space Veterinary Medicine Division, Bioastronautics and Space Mechatronics Research Group
Bogotá , COLOMBIA
adreinac@libertadores.edu.co

Adolfo Chaves-Jiménez
Space Systems Lab (SETEC lab), Instituto Tecnológico de Costa Rica
Cartago, COSTA RICA
adchaves@itcr.ac.cr

Gustavo Jamanca-Lino
Space Physics & Engineering Division, Bioastronautics and Space Mechatronics Research Group; Colorado School of Mines
Colorado, USA
gustavo.jamanca@community.isunet.edu

Juan Carlos Chávez
Space Physics & Engineering Division, Bioastronautics and Space Mechatronics Research Group; HUB Designers & Talent360
Washington DC, USA
contact@talent360.institute

***Abstract*— For many years, a manned mission to Mars has been a challenge for humanity. However, the recent technological advances in human factors and space systems engineering may overcome these limitations. Thus, there are some strategies to face various medical emergencies autonomously due to long distances and hostile conditions in order to develop healthcare monitoring and ensuring the safety of the astronauts. For this reason, an innovative research has been conducted from 2020 to 2021 under the supervision of the Bioastronautics & Space Mechatronics Research Group, resulting in this proposed project, which is a sensorized glove for medical emergencies controlling the rover developed by the Team "Tharsis" from the Universidad Nacional de Ingeniería, winner of the Technology Challenge Award for Wheel Design and Fabrication at the 7th Annual NASA Human Exploration Rover Challenge. In addition, the glove has a soft robotic hand exoskeleton with the purpose of fracture stabilization and to prevent future complications. This study presents a mechatronics conceptual design based on biomechanical fundamentals of the hand, where Fusion 360 was used for 3D mechanical systems development and Eagle for electrical and electronic circuit technical schematics, besides telecommunications and telerobotics protocols are analyzed. The "BIOX-GLOVE" is pretended to be applied on the Martian surface during extravehicular activities-EVA and also, on Earth in a Mars environment analogue and rehabilitation hospitals. In conclusion, favorable results were achieved; consequently, the next step of this project is to start the detailed engineering design in July 2021, and it is proposed to develop the prototype and perform the first test in a Martian analog.**

***Keywords*—Mars, Rover, Sensorized Glove, Medical Emergencies, Exoskeleton, Mechatronics, BIOX-GLOVE**

I. Introduction

This decade, 2020-2030, an age of space colonization will begin. The Artemisa project plans to explore the Moon to extract resources and establish a human base as a rehearsal for future Martian missions [1]. Mars human exploration is the next step; therefore, it is required to develop previous colonization activities [2]. The physical and psychological risks for the astronauts on this trip will be very crucial [3]. It is also essential to offer countermeasure alternatives to avoid body damage by the effects of microgravity, radiation [4] and the environmental conditions on the Martian surface [5]. (Fig.1)

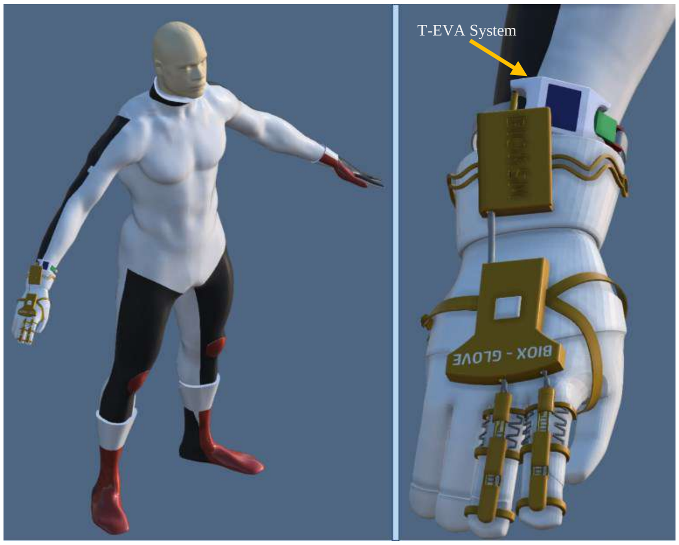


Fig.1. BIOX-GLOVE and Astronaut

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Working outside the spacecraft represents a health risk for an astronaut due to environmental factors such as the vacuum in space, radiation, extreme temperatures, and micrometeorites [6]. To solve this, Extravehicular Mobility Unit-EMU, was designed to provide the necessary functions to keep the user alive [7] during exploration activities-EVA.

Gloves are considered one of the most important and complex wearable accessories to operate because specific skills must be developed and in addition, when the suit is pressurized, the mobility of the hands becomes challenging due to the stiffness [8]. Previous studies state that wrist fractures could be caused due to the axial forces [9]; therefore, mechatronics engineering is currently working on designing robotic gloves and tools that allow the astronaut to enhance work performance [10]. Thus, National Aeronautics and Space Administration-NASA, through its “Evolvable Mars Campaign” program, is seeking to make human exploration on Mars surface pioneering and sustainable [11].

The long-term goal is to send astronauts to live on the red planet [12]. Currently, many satellites on Mars are analyzing the surface to find landing candidates places. The first human exploration missions will need to cover a big area for sampling in situ. In this scenario, the use of rovers will be necessary, same as the Apollo 17 mission on the Moon [13]. In this context, for NASA and manned space programs, the lives of astronauts are a priority considering that there is a permanent risk of suffering illnesses and accidents when exploring not-known areas such as the surface of Mars [14]. The gravity on Mars (1/3 g of the Earth) can be considered potentially dangerous, causing increasing object’s acceleration in the environment, and thus, producing injuries [15].

The estimated probabilities of an emergency outside the Earth are on average 0.06 per “crew member/year-of-space-flight-expedition” [14, 15]. For example, with a crew of 6 astronauts on the surface of Mars and a mission duration of 900 days, it would be expected that at least one emergency would occur [16]. EVA and continuous exposure to extreme environments add debilitating effects that predispose to severe traumatic injuries [17] such as bone fractures [18], hemorrhages, thoracoabdominal injuries [14], decompressive diseases [19] as well as contusions involving the digits of the hand, peripheral nerves and generalized fatigue due to the working hours [20, 21]. Prompt medical care for the injured astronaut could be problematic due to the distances between the exploration and Command Center. Therefore, the decisions must be taken quickly in situ because the communication time to the Earth is often between 8 to 56 minutes [19] [22].

Due to this context, BIOX-GLOVE is proposed as a medical wearable robot [23, 24] in order to activate an emergency system on the Martian surface [25, 26] during extravehicular activities, using telecommunications and control protocols applied in a rover (ambulance vehicle) to transfer the astronaut to the Command Center [27]. In addition, it has a hand exoskeleton designed with soft materials to improve the physical capabilities, this innovative technology will be used on Earth for patient’s hand rehabilitation [28]. The research is based on T-EVA project [29] developed by Bioastronautics and Space Mechatronics Research Group (BIO&SM). (Fig. 1)

1.1. Research Objectives

The research requirements are aligned with the NASA’s Journey to Mars [12] and the Agreement on the Rescue of Astronauts, the Return of Astronauts and the Return of Objects Launched into Outer Space [30]. One of the challenges is to integrate human and robotic missions to achieve a successful task [12]. The exploration needs tools to control rovers during accidents and emergency situations in order to come back to the Command Center for medical attention [31, 32, 33].

The BIOX-GLOVE is an accessory that will be placed above the EVA suit and must be comfortable, ergonomic, and able to withstand the environmental conditions on Mars, such as temperature, radiation and dust; therefore, it will be used in case of partial or total motor impairment of the lower extremities. This technology is suggested to be used during any Extravehicular Activities.

The astronaut will turn on the glove in 2 ways during emergency situations [34]:

a) First, by performing a sequence of programmed gestures to control the Rover.

b) Second, by pressing the activation button (located on the back of hand). The exoskeleton will stabilize the musculoskeletal system of the wrist and hand.

1.2 Research and Development Methodology

The steps where elaborated based on 3 key points: a) Define the problem and the scope of the objectives. b) Define the main requirements necessary for the project execution. c) Modeling through 3D design software and the proposed validation test of BIOX-GLOVE. (Fig. 2)

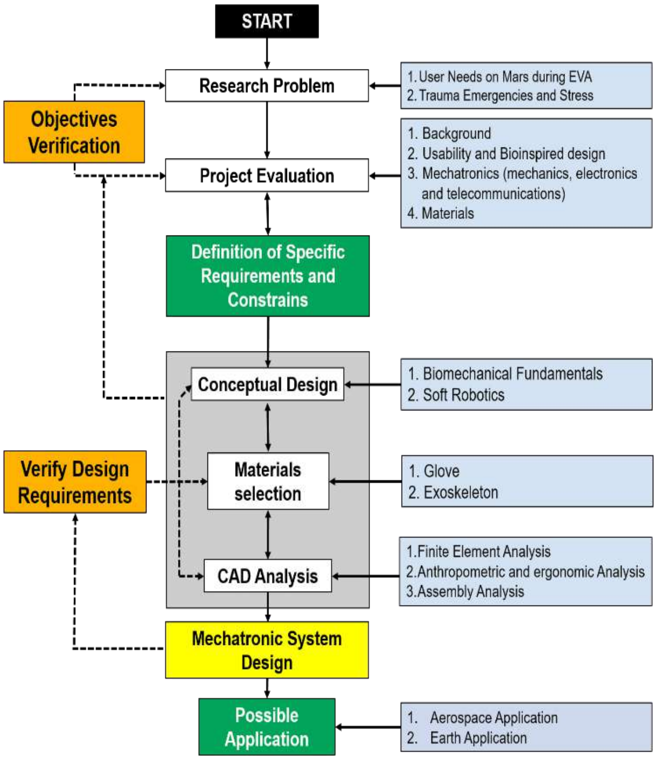

START
Research Problem
1. User Needs on Mars during EVA
2. Trauma Emergencies and Stress
Objectives Verification
Project Evaluation
1. Background
2. Usability and Bioinspired design
3. Mechatronics (mechanics, electronics and telecommunications)
4. Materials
Definition of Specific Requirements and Constrains
Conceptual Design
1. Biomechanical Fundamentals
2. Soft Robotics
Verify Design Requirements
Materials selection
1. Glove
2. Exoskeleton
CAD Analysis
1.Finite Element Analysis
2.Anthropometric and ergonomic Analysis
3.Assembly Analysis
Mechatronic System Design
Possible Application
1. Aerospace Application
2. Earth Application

Fig. 2. The methodology of the conceptual design of BIOX-GLOVE

II. Biomechatronics Foundations of Human Factors

Biomechatronic applications have been made for more than three decades to overcome several limitations; research related to biomechatronics and medical robotics cover a diverse aspect in interdisciplinary areas inspired by industrial, military, and medical applications. This area provides new methodologies and tools to design and build devices to cooperate with humans [35-38].

Gloves have long taken importance in the workplace, for the purposes of providing protection and safety against physical injury to the hands. However, they are even more important in space exploration when performing hand tasks in high-risk operations [39]. Over the years, the designs have been improving, and along with this, the most relevant points are that now it can resist cuts, breaks, perforations, slides, and abrasion [40]. At present, mechanical and electronic subsystems are being incorporated to improve the performance of the kinematic synergies of the hand, in order to provide a support that ensures the correct interaction of movements to counteract the rigidity of the pressurized suit during EVA [6].

These devices will make it easier for people to expand work areas and while guaranteeing their safety in highly dynamic hostile places such as space and planetary exploration [41]. However, these instruments are not only for hand protection and safety purposes, their sensors are capable of recognizing predetermined hand signals [42] are also added to BIOX-GLOVE to activate the medical emergency system [41].

2.1. Antropometric Analysis of the Hand

The hand is one of the most complicated biomechanical systems for study and application due to the multiple combinations in the reproduction of movements used to perform simple to complex activities. There are many kinematic interactions between the position of the wrist joint and the digits´ efficiency [8]. A detailed analysis of the pseudo-kinematic restrictions affects certain degrees of freedom due to the disposition of the tissues and tendons of their anatomical structure [8], [43].

For glove sizing, it is necessary to determine the hands' anthropometric dimensions to make a bio-inspired anatomical design with great precision of the movements [44], providing great flexibility and comfort when performing various tasks [39]. Likewise, it is recommended that these devices reflect the morphological characteristics of an astronaut's hand to achieve adequate performance without excessively affecting the control of force and holding of objects [45, 46].

The glove was designed considering morphological differences, ethnicity, and gender [47]. Due to the anthropometric variability and heterogeneity of the population groups, the measurement of the hands differs between Latin Americans [48], North Americans [43, 44], and Europeans [45]. These values are generally represented by percentiles; that is, for smaller people (5th percentile) and older people (95th percentile); therefore, it has been decided to obtain the estimated numerical average of the anthropometric measurements of the authors' hands in a referential way. (Fig. 3.a) For this analysis the following measurements have been noted: wrist circumference (WH), hand length (HL), palmar length (PL) hand width (HW), palmar width (PW), Wrist width (WW), and the total lengths for each digit (TLDi) in Table I.

TABLE I. Hand Anthropomorphic Measurements

Hand Measure (cm)	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**X**
Hand Length (HL)	19	19	18	16.5	18	17.9	19	18.8	18	18.5
Palmar Length (PL)	11	10.5	11	9	10.6	10.5	11	10.6	10	10.4
Hand Width (HW)	9.7	10	10	9	10.2	10.3	10	11	10	9.9
Palmar Width (PW)	7.4	8.8	8.5	7.5	8.9	7.5	8.8	10	8.5	8.6
Wrist Circumference (WC)	16.4	16	18	17	18	15.2	16	17.8	15.5	17.0
Wrist Width (WW)	6.1	6	7	6	6.4	6.6	7	6.9	6	6.3

Based on Authors Data (Ai: Random Sequence, X: Average of hand measurements)

The total lengths of each digit (TLDi) is obtained by calculating the lengths of the metacarpals (MCLi) "palm" and the lengths of the digits (DLi) according to the plane of the anatomical description (i = D-I for the first digit (thumb); D-II for the second digit (index); D-III for the third digit (middle); D-IV for the fourth digit (ring) and D-V for the fifth digit (pinky)). Before the measurement, 3 reference points were considered: first, the skinfold lines of flexion of the wrist on the palmar side; second, the digitopalmar skin fold lines [49]; and third, the edge of the distal end of each finger. Subsequently, it has been measured at the intersection of the points following the direction of each digit in extension. With this result the dimensions of LTDi are obtained, and it is represented mathematically with the following formula:

TLDi = MCLi + DLi. (Figure 3.b)

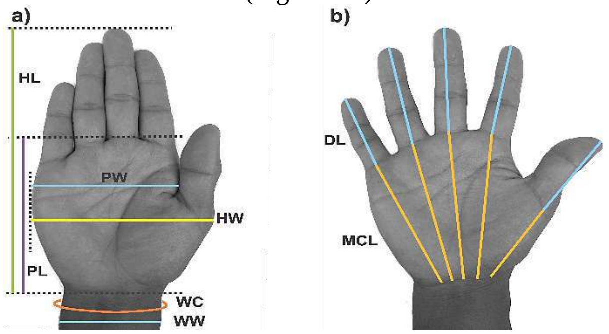


Fig. 3. Anthropometric Sizing: a) Measurements; b) Total length for each digit

Due to the complexity of the kinematic and dynamic model and the high number of degrees of freedom of the hands [8] [50], the gloves have been adapted trying to maintain similar characteristics by replicating hand movements with little interference. The movements have multiple functions such as ergotic, epistemic and semiotic [51]. For this case, the study is focused on the semiotic aspect to correctly size the sensors and characterize the hand gestures following a sequence of images called: M1, M2, ..., Mx, (M1 is the initial posture and Mx is the final posture). The sensors that BIOX-GLOVE has, allow it to capture information of the movement trajectories of the digits (dynamic gestures), characterized by the configuration of the hand digits. Gestures are contents of movements that start from a resting position (initial position), continue with a phase of substantial increase in speed and end by returning to the initial position [51].

2.2. Biomechanical Protocols

The device must follow protocols to verify the integrity and correct operation of the upper limb for the device handling (A); additionally, contain a series of pre-established characterizable movements to translate the biomechanical signals into vehicle movements (B). For this study, only D-II and D-III are characterized since their flexion and extension influence others [52]. Therefore, it is established the activation protocol and the usage protocol.

A) Activation Protocol

Based on the upper limb anatomical and biomechanical considerations, the activation protocol (M1 – M7) must be executed to ensure the correct state of this limb; that is to say, the nervous and muscular structures of the astronaut's upper limbs are capable of performing complex manipulation activities [53].

First, the hand must be extended (Fig. 4.a), avoiding the formation of the proximal and distal transverse and longitudinal arches (PTA, DTA and LA) [54]. This first step is to set up the hand and wrist in a neutral position; the device can be started from there. Second, a complete circumduction of the hand must be performed (Fig. 4.b). This movement, composed of flexion, extension, abduction, and adduction of the hand, verifies that the distal radiocarpal and radioulnar joints are functioning correctly [52]. Likewise, correct functioning of both the epitrochlear and epicondylar muscles and the forearm supinators and pronators is required [52] [54]; in this way, it is verified their full functionality. Third, the hand must return to the neutral position (Fig. 4.c.1), then clench the hand for 5 seconds (Fig. 4. c.2) and return to the neutral position (Fig. 4.c.3). With this movement, it is verified that the metacarpophalangeal (MCP), proximal, and distal interphalangeal joints (PIP and DIP) of all digits are capable of fully flexing and extending [52, 53], as well as the correct functioning of muscles activated by the radial and ulnar nerves [53]. Finally, active flexion at 70° (Fig. 4.d.1) and active extension 60 ° (Fig. 4.d.2) of the hand is executed, then return it to the neutral position. The device will be activated when it has verified the correct execution of the movements. For the user's safety, if any of these movements cannot be executed correctly, the device will not activate; since it would be a sign of a biomechanical malfunction in the upper limb, it would not be able to control the vehicle correctly.

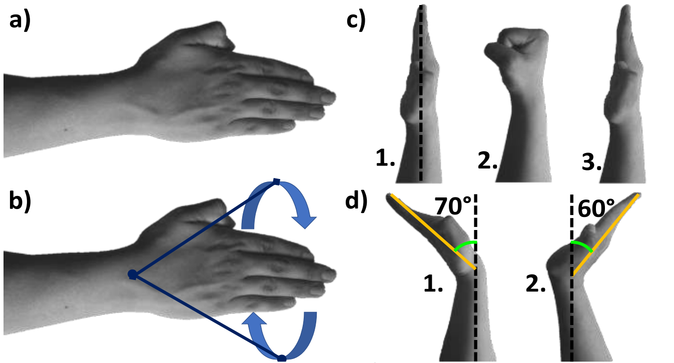


Fig. 4. Activation protocol. a) Top view of the neutral hand position with the fingers and palm extended (M1). b) Circumduction movement of the hand (rotation around the wrist) (M2). c) Lateral View of 1. Extension (M3) - 2. Flexion (M4) - 3. Full extension of the hand (M5). d) Side view of 1. Flexion (M6) - 2. Extension (M7).

B) Usage Protocol

The use of the device depends on a series of hand movements, which represents a specific action in the vehicle's control. These must be interdependent among themselves since their execution must not affect or be confused with another, but in turn, they must work together to execute the activities in the vehicle. Likewise, these movements are based on the upper limb's biomechanics, specifically on the angles formed by joints. The movements required in the vehicle are forward, backward, stop, turn right, and turn left. Therefore, the usage protocol consists of the following steps.

First, in the neutral position, start the device to begin remote control of the vehicle. To advance, flex the hand 70° [52], then close the hand into a fist and only leave D-II extended (Fig. 5.a). In this way, all DIP, PIP and MCP joints of each digit, except D-II, will be fully flexed, allowing them to be easily identified by the flexion sensors and avoiding some error in the algorithm. Likewise, as the wrist joint is flexed, it will be possible to ensure correct recognition of the vehicle's biomechanical gesture to move forward.

On the other hand, to move back, the hand is required to be hyperextended 60° [52], then all the PIP and DIP joints of each digit must be flexed (Fig. 5.b). In this way, it is verified by partial bending of D-II and D-III that the action is being executed correctly and would be free of confusion errors. Likewise, as the wrist joint is hyperextended, it will be possible to correctly recognize the biomechanical gesture for the vehicle to move backwards.

For left and right turns, the hand must be in the neutral position. Then, to turn to the right, radial deviation (abduction) must be performed at 20° [53] (Fig. 5.c.1); while to turn to the left, ulnar deviation (adduction) must be performed at 30° [53] (Fig. 5.c.2).

Eventually, to activate the braking system, it is required to close the hand into a fist, in such a way that all DIP, PIP and MCP joints are fully flexed (Fig. 5.d); while the wrist joint is at the neutral position (0°). This configuration was determined as it is quick to execute, requiring emergency braking.

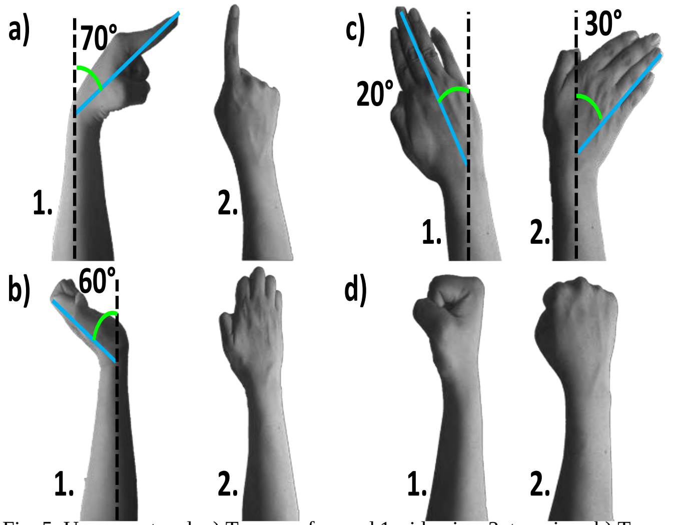


Fig. 5. Usage protocol. a) To move forward 1. side view 2. top view. b) To go back 1. side view 2. top view. c) Top view of turning movements to the 1. left 2. right. d) To brake 1. side view 2. top view.

III. Mechatronic System Design

This research involves fields of engineering such as mechanical, electrical, electronics, telecommunications and telerobotics. It consists of using a sensorized glove to control a manned rover in the case of a medical emergency (body harm) on Mars. BIOX-GLOVE has flex sensors on the fingers, and gyroscopes placed between the wrist and forearm to perform programmed hand movements. Besides, it has a button to activate the exoskeleton as an external fixator, stabilizing the radiocarpal joint in case of traumatic injuries due to falls, thus avoiding further complications.

3.1. Mechanical Design

A) Glove

The design was made using Fusion 360. The glove wraps around the back of the user's hand with a JST SMP connector to make sure it doesn't fall off. The emergency button is located right in the center of the main body, where the main board and a gyro box is located. As it was mentioned, on D-II and D-III, 2 flex sensors are located. Meanwhile, a second gyro box and the battery are located on the back side of the wrist, inside another box, held by a band (Fig. 6). For astronaut's glove manufacturing, it is recommended to use VECTRAN (aluminized and Teflon coated) because it's characteristics such as: abrasion resistance, high radiation exposure and property retention at high/low temperature [55, 56].

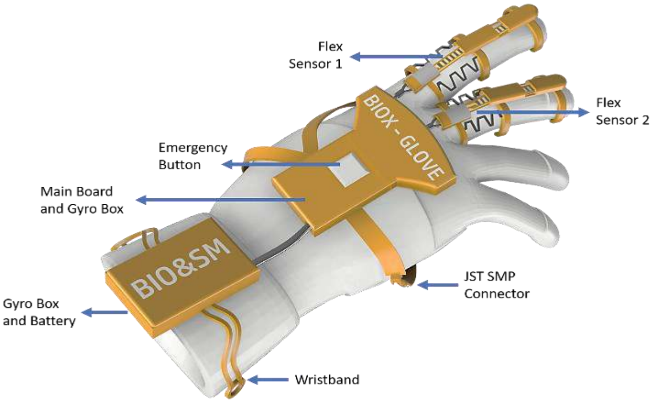


Fig. 6. Components of BIOX-GLOVE, Isometric View

The glove's bands and holders are meant to be made of 3 layers of materials, since it is one of the most vulnerable parts of the suit [40]. An inner bladder, a restraining polyester intermediate one and an outer thermal micrometeoroid garment. The boxes should be fabricated by 3D printing using Polylactic Acid (PLA), since it would make it easier to assembly and the material is durable and has a relatively low density.

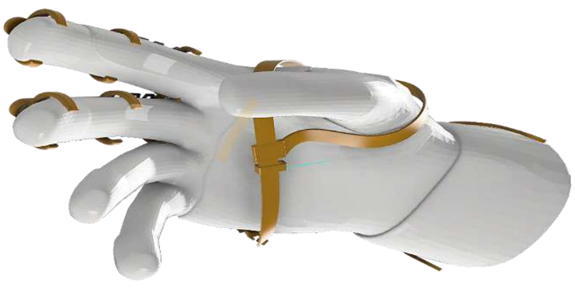


Fig. 7. BIOX-GLOVE, Bottom View

B) Exoskeleton

The exoskeleton [57, 58, 59, 60] is composed of 3 main modules, whose operation is interdependent. Module 1 (Fig. 8.a) has 12 submodules which are found between the PIP (a.1 - a.3), DIP (a.4 - a.7) and MCP (a.8 - a. 12) of each digit; reducing their degrees of flexion and extension. Module 2 (Fig. 8.b) has 4 submodules found in the palm of the hand to restrict the MCP joint flexion and to induce the formation of the DTA and LA (D-II and D-III mainly). Finally, Module 3 (Fig. 8.c) is located on both thenar and hypothenar eminence to ensure the formation of the PTA; likewise, this module completely covers the wrist joint, limiting its ranges of movement at all angles.

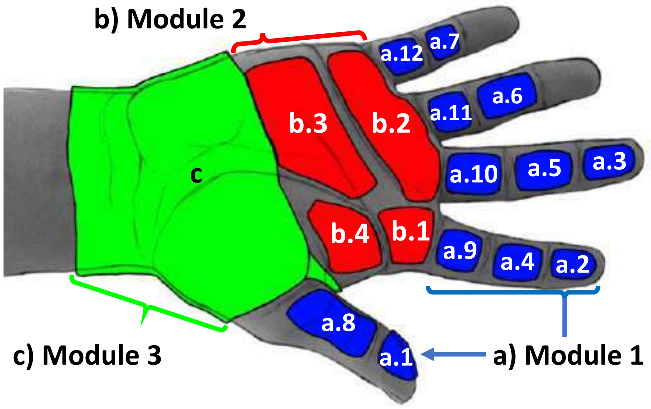


Fig. 8. Palmar View of the exoskeleton and its main modules

When the exoskeleton is activated, each of the sub-modules increase their volume through the inlet of pressurized air into their internal chambers; returning the injured hand to its natural configuration. Fig. 9 shows the hand with the 3 main modules of the exoskeleton activated; but for this conceptual design proposal, only the D-II sub-modules will be shown. The sub-modules of 1 and 2, working interdependently, will generate a controlled change in their volume, for which it will be possible to modify the angles of the MCP, PIP and DIP joints [60, 61, 62, 63], limiting their mobility. Besides, note that module 3 completely covers the wrist, partially restricting its mobility and keeping it parallel to the arm (0 °).

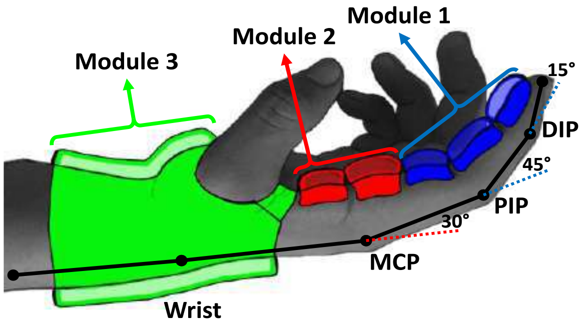


Fig. 9. Lateral view of the exoskeleton and its main modules activated

Therefore, the exoskeleton would not only partially restrict the mobility of the injured hand joints, keeping its natural relaxed configuration, but would also form an additional soft neumatic protection to ensure their safety, allowing the user to control the vehicle.

3.2. Electronics and Electrical Design

A) Glove Sensors and Circuits

The system's electronics are focused on the trajectory of the movement of the digits (from D-II, D-III): pronation, supination, and position of the hand and forearm. There are 2 signals for processing:

a) The first: SEN-FLEX 2P2I sensor (resistivity: 45k to 125k ohms), which works when the finger is flexing projected in motion and gestures for the activation and start-up protocols, sends a voltage range between 0V and 3.3V to the control unit (Fig. 10).

b) The second: MPU6050 gyro sensors (fig. 11a), which provide guidance and reference the Mars Rover's handling. It has the I2C protocol (inter-integrated circuits), which gives less use of pins and a faster data transfer with the control device; when using both same data bus, each one of the sensors will have an address.

The control unit is the STM32F103 ARM Cortex M3 microcontroller with a maximum CPU speed of 72 MHz, 128Kbyte of Flash, it has 26 inputs and 37 outputs. The programming will be carried out in C++ for sensor data processing algorithm, and with the SPI (serial peripheral interface) protocol to send data that contains the configuration, address decoder, and buffer controller, that will be received by the NRF2401L radio frequency module with 2.4GHz using the low power ISM band [64]. Glove-transmitter PCB size: Length 33.26mm, Width 43.31mm. (Fig. 11.b and Fig. 15)

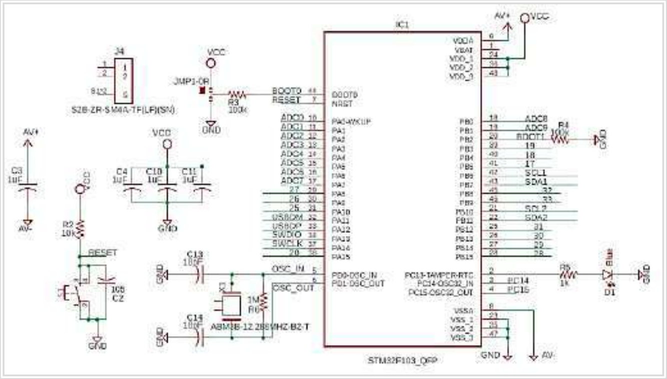


Fig. 10. Microcontroller STM32F103 (Transmitter) - Circuit Diagram

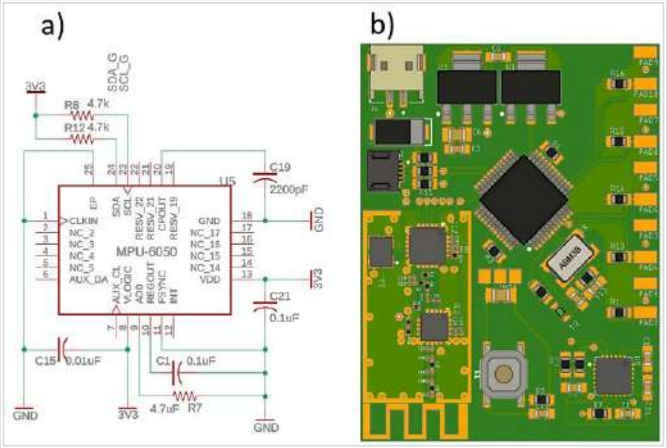


Fig. 11. a) MPU-6050 with connections. b) PCB 3D Design – BIOX-GLOVE

In order to design the circuits, it was required independent grounds planes due to eddy currents [65], where the following components: a) Analog: (SEN-FLEX 2P2I). b) Digital: (MPU6050, NRF2401L, RFX2401C, STM32F103 microcontroller (Fig. 12)), have both grounds planes connected by a 3.3mH inductor; Therefore, two 3.3V voltage regulator circuits will be used on the 1117AMS, delivering a maximum current of 500mA. Furthermore, to prevent EMI (Electromagnetic Interference) or RFI (Radio Frequency Interference), a shield EMC (Electromagnetic Compatibility) is applied on its ground plane [66], whose function is to protect the single-chip, NRF2401L and RFX2401C.

Likewise, the energy subsystem has been divided into 2 sources: a) 85.5mA analog circuit. b) 419.8mA digital circuit, with a total of 505mA, which is why a 3.7V 2S 750mA lithium battery is required.

B) Rover Sensors and Circuits

The receiver board has the following RFX2401C and NRF2401L radio frequency circuits (Figure 13.a) in slave mode; it uses the SPI protocol with the STM32F103 ARMcortex M3 microcontroller, which will process the data received from BIOX-GLOVE and generate the commands for the movement and control of the Mars Rover. The microcontroller will use 2 Serial communication pins (Tx and Rx pins) that will connect with a single-chip CP2102 converter USB (Universal Serial Bus) to TTL (Transistor-Transistor Logic). Rover- receiver PCB size: Length 75.55mm, Width 30.47mm. (Fig. 13.b and Fig. 16)

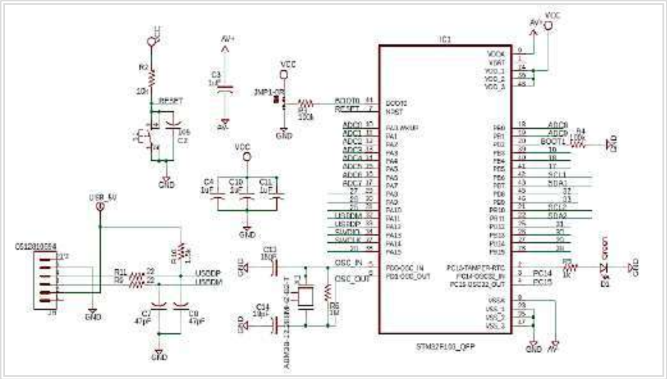


Fig. 12. Microcontroller STM32F103 (Receiver) - Circuit Diagram

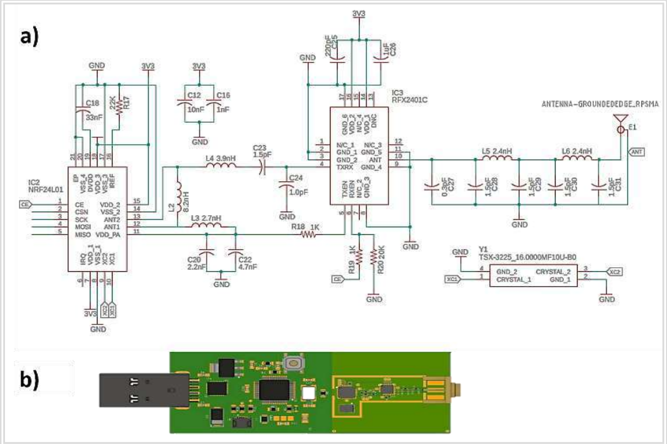


Fig. 13. a) NRF24L01 with connections. b) PCB 3D Design – Rover

3.3 Telecommunications and Telerobotics Technology

The glove's communication with the rover will be unidirectional (master - glove, slave - rover). This research proposes to use: A radio frequency module nRF24L01 Single Chip 2.4GHz Transmitter which operates the 2.4GHz ISM band, with a consumption 11.3mA TX at 0dBm output power and 12.3mA RX at 2Mbps data rate for work in a range of 500 meters in an open field. Also, the RFX2401C a fully integrated front-end module, is used to incorporate all the RF functionality needed for IEEE 802.15.4/ZigBee, with an Auto packet transaction handling wireless sensor network. This component has the lowest power consumption in its category, with only 10.5mA at an output power of -5 dBm and 18 mA in receive-mode.

In addition, the RFX2401C architecture integrates the PA (power amp) and LNA (low noise amp). It is coupled on the transmitter and receiver circuitry, additionally with efficiency +22 dBm output power, low noise with 2.5 dB, and small form factor package, 16-pin (3 x 3 x 0.55 mm).

As shown in the workflow diagram (Fig.14), when the astronaut suffers accidents, he has two ways to activate the glove depending on his health condition:

a) The first way is to perform the pre-established gestures. After the verification, the flex sensors and gyros will provide data according to the protocol of use to the controller to process and transmit it to the rover and command center.

b) The second way is to press the emergency button, which will activate an exoskeleton that works with a system of bags that will inflate and immobilize the hand. The control of the rover will be only with the two gyro sensors.

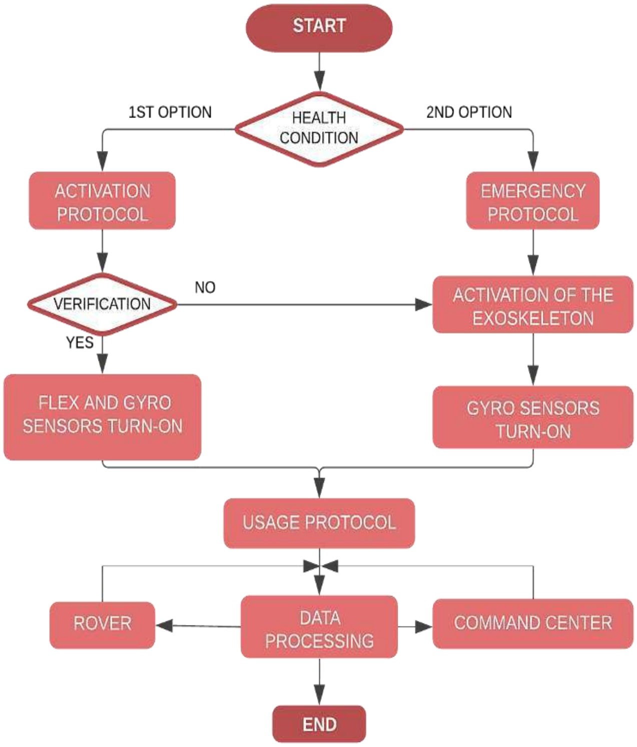


Fig. 14. Workflow Diagram of BIOX-GLOVE

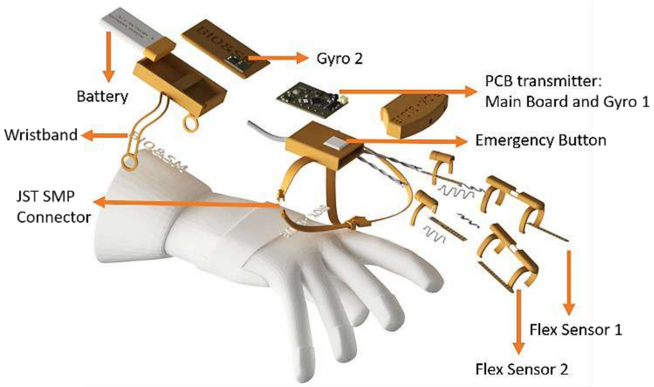


Fig.15. Exploded View of BIOX-GLOVE

When the components of the glove (Fig.15): for Activation Protocol (gyro sensor 1, gyro sensor 2, flex sensor 1 and flex sensor 2), or for Emergency Protocol (gyro sensor 1 and gyro sensor 2), perform the preestablished gestures, the Mars Rover (Fig. 16) will receive the movement order.

a) To start the movement, the user will have to flex the hand 70° [52], leave D-II extended (Fig. 5.a) and close the rest of the hand; the servomotors of the Mars Rover will turn on.

b) All the PIP and DIP joints of each digit must be flexed (Fig. 5.b) when the hand is hyperextended 60 ° [52], then the Mars Rover (Fig. 17a) will move back.

c) To turn left, ulnar deviation (adduction) must be performed at 30° [53] (Fig. 5.c.2); while to turn right, radial deviation (abduction) must be performed at 20° [53] (Fig. 5.c.2); while to turn right, radial deviation (abduction) must be performed at 20° [53] (Fig. 5.c.1) the Mars Rover will turn left or right.

d) All DIP, PIP and MCP joints are fully flexed (Fig. 5.d) when the hand is close into a fist, the Mars Rover (Fig. 17b) activates the emergency brake.

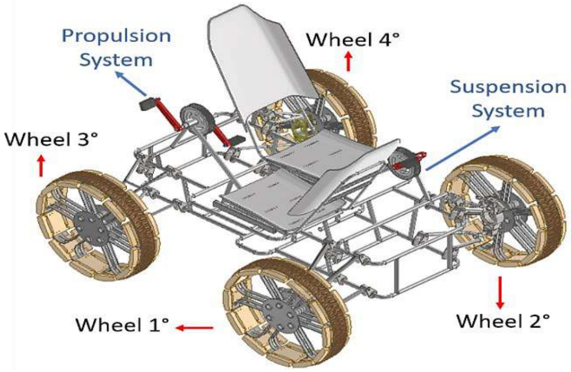


Fig.16. Isometric view of the Mars Rover.

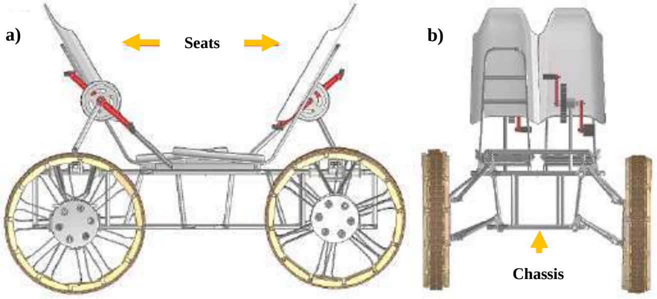


Fig. 17. Mars Rover a) Frontal View b) Back View

IV. Conclusion and Further Work

4.1 Aerospace Application

Mars has an irregular topography [67] which could generate risks for the operation in the moment of providing medical assistance in emergency situations (Fig.18). In fact, this action requires a series of strategic skills to mitigate existing risks, thus maintaining safety and improving the possibility of survival [68, 69]. These events constantly threaten the well-being of the crew and the success of the mission [68]. Therefore, enough medical and technological resources must be available to maintain good health.

Fig.18. Utopia Planitia on Mars. Date: September 5, 1976. Image Credit: NASA [70]

Gloves are one of the pieces of the spacesuit designed with the goal of providing astronauts with the manual agility to operate tools and interact at the Martian surface. Likewise, they provide insulating protection and are flexible, which allows working for long periods avoiding hand fatigue [40], [71].

The main application of BIOX-GLOVE is to activate an emergency plan on Mars. It is proposed to incorporate this biomechatronic innovation in future medical protocols [72] that activates an emergency system in accident cases, which will facilitate the immediate evacuation of the injured astronaut. In addition, this device, thanks to having strategically placed exoskeletons, is capable of providing stability in cases of bone fractures of the wrist and hand [9, 21].

Currently, unmanned rovers cooperate closely in planetary exploration; the operation of these vehicles is by remote control or semi-automatic [73, 74]. However, it is proposed that they will have a wider application for future missions, even as ambulance vehicles to transport astronauts with severe injuries on the Martian surface. Likewise, these means of transportation would be crucial to resolving various medical events more quickly, with good results and greater chances of rescue [68].

Finally, with this project, new adaptable interfaces could be created (T-EVA)[29] between the systems installed inside the astronaut suit and the base station, incurring in the development of developing new portable electronic technologies allowing assisted interaction and reducing the risk of injuries. Likewise, the glove is made by 3d printing technology for manufacturing in the command center or repair after an emergency.

4.2 Earth Application

A) Medical Background and Proposed Validation Test

Exoskeletons allow medical assistance in physical rehabilitation areas for people with motor impairment[26] by providing help to perform tasks and activities of daily living. These robotic-mediated therapies are helpful for specific and repetitive treatment in order to maximize the recovery process, especially in patients [75] with a history of stroke, associated with functional hand pathologies [76]. The main objective of these systems is to restore basic motor function and improve people's quality of life [77, 78].

Likewise, through artificial intelligence, gloves are implemented with sign language recognition systems based on flexible sensors, capable of obtaining data on the shape of the hand movement in an inertial measurement unit to recognize the gesture and transmit information through speech or text [26] [42] [79] making the most of non-verbal communication skills in patients with language aphasia.

The aerospace industry has provided significant knowledge and technology to improve our patient management on Earth [80, 81, 82]. Therefore, it is proposed that BIOX-GLOVE also have terrestrial applications in a hospital environment, especially in emergency medical services, because it has sensors and can activate a medical emergency system from the place where the event occurred. This action could anticipate the medical team to coordinate the transfer in an appropriate and timely manner. In another area, this device, due to its biomechatronic innovation, can be useful to implement physical medicine and rehabilitation services aiding patients with hand disabilities. This device has the great potential to be part of the complementary treatment of degenerative cartilage diseases such as rheumatoid arthritis.

It will be tested on Martian and Lunar analogs to validate this device's functionality since they have similar topographic conditions. La Joya Desert in Peru is considered a candidate for space exploration simulation due to its isolated location and geographic characteristics. [83-85] (Fig.19).

Fig.19. La Joya Desert, Arequipa, Peru. Image Credit: Saúl Pérez [86].

ACKNOWLEDGMENT

The Bioastronautics and Space Mechatronics Research Group (https://orcid.org/0000-0002-7173-3960) under the management, supervision, and support of Dr. José Cornejo, Juan Carlos Chavez and PhD Adolfo Chaves. Also, Gustavo Juarez and Martin Fuentes members of the Team "Tharsis" from the Universidad Nacional de Ingeniería for the support in the mechatronic design of the project. Also, Dr. Niels Castrejón for the research in biomechanical fundamentals. Finally, thanks to Daniel A. Angeles for the support and translation.

REFERENCES

[1] S. Loff, "Programa Artemis | NASA," 31 January 2021. [Online]. Available: https://www.nasa.gov/artemisprogram. [Accessed 31 January 2021].

[2] K. Szocik, T. Wójtowicz and M. Braddock, "The Martian: Possible Scenarios for a Future Human Society on Mars," *Space Policy,* vol. 54, p. 101388, 1 November 2020.

[3] M. V. Rivera, J. Cornejo, K. Huallpayunca, A. B. Diaz, Z. N. Ortiz-Benique, A. D. Reina, G. Jamanca Lino and V. Ticllacuri, "Medicina humana espacial: Performance fisiológico y contramedidas para mejorar la salud del astronauta," *Revista de la Facultad de Medicina Humana,* vol. 20, no. 2, pp. 131-142, 27 March 2020.

[4] V. Ticllacuri, G. J. Lino, A. B. Diaz and J. Cornejo, "Design of Wearable Soft Robotic System for Muscle Stimulation Applied in Lower Limbs during Lunar Colonization," *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, Lima, Peru, 2020, pp. 1-4, doi: 10.1109/INTERCON50315.2020.9220206.

[5] M. Bizzarri, M. G. Masiello, R. Guzzi and A. Cucina, "Journey to Mars: A Biomedical Challenge. Perspective on future human space flight," *Organisms. Journal of Biological Sciences,* vol. 1, pp. 15-26, 15 December 2017.

[6] Á. Villoslada, C. Rivera, N. Escudero, F. Martín, D. Blanco and L. Moreno, "Hand Exo-Muscular System for Assisting Astronauts During Extravehicular Activities," *Soft Robotics,* vol. 6, no. 1, 12 February 2019.

[7] N. C. Jordan, J. H. Saleh and D. J. Newman, "The extravehicular mobility unit: A review of environment, requirements, and design changes in the US spacesuit," *Acta Astronautica,* vol. 59, no. 12, pp. 1135-1145, 1 December 2006.

[8] A. Favetto, F. Chen Chen, E. P. Ambrosio, D. Manfredi and G. C. Calafiore, "Towards a hand exoskeleton for a smart EVA glove," *2010 IEEE International Conference on Robotics and Biomimetics*, Tianjin, China, 2010, pp. 1293-1298, doi: 10.1109/ROBIO.2010.5723515.

[9] S. Strauss, R. L. Krog and A. H. Feiveson, "Extravehicular mobility unit training and astronaut injuries," *Aviation, Space, and Environmental Medicine,* vol. 76, no. 5, pp. 469-474, 2005.

[10] M. C. Carrozza, F. Vecchi, S. Roccella, L. Barboni, E. Cavallaro, S. Micera and P. Dario, "The adah project: an astronaut dexterous artificial hand to restore the manipulation abilities of the astronaut," in *7th ESA Workshop on Advanced Space Technologies for Robotics and Automation 'ASTRA 2002*, ESTEC, Noordwijk, The Netherlands, 2002.

[11] D. W. Beaty, R. P. Mueller, D. B. Bussey, R. M. Davis, L. E. Hays, S. J. Hoffman and E. Zbinden, "Some strategic considerations related to the potential use of water resource deposits on Mars by future human explorers," 11-15 April 2016.

[12] NASA, "NASA's Journey to Mars - Pioneering Next Steps in Space Exploration," *National Aeronautics and Space Administration,* pp. 1-36, October 2015.

[13] C. L. Mansfield, "Apollo 17," NASA, 16T12:11-04:00 03 2015. [Online]. Available: http://www.nasa.gov/mission_pages/apollo/missions/apollo17.html. [Accessed 08 February 2021].

[14] A. W. Kirkpatrick, C. G. Ball, M. Campbell, D. R. Williams, S. E. Parazynski, K. L. Mattox and T. J. Broderick, "Severe traumatic injury during long duration spaceflight: Light years beyond ATLS," *Journal of Trauma Management & Outcomes,* vol. 3, no. 1, p. 4, 25 March 2009.

[15] B. A. Houtchens, "Medical-care systems for long-duration space missions," *Clinical Chemistry,* vol. 39, no. 1, pp. 13-21, 1 January 1993.

[16] M. Komorowski, S. Fleming and A. W. Kirkpatrick, "Fundamentals of Anesthesiology for Spaceflight," *Journal of Cardiothoracic and Vascular Anesthesia,* vol. 30, no. 3, pp. 781-790, 2016.

[17] G. Horneck and B. Comet, "General human health issues for Moon and Mars missions: Results from the HUMEX study," *Advances in Space Research,* vol. 37, no. 1, pp. 100-108, 1 January 2006.

[18] E. S. Nelson, B. Lewandowski, A. Licata and J. G. Myers, "Development and Validation of a Predictive Bone Fracture Risk Model for Astronauts," *Annals of Biomedical Engineering,* vol. 37, no. 11, pp. 2337-2359, 01 November 2009.

[19] P. D. Hodkinson, R. A. Anderton, B. N. Posselt and K. J. Fong, "An overview of space medicine," *BJA: British Journal of Anaesthesia,* vol. 119, no. suppl_1, pp. i143-i153, 2017.

[20] S. Atrauss, "Extravehicular mobility unit training suit symptom study report," *Johnson Space Center, Houston, TX,* 2004.

[21] R. A. Opperman, J. M. A. Waldie, A. Natapoff, D. J. Newman and J. A. Jones, "Probability of spacesuit-induced fingernail trauma is associated with hand circumference," *Aviation, Space, and Environmental Medicine,* vol. 81, no. 10, pp. 907-913, 2010.

[22] N. Sreekanth, A. Dinesan, A. R. Nair, G. Udupa and V. Tirumaladass, "Design of robotic manipulator for space applications," *Materials Today: Proceedings,* 7 December 2020.

[23] J. Cornejo, J. A. Cornejo-Aguilar and J. P. Perales-Villarroel, "Innovaciones internacionales en robótica médica para mejorar el manejo del paciente en Perú," *Revista de la Facultad de Medicina Humana,* vol. 19, no. 4, pp. 105-113, 2019.

[24] S. Lombardo, K. Duda and L. Stirling, "Evaluating the Effect of Spacesuit Glove Fit on Functional Tactility Task Performance," *2020 IEEE Aerospace Conference*, Big Sky, MT, USA, 2020, pp. 1-10, doi: 10.1109/AERO47225.2020.9172314.

[25] X. Yu, P. Wang and Z. Zhang, "Learning-Based End-to-End Path Planning for Lunar Rovers with Safety Constraints," *Sensors,* vol. 21, no. 3, p. 976, January 2021.

[26] S. Mugala, P. Jjagwe and J. Asiimwe, "Glove Based Sign Interpreter for Medical Emergencies," in *2019 IST-Africa Week Conference (IST-Africa),* Nairobi, Kenya, 2019.

[27] C. Chia-Ye and P. Rita M, "Soft robotic devices for hand rehabilitation and assistance: a narrative review," *Journal of NeuroEngineering and Rehabilitation,* vol. 15, no. 1, p. 9, 17 February 2018.

[28] P. Polygerinos, Z. Wang, K. C. Galloway, R. J. Wood and C. J. Walsh, "Soft robotic glove for combined assistance and at-home rehabilitation," *Robotics and Autonomous Systems,* vol. 73, pp. 135-143, 1 November 2015.

[29] P. Palacios, W. Castillo, M. V. Rivera and J. Cornejo, "Design of T-EVA: Wearable Temperature Monitoring System for Upper Limbs during Extravehicular Activities on Mars," *2020 IEEE Engineering International Research Conference (EIRCON),* Lima, Peru, 2020, pp. 1-4, doi: 10.1109/EIRCON51178.2020.9254027.

[30] UNOOSA, "https://www.unoosa.org/," 04 January 2021. [Online]. Available: https://www.unoosa.org/pdf/gares/ARES_22_2345S.pdf. [Accessed 04 January 2021].

[31] A. Nicogossian, "Medicine and space exploration," *The Lancet,* vol. 362, pp. s8-s9, 1 December 2003.

[32] J. Mindock, J. Reilly, D. Rubin, M. Urbina, M. Hailey, J. A. Cerro, A. Hanson, K. McGuire, T. Burba, C. Middour, M. Krihak and D. Reyes, "Systems Engineering for Space Exploration Medical Capabilities," in *AIAA SPACE and Astronautics Forum and Exposition*, USA, American Institute of Aeronautics and Astronautics, 2017.

[33] D. Hamilton, K. Smart, S. Melton, J. D. Polk and K. Johnson-Throop, "Autonomous medical care for exploration class space missions," *The Journal of Trauma,* vol. 64, no. 4, pp. S354-363, April 2008.

[34] D. R. Williams and M. Turnock, "Human Space Exploration The Next Fifty Years," *McGill Journal of Medicine : MJM,* vol. 13, no. 2, p. 76, June 2011.

[35] M. Vargas, J. Cornejo and L. E. Correa-López, "Ingenieria biomedica: La revolución tecnológica para el futuro del sistema de salud [Cartas al Editor]," *Revista de la Facultad de Medicina Humana,* vol. 16, no. 3, pp. 95-96, 2016.

[36] M. K. Habib, "Human adaptive and friendly mechatronics (HAFM)," *2008 IEEE International Conference on Mechatronics and Automation*, Takamatsu, Japan, 2008, pp. 61-65, doi: 10.1109/ICMA.2008.4798726.

[37] S. Najarian, D. Javad , G. Darbemamieh and S. H. Farkoush, "Sensing Technology," in *Mechatronics in Medicine: A Biomedical Engineering Approach*, McGraw-Hill Education, 2012.

[38] J. Cornejo, J. P. Perales-Villarroel, R. Sebastian and J. A. Cornejo-Aguilar, "Conceptual Design of Space Biosurgeon for Robotic Surgery and Aerospace Medicine," *2020 IEEE ANDESCON*, Quito, Ecuador, 2020, pp. 1-6, doi: 10.1109/ANDESCON50619.2020.9272122.

[39] M. M. S. Mousavi, A. Favetto, F. C. Chen, E. Ambrosio, S. Appendino, D. Manfredi, F. Pescarmona and A. Soma`, "41st International Conference on Environmental Systems," Portland, Oregon, American Institute of Aeronautics and Astronautics, 2011.

[40] D. A. Shockey, R. S. Piascik, B. J. Jensen, L. S. Hewes and J. K. Sutter, "Textile Damage in Astronaut Gloves," *Journal of Failure Analysis and Prevention,* vol. 13, no. 6, pp. 748-756, 1 December 2013.

[41] V. P. Katuntsev, Y. Y. Osipov and S. N. Filipenkov, "Biomedical problems of EVA support during manned space flight to Mars," *Acta Astronautica,* vol. 64, no. 7, pp. 682-687, 1 April 2009.

[42] C. Wei-Chieh, H. Wen-Jyi, T. Tsung-Ming, H. De-Rong and J. Yun-Jie, "Continuous Finger Gesture Recognition Based on Flex Sensors," *Sensors,* vol. 19, no. 18, p. 3986, 15 September 2019.

[43] A. Jurine, "Anthropometry and Biomechanics," National Aeronautics and Space Administration, 27 August 2020. [Online]. Available: https://msis.jsc.nasa.gov/sections/section03.htm#_3.2_GENERAL_ANTHROPOMETRICS. [Accessed 25 January 2021].

[44] O. Kwon, K. Jung, H. You and K. Hee-Eun, "Determination of key dimensions for a glove sizing system by analyzing the relationships between hand dimensions," *Applied Ergonomics,* vol. 40, no. 4, pp. 762-766, 1 July 2009.

[45] M. Vergara, M.-J. Agost and V. Bayarri, "Anthropometric characterisation of palm and finger shapes to complement current glove-sizing systems," *International Journal of Industrial Ergonomics,* vol. 74, p. 102836, 2019.

[46] A. Yu, K. L. Yick, S. P. Ng and J. Yip, "2D and 3D anatomical analyses of hand dimensions for custom-made gloves," *Applied Ergonomics,* vol. 44, no. 3, pp. 381-392, 1 May 2013.

[47] H. Hsiao, J. Whitestone, T.-Y. Kau and B. Hildreth, "Firefighter Hand Anthropometry and Structural Glove Sizing: A New Perspective," *Human Factors,* vol. 57, no. 8, pp. 1359-1377, 1 December 2015.

[48] O. Oviedo-Trespalacios, L. M. Buelvas, J. Hernández and J. Escobar, "Hand anthropometric study in northern Colombia," *International Journal of Occupational Safety and Ergonomics,* vol. 23, no. 4, pp. 472-480, 2017.

[49] H. Bianchi, D. Saravia and N. E. Ottone, "Unusual pattern of the first dorsal metacarpal artery," *Surgical and Radiologic Anatomy,* vol. 39, no. 7, pp. 799-802, 01 July 2017.

[50] A. F. García G and A. J. Becerra, "Prototipo de mano robótica inspirada en la mano humana," *Revista Tekhnê,* vol. 13, no. 2, pp. 27-42, Julio - Diciembre 2016.

[51] F. Flórez-Revuelta, Modelo de representación y procesamiento de movimiento para diseño de arquitecturas de tiempo real especializadas, Alicante : Biblioteca Virtual Miguel de Cervantes: Tesis Doctoral, 2002.

[52] S. J. Hall, Basic Biomechanics, 8th ed., McGraw-Hill, 2018.

[53] J. Hamill, K. M. Knutzen and T. R. Derrick, Biomechanical Basis of Human Movement, 4th ed., Lippincott Williams & Wilkins, 2014.

[54] M. Nordin and V. H. Frankel, Basic Biomechanics of the Musculoskeletal System, 4th ed., Lippincott Williams y Wilkins, 2012.

[55] G. A. Smithers, M. K. Nehls, M. A. Hovater, S. W. Evans, J. S. Miller, R. M. Broughton, D. Beale y F. Kilinc-Balci, “A One-Piece Lunar Regolith Bag Garage Prototype,” NTRS - NASA Technical Reports Server, USA, 2007.

[56] M. M. Peer Mohamed, M. Holynska, P. Weiss, T. Gobert, Y. Chouard, N. Singh, T. Chalal, N. Sejkora, G. Groemer, S. Schmied, M. Schweins, T. Stegmaier, G. T. Gresser y S. Das, “Planetary exploration textiles (pextex) - materials selection procedure for surface EVA suit development,” de 71 nternational Astronautical Congress (IAC) , 2020.

[57] A. Stilli *et al*., "AirExGlove — A novel pneumatic exoskeleton glove for adaptive hand rehabilitation in post-stroke patients," *2018 IEEE International Conference on Soft Robotics (RoboSoft)*, Livorno, Italy, 2018, pp. 579-584, doi: 10.1109/ROBOSOFT.2018.8405388.

[58] H. K. Yap *et al*., "A Fully Fabric-Based Bidirectional Soft Robotic Glove for Assistance and Rehabilitation of Hand Impaired Patients," in *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1383-1390, July 2017, doi: 10.1109/LRA.2017.2669366.

[59] A. S. Gorgey, "Robotic exoskeletons: The current pros and cons," *World Journal of Orthopedics,* vol. 9, no. 9, pp. 112-119, 18 September 2018.

[60] E. Matheson and G. Brooker, "Augmented robotic device for EVA hand manoeuvres," *Acta Astronautica,* vol. 81, no. 1, pp. 51-61, 01 December 2012.

[61] L. Randazzo, I. Iturrate, S. Perdikis and J. d. R. Millán, "mano: A Wearable Hand Exoskeleton for Activities of Daily Living and Neurorehabilitation," in *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 500-507, Jan. 2018, doi: 10.1109/LRA.2017.2771329.

[62] J. Kawashimo, Y. Yamanoi and R. Kato, "Development of easily wearable assistive device with elastic exoskeleton for paralyzed hand," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Lisbon, Portugal, 2017, pp. 1159-1164, doi: 10.1109/ROMAN.2017.8172450.

[63] D. Popov, I. Gaponov and J. Ryu, "Portable Exoskeleton Glove With Soft Structure for Hand Assistance in Activities of Daily Living," in *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 2, pp. 865-875, April 2017, doi: 10.1109/TMECH.2016.2641932.

[64] I. M. M. Yusoff, N. M. Salleh, A. A. Azlan, R. A. Awang and M. T. Ali, "Radiated Electromagnetic Bad Gap Antenna for ISM Band in Medical Application," *2018 IEEE International RF and Microwave Conference (RFM)*, Penang, Malaysia, 2018, pp. 103-106, doi: 10.1109/RFM.2018.8846532.

[65] S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. v. Molnár, M. L. Roukes, A. Y. Chtchelkanova and D. M. Treger, "Spintronics: A Spin-Based Electronics Vision for the Future," *Science,* vol. 294, no. 5546, pp. 1488-1495, 2001.

[66] C. Sarkar, D. Guha and C. Kumar, "Compound ground plane designed for probe-fed dielectric resonator antennas," *2017 IEEE International Conference on Antenna Innovations & Modern Technologies for Ground, Aircraft and Satellite Applications (iAIM)*, Bangalore, India, 2017, pp. 1-4, doi: 10.1109/IAIM.2017.8402633.

[67] M. H. Carr and J. F. Bell, "Chapter 17 - Mars: Surface and Interior," in *Encyclopedia of the Solar System (Third Edition)*, T. Spohn, D. Breuer and T. V. Johnson, Eds., Boston, Elsevier, 2014, pp. 359-377.

[68] J. M. Robertson, R. D. Dias, A. Gupta, T. Marshburn, S. R. Lipsitz, C. N. Pozner, T. E. Doyle, D. S. Smink, D. M. Musson and S. Yule, "Medical Event Management for Future Deep Space Exploration Missions to Mars," *The Journal of Surgical Research,* vol. 246, pp. 305-314, 12 November 2020.

[69] J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque and M. Winckler, Human-Computer Interaction – INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part IV, vol. 9299, Germany: Springer, 2015, pp. 18-36.

[70] NASA, "Mars - Sparsely cratered plains," Encyclopedia Britannica, 09 February 2021. [Online]. Available: https://www.britannica.com/place/Mars-planet. [Accessed 09 February 2021].

[71] A. Favetto, F. C. Chen, E. P. Ambrosio, D. Manfredi and G. C. Calafiore, "Towards a hand exoskeleton for a smart EVA glove," in *2010 IEEE International Conference on Robotics and Biomimetics*, Tianjin, China, 2010.

[72] I. Cinelli, "Space Medicine Requirements open to Innovation in Biomedical Engineering," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, 2018, pp. 973-976, doi: 10.1109/EMBC.2018.8512326.

[73] B. H. Wilcox, "Robotic vehicles for planetary exploration," *Applied Intelligence,* vol. 2, no. 2, pp. 181-193, 01 August 1992.

[74] P. S. Schenker, T. L. Huntsberger, P. Pirjanian, E. T. Baumgartner and E. Tunstel, "Planetary Rover Developments Supporting Mars Exploration, Sample Return and Future Human-Robotic Colonization," *Autonomous Robots,* vol. 14, no. 2, pp. 103-126, 01 February 2003.

[75] D. Simonetti, N. L. Tagliamonte, L. Zollo, D. Accoto and E. Guglielmelli, "Chapter 3 - Biomechatronic design criteria of systems for robot-mediated rehabilitation therapy," in *Rehabilitation Robotics*, R. Colombo and V. Sanguineti, Eds., Academic Press, 2018, pp. 29-46.

[76] M. Borghetti, E. Sardini and M. Serpelloni, "Sensorized Glove for Measuring Hand Finger Flexion for Rehabilitation Purposes," in *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 12, pp. 3308-3314, Dec. 2013, doi: 10.1109/TIM.2013.2272848.

[77] H. Cao and D. Zhang, "Soft robotic glove with integrated sEMG sensing for disabled people with hand paralysis," *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Qingdao, 2016, pp. 714-718, doi: 10.1109/ROBIO.2016.7866407.

[78] Z. Ma, P. Ben-Tzvi and J. Danoff, "Sensing and Force-Feedback Exoskeleton Robotic (SAFER) Glove Mechanism for Hand Rehabilitation," in *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Boston, Massachusetts, USA, 2015.

[79] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih and M. M. b. Lakulu, "A Review on Systems-Based Sensory Gloves for Sign Language Recognition State of the Art between 2007 and 2017," *Sensors,* vol. 18, no. 7, p. 2208, 2018.

[80] G. J. Lino, P. Palacios, J. L. Napán and J. Cornejo, "Conceptual Design of ARIES Rover for Water Production and Resource Utilization on the Moon," *2020 IEEE Engineering International Research Conference (EIRCON)*, Lima, Peru, 2020, pp. 1-4, doi: 10.1109/EIRCON51178.2020.9253766.

[81] I. Cinelli and L. Brown, "Innovation in Medical Technology Driven by Advances in Aerospace," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, 2018, pp. 941-944, doi: 10.1109/EMBC.2018.8512444..

[82] V. M. Gonzales, J. Cornejo and R. Palomares, "Mechatronics Design of High-Altitude Balloon Paulet-1 for Peruvian Aerospace Monitoring," *2020 Congreso Estudiantil de Electrónica y Electricidad (INGELECTRA)*, Santiago, Chile, 2020, pp. 1-6, doi: 10.1109/INGELECTRA50225.2020.246965.

[83] J. E. Valdivia-Silva, R. Navarro-González, F. Ortega-Gutierrez, L. E. Fletcher, S. Perez-Montaño, R. Condori-Apaza and C. P. McKay, "Multidisciplinary approach of the hyperarid desert of Pampas de La Joya in southern Peru as a new Mars-like soil analog," Geochimica et Cosmochimica Acta, vol. 75, no. 7, pp. 1975-1991, 01 April 2011.

[84] S. Engler, J. Hunter, K. Binsted and H. Leung, "Robotic Companions for Long Term Isolation Space Missions," in *2018 15th International Conference on Ubiquitous Robots (UR)*, Honolulu, HI, EE. UU, 2018.

[85] J. E. Valdivia-Silva, R. Navarro-González, L. Fletcher, S. Pérez-Montaño, R. Condori-Apaza, F. Ortega-Gutiérrez and C. McKay, "Martian Analogue of Pampas de La Joya: an update and future implications," International Journal of Astrobiology, vol. 11, no. 1, pp. 25-35, 2012.

[86] C. Becerra Gutiérrez and F. A. Paz Quiroz, "Peru: NASA-backed mission sets up research camp in southern desert," Empresa Peruana de Servicios Editoriales S. A. EDI, 22 Setiembre 2017. [Online]. Available: https://andina.pe/ingles/noticia-peru-nasabacked-mission-sets-up-research-camp-in-southern-desert-683936.aspx. [Accessed 27 February 2021].

Hybrid Energy Storage for Underground Mining EV

Ashleigh K. Townsend[1], Immanuel N. Jiya[2] and Rupert Gouws[1]
[1]*School of Electrical, Electronic and Computer Engineering, North West University*
Potchefstroom 2520, South Africa
[2]*Department of Engineering and Science, University of Agder,*
Grimstad 4879, Norway
ashleighktownsend2@gmail.com

***Abstract*—One of the main concerns when it comes to contributions to the carbon footprint is the consumption of fuel by vehicles, more specifically that of mining vehicles. One solution to this problem is replacing the combustion engine vehicles with electric vehicles. This also poses the problem of low power density. This paper proposes a solution of using supercapacitors and hybrid capacitors with the aim of investigating how they can affect the performance of electric vehicles. For this purpose, a background of fuel consumption and alternative solutions to the problem are discussed, a prototype was designed and built. It was found that when supercapacitors and hybrid capacitors were used in combination with the battery the system had a notably longer endurance.**

Keywords—Electric vehicle (EV); hybrid energy storage; supercapacitor (SC); hybrid capacitor (HC)

I. Introduction

In recent years it has been noticed that the world is moving into an environmental era where more eco-friendly choices are favoured above the conventional choices that increase the overall carbon footprint [1]. South Africa along with many other developing and developed countries, has signed the Paris Agreement. With this in mind, the DMR (Department of Mineral Resources) of South Africa has increased the importance of each mine decreasing their carbon footprint, as the coal mines in South Africa contribute to 9% [2] of the countries greenhouse emissions alone. The large number of emissions due to the mines is mainly accounted for by the large number of vehicles and the size of these vehicles in operation there. Two additional concerns due to the vast amount of emissions were raised, namely, safety and cost concerns. The DMR raised the concern of safety for the workers, as the magnitude of fumes around these vehicles causes the ventilation to become insufficient, increasing effects of fume inhalations and temperature increases in the vicinity of these vehicles. The cost concerns were raised by the chamber of mines, which include concerns due to increased ventilation costs, increasing fuel price, increasing salary demands and costs to reduce the carbon footprint [3].

The mines are constantly expanding leading to an increased demand for fuel, which has a limited supply. The vehicles that are currently in use in the mining industry, mainly Load Haul Devices (LHDs) and Utility Vehicles (UVs) [2], [4] use either diesel or petrol to operate, which although initially cheaper and more readily available, pose several problems. Firstly, the fuel consumption is too high, and secondly the fuel consumption causes the vehicles to emit extremely poisonous fumes both above and below ground. The ventilation underground is not sufficient for the volume of fumes concentrated in the more confined working areas [5]. This not only poses a threat to the workers' health but also causes the temperature in the vicinity of these vehicles to be much higher than optimal [6]. Therefore, if the mine health and safety act and the fuel costs are taken into consideration, it is beneficial to implement an alternative energy source that will reduce fuel consumption and therefore costs, as well as increasing the health status of the workers.

Although the electric vehicles (EV) do not emit exhaust fumes, they do require electricity to operate, which in itself is also a massive greenhouse gas contributor, as coal is burnt to produce electricity in South Africa [7]. They can be more economically friendly in the long run once solar panels are more widely implemented for electricity production, but this is not in the scope of this research. The electrical vehicles also create a safer environment for the workers in the immediate vicinity. Another huge advantage of the electrical vehicles is that the unnecessary fuel expenses will be reduced [8].

As previously mentioned there are currently multiple solutions available and in place to solve the problem of a too large fuel consumption [9]. Hybridization is currently one of the more preferred solutions to any of these fuel consumption issues as it decreases the consumption significantly with the bonus of adding a couple other advantages to the existing system [10]. The idea of combining batteries with supercapacitors (SCs) is not a new one, it has been researched in depth in a previous research [11], however, it is yet to be implemented to see how it functions in an application.

EVs have emerged as an alternative to combustion engines as they do not emit carbon dioxide, however the extended charge period of the battery packs and the short drive range thereof lead them to still be under-utilized. One option to improve on these downfalls would be to integrate the EV with a hydrogen fuel cell (FC) [12]. The FC has a high energy density but a low power density, allowing it to be able to supply average power for an extended period of time if that time does not include power requirement spikes [13]. Power spikes occur when the load requires peak power, which in turn drains the energy source, to overcome this issue a buffer battery is used to supplement the FC during the spikes, increasing the operational period of the FC. Just like EVs, FC-EVs produce no pollutants during operation, however the production of hydrogen causes its availability to be quite low [14], [15]. The proton exchange membrane fuel cell (PEMFC) is preferred for vehicular use as it offers many advantages over the general FC, including the ability to tolerate air – it does not require pure oxygen.

Hydrogen FCs have their advantages and disadvantages that determine their application. Some advantages include: a driving range and refuelling time comparable to that of conventional vehicles (320-480 km), no tailpipe pollution, hydrogen is a by-product of mining for platinum and therefor relatively freely available, refuelling is limited by the speed of

978-1-6654-4067-7/21/$31.00 ©20XX IEEE

replacing the hydrogen gas, regenerative braking can be utilized to capture lost energy and charge the battery, it is highly efficient with a high energy density and low operating temperature and has a long life span [16]. Hydrogen FCs are however not without their disadvantages, namely, the size and weight of the system can pose a problem when replacing existing systems, the performance is degraded by carbon monoxide [16], the system is currently expensive and the storage of the hydrogen on the vehicle currently poses a challenge [17].

Hybrids use many variations of combinations of alternative power sources with the electric system [18]. Some alternative sources include combustion engines, solar, wind or ultra-capacitors. These sources are combined in many variations (that are sometimes user defined) in order to achieve optimal efficiency [19], [20] . Battery powered EVs use batteries to power an electric motor that puts the vehicle in motion. Batteries are recharged using grid electricity from wall socket or charging unit. These vehicles do not produce pollution themselves, but the production of the electricity required does cause pollution if renewable energy sources are not used [21]. The most common battery used in EVs is the lead acid battery as it is more robust, has comparably high recharge cycles, is generally safer than the other types and has good energy and power characteristics as compared to the other types [22].

When using batteries the advantages associated with them include, increased efficiency compared to combustion engines, no emissions, the possible use of renewable energy, pollution caused by production of electricity is much less than that caused by combustion engines, the long term expenses decrease, regenerative braking and idle-off can be utilized to increase efficiency, the system can be recharged at home, has near instant torque and an extremely fast acceleration [21]. More so, the system is smaller than the others and has an overall decrease in fuel consumption [23]. Although batteries tend to be the more viable option for energy storage in EVs [24] they do however have their downfalls including, an extended charge period, a short driving range (110 – 160 km) [21], low power density and the fuel consumption is heavily reliant on the powertrain characteristics [23].

II. Proposed Topology of Mining EV.

Pure EVs make use of grid power to charge the batteries that then store the power until the vehicle is required. The battery power is used to supply an electric motor, as well as all auxiliary control systems; the electric motor is then used to propel the vehicle. Although this system can sustain moderate power requirements, the battery essentially has a high energy density and low power density. The latter means that the energy of the vehicle will quickly be depleted if high power demands are persistently required. Therefore, another power source is required that will be able to supplement the system during the peak power periods allowing the batteries to have a longer operational period. There are many power sources that can be utilized for this purpose, but ultra–capacitors offer some advantages over the others – they recharge faster, they have much higher power densities and can be used in a variety of kinetic energy storage applications.

For this paper, the focus is on investigating how ultra-capacitors (supercapacitors – SCs, and hybrid capacitors – HCs) affect the efficacy of a battery powered EV. The system that was altered and tested consisted of an electric DC motor powering the vehicle and a windscreen wiper motor powering the steering system of the vehicle. All these components and the control system thereof were powered by the battery, in this case a 12 V, 7 Ah Lead acid battery. The layout of the proposed system is seen in Fig. 1. In altering the pre-existing system, a power switch was used to switch between the power sources when required, this was done using sensors that determined the current requirements of the motor at each stage of a test path. Fig. 1 shows the functional operation of the proposed solution.

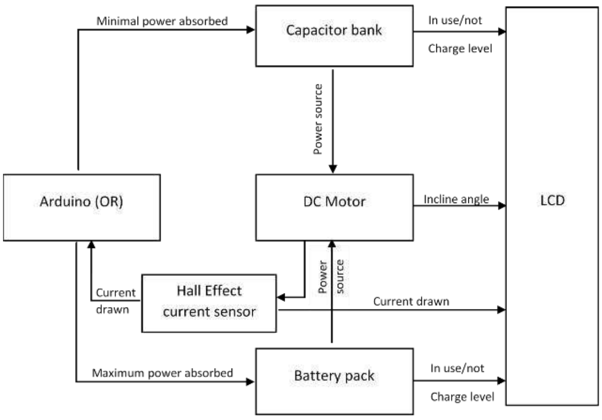


Fig. 1. Functional operation of system.

The power source consisted of a microcontroller that received information from the DC motor, via a sensor, that was used to control the power switch. The information from the DC motor sensors consisted mainly of the power absorption of the motor. When the motor absorbed significantly more power, it was assumed that the vehicle was traversing an incline, or that the motor was starting up and therefore required more power. This need was conveyed to the power switch that then chose the capacitor bank as the main power source. When the DC motor power absorption stabilized, the power switch chose the battery as main power source. When the power absorption remained low or disappeared entirely the capacitor bank stored any energy that the battery would normally waste. When the vehicle was again in motion, the capacitor bank was favoured to allow the vehicle to have a smoother acceleration until the vehicle reached a stable power absorption in which it favoured the battery as main power source. For convenience of the user, a Bluetooth module received the sensor values (current drawn by the motor, state of charge of current power source, incline level) and stored them to be analysed when desired. Different combinations of the power sources were used in order to compare the results and the results were used to determine any changes in efficiency of the vehicle.

For the integration of the power source with the vehicle and the remainder of the circuit, a DC/DC buck-boost converter was designed, a buck for the battery (12.5 V< battery voltage < 13.8 V) and a buck-boost for the capacitor banks (4 V < capacitor bank voltage < 16 V). This ensured that the motor received a continuous 12 VDC supply. Only one power source was used to power the drive motor at a time, therefore only one DC/DC circuit was required. The same h-bridge design that was used to control the servo motor and wiper motor was used for the buck-boost DC/DC converter for the energy storage hybridisation.

III. Results

The final circuit built and used can be seen below in Fig. 2, the design of the vehicle in Fig. 3(a), and the final vehicle implemented in Fig. 3(b). The SCs and HCs used can be seen in Fig. 4. A specific path was chosen with a specified length and varying levels of incline to compare the various combinations. First, the battery was connected as the sole power source, second was the SC bank, third was the HC bank and finally all three were used as a combination. Fig. 5 shows the incline of the path that was used to test the vehicle with the respective power sources and the amount of current drawn by the motor at each point.

To compare the energy sources the distance and time it took each source to traverse along the path until they were depleted was recorded. The discharge period of the batteries was recorded for individual use and when combined with the capacitor bank. Fig. 7 shows the results obtained from these respective tests. From the figure when the battery was used in combination with the capacitor bank the voltage thereof dropped at a slower rate than when used on its own. This allowed the vehicle to become more efficient over the long term. The discharge period of the capacitors was recorded for individual use and when combined with each other and the battery, this is shown in Fig. 7. The combination of the capacitors causes the discharge rate to be slightly higher than the SCs but significantly lower than the HCs. This caused the system to, again, be more efficient than when the sources were used individually.

In Fig. 8, there are two graphs, the current drawn by the EV and the state of charge of the EV (showing the state of charge of the power source in use), thus also showing when the sources were switched.

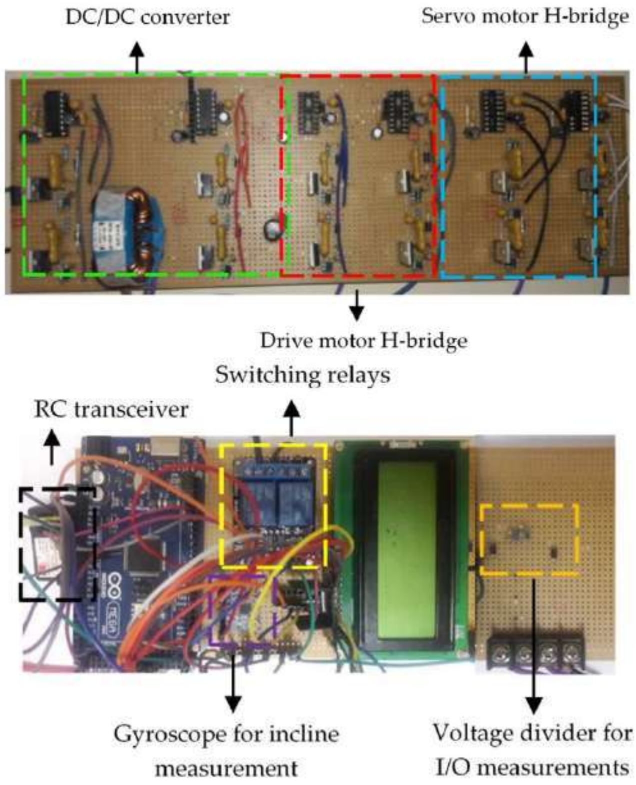
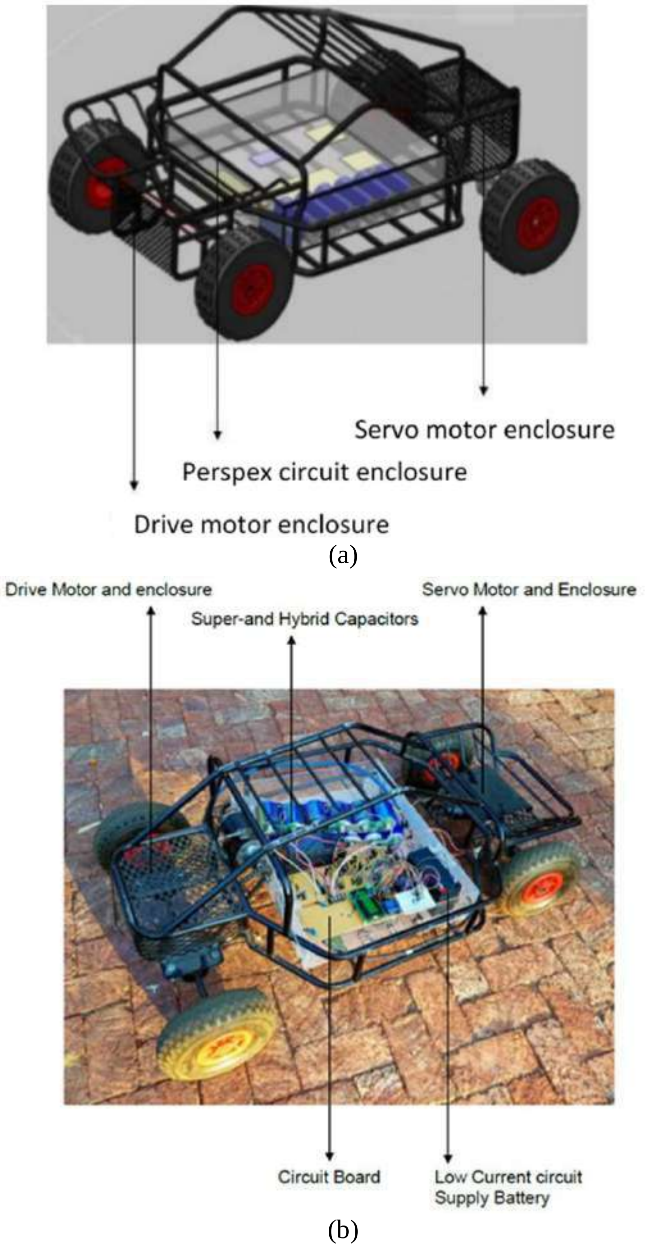


Fig. 2. Final circuit design.

The red blocks represent the sections where the capacitors were in use and the sections in between represent the use of the battery as primary source of power. The system was set to switch to the capacitor bank when the motor would draw current above 7.5 A. This was chosen after evaluating the results obtained from the individual use of power sources on an incline. After the measurements were taken the runtime of the vehicle with the individual power sources and the combination of the sources was documented. The battery lasted 2134 seconds, the SCs lasted 473 seconds, the HCs lasted 100 seconds and finally the combination of all three sources lasted 2849 seconds.


Servo motor enclosure
Perspex circuit enclosure
Drive motor enclosure
Drive Motor and enclosure
Servo Motor and Enclosure
Super-and Hybrid Capacitors
Circuit Board
Low Current circuit Supply Battery

(a)

(b)

Fig. 3. (a) Final vehicle design and (b) final constructed vehicle.

(a)

(b)

Fig. 4. (a) Supercapacitor bank and (b) hybrid capacitor bank.

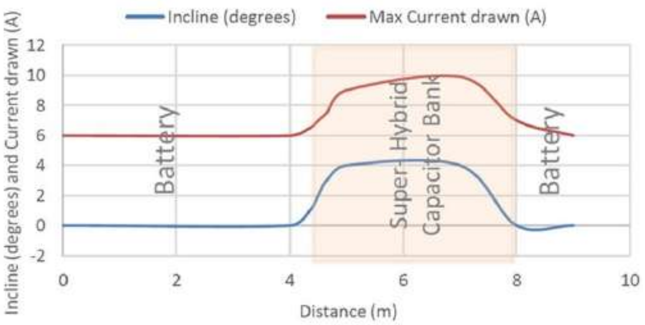


Fig. 5. Test path for vehicle.

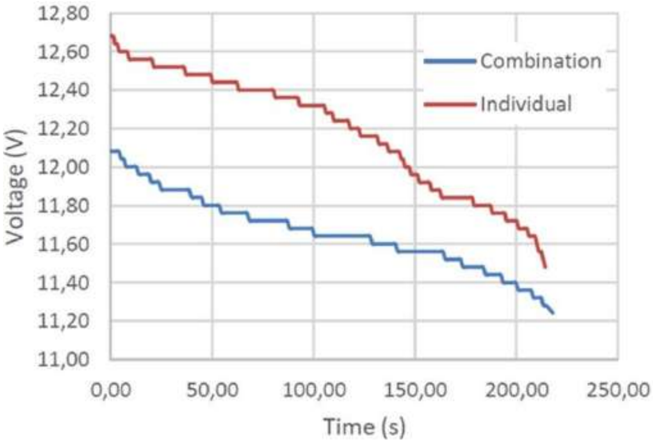


Fig. 6. Battery discharge with and without capacitor bank.

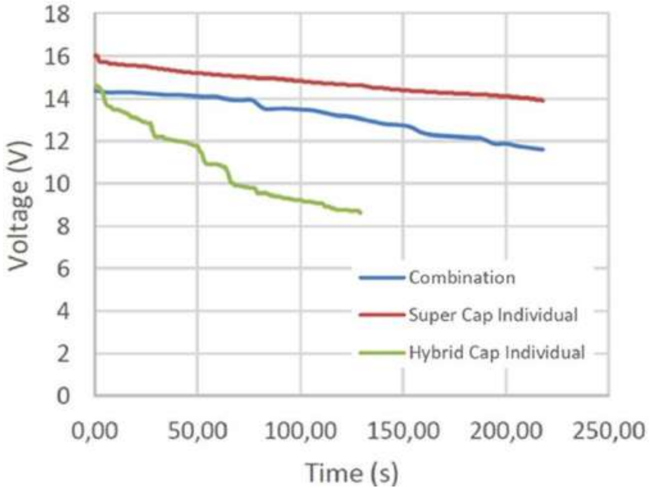


Fig. 7. Supercapacitor, hybrid capacitor an capacitor combination discharge on test path.

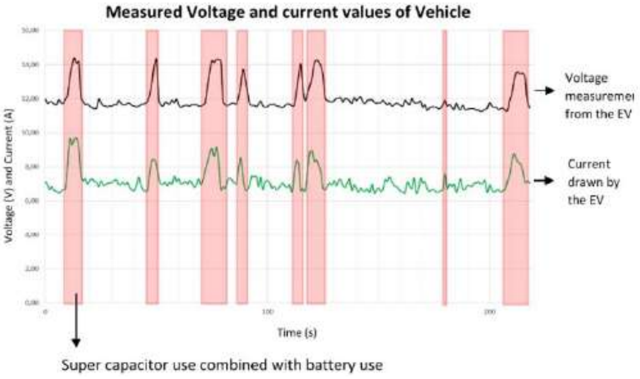


Fig. 8. Battery, supercapacitor and hybrid capacitor combination performance on test path.

Summarily, it was seen that the parallel combination of the capacitors in conjunction with the battery did result in a more efficient system. However, the benefit of the use of HCs was silenced due to the switching system not switching between the capacitor banks. It is therefore suggested to implement the switching between the SCs, the HCs, and the battery. For this research, the motor on the vehicle was replaced with a more efficient version, but the motor used could be further improved if it were to have a larger torque ability which would allow for the traverse of a steeper incline as well as giving the system more power to allow for a more inclusive testing environment. As a Bluetooth module had already been implemented in the circuit to increase ease of use the next step would/could be to implement a Neural Network onto the vehicle. This would further optimize the system and reduce complications due to human error.

IV. Conclusion

The super- hybrid capacitor EV was initiated to research the combined use of SCs and HCs along with batteries to reduce the fuel consumption and increase the efficiency of mining vehicles. The mining vehicles currently in use consume extreme amounts of fuel and emit equally extreme amounts of hazardous fumes into the environment. This vehicle would allow the eradication of the need for fossil fuels on the mines and create a healthier environment for its workers. The proposed solution, a vehicle that combined the use of a battery with SCs and HCs, was evaluated. After these elements were examined the design of the solution was simulated, built, and tested. It was found that the vehicle required a battery as the capacitors could not supply a constant voltage over an extended period. However, the battery lasted much longer with the use of the capacitors, as they supplemented the power source when the current drawn exceeded a certain current. The capacitors were used in this way as they provided a large current (and therefore power) burst, when required. This knowledge allowed for the optimal use of the capacitors and the maximum amount of operational time. Further research and testing would focus on the switching method between the sources, the construction of the vehicle used and the use of an intelligent controller, which could extend the endurance further.

References

[1] F. Berthold, A. Ravey, B. Blunier, D. Bouquain, S. Williamson, and A. Miraoui, “Design and Development of a Smart Control Strategy for Plug-In Hybrid Vehicles Including Vehicle-to-Home Functionality,” *IEEE Trans. Transp. Electrif.*, vol. 1, no. 2, pp. 168–177, 2015, doi: 10.1109/TTE.2015.2426508.

[2] A. P. Cook and P. J. D. Lloyd, “The estimation of greenhouse gas emissions from South African surface and abandoned coal mines,” *J. South. African Inst. Min. Metall.*, vol. 112, no. 12, pp. 1087–1090, 2012.

[3] M. J. Mabiza, C. Mbohwa, and M. Mutingi, “Life cycle inventory analysis and equivalent carbon dioxide emissions calculation of the mining and ore concentration processes of PGM at the Anglo American Platinum Ltd, South Africa,” in *2014 IEEE International Conference on Industrial Engineering and Engineering Management*, Dec. 2014, pp. 1018–1022, doi: 10.1109/IEEM.2014.7058792.

[4] W. A. Bisschoff, O. Dobshanskyi, and R. Gouws, “Integration of battery and super-capacitor banks into a single-power system for a hybrid electric vehicle,” in *2016 II International Young Scientists Forum on Applied Physics and Engineering (YSF)*, 2016, pp. 10–13, doi: 10.1109/YSF.2016.7753749.

[5] T. H. Kim, B. K. Lee, H. S. Song, S. M. Shin, and B. H. Lee, “HILS-based analysis of characteristics and performance of internal combustion engine vehicles with varying battery types,” *2016 IEEE Transp. Electrif. Conf. Expo, Asia-Pacific, ITEC Asia-*

Pacific 2016, pp. 326–331, 2016, doi: 10.1109/ITEC-AP.2016.7512972.

[6] J. O'Brien, "Draft Environmental Management Programme Report for the Klipfontein Concentrator Decommissioning Project Anglo American Platinum Limited : Rustenburg Platinum Mines - Rustenburg Section," Rustenburg, 2014.

[7] I. T. Vadium, R. Das, Y. Wang, G. Putrus, and R. Kotter, "Electric vehicle Carbon footprint reduction via intelligent charging strategies," in *2019 8th International Conference on Modern Power Systems (MPS)*, May 2019, pp. 1–6, doi: 10.1109/MPS.2019.8759783.

[8] W. Jacobs, M. R. Hodkiewicz, and T. Bräunl, "A Cost-Benefit Analysis of Electric Loaders to Reduce Diesel Emissions in Underground Hard Rock Mines," *IEEE Trans. Ind. Appl.*, vol. 51, no. 3, pp. 2565–2573, 2015, doi: 10.1109/TIA.2014.2372046.

[9] E. Vinot, R. Trigui, Y. Cheng, C. Espanet, A. Bouscayrol, and V. Reinbold, "Improvement of an EVT-based HEV using dynamic programming," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 40–50, 2014, doi: 10.1109/TVT.2013.2271646.

[10] V. Sreedhar, "Plug-In Hybrid Electric Vehicles with Full Performance," in *2006 IEEE Conference on Electric and Hybrid Vehicles*, Dec. 2006, pp. 1–2, doi: 10.1109/ICEHV.2006.352291.

[11] I. N. Jiya, N. Gurusinghe, and R. Gouws, "Combination of LiCs and EDLCs with Batteries: A New Paradigm of Hybrid Energy Storage for Application in EVs," *World Electr. Veh. J.*, vol. 9, no. 4, p. 47, Nov. 2018, doi: 10.3390/wevj9040047.

[12] F. Ognissanto, T. Landen, A. Stevens, M. Emre, and D. Naberezhnykh, "Evaluation of the CO2 emissions pathway from hydrogen production to fuel cell car utilisation," *IET Intell. Transp. Syst.*, vol. 11, no. 7, pp. 360–367, 2017, doi: 10.1049/iet-its.2016.0210.

[13] B. Lee, S. Kwon, P. Park, and K. Kim, "Active power management system for an unmanned aerial vehicle powered by solar cells, a fuel cell, and batteries," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 4, pp. 3167–3177, 2014, doi: 10.1109/TAES.2014.130468.

[14] N. S. Mubenga and T. Stuart, "A case study on the hybridization of an electric vehicle into a fuel cell hybrid vehicle and the development of a solar powered hydrogen generating station," *IEEE Power Energy Soc. Gen. Meet.*, pp. 1–8, 2011, doi: 10.1109/PES.2011.6039198.

[15] K. Okura, "Development and future issues of high voltage systems for FCV," *Proc. IEEE*, vol. 95, no. 4, pp. 790–795, 2007, doi: 10.1109/JPROC.2006.890111.

[16] L. C. Iwan and R. F. Stengel, "The application of neural networks to fuel processors for fuel-cell vehicles," *IEEE Trans. Veh. Technol.*, vol. 50, no. 1, pp. 125–143, 2001, doi: 10.1109/25.917898.

[17] G.-E. Oscar, L. Teresa, and N.-A. Emilio, "Fuel cells: A real option for unmanned aerial vehicles propulsion," *Sci. World J.*, vol. 2014, pp. 1–12, 2014, doi: http://dx.doi.org/10.1155/2014/497642.

[18] B. Frieske, M. Kloetzke, and F. Mauser, "Trends in vehicle concept and key technology development for hybrid and battery electric vehicles," *World Electr. Veh. J.*, vol. 6, no. 1, pp. 9–20, 2013, doi: 10.3390/wevj6010009.

[19] A. Emadi, K. Rajashekara, S. Williamson, and S. M. Lukic, "Topological Overview of Hybrid Electric and Fuel Cell Vehicular Power System Architectures and Configurations," *Veh. Technol. IEEE Trans.*, vol. 54, pp. 763–770, Jun. 2005, doi: 10.1109/TVT.2005.847445.

[20] B. Chen, X. Li, and S. Evangelou, "Comparative Study of Hybrid Powertrain Architectures from a Fuel Economy Perspective," in *14th International Symposium on Advanced Vehicle Control, AVEC'18*, May 2018, pp. 1–6.

[21] P. Thounthong, V. Chunkag, P. Sethakul, B. Davat, and M. Hinaje, "Comparative study of fuel-cell vehicle hybridization with battery or supercapacitor storage device," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 3892–3904, 2009, doi: 10.1109/TVT.2009.2028571.

[22] R. Xiong, Y. Zhang, J. Wang, H. He, S. Peng, and M. Pecht, "Lithium-Ion Battery Health Prognosis Based on a Real Battery Management System Used in Electric Vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4110–4121, 2019, doi: 10.1109/TVT.2018.2864688.

[23] R. Ghorbani, E. Bibeau, and S. Filizadeh, "On conversion of hybrid electric vehicles to plug-in," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 2016–2020, 2010, doi: 10.1109/TVT.2010.2041563.

[24] X. Gong, R. Xiong, and C. C. Mi, "Study of the Characteristics of Battery Packs in Electric Vehicles with Parallel-Connected Lithium-Ion Battery Cells," *IEEE Trans. Ind. Appl.*, vol. 51, no. 2, pp. 1872–1879, 2015, doi: 10.1109/TIA.2014.2345951.

A Hybrid Framework for Securing Data Transmission in Internet of Things (IoTs) Environment using Blockchain Approach

1st Mohammed Hayman Salih Mohammed
Department of Computer Control and Management Engineering
Sapienza University of Rome
Rome, Italy
hayman.salih@uniroma1.it

***Abstract*—Internet of Things (IoT) is a relatively new concept in computer science that connects devices with constrained resources to insecure internet through different technologies. IoT's fundamental components, including the wireless sensor networks and the internet, have an unsecured foundation that leads to DoS attacks, namely, sinkhole, blackhole, and grey hole attacks. To maintain the integrity and security of the IoT networks, many researchers implemented distributed ledgers in IoT environments. In this paper, we designed a hybrid framework between blockchain and IoT devices; the aim is to secure data transmission in IoT networks using blockchain technology to defend against DoS attacks. Our framework is called HFSDT-IoT. The method proposed in this paper consists of two phases to maintain security in the IoT. In the first phase, both a list of attackers and a safe list based on the Ethereum Proof-of-Stake (PoS) protocol is used, which complemented by utilizing the proposed IDSs Intrusion Detection System to discover malicious things. In the second phase, to decrease the obstacles in facilitating communication between blockchains, the inter-blockchain communication model creates a network of multiple blockchains with secp256k1 encryption used for heterogeneous blockchains. The experimental results of simulated scenarios show the HFSDT-IoT strategy can achieve better results when DoS attacks were launched compared to other blockchain-based methods, namely Bubble of Trust and Credibility Verification Method.**

***Index Terms*—Blockchain, Security attacks, Proof-of-Stake, Ethereum, Internet of Things (IoTs), HFSDT-IoT, Intrusion Detection System IDS**

I. Introduction

INTERNET OF THINGS (IoTs) is a network infrastructure that connects multiple computers, machines, and software services. IoT enhances human life and represents an essential part of today's urban lifestyle, packed with mobile devices and technologies [1]. Since IoT systems are having resource-constrained and ad-hoc in design, cyber-attackers may infiltrate them [2].This emphasises the need for research and development to design a safe and stable IoT system to tackle the problems and prevent attacks. Several attacks, including DoS attacks, are entering the machine illegally [3]. Since the attack affects IoT, it is a laborious task to eliminate the threat and get the device back online. Arraising the DDoS threats in the IoT environment makes the traditional information security measures, such as encryption or detection of intrusion, not enough for the system protection [4]. To elaborate, encryption and intrusion detection do not consider the sensor and actuator measurements or its compatibility factor with the IoT's physical process and control mechanism, and these process and control mechanisms are substantial to the protection scheme. Previously, IoT systems' problem was the mere attempt to eliminate a single type of attack and were only resistant to protect themselves against such an attack. If the system were subject to a combined attack, it would be practically inoperable, and the act of the intrusion operation would fail the IoT system quickly. Due to the vast network and sensing data produced by IoT devices and systems, cryptographic authentication is highly effective in continuous monitoring and analysis for IoT systems' security. The blockchain technology was proposed by Nakamoto Satoshi in 2009 [5] [6].

To maintain integrity and provide a more robust security propensity to the IoT environment, we designed a hybrid framework using a blockchain-based approach. The designed security framework uses ethereum smart contracts alongside the Intrusion Detection System(IDS). The designed framework contains two phrases, the first of which will establish a list of safe systems and attacker systems, while the second phase relies on blockchain network encryption. In the first phase, we designed a discovery and detection system using the PoS protocol of Ethereum alongside IDSs. The system divides the areas into different regions, and each region has its IDS. The IDS nodes are responsible for categorizing all the things or nodes into specified lists of attack and safe and updating these lists into a smart contract. The second phase of the system aims to decrease the obstacles in facilitating blockchain connections by employing an inter-blockchain connection model for heterogeneous blockchains to create a network of multiple blockchains with secp256k1 encryption. In this model, the blockchain system can communicate with other blockchain systems, and once the two systems are connected, the data and the messages get shared.

For method evaluation, a list of attackers, along with a safelist, is used to evaluate the process of discovering new

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

attacks and internal protection systems that have been deployed. The result has been compared with the other IoT-blockchain based systems, namely Bubbles of Trust proposed in “A decentralized blockchain-based authentication system for iot” [7] Furthermore, the Credibility Verification system proposed in “Blockchain based credibility verification method for IoT entities” [8]. The simulation was carried out using an NS2 environment with the same simulation scenarios. The results showed that our proposed method could detect 93 percent of the attacked nodes while the bubble of trust method can detect only 85 percent, and the credibility verification method is 78 percent.

The paper presented here is organized as follows: Section 2 introduces some relevant terms regarding application scenarios, security attacks, and detection schemes, section 3 discusses the proposed HFSDT-IoT security strategy and how it works under such circumstances. Section 4 focuses on performance evaluation parameters, investigations, and simulation of the results, while section 5 concludes the paper.

II. Relevant Terms

This section provides an overview of this research work’s fundamental parameters: application scenarios, security threats targeting IoT, and detection schemes to protect IoT.

A. Security Attacks

As IoT systems rely on wireless networks for connectivity, they are vulnerable to active and passive security threats and functional decomposition. Figures 1 and 2 show a vulnerability threat to IoT. The following vulnerabilities are of concern in this paper:

Sinkhole Attack: One of the primary attacks threatening the user agent server (UAS) is a type of attack known as the Sinkhole (SH) attack. In this attack, a malicious node broadcasts illusory information to the IoT about the routing path to impose itself as a legitimate route towards specific nodes for incoming connections and thus attract all incoming and outgoing traffic data. The Sinkhole attack’s objective is to draw all the traffic on the network towards the sinkhole node and alter the data packets or silently drop them altogether. Sinkhole attacks can increase the network overhead and consumption of energy while decreasing the functionality and lifetime of the network before ultimately annihilating the network system all together [9].

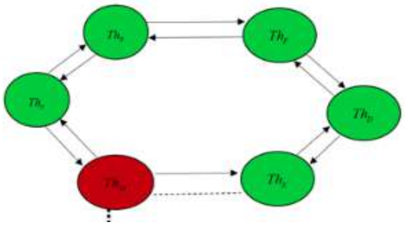


Fig. 1. Simple gray hole attack.

Black Hole Attack: This type of attack acquires when a Routing Request (RREQ) packet is sent, and as a response, the black hole node transmits a forged routing response (RREP) with a claim that it is a short and unexpired path even when the destination entry from the routing table is absent. When the constructed RREP packet enters the source node, it is possible to delete all valid RREP messages sent from other mediums and destination nodes from this mishap intermediate node. The black hole node effectively attracts the traffic to the target destination by misleading a source node, and instead of sending inbound RREP messages, all the data packets are dropped in the black hole node to prevent detection. The black hole node resets the hop count to a low value when a propagation path is forged, and the number of destination sequences to a high value maximises its probabilities to appear at the source node. The black hole attack also appears from the source node by forging fields-source sequence numbers in RREQ packets and hop counts, resulting in the poisoning of routing tables in intermediate and destination nodes [9].

Simple gray hole attack: In this kind of gray hole attack, a malicious node foists itself as a medium node that belongs on the shortest route to the final destination. Regardless of the routing table, the gray hole node is always accessible to reply route requests and it may receive data packets but will drop the packets instead of forwarding them as directed by the RREQ. In the flooding-based protocols, before sending a reply by healthy nodes, the gray hole node sends a reply to the requesting node. In this way, the selected route will contain a malicious node which drops the packets or send them to incorrect nodes. The process of a gray hole attack is shown in Fig. 2. As seen in this figure, the data packets should be transferred from a source $Th_S$to a destination Th_D. For this purpose, a proper route from the source to the destination should be detected. So, if Th_M is malicious, it fakes itself as a node present in the shortest path to the destination. Then, it will respond to the request by sending a reply to the Th_S sooner than other nodes. This way, the Th_S will send the data packets to Th_M and discard the replies received from other nodes [9].

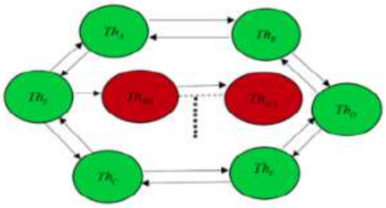


Fig. 2. Cooperative gray hole attack.

Cooperative gray hole attack: another way to implement a gray hole attack is known as cooperative gray hole attack in which multiple malicious nodes cooperate together in the attack to violate the routing protocol and security system. As depicted in Fig. 3, when malicious Th_{M1} and Th_{M2} act together, Th_{M1} refers to Th_{M2} as its next hop. Following the scenario of the source $things_S$ sends a Further Request packet $FReq-packet$ to Th_{M2} through another route other than Th_{M1} (e.g.$Th_S - Th_C - Th_E - Th_{M2}$). The $things$ asks Th_{M2} if it is the next hop ofTh_{M1} and if it has a valid route to the destination Th_D . Since the Th_{M2} is cooperating with Th_{M1}, following Reply $FReq-packet$ will be positive. Consequently, the source Th_s assumes that the route i$Th_S - Th_{M1} - Th_{M2}$s is secure, and starts sending the data packets. Once intercepted, the packets will be dropped by Th_{M1} . [9]

B. Detection schemes

Several studies developed and employed various security metrics to detect security attacks and secure IoTs from misbehaviour attacks. This is not a recent problem and has been previously studied extensively. Researchers currently propose several techniques to deal with misbehaviour attacks. Below, we focus on some recent studies and research on blockchain and IoT networks to enhance security. In the proposed framework in [5], blockchain stores all IoT devices communications as transactions, resulting in a more straightforward and more robust chain of custody process. Blockchain ensures the integrity of the data analyzed and strengthens the security while preservation is more reliable through a decentralized method. Furthermore, the forensic investigation participants (e.g., the device users, the manufacturers, the investigators, and service providers) can transparently confirm the investigation process as the ledger is publicly distributed. Paper [7] proposed a decentralized system named bubbles of trust using Etherum blockchain for IoT networks. The approach creates virtual secure zones to authorize the devices to be a part of the network, and they communicate securely. The system's security parameters are designed in a secure way where data integrity and availability considered for different scenarios. Besides, the authentication scheme in the system and security properties are resilient toward incoming attacks. Paper [10] invented a framework for IoT networks using a blockchain approach where the new framework improved the security and self-defence capability against cyber attacks. Moreover, the study provides an intensive discussion to enhance the power grid's security by using meters as nodes in a distributed network. The authors also claim that their proposed method can be considered a promising solution for data security in modern power systems. In the research [11], due to the limited security prosperities of IoT devices, the authors used the blockchain Ethereum based algorithm to enhance security. They implemented their idea using the whitelisting application alongside with consensus protocol of Ethereum. In this paper [12], the design of a novel architecture is proposed by combining these technologies introducing new opportunities for flexible and efficient DDoS mitigation solutions across multiple domains. The main advantages are the deployment of an already existing public and distributed infrastructure to advertise white or blacklisted IP addresses and the usage of such infrastructure as an additional security mechanism to existing DDoS defence systems, without the need to build specialized registries or other distribution mechanisms, which enables the enforcement of rules across multiple domains. In [13], the authors investigated the issues related to data integrity in wireless networks and the challenges to design a secure system. In their study, a new method has been proposed using the filtering techniques in public wireless networks. Before starting the transmission process, the safety devices register the IP and MAC address in the router for security and a safe network design. This way, it protects the system's information, and unsecured devices can not be a part of the communication process. Software-Defined Network (SDN) as a new centralized network controller has been recently applied using Openflow protocol. In the study [14], the authors investigated the SDN controller and the policy of filtering the MAC and IP address to give stronger security proprieties to the system. Besides, they proposed the use of a firewall for the SDN networks for MAC filtering. Their study checked the performance in terms of packet delay for MAC and IP filtering technique. In the paper [15], the authors proposed a discovery tool to detect spoofed Mac attacks. These kinds of attacks appear in the system when the address resolution protocol ARP pairs false IP/MAC address. Moreover, the proposed discovery tool the authentication to the new pairs confirms only when the proposed method checks all possible spoofing causes. This way, the checking process guarantees security before paring. In the research [16], the authors investigated the security issues available in IoT networks. In order to enhance the security parameters and give more robust prosperity to IoT networks, SDN Software Defined Network) Edge Computing has been used. Using this combination is to benefit from the SDN filtering mechanism and enable the IoT networks for better security and access control proprieties. To enhance the security in blockchain nodes, the authors in [17] used SDN to filter network traffic. For that, a firewall and ChainGuard utility for blockchain applications has been used. This change guarantees the packets from illegitimate sources cannot affect the blockchain. ChainGuard provides access control functionality and can effectively mitigate flooding attacks from several sources at once.

III. Proposed hybrid framework

In the following section, we design a DoS-security threats-immune schema by employing the blockchain approach. The HFSDT-IoT consists of three sections, such as the overview of the proposed framework is discussed in Sect, 3.1, architecture of the proposed hybrid framework is discussed in Sect, 3.2, and encrypted connection between different blockchains is discussed in Sect, 3.3.

A. Overview of proposed HFSDT-IoT framework

The internet of things (IoT) has a big advantage in various fields and industries which is that its growth and adaptation speed is very high. Very soon billions of connectable devices will spread through smart homes and cities, collect data, transmit them to huge databases for analysis and processing, and run the commands sent from smart applications and learning-device-based systems. These technologies have been proven for connections among generic computer devices for years and already answer the needs of small IoT ecosystems like smart homes. However, with the growth of IoT, centralized networks will soon become the system bottleneck and because of the network traffic overload and latency and sudden interruptions in sensitive transactions, they need hubs and communication hardware suitable to these situations for more investment. A solution is decentralized IoT networks for improving the speed and connectivity. In many cases, replacing the internet connection with local inter-device connections leads to increases speed and efficiency. Fortunately, the lack of decentralization has been solved in another popular technology: Bitcoin. The famous encrypted currency which is designed using the less known (but very exciting) blockchain technology. Blockchain is a data structure which allows for the creation and maintenance of a ledger of transactions which is shared among the distributed network nodes. Blockchain through encryption allows the participants to skillfully manage the ledger without having a central reference. When blockchain gets adapted for IoT, it will use the same mechanism used in Bitcoin transactions for creation of an unchangeable record of the smart devices and the transactions between them. Doing so enables independent smart devices to connect directly and check the transaction credit without the need for a centralized force. The devices are registered in the blockchain once they enter IoT networks and can process the transactions after that. The transactions are signed digitally using public key encryption which uses two keys including the public key and the private key. These two keys are mathematically related to each other. Considering the utilized mathematical complexity, it is almost impossible for these keys to be guessed and this makes hacking the transactions almost impractical. The public key is used to sign and encrypt a message for sending and the specific receiver can decrypt the message using its private key. In fact, it can be stated more simply that blockchain is a distributed database which is identified as a distributed ledger. In this technology, the network is able to preserve the information integrity of the database content despite the existence of several users which record and correct data simultaneously and the possibility for the data to interfere with each other. Considering the encrypted data structure that the blockchain has, integrity is preserved without any central controller. Although the popularity of blockchain is rising by the day, the increase in the number of cyber-attacks might damage this process. Since perpetrators seek to attack and perform malicious operations in all networks, due to the exchange of data between the things these perpetrators also exist in the internet of things in the present world. One cyber-attack is the DoS attack which causes huge computing resources to be wasted by sending lots of fake requests. In these attacks where one or two main computers get attacked, the service becomes unavailable by sending fake requests and extensive attacks. Performing the DoS attack on a blockchain-based network is difficult as the attacker need to reach consensus or agreement in more than half of the nodes. Therefore, using blockchain to increase the security in the internet of things is addressed. When blockchain adapts to IoT, it will use the same mechanism used in Bitcoin or Ethereum transactions for creating an immutable record of the smart devices and the transactions between them. Doing so enable the smart devices to connect directly and verify the validity of transactions without the need for a centralized force. The devices are registered in the blockchain once they enter IoT networks and can process the transactions after that. The process of a DoS attack is shown in Figure 3

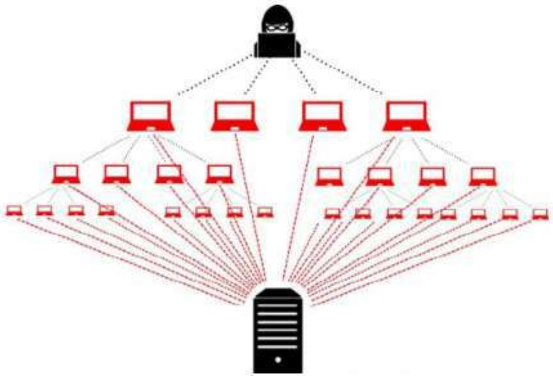


Fig. 3. Cooperative gray hole attack.

In the following section, we study a method based on blockchain for detecting and preventing DoS attacks in the internet of things. In the proposed method, a list of attackers along with a safe list for discovering attackers based on the smart contracts in the PoS protocol of Ethereum which is completed by utilizing IDSs. Also, in order to decrease the obstacles in facilitating blockchain connections, an inter-blockchain connection model is presented for heterogeneous blockchains by creating a network of multiple blockchains with secp256k1 encryption. In this model, the blockchain system is able to communicate with other blockchain systems. After two systems are connected, the data and the messages get shared.

B. Phase 1: Architecture of the proposed hybrid framework

In the proposed method, the PoS protocol of Ethereum is used. Ethereum is based on a smart contract where the smart contracts are applications that run exactly as specified by the creators and Ethereum allows the developers to create various smart contract applications. The architecture of the proposed system is presented in Figure 4. The web server might undergo a DoS attack by the hosts in several domains. These attacks are identified in the proposed method using the Intrusion detection system IDS and the considered list of attackers which are placed in the smart contract. The users in the proposed system first need to create a safe list which is of the smart-contract-based type. However, once the attackers affect the web server, the under-attack users receive the IP address of the attackers from the IDS and save them in the list of attackers. In the Ethereum blockchain, once every 14 seconds a new block is created. Therefore, the common domain of the updated lists receives the addresses which must be blocked and verify the attack authenticity by analyzing the traffic statistics and validity of the target address. Each IDS node covers the region in the network that it watches. Different number of IDSs affects the region being covered and also the attack identification area.

There are three hypotheses regarding this matter:
• The IDS node is placed in an area where it can put the BLOCK messages in the list of attackers in order to identify the DoS attack.
• The authentication mechanism has set of rules applied by IDS in order to detect or delete malicious things.
• Each intrusion detection system (IDS) must check its own area.

As soon as other domains receive the list of attackers from the IDS, retrieved it, and verified the attack, various reduction strategies can be employed based on the existing policies and mechanisms in the area. Also, it can block the malicious traffic near its source.

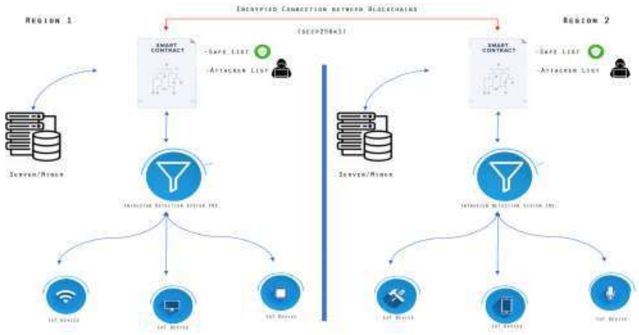


Fig. 4. HFSDS-IoT system model

The architecture and system model of the proposed system shown in figure 5. The system consists of the following three components:
Users : they might update the IP addresses located in the safe list or the list of attackers in the Ethereum blockchain.
IDSs: They propagate the IP addresses located in the list of attackers, retrieve the lists containing the propagated IP addresses, and execute the DoS reduction mechanisms. In the proposed method, the things participating in the blockchain network have IDS units for monitoring and evaluation that perform one or both operations at the same time (which usually perform both simultaneously). Furthermore, the raw warning data generated by the IDSs is saved in a blockchain through a series of transactions which are repeated among the things participating in the network.
Smart contract: This is the regulatory body of the system which is responsible for authorizing the devices and ensuring that they do not operate beyond their gas limit.
Server/Miner: The IoT-Ethereum network includes multiple volunteers acting as the servers/miners. These entities are responsible for verifying the transactions and data exchange through smart contracts using high computational and processing capabilities.

1) Detection of the IP address of the malicious thing by the IDS: Detection of a malicious thing is carried out based on the requested transactions of a thing. The transactions in Ethereum are includes the following important items:
Signature: signature that authenticates the sender and specifies their intention for sending the message for the receiver through the blockchain.
Value: specifies the value intended to be sent by the sender to the receiver.An optional option for information; which can contain a message that is sent by convention.
Gas limit; specifies the maximum number of computational steps that the transaction execution is allowed to pass.
Gas price; specifies the fee that the sender expects to be paid for each fuel unit.
The IDS in the proposed method checks each one of these items and in case these items do not match in a transaction, it introduces the transaction as malicious and adds it to the list of attackers in the smart contract. The communication among the things takes in two separate logical layers, namely the layer for alert exchange and the layer for consensus. In the layer for alert exchange, our proposed method carries out the process to propagate the warning data, which includes the malicious IP addresses. In specific, the IDS nodes are responsible for exchanging and gathering the data for the alerts data according to their role as units for monitoring or analysis. Moreover, the operation to evaluate the data is carried out using IDSs via running an agreement protocol and reaching a consensus on the transactions that have to be present in the ledger. What results in intensifying the extreme participants' responsiveness and guaranteeing the data integrity, is the connection existing between the two layers, the result of the layer for consensus, and the fundamental blockchain structural properties. The engaged things run a consensus protocol to guarantee the validity of the transactions before adding them to a block. This process, which will be introduced in the next phase, guarantees that only the alerts that are formed correctly and soundly are included in the intended blockchain, the alert data transactions are resistant to any manipulation, and each participating entity has a global perspective of the alerts. List of attackers: the general Ethereum blockchain is the executor of smart contracts that use logic to report the IP addresses of the attackers and the alert data obtained by the IDSs in the blockchain. The considered architecture has been built according to the following principles:

1) The DoS discovery and reduction countermeasures which have been presented by the attackers or the users.
2) The attack information reception report which is discovered by the IDSs. Also, the IDSs can encrypt the special alert data in two layers and also the distributed ledger and the keys will only be available to the participating things. This way, the things or other devices which are not a member of this collaborative group and do not have the specific private key, cannot see the alert information exchanged and processed among the

verified participants. Therefore, the peers can cooperate in multiple groups without revealing any confidential alert information to external things while remaining in the same network.

3) In order to prevent the attack, the blockchain DoS reduction modules are run on the users and report their IP addresses and listen to the blockchain.
4) The IDSs report the malicious addresses to the list of attackers by proving the IP address ownership of the users.
5) Different areas, different security policies, and also different fundamental management systems are run. Immediately after being notified of a DoS attack where a customer verifies its validity, some countermeasures get defined based on the security policies of the domain and the existing actions.
6) Consensus is controlled by a pre-selected set of things. In this case, the verifiers are known and the risk of several malicious peers connecting to the system simultaneously (DoS) and eliminating the validity by the IDSs is decreased. Furthermore, if a peer has malicious behavior, i.e. starts sending fake alerts, it will be discovered by the IDSs and added to the list of attackers. Then, the blockchain rules will be altered and the fake transactions will be undone which will be described in length in the next phase.
7) The next important item is related to the alert data and malicious IPs during the sharing process at each one of the two communication layers (alert exchange and blockchain consensus). For this purpose, our suggestion is to use a compressed representation, using bloom filters for example, for communications in the consensus layer in order to decrease the overload and the size of the blockchain. On the other hand, in the alert exchange layer, where the data exchange can potentially occur immediately after the request, raw data exchange presents the highest validity level to the system.
8) In the proposed method, regarding the alert information about the IP of the malicious things and confidential information (exchanges in the consensus layer), it is important to provide a mechanism that keeps the privacy of the participating sides. One of the solutions can be the encryption of alert data using symmetric-key encryption and creating keys be only for participants that have the right to read them. This allows each peer to remain in the same network but be able to decrypt and only check the alert data which they have access to. However, this creates overload in the form of key management and distribution. As mentioned before, in the proposed method, the keys are only available to participating things. This way, the peers which are not a member of this collaborative grout and do not have the specific key cannot see the alert information exchanged and processed among the verified participants.

Therefore, in the proposed method, at first each domain, the participating users or customers for example, need to create a safe list containing the IP address or an address range verified by a source. Then, the discussed safe list is registered as a register-based contract such that the participation can be easily tracked. Therefore, the corresponding safe list can be identified. Then, the input traffic in the customer and user area is analyzed by IDSs and malicious IPs are filtered using this monitoring tool. In the proposed method, the completed blockchain smart contract, which can be deployed as a complementary security features for any system in danger of malicious DoS attacks, uses the IP addresses placed in the list of attackers or the safe list. Since the connection might be among multiple blockchains, in the next phase of the proposed method, an inter-blockchain communication model for heterogeneous blockchains through the creation of a network of multiple blockchains is introduced to decrease the obstacles of facilitating blockchain communications. In this model, the blockchain system is able to establish connection with other blockchain systems. After the two systems are connected, the data and the messages get shared.

C. Phase 2: Encrypted connection between different blockchains

When a blockchain transaction from a user is transmitted to the blockchain system, the system concludes the received transactions and stores the results in the ledger. The agents transfer among the accounts existing in the system as well. Managing the transactions that need asset transfer among two different blockchains is not the same, however. First, the trajectory for delivering the transactions to the target blockchain system should be identified by the source system. Second, there have to be similar answers from the two chains participating in the transaction, following the completion of the cross-chain transaction. The method to move cross-chain assets presented here not only provides the account balance on two chains for having reliable transfers (i.e. resulting in an attack-free method) but also considers the account balance for transfers within the blockchain. In our proposed method, the information on the blockchain address is recorded in a standard format. During the cross-chain transactions execution process, once the decryption is concluded, to ensure the stability of both considered systems, three steps are employed. Guaranteed transfer restricts unreliable participants from making payments. In each blockchain, there exists a public guaranteed address that functions and a reliable mediator for payment within the chain. Based on the routing table available in the router blockchain, the transactions are then transferred. We define the transformation process between a local transaction and a standard cross-chain transaction as two functions. These functions consist of the package and the package removal functions. The responsibility of the package functions is to wrap up a local transaction into a standard cross-chain transaction while the package removal function is responsible for parsing a standard transaction to a local transaction. The assets can be transferred from a chain to another chain through transfers using the connection between blockchain systems. In the inter-blockchain connection model of the proposed method, the transaction needs to be a series of atomic, consistent, durable, and isolated operations. In order to ensure system durability, after a transaction is completed, it is recorded in the ledger and will be immune against system failures. For separation, the writing and reading operations in a transaction need to be serialized, meaning that the transactions are completely separate from each other. In order to guarantee transaction consistency, the cross-chain blockchain method with elliptic curve cryptography is proposed. In our proposed method, three steps are carried out which lead to an agreement between two chain sets. This way, the receiver can make the final decision on whether to perform the extra phase or disconnect. Specially, the ledger is also used in order to eliminate the need for a third party.

1) Performing inter-blockchain transactions based on the secp256k1 encryption: After discovering the malicious thing and removing it from the network which is done in the first phase using the IDSs where each malicious and safe IP becomes a part of the smart contract by being placed in the list of attackers or the safe list, the transmission of the assets is carried out using the elliptical curve cryptography (ECC), which is known as secp256k1 in Ethereum and through a crossover blockchain in order to send the assets safely and as a correct transaction from a blockchain to another. The security of the transactions based on secp256k1 is used in order to ensure that the sent transaction is safe. The Elliptical Curve Cryptography (ECC) is a public key encryption method which is designed based on an algebraic structure of elliptical curves on finite fields. The Elliptical Curve Cryptography (ECC) is used to ensure complete immunity against security risks like confidentiality, integrity, privacy, and authentication. An elliptical curve E is described as presented in Eq. (1).

$$Y^2 = \left(X^3 + AX + B\right) \tag{1}$$

The highest degree in this equation is 3. In order to perform higher levels of encryption and decryption, the standards proposed by NIST must be met. If the root of both sides is taken, the equation will become the as presented in Eq. (2).

$$Y = (\pm\sqrt{X^3 + AX + B}) \tag{2}$$

The elliptical curve cryptography algorithm used by Ethereum is called secp256k1. The secp256k1 equation is as presented in Eq. (3).

$$Y^2 = \left(X^3 + 7\right) \tag{3}$$

This cryptography needs a smaller key compared to other cryptography methods based on Galois fields.When a blockchain wants to transfer some transactions to the next blockchain, we use the secp256k1 algorithm in order to protect the information. Using this algorithm, the transactions get encrypted before the actual routing process. After encryption, the main data are protected against malicious and unauthorized nodes. The SHA-1 method is a hashing method which is used to verify the validity of the data. The SHA-1 algorithm is considered a safe algorithm for the hashing operation. In fact, Hash is a unique string where based on an input string data, an output hash is created which is unique. The transactions created in a blockchain are encrypted using the secp256k1 cryptography algorithm. The hash value used to encrypt the generated transactions is created by the SHA-1 hashing algorithm. This way, the network sends the encrypted transactions and the corresponding hash value through the next chosen blockchain and each blockchain has its own safe list and list of attackers in its Ethereum smart contract. The receiving block in the next blockchain calculated the hash value for the received encrypted packet. If the new hash value is equal to the received hash, the data were protected which demonstrates that the previous block was not an attacker. If the hash value is conflicting, the thing sends the IP of the sender block to its nearest IDS. Then, the IP address of the malicious node is added to the list of attackers found in the smart contract and the malicious node gets separated from the network. In case there are no malicious nodes, the packets get sent all the way to the destination this way and no damage occurs to the transmitted data because they are encrypted. The exchange process in the cross-blockchain in the proposed method is described briefly in Figure 5:

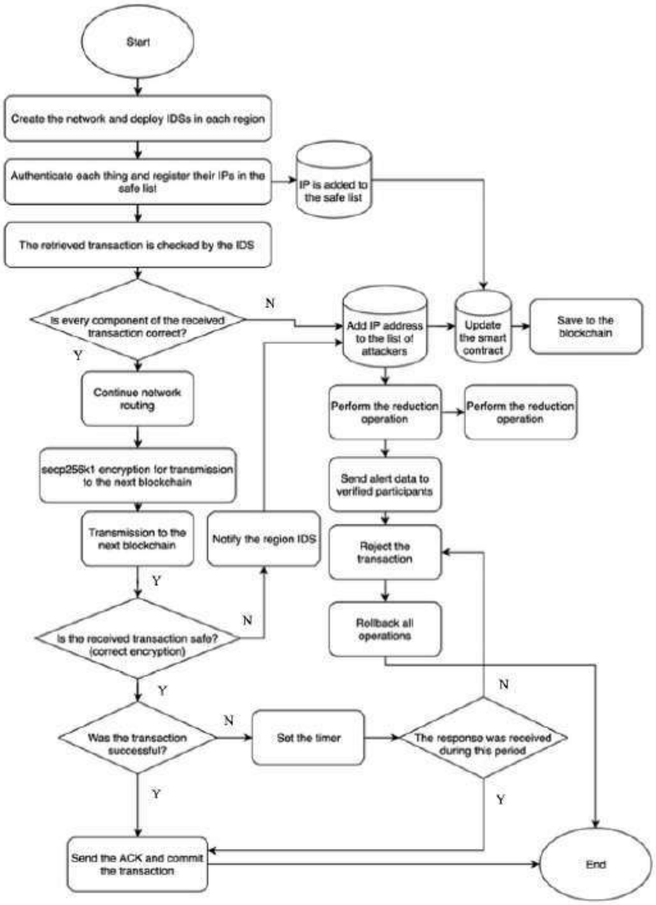

Start
Create the network and deploy IDSs in each region
Authenticate each thing and register their IPs in the safe list
IP is added to the safe list
The retrieved transaction is checked by the IDS
Is every component of the received transaction correct?
N
Y
Add IP address to the list of attackers
Update the smart contract
Save to the blockchain
Continue network routing
Perform the reduction operation
Perform the reduction operation
secp256k1 encryption for transmission to the next blockchain
Send alert data to verified participants
Transmission to the next blockchain
Notify the region IDS
Reject the transaction
Y
N
Is the received transaction safe? (correct encryption)
Rollback all operations
Y
N
Was the transaction successful?
N
Set the timer
The response was received during this period
Y
Y
Send the ACK and commit the transaction
End

Fig. 5. Flowchart of the HFSDT-IoT

Parameter	Value
Coverage area (m x m)	First scenario: 3000 x 3000 Second scenario: 4000 x 4000
Simulation tool	NS-2
MAC	IEEE 802.11
Transport	UDP/IPv6
Range of communication	220 m
Bandwidth	2 Mbps
Traffic type, rate	CBR, 25 packets/sec
Model of mobility	Random way point
RX and TX ratio	95%
Number of nodes, and Packet size	500, 256 Kbps
Number of connections, and Pause time	50, 100 sec
Maximum mobility (varying)	5 m/sec - 25 m/sec
Percentage of malicious nodes	0% - 30%
Simulation time (varying)	500-2000

TABLE I
SETTING OF SIMULATION PARAMETERS.

IV. Evaluating the Performance

In the following section, the performance of the HFSDT-IoT is evaluated to prevent DDoS attacks.

A. Performance metrics

The feasibility and efficiency of our proposed HFSDT-IoT solution were carefully tested in this section using extensive simulations. The results are compared with Bubble of Trust and Credibility Verification Method, approaches proposed in [7], and [8] respectively. The detection ratio, and packet delivery rate are evaluated.

Detection rate:Detection ratio is the division of the number of IoT things that have been misbehaved into the cumulative number of real things. This means that it is likely that all the security threats will be identified effectively. The formula is then shown in the below detection rate of Eq 4.

$$DR = \left(\frac{TPR}{TPR + FNR}\right) * 100 \qquad (4)$$

PDR is the percentage division of total data packets collected at the destination Thing to total data packets sent by the source Thing. Eq. 5 shows the average PDR collected for tests.

$$PDR = \frac{1}{n} * \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} * 100\% \qquad (5)$$

B. Simulation setup and comparing algorithms

The key benefit of simulation is that it simplifies research and protocol verification, particularly in large-scale systems. Our proposed solution's efficiency is evaluated using NS-2 as the simulation method in this section, and the findings are discussed further. It is worth noting that both HFSDT-IoT, Bubble of Trust, and Credibility Verification Method parameters and settings are treated the same.

C. Simulation results and Analysis

This section analyses the security performance of HFSDT-IoT under the two DoS attacks scenarios outlined in Table 1. The simulations carried out using 500 IoT things consistently expanded in the network area initially.

The detection verification rate has been examined for HFSDT-IoT compared to two other methods, namely, Bubble of Trust and Credability Verification Method. Figure 6. represents the summarized results for the number of things examined. The represented scenarios in Figure 6 (A, B and C) are the number of malicious things set as 10%, 20% and 30%, respectively. The detection performance for both compared methods is reduced, mainly when the number of attacks is high, as seen in the diagrams. This decrease is even greater in creditability verification comparing to the Bubble of Trust method. The detection performance in our proposed algorithm is much higher than the other algorithms, and the detection rate can reach up to 93.5 %.

The obtained results show that when the number of normal things and the rate of malicious things are equal to 500 and 20, respectively. When there are 30 malicious nodes in the IoT network using HFSDT-IoT, the detection rate decreases to 87.13%. Employing the same scenario for the Bubble of Trust and Credability Verification Method decreases the detection rate to 75% and 70%, respectively.

Figure 7 demonstrates the node's ratio of detection against DoS attack with differing misbehaving ratio. As can be seen, there is an inverse proportional relationship between the node's detection rate and the misbehaving ratio. With a misbehaving ratio of less than 0.1, the detection rate exceeds 94, while in the misbehaving ratios of greater than 0.1, the node's detection rate surpasses 93 compared to the other two methods.

The comparison results of the HFSDT-IoT proposed scheme, with the Bubble of Trust, and Credability Verification Method in terms of packet dilevery rate PDR at different percent of DoS attacks is provided in Figure 8.

V. Conclusion

With an expansion in IoT and frank implementation of these networks, a need for securing the communications among IoT nodes recognised as essential. A stable, intelligent agent-based strategy must be developed to meet the challenges, which achieves both immense safe mode and the efficiency of the desired networks. In this paper, a hybrid method consisting of two phases was proposed for maintaining security in the IoTs systems using blockchain technology. In the first phase, the Ethereum PoS protocol was used to discover DoS attacks. In this phase, the smart contracts found in the Ethereum protocol improved using safelists and the list of attackers that the IDSs discover. In the second phase, secp256k1 encryption with cross-blockchain transmission has been used to execute inter-blockchain transactions safely.

The performance HFSDT-IoT investigated using NS-2. The obtained results in the simulation demonstrated that the HFSDT-IoT was very robust against DoS attacks, and It presented that it has a high level of security. The detection rate is more than 93.50 %, and the packet delivery rate is more than 95.76 % compared to current approaches.

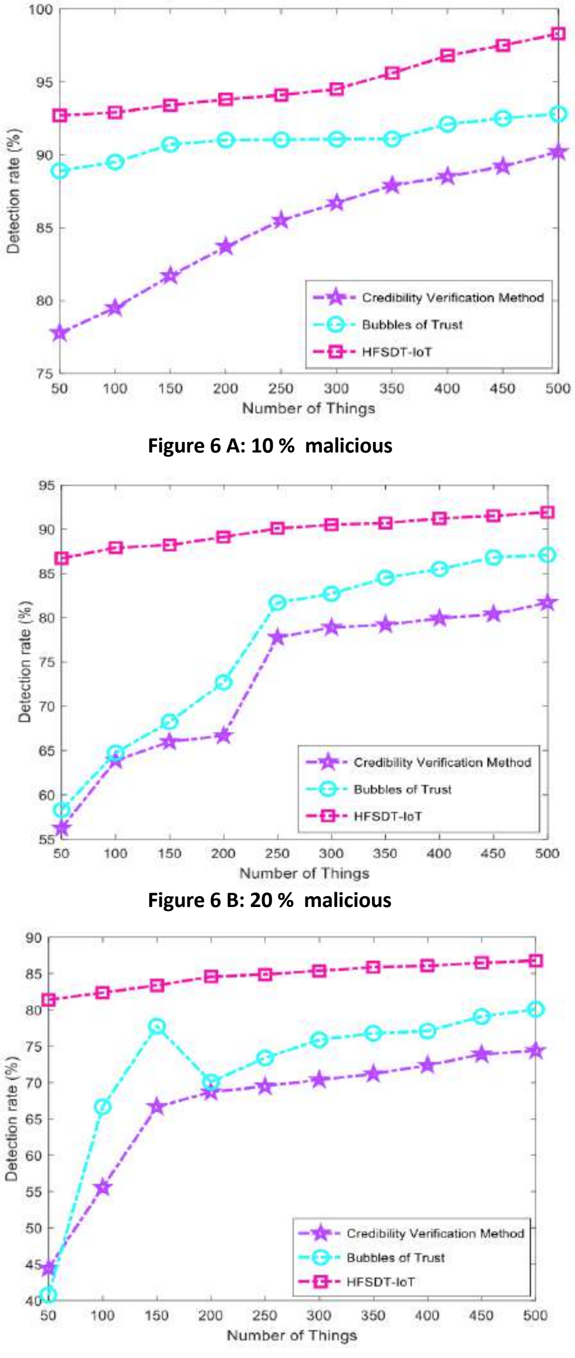


Figure 6 A: 10 % malicious

Figure 6 B: 20 % malicious

Figure 6 C: 30 % malicious

Fig. 6. Comparison of the HFSDT-IoT proposed scheme, Bubble of Trust, and Verification Method approaches in term of DR.

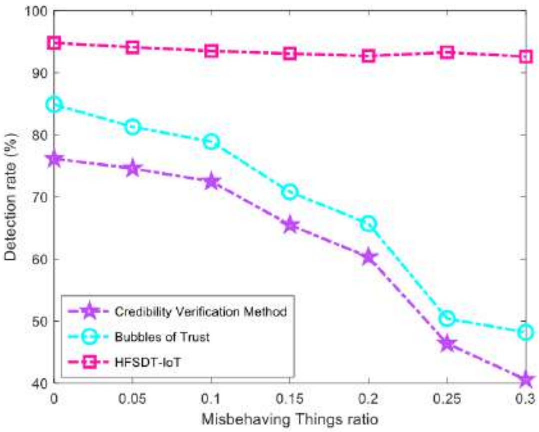


Fig. 7. Comparison of the HFSDT-IoT proposed scheme, Bubble of Trust, and Credability Verification Method approaches in term of detection rate DR. Number of Things vs Misbehaving ratio

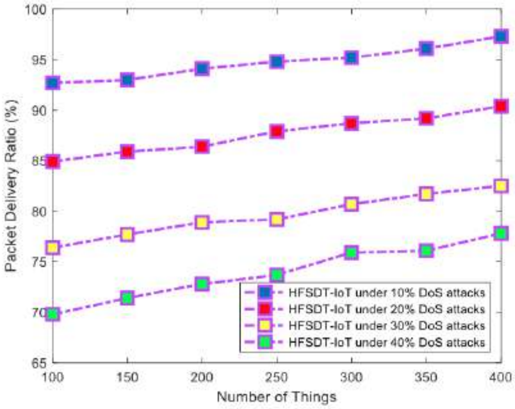


Fig. 8. Comparison of the HFSDT-IoT proposed scheme, in terms of PDR at different percent of DoS attacks

REFERENCES

[1] O. Vermesan and P. Friess, *Internet of things: converging technologies for smart environments and integrated ecosystems*. River publishers, 2013.

[2] Y. Qian, Y. Jiang, J. Chen, Y. Zhang, J. Song, M. Zhou, and M. Pustišek, “Towards decentralized iot security enhancement: A blockchain approach,” *Computers & Electrical Engineering*, vol. 72, pp. 266–273, 2018.

[3] F. C. Delicato, P. F. Pires, and T. Batista, “The resource management challenge in iot,” in *Resource management for Internet of Things*. Springer, 2017, pp. 7–18.

[4] D. Li, Z. Cai, L. Deng, X. Yao, and H. H. Wang, “Information security model of block chain based on intrusion sensing in the iot environment,” *Cluster Computing*, vol. 22, no. 1, pp. 451–468, 2019.

[5] J. H. Ryu, P. K. Sharma, J. H. Jo, and J. H. Park, “A blockchain-based decentralized efficient investigation framework for iot digital forensics,” *The Journal of Supercomputing*, pp. 1–16, 2019.

[6] S. Nakamoto, “Bitcoin white paper,” 2008.

[7] M. T. Hammi, B. Hammi, P. Bellot, and A. Serhrouchni, “Bubbles of trust: A decentralized blockchain-based authentication system for iot,” *Computers & Security*, vol. 78, pp. 126–142, 2018.

[8] C. Qu, M. Tao, J. Zhang, X. Hong, and R. Yuan, “Blockchain based credibility verification method for iot entities,” *Security and Communication Networks*, vol. 2018, 2018.

[9] D. Airehrour, J. Gutierrez, and S. K. Ray, "Secure routing for internet of things: A survey," *Journal of Network and Computer Applications*, vol. 66, pp. 198–213, 2016.

[10] G. Liang, S. R. Weller, F. Luo, J. Zhao, and Z. Y. Dong, "Distributed blockchain-based data protection framework for modern power systems against cyber attacks," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3162–3173, 2018.

[11] D. M. M. Mena and B. Yang, "Blockchain-based whitelisting for consumer iot devices and home networks." in *SIGITE*, 2018, pp. 7–12.

[12] B. Rodrigues, T. Bocek, A. Lareida, D. Hausheer, S. Rafati, and B. Stiller, "A blockchain-based architecture for collaborative ddos mitigation with smart contracts," in *IFIP International Conference on Autonomous Infrastructure, Management and Security*. Springer, Cham, 2017, pp. 16–29.

[13] R. D. Sari, A. Supiyandi, M. M. Siahaan, and R. B. Ginting, "A review of ip and mac address filtering in wireless network security," *Int. J. Sci. Res. Sci. Technol*, vol. 3, no. 6, pp. 470–473, 2017.

[14] P. Rengaraju, S. S. Kumar, and C.-H. Lung, "Investigation of security and qos on sdn firewall using mac filtering," in *2017 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2017, pp. 1–5.

[15] D. P. Mishra, P. P. Satapathy, and B. Mishra, "Designing a secure network interface by thwarting mac spoofing attacks," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. ACM, 2016, p. 102.

[16] C. Aggarwal and K. Srivastava, "Securing iot devices using sdn and edge computing," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, 2016, pp. 877–882.

[17] M. Steichen, S. Hommes, and R. State, "Chainguard—a firewall for blockchain applications using sdn with openflow," in *2017 Principles, Systems and Applications of IP Telecommunications (IPTComm)*. IEEE, 2017, pp. 1–8.

Understanding the Information Disseminated Using Twitter During the COVID-19 Pandemic

Jorge Torres[1], Vaibhav Anu[2], Aparna S. Varde[3]
Department of Computer Science
Montclair State University, Montclair, NJ, USA
torresj44@montclair.edu[1], anuv@montclair.edu[2], vardea@mail.montclair.edu[3]

***Abstract*—Twitter, with its ever-growing influence, has continued to serve as a means of spreading information and often providing early warnings to the situations that the world is encountering. The COVID-19 pandemic is no exception. With this disease resulting in hundreds of thousands of deaths, it is valuable that an analysis is conducted regarding the source of information posted on social media sites such as Twitter. In this study, we specifically analyze the source-URLs being posted by influential Twitter accounts. Our main goal in this study is to understand the kind of online materials, i.e., external weblinks that Twitter users prefer to promote/share about COVID-19.**

Keywords—Data analytics, Coronavirus, COVID-19, Pandemic, Social media, Twitter informatics

I. Introduction

With the rise of COVID-19 many people across the globe are looking to the Internet for information related to aspects such as symptoms, remedies, and precautions. As in many of the global disasters of the past, this pandemic came with early warnings. News of the virus was believed to have started late December 2019 and by late January 2020 it had already reached the United States and had become recognized as an emergency by the World Health Organization (WHO) [1][2]. It still had not become a pandemic until March 11, 2020 thus giving the world about 90 days of preparation. While governments across the globe were preparing their own approach to deal with the pandemic, their citizens were dependent on the global coverage of COVID-19 and the experiences from the pandemic's past victims to find more information.

One such approach used by the general public to discover COVID-19 information was through visiting the social media site Twitter. Despite its 280-character limit, Twitter has been used to spread information of all kinds ranging from what someone has eaten today to the personal opinions of those who have a large Twitter following. In addition, *Twitter also allows users to provide external links (URLs - Universal Resource Locators)* which connect to websites outside of Twitter. These links can be used to direct a Twitter user's followers to a website more relevant or useful to the topic at hand. In the case of the pandemic, these links were used to provide information about COVID-19 directly from government websites, webpages of medical organizations (e.g., Johns Hopkins), and informational websites maintained by other organizations/individuals across the globe. A pertinent question that needs to be asked however is: how much of that data is safe to use (i.e., whether the data is trustworthy or not). After all, when used incorrectly, the wrong information could result in disastrous and in the worst case fatal situations for many people across the globe [3].

To that end, the intention of this article is to investigate what kind of information has been spread through social media regarding COVID-19. Twitter is a popular resource for such a data informatics task due to its easy-to-use application programming interface (API) which allows for researchers to easily collect data from Twitter without endangering the users who frequent Twitter. Such data can then be grouped and analyzed depending on the researcher's intentions. While simply viewing tweets might not seem significant on its own, it is the sheer quantity and influence of tweets collected that results in analyzations being vast and diverse. Via this data we can now observe the websites (URLs) that were promoted by Twitter users during the first days of the pandemic announcement. This allows for further analysis of whether these websites contain misinformation and if that misinformation can potentially harm others. Unfortunately, being able to determine if a tweet is indeed factual requires significant analysis and will not be the focus of this paper. Instead, this article is created to ask and answer two significant questions.

1. RQ1: What kind of website links (i.e., URLs), related to COVID-19, were shared most frequently by verified Twitter users during the start of the pandemic?
2. RQ2: What type of Twitter messages/posts related to COVID-19 garnered the most social media attention during the initial days of the pandemic?

These two questions represent two distinct viewpoints on Twitter. These include: (1) the viewpoint of the "99%" which carry very little influence individually, but through sheer numbers can bring certain information into focus, and (2) the viewpoint of the "1%" which already have a large influence but contain less posts overall due to the small number of these individuals that exist. Both these viewpoints are significant.

We believe this study is valuable since using Twitter to predict flu trends and tracking public reactions has been done in the past with other sicknesses like the swine flu [4]. However, this study is unique given COVID-19's largely divided opinions regarding its threat potential. Therefore, it is beneficial to find people's information sources.

The rest of this paper is organized as follows. Section II provides a discussion on related work, followed by Section III

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

that provides our research methodology and Section IV that provides data analysis and results. Section V provides a discussion on the implications of our results, followed by a brief conclusion in Section VI.

II. Related Work

This is not the first time that Twitter has been used in such a prolific manner. Prior to COVID-19 there was the swine flu which also circulated through various social media websites [4]. During this time an important method that researchers deployed was to use Twitter as a live-tracking tool by having programs detect when certain keywords were frequently used and what terminology was used. In the study by Chew et al. [5], we see that as time progressed Twitter authors gradually began to use WHO terminology regarding H1N1 which implies that as awareness created by the WHO increased, so did the correct usage of terminology related to the disease.

Such facts were taken from earlier years, but are relevant with respect to COVID-19 data today, e.g., with misinformation. During the pandemic, countless tweets were created with positive intentions however not all of that data was correct. As per Kouzy et al. [7] in a study of about 650 tweets, 1 in 4 tweets contain some form of misinformation. This could include data such as whether the pandemic has spread to a given local area, but it can also include data such as wearing the wrong type of mask. Both types of information can cause direct and indirect harm as that knowledge can now be spread by the readers.

Additionally, Twitter can also be used to bring attention to conspiracies spread by users. During COVID-19 pandemic, a popular conspiracy was how 5G towers were contributing to the pandemic. According to research we can see that these users who tweeted about the subject consisted of isolated groups lacking authoritative figures (e.g. government organizations) to prevent misinformation [3]. This shows that there are several other methods of analysis used to see how information is spread among Twitter users and the general public.

While the location detection feature from Twitter is only viable when posting from a mobile phone, Twitter also records other important data. The data attributes most valuable to this study for filtering purposes are the tweet dates and the hashtags pertaining to removing unrelated or old tweets that might have created noise in our dataset. In addition, the author of each tweet, the number of followers of the author, whether the author is verified by Twitter, and the link to the tweet itself is provided through Twitter's API. This enriches the dataset which allows for researches to find useful information.

III. Methodology

This section provides the details about the data source from which we obtained tweets related to COVID-19, the data analysis tools/instrumentation used in this study, and the data filtration and cleaning process used during this study.

A. Data Source

We first started by accessing the tweets stored as a dataset in the George Washington University (GWU) Libraries Dataverse [8]. When we accessed this dataset, it contained 239,861,658 (~240 million) tweet-ids related to COVID-19. Please note that the dataset has been updated since then and more tweet-ids have been added. The authors of this dataset have collected the tweet-ids using Twitter Stream API's POST filter/statuses method. Furthermore, the dataset authors have used the track parameter with the following three keywords: #COVID19, #Coronavirus, and #Coronaoutbreak.

Please also note that the dataset from GWU Dataverse contains only tweet-ids, which are unique ids that Twitter assigns to every tweet. We first downloaded the tweet-ids from above-mentioned dataset [8], and next used a process called rehydration [10]. Rehydrating tweet-id dataset involves sending the tweet-ids up to Twitter's live site to retrieve the full tweets.

B. Data Analysis Instruments

We used the following tools for rehydrating, viewing, and analyzing the data:

DocNow Hydrator: Hydrator is a Python-based, open-source desktop application that provides a user-interface to rehydrate tweet-id datasets [10]. Hydrator accepts tweet-ids and

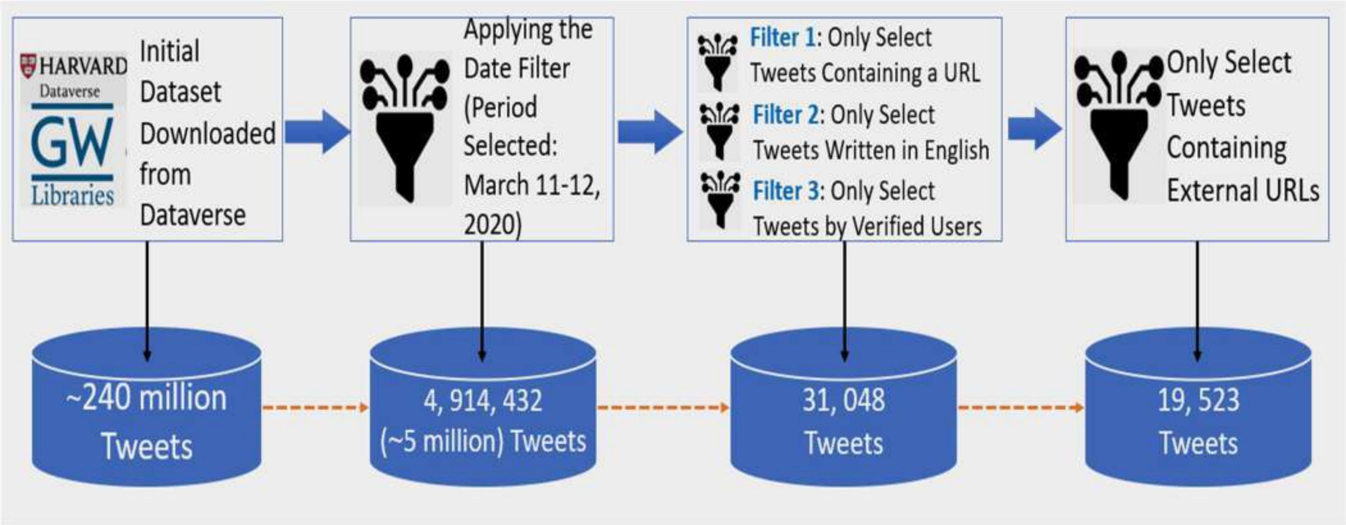

HARVARD Dataverse
GW Libraries
Initial Dataset Downloaded from Dataverse
Applying the Date Filter (Period Selected: March 11-12, 2020)
Filter 1: Only Select Tweets Containing a URL
Filter 2: Only Select Tweets Written in English
Filter 3: Only Select Tweets by Verified Users
Only Select Tweets Containing External URLs
~240 million Tweets
4, 914, 432 (~5 million) Tweets
31, 048 Tweets
19, 523 Tweets

Fig. 1. Data Filtering and Cleaning Process

returns the corresponding data in the form of JSON files, which can be converted to CSV format (the conversion to CSV format is an added feature that Hydrator provides to its users).

Tad CSV Viewer: Tad is a powerful data engineering tool that can be used to open large CSV files [9]. We selected Tad because it is an open-source tool (MIT Licensed). Tad provides a robust filtering and sorting system when dealing with large amounts of data. After applying the necessary filters on the dataset using Tad, we then transferred the filtered data from Tad to MS Excel for further analyses.

Authors note that, apart from above-mentioned tools we used, there are several big data analysis techniques and machine learning tools [17-21] that have been used successfully for analyzing large amounts of data.

C. Data Filtering and Cleaning

In this sub-section, we describe our data cleaning and filtration procedure, an overview of which is provided in Fig. 1. As previously mentioned, our initial dataset contained 239,861,658 (~240 million) tweets related to COVID-19. Of the 240 million tweets, we narrowed the search down to the tweets posted on March 11th and 12th (of the year 2020), which were the first two days when the World Health Organization (WHO) declared COVID-19 as a pandemic. This, in addition to narrowing the tweet count to 4,914,432, also allowed us to focus on the particular days when the pandemic was at its objectively most critical point.

Furthermore, we were only interested in those tweets that contained links (i.e., URLs), since those links can support any claims the tweet author might have mentioned in their tweet. For example, if the author believes their followers should wash their hands, they will most likely link a website that supports their beliefs which improves the credibility of the tweet author. Additionally, we also limited the analysis in this study to tweets that were in English language. The easiest way to achieve this was to filter by the language in which Twitter classified their tweets and select only the English language tweets. Next, we narrowed the tweets again by considering the *verified Twitter users only*. This ensured that the users considered in this study are at least somewhat credible (as deemed by Twitter), which lowers the chance of the researchers clicking on malicious links or having our research data be skewed by users spamming tweets. These three restrictions narrowed our tweet count from 5 million to 31,048. The next issue we faced when analyzing this data was that a large percentage of users would post links that were still within the Twitter domain (i.e., the source-URL was www.twitter.com and not an external URL). In the current study, we specifically focused on external URLs posted by the verified Twitter users. Please note that, while posts containing links within the Twitter domain were not harmful in any noticeable way, they were not useful for our data analysis and thus these links were removed. This exercise brought the tweet count further down to 19,523.

IV. Data Analysis and Results

This section is organized around the two research questions (RQs) that we described in Section I.

RQ1: What kind of website links (i.e., URLs), related to COVID-19, were shared most frequently by verified twitter users during the start of the pandemic?

The objective for this research question (RQ) was to determine the primary source from where users were getting their COVID-19 related information (and sharing it with their social media audience). In order to answer this RQ, we started by evaluating domain-names of URLs found in the 19523 tweets that we identified during our data filtration and cleaning process (see Fig. 1). As an example, a URL posted in a tweet might be: "https://www.cdc.gov/coronavirus/2019-ncov/index.html". The URL provides the link to a unique/specific page and such a URL is called an *absolute-URL*. However, for our initial analysis, we were just interested in the domain-name (i.e., source) of this URL, which in this case is "www.cdc.gov".

In order to extract only the domain-names from the 19523 URLs, we used an Excel formula on our tweet dataset. After

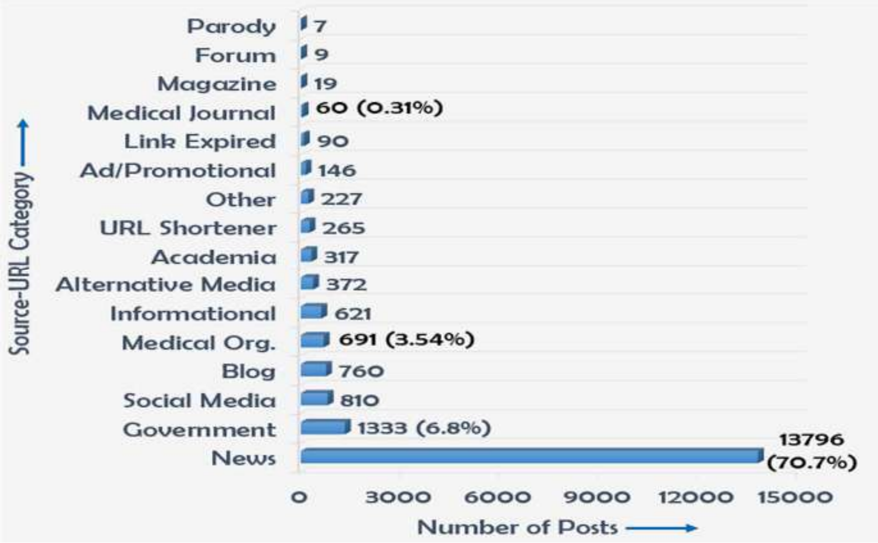


Fig. 2. Domain-name (or Source-URL) Distribution by Category

extracting the domain names, the next step was to classify the domain-names into intuitive categories so as to help enable us in answering RQ1. The researchers agreed upon 16 categories that best described the various domain-names. The distribution of source-URLs (or domain-names) appear in Fig. 2.

As observed in Fig. 2, many of the 16 categories are self-explanatory, e.g., social media links directing to Facebook or Instagram. We describe some of the categories here for reader's reference. The first category of note is Informational: this covers any website that is not directly related to COVID-19 but would like to spread awareness of their business status or general safety tips to the public as relevant to the pandemic. This also includes statistical websites that allow for viewing of the total case count of COVID-19. The next is Alternative Media which covers any form of audio or video media. Another category is 'Other' which is worth mentioning as this fits everything else that cannot be categorized properly, e.g., forms and surveys.

Fig. 2 provides several interesting observations on RQ1:

- Out of the 19523 URLs, most (about 70%) belonged to the category *News* outlets, i.e., websites such as CNN, ABC News, BBC, etc.
- A very small number (only 60, i.e. around 0.31%) was associated with the *Medical Journal* category. This was interesting as we anticipated that a significantly larger number would be associated with this category. It was concerning as it showed that users were not relying on scientific journals for their information about COVID-19.
- Another category titled, *Medical Organizations*, where researchers expected to do well, also had a very small number of posts (only 3.54% of the total posts). The *Medical Organizations* category is related to healthcare facilities such as Johns Hopkins, Mayo Clinic, etc.

The aforementioned analysis (depicted in Fig. 2) specifically looked at the source-URLs (i.e., domain-names) of the links posted in the 19523 tweets we analyzed. We performed another analysis, wherein we focused on the *absolute-URLs*. This secondary analysis was conducted on the complete URLs posted on Twitter. Absolute-URLs are quite literally the addresses of the Internet. Just as providing our address can be useful to find our home, clicking on an absolute-URL is useful to find the very page that is linked or referenced.

Table I provides a list of those absolute-URLs that were found to be most frequently shared. We have only provided the top 10 most popular absolute-URLs (related to COVID-19) posted by users. After ranking the absolute-URLs (as shown in Table I), we proceeded to check whether these links provided accurate and/or valuable information about COVID-19. Please note that Table I was created based on individual tweets containing specified URL and does not account for the number of people who viewed a specific tweet. This could, for example, imply that while 192 people tweeted a URL to medium.com (which is ranked #1 in Table I), it is possible that thousands of users just viewed the cdc.gov website (which is ranked #8) while the table does not account for that.

Table I. Top 10 Absolute-URLs Posted by Verified Twitter Users

Rank	URL	Count
1	https://medium.com/@who/seven-simple-steps-to-protect-yourself-and-others-from-covid-19-83898c2eb972	192
2	http://nhs.uk/coronavirus	68
3	https://www.pscp.tv/w/cTe04TI2MTAyMHwxZGp4WFFrcUFwVktapJGVdYKprVKmQX0qL9KIdsnjZyP0uDDSTx5lwsBm8uQ=	57
4	http://dawn.com/live-blog/	45
5	https://www.mirror.co.uk/news/uk-news/coronavirus-outbreak-live-updates-advice-21657837	43
6	https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca	41
7	https://pm.gc.ca/en/news/news-releases/2020/03/11/prime-minister-outlines-canadas-covid-19-response	33
8	https://www.cdc.gov/coronavirus/2019-ncov/index.html	29
9	http://bit.ly/COVID19Mythbus	27
10	https://www.thequint.com/news/breaking-news/coronavirus-covid-19-latest-news-live-updates-2	25

RQ2: What type of Twitter messages/posts related to COVID-19 garnered the most social media attention during the initial days of the pandemic?

The data analysis for RQ2 provided insights related to the question: "what type of messaging (i.e., information) about COVID-19 was trending during the initial days of the pandemic and what type of users were behind this messaging?"

Please note that when performing the data analysis for RQ2, we did not focus on the tweet containing URLs. Instead, the focus of this analysis was more on the users and the content they posted. A major challenge here was to make sure that we separate a user's popularity during the pandemic and outside of the pandemic. In order to ensure this, we could not rely on users' followers, favorites, and friend counts as those could be attributed to having been earned prior to COVID-19 and thus would skew the data. It was decided that one way of preventing the skew of the data was to simply view the retweet counts and see what was being retweeted most frequently (as shown in Table II). The reason for this strategy was that it accounts for both the majority of users with very little influence and the minority of users with a lot of influence as the majority can still increase the retweet count through their sheer quantity and the minority can use their influence to draw users to the tweet thus increasing the retweet count further.

In order to understand the type of messages related to COVID-19 that were retweeted most frequently, we observed the top 7 most retweeted messages. Table II summarizes this data on COVID-19. From this table, we observe that the top 7 most frequently retweeted posts amounted to 85581 retweets. We find this to be a sizable count for our data analysis. Likewise, Table II provides several interesting observations on RQ2:

- The World Health Organization (WHO) appears twice in the list, which signifies that, in general, people trusted WHO to provide valuable or useful information about the COVID-19 pandemic.

- The tweets posted by medical practitioners: Dr. Trish Greenhalgh and Dr. Dena Grayson, as well as the well-known epidemiologist Dr. Anthony Fauci, appear in this list. This could be because people look towards qualified individuals for getting their COVID-19 information.

V. Discussion and Implications

This section provides a discussion on the implications of the data analysis shown in the previous section.

Considering RQ1, we can use data from Fig. 2 and Table I to view the kind of URLs (related to COVID-19) that users frequently shared on Twitter. As can be seen in Fig. 2, out of 19,523 tweets containing URLs, 70% of the tweets pointed to News websites, 4% pointed to Blogs, and 4% pointed to social media websites excluding Twitter. This might create the notion that most people are individually tweeting News websites over websites like blogs, government websites, and other social media websites. However, Table I highlighted that the top most frequently shared absolute-URL was a website by the name of medium.com which is a blog (an absolute-URL from medium.com was shared 192 times, as shown in Table I).

Furthermore, in Table I, the second most frequently shared absolute-URL was from a government website (nhs.uk), whereas the Government category (see Fig. 2) represented only about 6.8% of the total websites posted. Although Fig. 2 showed that URLs belonging to the Government category or the Social Media category were not popular, data in Table I highlighted that these categories still did get enough attention from the public. When we visited the top three absolute-URLs shown in Table I, we found that the information presented in these URLs was backed by or linked to the Centers for Disease Control and Prevention (CDC) or the World Health Organization (WHO) implying that they had a large influence when America announced COVID-19 as a pandemic. It is also noticeable that a majority of the frequently shared absolute-URLs in Table I are from News websites, which was clearly predicted by Fig. 2.

With regards to RQ2, we refer to Table II, which contains the most frequently retweeted (i.e., re-shared) messages/posts during the time of the pandemic announcement in America. It is clear from the list shown in Table II that the most retweeted messages included either statistical information about COVID-19, region specific news, or safety tips. Data in Table II also showed that, in some cases, the most frequently retweeted messages also contained political commentary or information related to political events (which we have hidden for the purposes of this study). It also included posts from doctors.

VI. Conclusion

This study presented a systematic analysis of a sample of Twitter data related to COVID-19 during the start of the pandemic. Our analysis showed that Twitter users had a clear preference promoting online materials (i.e., URLs) surrounding COVID-19. Users promoted COVID-19 articles published in quality News websites such as CNN and BBC. They also preferred to promote materials from well-known healthcare agencies such as WHO, CDC, and NHS. It is worth highlighting that users generally were not inclined to promote/share COVID-19 materials directly from Medical Journals such as The Lancet or British Medical Journal (BMJ). Interestingly, only about 0.31% of the URLs shared by Twitter users were scientific journals (belonging to the *Medical Journal* category).

TABLE II. MOST FREQUENTLY RETWEETED COVID-19 POSTS

Retweet Count	Message (i.e., the text in the post)	User Screen Name
34153	These are 7 simple steps to protect yourself and others from #COVID19. #coronavirus	WHO (World Health Organization)
13423	This is the best article I've read on #COVID19. "Countries that are prepared will see a fatality rate of 0.5-0.9%. Countries that are overwhelmed will have a fatality rate of 3-5%". Let's act to flatten the curve.	trishgreenhalgh (Dr. Trish Greenhalgh)
9403	AWESOME #Minnesota now offers curbside #coronavirus testing! Call ahead, drive up, and your samples will be taken in just a few minutes with results provided the next day. EVERY state should offer this. #COVID19 #CoronavirusPandemic h/t @kris_lovaas	DrDenaGrayson (Dr. Dena Grayson)
8732*	[redacted]	TeamTrump
7930	RT @WHO: Media briefing on #COVID19 with @DrTedros. #coronavirus	WHO
7282*	[redacted]	tedlieu (Ted W. Lieu)
4658	Dr. Anthony Fauci: "It is ten times more lethal than the seasonal flu." Watch full #Coronavirus hearing here: cs.pn/2W3IjV2	cspan (Cable-Satellite Public Affairs Network)

******Please note that any messages (i.e., posts) that provided political commentary have been hidden / masked in the above table.*

Since public opinion on social media is significant as found in many works, e.g. [11] – [15], and since COVID-19 receives much attention in computational research, e.g. [2], [3], [16], [22], [23] our study in this paper contributes to such paradigms. Overall, this study has revealed some interesting insights into the social media psyche of the public about the COVID-19 pandemic. We intend to extend this study by considering a larger sample of Twitter data and addressing other research questions, e.g. on epidemiology and vaccinations, in our future work.

References

[1] J. Fan, X. Liu, W. Pan, M. W. Douglas, S. Bao, “Epidemiology of coronavirus disease in Gansu Province, China, 2020,” *Emerging Infectious Diseases,* 2020, 26(6), ISSN: 1080-6059.

[2] E. Chen, K. Lerman, E. Ferrara, “Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set,” *JMIR Public Health and Surveillance*, 2020, 6(2)e19273.

[3] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, D.G.Rand, “Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention” *Psychological Sc*. 2020, 31(7).

[4] M. Szomszor, P. Kostkova, & C. St Louis, “Twitter informatics: Tracking and understanding public reaction during the 2009 Swine Flu pandemic,” *Proc. IEEE/WIC/ACM Intl. Conf. Web Intelligence*, 2011, pp. 320-323.

[5] C. Chew, G. Eysenbach, “Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak,” *PLoS One*, 2010, doi: 10.1371/journal.pone.0014118.

[6] H. Achrekar, A. Gandhe, R. Lazarus, S.H. Yu, B. Liu, “Predicting flu trends using Twitter data,”. *Proc. IEEE Intl. workshop on Cyberphysical Networking Systems,* 2011, pp. 713-718/

[7] R. Kouzy et al., “Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter,” *Cureus*, 2020, 12(3)e7255.

[8] D. Kerchner, L. Wrubel, "Coronavirus Tweet IDs", https://doi.org/10.7910/DVN/LW0BTB, *Harvard Dataverse*, v1, 2020.

[9] Tad CSV Viewer, https://www.tadviewer.com/, 2018.

[10] E. Summers, N. Ruest, “DocNow/hydrator: Turn tweet IDs into Twitter JSON & CSV”, https://github.com/DocNow/hydrator, 2020.

[11] X. Du, M. Kowalski, A. Varde, G. de Melo, R. Taylor, “Public opinion matters: Mining social media text for environmental management”, *ACM SIGWEB*, 2019 (Autumn) 5:1-5:15.

[12] M. Puri, A. Varde, B. Dong, “Pragmatics and semantics to connect specific local laws with public reactions” *Proc. IEEE Big Data* 2018, pp. 838-845.

[13] H. Chen, D. Zimbra “AI and Opinion Mining”, *IEEE Intelligent Systems*, 2010, 25(3):74-80.

[14] K. Gandhe, A. Varde, X. Du, “Sentiment analysis of Twitter data with hybrid learning for recommender applications”, *Proc. IEEE UEMCON*, 2018, pp. 57-63.

[15] X. Ding, B. Liu, “The utility of linguistic rules in opinion mining”, *Proc. ACM SIGIR,* 2017, pp. 811-812.

[16] D. Karthikeyan, A. Varde, W. Wang, “Transfer learning for decision support in Covid-19 detection from a few images in big data”, *Proc. IEEE Big Data,* 2020, pp. 4873-4881.

[17] M. Singh and V. Anu, "Graph Based CIA in Requirements Engineering," *in 2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, 2020 pp. 5828-5830.

[18] M. Singh, V. Anu and G. S. Walia, “A vertical breadth-first multilevel path algorithm to find all paths in a graph”, *In: Alhajj R., Moshirpour M., Far B. (eds) Data Management and Analysis. Studies in Big Data*, vol 65. Springer, Cham, 2020.

[19] M. Singh, V. Anu, G. S. Walia and A. Goswami, "Validating Requirements Reviews by Introducing Fault-Type Level Granularity: A Machine Learning Approach", *In Proceedings of the 11th Innovations in Software Engineering Conference (ISEC)*, pp. 1-11, 2018.

[20] K. Z. Sultana, V. Anu and T.Y. Chong, "Using software metrics for predicting vulnerable classes and methids in Java projects: A machine learning approach", *Journal of Software Evolution and Process,* 2021, 33(3):e2303.

[21] T. Chong, V. Anu and K. Sultana, "Using software metrics for predicting vulnerable code-components: a study on java and python open source projects," *in 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, New York, NY, USA, 2019 pp. 98-103.

[22] B. Ghoshal, A. Tucker, “Estimating uncertainty and interpretability in deep learning for coronavirus (Covd19) detection”, 2020, arXiv:2003.10769.

[23] M.N.K. Boulos, E.M. Geraghty, “Geographical tracking and mapping of coronavirus disease COVID-19, *Intl. J. Health Geographics*, 2020, 19(1).

Collaborative Global Impact Cloud Computing Risk Assessment Framework

Lanier Watkins
Johns Hopkins University
Information Security Institute

Cheng-Hao Cho
Johns Hopkins University
Information Security Institute

John Hurley*
Defense Intelligence Agency

Aviel Rubin
Johns Hopkins University
Information Security Institute

***Abstract*—Currently, risk assessment is more of an art than a science. In this paper, we demonstrate how a previously proposed theoretical risk assessment framework that focuses on making risk assessment more scientific can be implemented and applied to cloud networks. Our work offers a practical risk assessment implementation for cloud networks where disparate network owners can directly measure network risks in an objective, uniform and repeatable manner across networks by allowing the network owners to collaboratively agree on risk metrics and continuously monitor their cloud networks with the same tool, which employs these agreed upon metrics. The end result is a science-based risk metric (SBRM) based on the global impact of the vulnerabilities in the network. Specifically, we demonstrate the feasibility of implementing this SBRM by: (1) imposing real vulnerabilities on realistic cloud networks built from real hardware (multi-core 3 Ghz Xeon Deterlab servers), (2) demonstrating the operational use of our risk assessment implementation to measure risk across several emulated cloud networks, and (3) describing our approach.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Cloud networks face the same risks as traditional networks and even more. Typically, in cloud networks, enterprises share physical machines with many different types of users. This means that even if one user or enterprise is security conscious, other users may not be, and thus pass along their risk to the entire cloud network. Therefore, in addition to the risks that traditional enterprise networks face, cloud networks also have to defend against network virtualization risks, which target the virtualization layer between users in cloud computing.

Furthermore, because enterprises no longer have full control of their resources in a cloud network, they must set up controls to ensure that the cloud providers do not somehow gain access to their data. Additionally, they will have to rely on the cloud network provider for availability. Thus, part of the enterprise must incorporate the cloud network provider as a factor when managing risk. Despite these extra risks, an increasing number of enterprises are switching to cloud networks, because of their flexibility, reliability, and cost savings, which make the need for risk management greater than ever.

Our contributions include: (1) designing and implementing a science-based risk assessment implementation for cloud networks and (2) demonstrating the operational benefit and applicability of the implementation to cloud networks. The rest of the paper is organized as follows. In Section 2 we discuss the motivation for our work, and in Section 3 we detail related work. Next, in Section 4 we describe risk management in cloud networks, and in Section 5 we layout the need for a Global Impact Database. In Section 6 we discuss the vulnerabilities used in our experiments. Then, in Section 7 we detail our experimental evaluation, and in Section 8 we discuss results. Finally, in Section 9 we conclude and discuss future work.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. Motivation

Unfortunately, many enterprises have come up with their own methods for measuring risk; therefore, there is no established standard. One way of measuring risk is the frequently used Common Vulnerability Scoring System (CVSS). Even though frequently used, this system has deficits. One example was pointed out by Watkins et al. [6], where they compare ShellShock (CVE-2014-6271), a widely publicized vulnerability, and an unspecified Java vulnerability (CVE-2014-4227). The highly impactful Shellshock affected 20% to 50% of global web servers [8], and the Java vulnerability was only exploitable in a sandbox. Yet both have scores of 10.0 under CVSS. Clearly, this presents an issue when determining which vulnerability is a greater threat. This is just one example, among many others. There must be some way of distinguishing the differences between the two vulnerabilities when assessing risk.

Watkins et al. [6] propose a method of modifying the CVSS score based on additional criteria, including creating an Overlay Score and allowing for a more nuanced view of the risk presented by vulnerabilities. Rather than having a static score like CVSS, the Overlay Score is more dynamic and is reactive, as some vulnerabilities take the forefront relative to impact while others fade into the past. The Overlay Score is derived from the collaborative 5 Layer Cyber Maturity Model as explained by Watkins et al. [6], (1) collaborative management leads meet to agree on the criteria to rank vulnerabilities, (2) collaborative management leads agree on the impact the criteria will have on their risk model and thus identify criteria weights, (3) technical collaborative leads identify a scientific approach to assessing the global severity of vulnerabilities identified in the network, (4) technical collaborative leads agree on the implementation details of this scientific risk assessment approach, and (5) collaborative management leads agree on the minimum risk-level necessary for their continued collaboration. In a case-study, the authors apply this model by choosing a set of example criteria, as required by the the Analytic Hierarchy Process (AHP) to determine a ranking for these criteria. They also propose the notion of a vulnerability global impact database (with impacts of high-1.00, medium-0.05, or low-0.01). Finally, the Overlay Score for a specific vulnerability results from multiplying each criteria ranking by the vulnerability's global impact database score for that criteria and adding up the products (which yield a real number no greater than 1). This Overlay is then multiplied by the CVSS base score, which essentially serves as a way of scaling down the CVSS base score for less impactful vulnerabilities.

Our work puts these concepts from Watkins et al. [6] into practice and demonstrates their feasibility and practicality. We use actionable resources like yearly malware reports (e.g., Symantec, Mcafee) to extract customizable criteria like age of exploit, exploit availability, and global impact. We use this approach to build a database, which can be queried to enhance risk analysis. The full details of our approach are given in Section 5.

III. Related Works

There are many tangentially related experimental works that speak to cloud network risk metrics. We believe our work is unique in that our approach is based on dynamic science-based risk metrics that are customizable. We also demonstrate the feasibility of our approach through experiment on real hardware.

A. *Risk Management Suite of Tools*

Cloudcover is a suite of tools that can be used for risk management in cloud networks. It provides a cloud-diagnostic tool that uses proprietary analytics to determine weaknesses in security. It also presents real-time findings based on Confidentially, Integrity, and Availability (CIA) and analysis using continuous monitoring [5] [4]. Similarly, MetricStream provides software for enterprise governance, risk management, and compliance (GRC). Essentially, risk assessment is accomplished via continuous monitoring for internal vulnerabilities coupled with external data feeds, which are combined via key performance and risk indicators [2] [3]. Although our work is similar in that we perform continuous monitoring of vulnerabilities, we instead base our effort on augmenting CVSS scores such that they are more representative of the impact of the vulnerability and operationalizing to demonstrate the practicality of our approach.

B. *Qualitative Risk Management*

The authors in [9] introduce a qualitative risk assessment approach for cloud networks based on ISO 31000 for logistics companies. The basis of the measured risks are responses to a questionaire given by human employees. The authors in [15] base their risk assessment framework on ISO27005. They modify the risk assessment step in ISO27005 to allow the cloud client to evaluate the risk factors. Also, they limit their scope to Software-as-a-Service models only. Similarly, the authors in [11] base their risk management on ISO 27001 cloud computing risk assessment use cases. Specifically, the authors use 7 ISO 27001 compliant risk assessment use cases to help answer the question ”What are the risks of deploying my software (Software-as-a-Service), tools (Platform-as-a-Service), and/or infrastructure (Infrastructure-as-a-Service) in the cloud?” Our work is similar in that it is based on the analytical hierarchical process (AHP), which allows for the fusion of qualitative and quantitative information. However, the end result of our work is a numerical science-based risk metric that demonstrates the deployment feasibility of this approach.

C. *Artificial Intelligence and Risk Management*

The authors V. Malik et al. [10] see risk management as identifying several risk factors across many different risk categories and using statistical or artificial intelligence algorithms to minimize these risks. However, the authors do not mention the data sources that should be used to derive training data for the above mentioned algorithms or the feature vectors that should be used to extract patterns out of this data. Similarly, A.D. Kozlov et al. [12] developed a cloud risk management system based on fuzzy logic. Specifically, the authors use MATLAB to develop a fuzzy inference system based on identifying assets, identifying threats, defining vulnerabilities, analyzing risks, and processing risks. Our work is different in that we show the operational practicality of a method that uses the AHP to fuse statistical and qualitative data into a repeatable risk metric.

IV. Cloud Networks and Risk Management

Years ago when the authors in [6] developed their approach, there was need for enterprises with their own private networks to collaborate. Though this requirement still exists, networks are now mostly cloud-based. In fact, most companies use cloud networks in many aspects of their business [13], whether they be private (enterprise owned infrastructure, managed by provider like like HP or IBM), public (infrastructure totally owned and managed by provider like Amazon or Google), hybrid (enterprise and cloud provider share aspects of hosting network) network. Even though these networks are now likely cloud networks, the collaborative approach mentioned in [6] is still applicable. However, this approach is most likely only applicable to private and hybrid cloud network providers, since the necessary hardware vulnerability monitoring would likely only be possible in these types of cloud networks. Below in Section 5 we discuss how we implement the science-based risk management theory from [6] and demonstrate its feasibilitiiy in real networks.

V. Science-based Risk Assessment Design

Our design is based on 4 key aspects taken from Watkins et al. [6] First, the collaborative design of risk assessment metrics. Second, the vulnerability monitoring of the nodes within the network. Third, the creation of a global impact database based on global assessment data. Fourth, the continuous vulnerability monitoring of the cloud network and calculation of the Overlay score.

A. *Global Impact Database*

The global impact database serves as a way to make the chosen criteria actionable. It draws its data from published annual threat data (i.e., Symantec threat reports), which we use as an authoritative source of information on the global impact of certain vulnerabilities. Actually, the global impact database has a rating for each of the criteria, i.e., global impact, exploit availability, and age of exploit. This information was manually extracted from published annual threat data.

For global impact rating, we sifted through the report and assigned four levels of impact. A vulnerability would have a global impact rating of 1.00, if it was deemed to have high global impact. An example of such a vulnerability would be ShellShock. The vulnerability would be given a rating of 0.50, if it had moderate impact, such as some of the vulnerabilities in Adobe Flash Player. A lower rating of 0.25 would be given, if the vulnerability was mentioned in the report, but stated as not having much impact. Finally, if a vulnerability was not in the report at all, it would be given a score of 0.10. We chose these impact ratings based on the approach in Watkins et al. [6]. However, to actually be able to implement this approach, we made modifications (e.g., added the option of making a low rating 0.25 or 0.01 based on if it were mentioned in the report or not).

We used a 3 level rating for exploit availability: (1) an exploit would receive a score of 1.00, if there was an exploit readily available (e.g., exploits readily available for usage in Metasploit or some other scripting tool), (2) a vulnerability would be given a score of 0.50, if an exploit required other vulnerabilities to be present on the target system, and (3) all other vulnerabilities would be given a score of 0.25.

In terms of age of exploit, the scores were calculated in a cumulative manner. The score would increase by 0.10 for each year the exploit existed, with a maximum score of 1.00 for any exploit 10 years or older.

After the criteria rankings were calculated from the analytic hierarchical process, we then multiplied these rankings by the scores for each factor and added the resulting numbers to obtain a unit-less weight with which to multiply the CVSS score. The end result is the Overlay score.

To illustrate the the influence of the global impact database on the Overlay score, we offer the following example: consider the ShellShock (CVE-2014-6271) versus Oracle database (CVE-2010-0071) vulnerabilities, the CVSS base scores for each is 10.00. However, the Overlay scores for ShellShock is 8.91 since there is an exploit available along with a high global impact score, and an exploit age of 3 years old. The Oracle vulnerability has an Overlay score of 2.84, mostly since it does not appear in the global impact database and thus has a minimum rating for global impact.

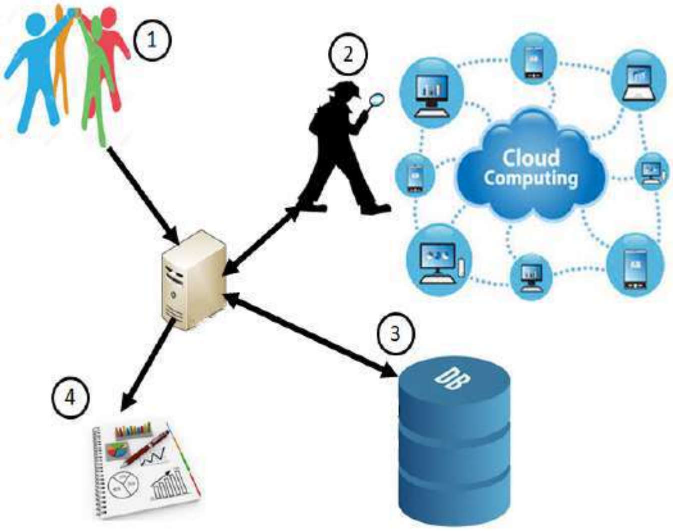


Fig. 1. Science-based Risk Metric Implementation Design: (1) accepts as input the relative importance of one criteria over another, (2) feeds in vulnerability information from each node in each network, (3) queries from the global impact database per vulnerability, uses all of this information to calculate the Overlay score, and finally uses the Overlay score to calculate the Overlay Science-based Risk Metric

B. *Science-based Risk Metric Implementation*

The SBRM implementation design is illustrated in Figure 1. In this section, we elaborate on this design. The python code written to calculate the SBRM takes as input, the relative importance of one criteria over another, which then is used to calculate the criteria ranking. The product of the criteria ranking and the database score (queried from the global impact database per vulnerability) is used to calculate a unit-less weight. The Overlay score is derived from the product of this unit-less weight and the CVSS score. Finally, the Overlay score is used to calculate the science-based risk metric (SBRM) score as specified by [6]. Note, we also calculate the SBRM using the CVSS score just as a comparison to the Overlay SBRM score. The SBRM defines the number of nodes expected to be compromised in the next year.

VI. Emulating Security Risks in Cloud Networks

We chose 5 different types of vulnerabilities to induce on the nodes in our test networks. All of these software packages could realistically be used in a cloud network. These vulnerabilities are invoked by either installing vulnerable software or by using an old Linux image. The Apache Server 2.4.18 was chosen to emulate the hosting of a webserver on the cloud network. The vulnerability associated with this is CVE-2016-8743. This is essentially a whitespace parsing defect. Oracle MySQL is another piece of software used, which could mimic database hosting in a cloud network environment. This software has several vulnerabilities associated with it such as: CVE-2016-5584, which allows remote administrators to affect confidentiality, CVE-2016-6662, which allows local users to bypass certain protection mechanisms, and CVE-2017-3331, which allows low privileged attackers to essentially crash the server. Adobe Flash Player is software that could also be found in a cloud network to stream video. This software is associated with vulnerabilities CVE-2015-0311 and CVE-2015-0312. Java is a commonly used programming language in software development. As software engineering companies offload their processing power onto cloud networks, it is common to find packages like Oracle's JDK. There could be a number of vulnerabilities associated with this software. Similarly, there could be quite a few vulnerabilities associated with the old Linux image Ubuntu 8.02 that we used. Some of the packages that were included were: OpenSSH 4.7p1, Apache Server 2.2.8, MySQL 5.0.51a, Samba 3.0.20, PostgreSQL 8.3.1, ProFTPD 1.3.1, and VNC 1.2.9. These packages could emulate the existence of legacy systems that may be used to support older products.

VII. Experimental Evaluation

The nodes used in our experiments were hosted in DeterLab [14], which allows users to reserve tens of physical nodes for several hours at a time. We used this environment to automate our experiments.

A. Experimental Setup

We setup 3 Linux networks composed of 25 nodes each using 3 GHz Xeon processors with 4 GB of RAM. To emulate vulnerabilities in these networks each node either had a vulnerable application installed or executed an Ubuntu 8.02 image with old, vulnerable software. The nodes running the Adobe Flash Player ran Ubuntu 14.04. The remaining nodes ran Ubuntu 16.04. The vulnerabilities were randomly distributed across the nodes in the 3 networks. In this paper we focus on 1 run where the first network had a total of 64 vulnerabilities, the second network had 323 vulnerabilities, and the third network had 529.

B. Procedure

Python scripts were used to randomly distribute vulnerable software on each network. Specifically, the vulnerable software chosen were: (1) Adobe Flash Player 11.2.202.438, (2) Oracle Java Development Kit 8u74, (3) Oracle MySQL Server 5.7, (4) Apache 2.4.18 and (5) an old Linux image containing several vulnerable software packages (i.e., Apache 2.2.8, Oracle MySQL 5.0.51a, Samba 3.0.20, PostgreSQL 8.3.1, Proftpd 1.3.1, and VNC 1.2.9). The selection of these vulnerabilities was done to reflect software that would be found on an enterprise cloud network. Next, the nodes were scanned using the Vuls vulnerability scanner. Then vulnerability information for each vulnerability in each network was fed into the python risk assessment implementation along with the inputs from the user on the criteria weighting, and the database scores from the global impact database to calculate the Overlay score and then the science-based risk metric (SBRM).

TABLE I
Total Vulnerabilities Per Network

Cloud Network	Total Number of Vulnerabilities
Network 1	64
Network 2	323
Network 3	529

TABLE II
Distribution of Vulnerability Categories Networks

Cloud Network	Apache Server	Oracle MySQL	Adobe Flash Player	Vulnerable Linux Image	Oracle JDK
Network 1	8	12	5	0	0
Network 2	0	11	12	2	0
Network 3	0	12	0	4	9

VIII. Results and Discussion

We demonstrated the feasibility of our implementation by arbitrarily choosing exploit availability as 5 times as important as global impact, 3 times as important as the age of exploit, and the global impact as 2 times as important as the age of exploit. Recall from the Cyber Maturity Model proposed by Watkins el al. [6], this would be agreed upon by the leaders of the organizations that want to collaborate and connect their networks. Next, we ran our scripts to randomly distribute vulnerabilities on the 3 networks, and the results are listed in Tables 1 and 2. The data in Table 1 illustrates the total number of vulnerabilities across all 3 networks, and Table 2 lists a breakdown of vulnerability categories across networks. The results of the SBRM calculations for both the CVSS scores and the Overlay scores are listed in Table 3. Note that even though network 3 has the most vulnerabilities, both the CVSS and Overlay SBRM state that network 2 is the most vulnerable. This particular run of our risk implementation yields results where both the CVSS and Overlay SBRM agree; however, this does not always happen. The main point here is that the Overlay SBRM reflects the global impact of the vulnerabilities in the network and is more informative than the SBRM calculated using the CVSS scores. This point is made even clearer in Tables 4 - 6. For instance, in Table 4, the top 5 most impactful vulnerabilities in network 1 are identified by their CVSS scores, but only yield a score of 10 for each of them. However, the corresponding Overlay score identifies CVE-2015-0311 as the most impactful and CVE-2016-6662 as the least impactful in the table. The scaling of the vulnerabilities based on their global impact by the Overlay score clearly provides the necessary information for network owners to be able prioritize the vulnerabilities that should be patched in their network. Not only does the Overlay score provide vulnerability priortization information, but it also allows owners to directly compare risks (via the Overlay SBRM) across networks. We believe that these results support our premise and demonstrate that it is feasible and practical to implement the SBRM as a risk assessment resource for cloud networks.

TABLE III
TOTAL NUMBER OF NODES EXPECTED TO BE COMPROMISED

Cloud Network	CVSS SBRM	Overlay SBRM
Network 1	17.08	4.62
Network 2	25.00	12.27
Network 3	16.09	4.34

TABLE IV
TOP 5 MOST IMPACTFUL VULNERABILITIES IN NETWORK 1

CVE	CVSS	Overlay
CVE-2015-0311	10.00	8.76
CVE-2015-5122	10.00	4.21
CVE-2015-0312	10.00	2.07
CVE-2015-5123	10.00	2.07
CVE-2016-6662	10.00	1.91

TABLE V
TOP 5 MOST IMPACTFUL VULNERABILITIES IN NETWORK 2

CVE	CVSS	Overlay
CVE-2015-0311	10.00	8.76
CVE-2015-5122	10.00	4.21
CVE-2007-2446	10.00	3.31
CVE-2010-0425	10.00	2.84
CVE-2012-1182	10.00	2.53

TABLE VI
TOP 5 MOST IMPACTFUL VULNERABILITIES IN NETWORK 3

CVE	CVSS	Overlay
CVE-2007-2446	10.00	3.31
CVE-2010-0425	10.00	2.84
CVE-2012-1182	10.00	2.53
CVE-2013-1902	10.00	2.38
CVE-2013-1903	10.00	2.38

IX. SUMMARY AND FUTURE WORK

We have demonstrated via a python implementation, the efficacy and usefulness of a theoretical risk metric system proposed by Watkins et al. [6]. We have shown that it is practical for cloud network owners wishing to collaborate to be able to directly compare their network risks and that the information used for this comparison can also be used to prioritize vulnerability patching. Although this implementation serves as a good proof of concept, it is limited in its capability and could be improved. The major limitation is the global impact database has to be updated manually. To improve this, natural language processing could be used to extract the necessary information out of the annual malware reports. In future work, we would like to extend this implementation beyond enterprise networks to ICS networks, and also automate updates to the global impact database.

ACKNOWLEDGEMENT*

The views expressed are the author's alone and do not represent the official policy or position of the National Intelligence University, the Department of Defense or any of its components, or the U.S. Government.

REFERENCES

[1] I. Vacas et al., ”Detecting Network Threats using OSINT Knowledge-based IDS”, In IEEE European Dependable Computing Conference, 2018.

[2] MetricStream Product Review, SC Media Magazine Online, Available At: https://www.scmagazine.com/metricstream-risk-management-solution-v60/review/6607/

[3] Metricstream Website, Availabe Online At: https://www.metricstream.com/platform/metricstream-cloud.htm

[4] Cloud Cover Website, Available At: http://www.cloudcover.cc/?page_id=137

[5] M. Aydin, ”Cloud-COVER: Using User Security Attribute Preferences and Propagation Analysis to Prioritise Threats to Systems”, In IEEE European Intelligence and Security Informatics Conference, 2015.

[6] Lanier Watkins and John Hurley, “Cyber Maturity as Measured by Scientific-based Risk Metrics”, In the Journal of Information Warfare (JIW), October 2015

[7] Symantec 2015 Internet Security Threat Report, Volume 20, Available At: https://docs.broadcom.com/doc/istr-15-april-volume-20-en

[8] D. Yadron, ”Google and Amazon Respond to Shellshock Security Flaw”, Wall Street Journal Online, 2014, Available At:https://www.wsj.com/articles/BL-DGB-37893

[9] Maniah et al., ”Risk analysis of Cloud Computing in the Logistics Process”, In IEEE International Conference on Vocational Education and Electrical Engineering, 2020

[10] V. Malik et al., ”Cloud, Big Data & IoT:Risk Management”, In IEEE International Conference on Machine Learning, Big Data and Parallel Computing, 2019.

[11] T. Weil, ”Risk Assessment Methods for Cloud Computing Platforms” In IEEE Annual Computer Software and Application Conference, 2019.

[12] A.D. Kozlov et al. ”Risk Management for Information Security of Corporate Information Systems Using Cloud Technology”, In IEEE International Conference on Management of Large-Scale System Development, 2018.

[13] C. Buhrkuhl, ”Private Cloud vs Public Cloud vs Service Providers” https://www.office1.com/blog/private-cloud-vs-public-cloud-vs-service-providers-office1

[14] Deterlab Website, Available At: https://www.isi.deterlab.net/index.php3

[15] S. Hassan ”Security risk assessment framework for cloudcomputing environments”, Security Comm. Networks 2014; 7:2114–2124, 2014.

Execution of Hybrid NOMA Schemes Concerning Outage Performance and Sum Rate Interplay

Anindya Bal
dept.of Electrical and Electronic Engineering
BRAC University
Dhaka, Bangladesh
anindyabal007@gmail.com

Rahat Tajwar
dept.of Electrical and Electronic Engineering
BRAC University
Dhaka, Bangladesh
rahat.tajwar@gmail.com

***Abstract*—The steadily developing number of users has achieved the test to distinguish an advantageous strategy for different access in upcoming wireless communication. It has been seen that non-orthogonal multiple access (NOMA) method gives far superior execution than orthogonal multiple access (OMA) procedure while exploiting the accessible assets. Interferences can be limited enough using successive interference cancellation (SIC) when actualized broadly, however it is normal situation in NOMA. But enormous usage increments the complexity of the framework and as a result energy effectiveness is decreased. To balance the detriments of NOMA procedure, interest in hybrid NOMA strategies has developed. Numerous new papers have proposed hybrid NOMA schemes as a reasonable alternative to moderate the disservices of NOMA. This paper plays out an investigation of the exhibition between hybrid NOMA systems such as CR-NOMA, U2X-NOMA, RIS-NOMA, HS-UAV NOMA, SCMA-NOMA and GFDM-NOMA. An examination of every one of these methods is led to recognize the best hybrid NOMA procedure for 5G and beyond wireless communication. The articulations for sum rate and outage probability have been determined for all the procedures. Through examination and reenactment results it is found that U2X-NOMA gives the best exhibition regarding sum rate and outage probability among the six discussed schemes.**

***Index Terms*—CR-NOMA, U2X-NOMA, RIS-NOMA, HS-UAV NOMA, SCMA-NOMA, GFDM-NOMA, Outage Probability, Sum Rate**

I. Introduction

Non-orthogonal multiple access (NOMA) has gained a prominent impact in establishing wireless networks of the next generation 5G and beyond wireless communication. To support various applications, a high data rate and stable communication would therefore be needed [1],[2]. The main aim is to be able to extend the channel capacity in the architecture of the cellular networks [3]. For instance, multiple access is an useful approach for next generations to accommodate as many cells as feasible at a given bandwidth with a fair extent of service quality [4],[5]. Non-orthogonal multiple access (NOMA) is considered to be a revolutionary multiple access framework for fifth generation (5G) and beyond 5G (B5G) cellular networks that can significantly improve the efficiency of application distribution [6]-[8]. In cooperative networks with buffer-aided power control, hybrid NOMA-OMA has been discussed in [9],[10] and the machine productivity of the computer is increased relative to other existing state-of-the-art heterogeneous network bandwidth management protocols [10]-[13]. SIC is used at the ends of the receiver for inter-user interference cancellation [14],[15]. In[16]-[18], several researchers proposed different but rather powerful methods to power distribution in order to efficiently disperse power to NOMA structures. For full duplex (FD) and half duplex (HD) switch supported cognitive radio NOMA (CR-NOMA) networks, the author provided a performance analysis in[19]. A NOMA improved unnamed aerial vehicle to everything (U2X) network, where stochastic geometry instruments were used to model the spatial randomness of several receivers, $R_x s$ is discussed in [21]. To establish reconfigurable intelligent surface NOMA (RIS-NOMA) in [24]-[26], the framework of a high-speed internet infrastructure and cost-effective data coverage offshore is shown. The efficiency of the hybrid satellite unnamed aerial vehicle NOMA (HS-UAV NOMA) infrastructure over the Rician fading channel is studied in [27]. How user message is stored using sparse code multiple access NOMA (SCMA-NOMA) is supplemented with encoder access and various power levels are assigned depending on channel conditions in[28]. In[29], they collaborative generalized frequency division multiplexing with NOMA (GFDM-NOMA) strategies to expand the multi-carrier NOMA network framework. The survey was carried out in the aforementioned paper strongly inspired us to make more effort on the hunt for an optimal NOMA hybrid system to significantly improve efficiency and verify its quality with extant NOMA hybrid systems. In this paper, six exceptional hybrid NOMA systems are analyzed and their usefulness in the outage and sum rate interplay. The aim of the investigation is to demonstrate intuitive perspectives in order to highlight the best hybrid NOMA system to meet the need for 5G and beyond wireless communications.

II. System Model

The system model of six most modern hybrid NOMA schemes such as CR-NOMA, U2X-NOMA, RIS-NOMA, HS-UAV NOMA, SCMA-NOMA and GFDM-NOMA has been clarified in this section.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

A. CR-NOMA

It takes into consideration a modern uplink/downlink relay-assisted method developed for CR-NOMA as in [19]. The restricting influence of interference from the starting point to the relays during the relay recruitment process is known to make the model more realistic. To pick the relay for transmitting the message to the destination, the partial relay selection (PRS) technique was being used. A relay-aided CR-NOMA network with a main source MS and multiple distribution switches $\{R_k|k = 1, 2, \ldots, K\}$ using conveying protocols as shown in Fig. 1. In just the same frequency spectrum, there are two separate pairs, $G_1 = \{S_1, D_1\}$ and $G_2 = \{S_2, D_2\}$ using the NOMA technique [19]. For the reliable information S_1 and S_2, we assume a worst-case communication case. In addition, we presume that there is a deep fading of the direct connection between secondary data pairs and that communication is only possible through intermediary relays. Quite importantly, uplink-downlink CR-NOMA communication has two stages. First one is the uplink process suggested that S_1 and S_2 could create a pair of NOMA [20] to relay the R_k simultaneously.

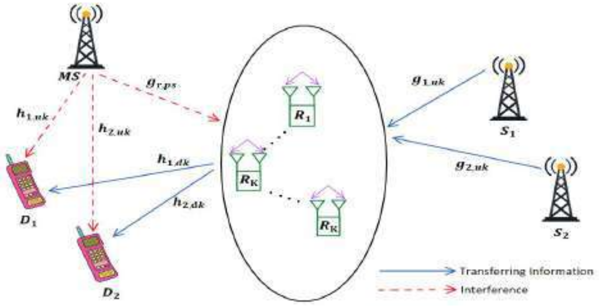


Fig. 1. Uplink Downlink Framework of CR-NOMA

The other one is R_k will simultaneously send an overlaid composite signal composed of transmitted patterns sent to D_1 and D_2 by sources in the downlink process. Restricting interference in CR-NOMA is visible because MS interferes with numerous R_k relays and two routes. In this case, with respect to full achievable device power, both uplink and downlink require SIC to work equally at both source destination pairs as in [19]. Both S_1 and S_2 relay signals x_1 and x_2 periodically during the designated time slots, following the concept of uplink NOMA. In NOMA, $\alpha_1 P_s$ and $\alpha_2 P_s$ are part of the assigned capacity for two x_1 and x_2 signals, respectively. Here, $\alpha_1$and α_2 are the power allocation coefficient and P_s is the total transmit power.

B. U2X-NOMA

We suggest an upgraded NOMA downlink U2X network, where several $R_x s$ installed with a single omni receiving antenna are communicated by a UAV installed with a single omni transmitting antenna as in [21]. With only a single UAV, Fig. 2 shows the wireless networking framework. The UAV-T_x is located in the middle of a circle with a distance radius of D [21]. In conjunction with the homogeneous poisson point method (HPPP) with the density, two $R_x s$, the close R_x w and the far R_x v, are distributed. Therefore, two $R_x s$, one close R_x and one outside R_x, are paired in each cluster to execute NOMA. From [22],[23], we assume the circular area is divided into two power regions. One is a tiny domain (Blue zone) and the other is a big hollow domain(Green Zone). The close $R_x s$ are believed to be located with radius R in the tiny domain, while the far $R_x s$ are located with radius R to D in the big hollow domain. d_w denotes the length between the T_x and the nearest R_x, and d_v denotes the range between the T_x and the nearby R_x. The variance in received signal of the paired NOMA $R_x s$ is needed to ensure the consistency of decoding [21]. Therefore, one close R_x and one outside R_x, are coupled in each cluster to execute NOMA. In reality, large-scale fading and small-scale fading are commonly found in the fading model. The path loss in between UAV-T_x and R_x is expressed by large-scale fading. Large-scale fading and small-scale fading are considered to be determined simultaneously and completely identical. Since the conventional Rayleigh fading channels can only model powerful scattering situations, which is not a possible option for the communication of UAV [22]. The small-scale channel benefits for R_x w and R_x v, respectively, are shown by h_w and h_v for terminology convenience.

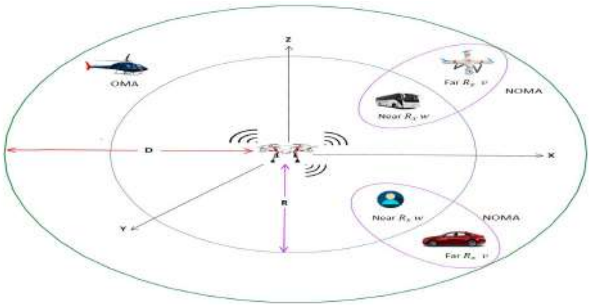


Fig. 2. U2X-NOMA Model

C. RIS-NOMA

As seen in Fig 3, we interpret the RIS-NOMA, in which the base station aims to connect with two locations. Location users are identified as the close user (CU) and the faraway user (FAU). We presume that all nodes have a single-antenna in order to achieve low complexity and low cost because even a RIS contains of N meta-surfaces as in [24]. The baseband identical fading channels from BS and the ith component in

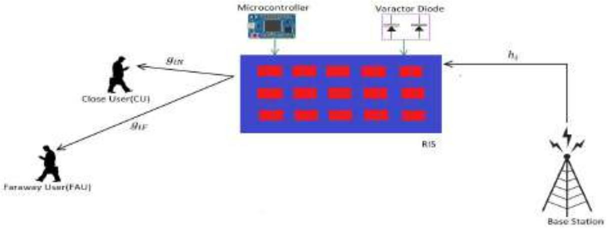


Fig. 3. System Representation of RIS-NOMA

the RIS are h_i, g_{iN} and g_{iF}, signals from the ith component in the RIS to CU and FAU, respectively. These pathways are supposed to be distinct, equivalent, changing smooth, dwindling. In this context, we considered the framework of two FAU and CU users that will obey Rayleigh correlations of various parameters of size. The signal reproduces through the RIS to two types of users from the transmitter, which helps to increase the efficiency of the signal at these locations as in [24]. Ideally, we expect that the channel state information (CSI) of such users can be accomplished by the RIS [25]. Surprisingly, such a CSI may be used by the RIS to optimize the SNR obtained at these locations. The BS sends an overlaid signal in the form of NOMA containing signals s_1, s_2 that are aimed at the user CU and FAU, respectively. Higher power level a2 is allocated to user FAU in order to have separate quality of service (QoS) for applications, such allocation scheme must be fulfilled, $\alpha_1 + \alpha_2 = 1$.

D. HS-UAV NOMA

In Fig. 4, the entire interaction is broken into two schedules. The first one is satellite to UAV and the other one is UAV to mobile users. We suggested an HS-UAV NOMA downlink network, where one satellite as seen in Fig. 4, aims to connect with terrestrial NOMA users with the assistance of a DF UAV relay U as in [27]. A consistent height h is used to install the UAV. The UAV and all users are believed to be running in half-duplex mode and all nodes are fitted with individual antennas. All served users are divided into two categories, A_1 and A_2, depending on the ranges between the UAV and the NOMA mobile users. We claim that in group $A_1$the close user D_n is located in the loop with both the radius R_n and in group A_2 the far user $D_f$is situated in the inner radius R_n and outer radius R_f loop $(R_f > R_n)$ [27]. We also assume that due to the barrier or extreme large-scale fading, the direct connections between the satellite and the users are missing. Random user matching is known to achieve greater efficiency gains. We will concentrate on the paired user's results in the following, and using the same approach, other pair performance can be extracted.

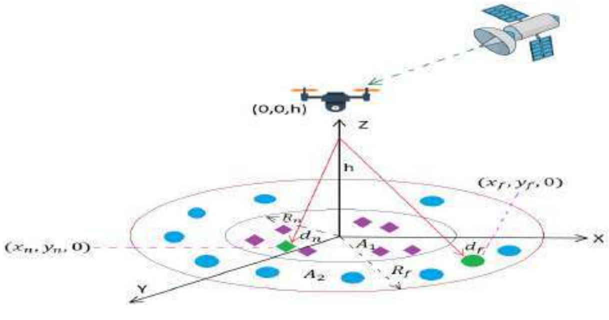


Fig. 4. Framework of HS-UAV NOMA

E. HS-UAV NOMA

We assumed one base station and J users as a downlink system. Suppose users of J_1 face stronger channel conditions than other users of $J_2 = J - J_1$, as seen in Fig. 5. The strongest and weakest users are respectively represented by $\{s_1, s_2,,.... s_{j1}\}$ and $\{w_1,w_2,,....w_{j2}\}$ as in [28]. The strong and the weak users are also attributed to as the near and far users, respectively, without damage of generalizability. The channel dimensions encountered by the strong and weak users respectively are denoted by $^{s}_{i}\}^{j_1}_{i=1}$ and $\{h^w_j\}^{j_2}_{j=1}$. Assume the number of orthogonal asset components like K Consider a solid user's $J_1 \times KSCMA$ algorithm focuses on codebookC_1. For weak users $J_2 \times KSCMA$ algorithm focuses on codebook C_2. MPAD1 and MPAD2 signify an MPA(Message passing algorithm) detector of weak and strong user respectively. The detection is exemplified in the second part of Fig. 5 of the ith strong user s_i. Before extracting the data from s_i, the SIC of the far users is carried out as in [28]. Initially, MPAD2 generates the approximate code words $\{x^{w_i}_j\}^{j_2}_{j=1}$ with y^s_i as input. The superscript 'i' here indicates that the detection operation is done at the position of s_i by the ith receiver [28]. These projections are only used for the identification of s_i data.

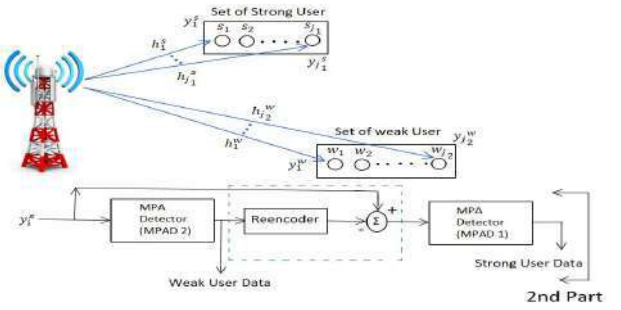


Fig. 5. System Model of SCMA-NOMA

F. GFDM-NOMA

In Fig. 6, we integrate GFDM waveforms with NOMA schemes to broaden the multi-carrier NOMA system model. In NOMA systems enabled by downlink GFDM, the situation with one BS and four users is presumed. Each conversion efficiency block is shared by four users, unlike conventional GFDM systems as in [29]. It is assumed that a close user with less assigned power and a much more transmitted power for far user is assumed. Due to the larger share of total power in NOMA networks, U_2 is first decoded by comparison to the U_1 signal as interference. The received U_1 signal on the

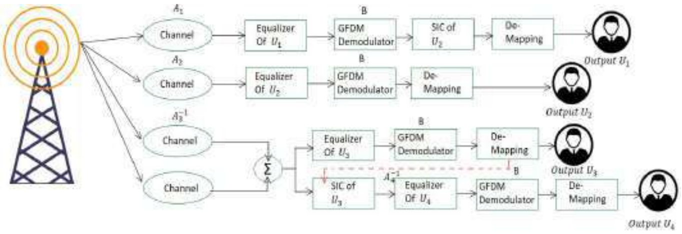


Fig. 6. Downlink Scenario of GFDM-NOMA

U_2 side is in most situations, entirely regarded as interference. Nevertheless, whenever two users endure from the same fading channels and decoding processes, from the point of view of theoretical study, the signal power of U_2 is enhanced or diminished by that of U_1, with a 50 percent chance. For U_3 and U_4 case, we use the same technique as hold on U_1 and U_2.

III. Numerical Analysis

In this section we will discuss among predefined six hybrid NOMA schemes where we calculated the outage probability and sum rate arrangements.

A. Outage Probability Analysis

For CR-NOMA, R_1 and R_2 are considered to be the specified target rate thresholds of G_1 and G_2, respectively, according to the necessary level of operation. Thus, the signal to noise ratio (SNR) thresholds are γ_{th1}^{FD}, γ_{th2}^{FD}. As in subsequent sections, the precise outage probability of G_1 and G_2 is determined as in [19].

$$\begin{aligned} OP_{G_1}^{FD} = 1 - &\sum_{i=1}^{K}\sum_{j=1}^{K}\binom{K}{i}\binom{K}{j}(-1)^{i+j-2} \\ &\times \frac{j\rho R\lambda_{1,\mu k}\lambda_{1,\mu k}\lambda_{1,dk}}{\left(i\alpha_2^{FD}\lambda_{2,\mu k}+j\lambda_{1,\mu k}\right)\left(i\alpha_3^{FD}\lambda_{r,PS}+\lambda_{1,\mu k}\right)} \\ &\times \frac{\left(a_3-\gamma_{th1}^{FD}a_4\right)}{\left(\gamma_{th1}^{FD}\rho P\lambda_{1,PS}+\left(a_3-\gamma_{th1}^{FD}a_4\right)PR\lambda_{1,dk}\right)} \\ &\times exp\left(\frac{-i\alpha_4 FD}{\lambda_{1,\mu k}}-\frac{\gamma_{th1}^{FD}}{\left(a_3\rho R-\gamma_{th1}^{FD}\alpha_4\rho R\right)\lambda_{1,dk}}\right) \end{aligned} \tag{1}$$

In which, $\gamma_{th1}^{FD}=2^{R_1}-1$, $\alpha_1^{FD}=\gamma_{th1}^{FD}\left(I_R+1\right)$, $\alpha_2^{FD}=\frac{\gamma_{th1}a_2}{a_1}$, $\alpha_3^{FD}=\frac{\gamma_{th1}^{FD}\rho P}{a_1\rho S}$ and $\alpha_4^{FD}=\frac{\alpha_1}{\alpha_1\rho S}$.

$$\begin{aligned} OP_{G_2}^{FD} = 1 - &\sum_{i=1}^{K}\sum_{j=1}^{K}\binom{K}{i}\binom{K}{j}(-1)^{i+j-2} \\ &\times \frac{j\lambda_{2,\mu k}\lambda_{2,\mu k}}{\left(i\beta_2^{FD}\tau_1\lambda_{1im}+j\lambda_{2,\mu k}\right)\left(i\beta_3^{FD}\lambda_{r,PS}+\lambda_{2,dk}\right)} \\ &\times \frac{\lambda_{2,dk}\lambda_{2,dk}}{m_1^{FD}\tau_2\lambda_{2im}\lambda_{2,dk}} \\ &\times \frac{\left(a_3\rho R-\gamma_{th1}^{FD}a_4\rho R\right)\lambda_{2,dk}}{\left(\gamma_{th1}^{FD}\rho P\lambda_{2,PS}+\left(a_3\rho R-\gamma_{th1}^{FD}a_4\rho R\right)\lambda_{2,dk}\right)}\times \\ &exp\left(\frac{-i\beta_4^{FD}}{\lambda_{2,\mu k}}-\frac{m_3^{FD}}{\lambda_{2,dk}}-\frac{\gamma_{th1}^{FD}}{\left(a_3\rho R-\gamma_{th1}^{FD}\alpha_4\rho R\right)\lambda_{2,dk}}\right) \end{aligned} \tag{2}$$

In which, $\gamma_{th2}^{FD}=2^{R_2}-1$, $\beta_1^{FD}=\gamma_{th2}^{FD}\left(I_R+1\right)$, $\beta_2^{FD}=\frac{\gamma_{th2}a_2}{a_2}$, $\beta_3^{FD}=\frac{\gamma_{th2}^{FD}\rho P}{a_2\rho S}$, $\beta_4^{FD}=\frac{\beta_1}{\alpha_2\rho S}$, $m_1^{FD}=\frac{\gamma_{th2}^{FD}a_3}{a_4}$, $m_2^{FD}=\frac{\gamma_{th2}^{FD}\rho P}{a_4\rho P}$, $m_3^{FD}=\frac{\gamma_{th2}^{FD}}{a_4\rho R}$.
For U2X-NOMA, first, we concentrate on evaluating the Outage probability (OP) of the faraway R_x v. At the UAV-T_x , where the spectrum sharing variables α_w^2 and α_v^2 are unchanged during transmission, the fixed channel assignment technique is configured. R_w and R_v, respectively, are considered to be the target levels of R_x w and v. Outage probability (OP) is defined as the tendency that the transmit power is lower than the aim rate needed, therefore given to the OP of v$^{\text{th}}$ close R_x is as [21], where, $\alpha_v^2-\varepsilon_v\alpha_w^2>0$.

$$\begin{aligned} OP_v = 1 - &\frac{3\left(mM_v\sigma^2\right)^{\frac{-3}{2}}}{\alpha\left(D^3-R^3\right)}\sum_{n=0}^{m-1}\frac{1}{n!} \\ &\gamma\left(n+\frac{3}{2}+1, mM_v\sigma^2 D^\alpha\right)-\gamma\left(n+\frac{3}{2}+1, mM_v\sigma^2 R^\alpha\right) \end{aligned} \tag{3}$$

Here, $M_v=\frac{\varepsilon_v}{P_u\left(\alpha_v^2-\varepsilon_v\alpha_w^2\right)}$, $\varepsilon_v=2^{R_v}-1$, γ is the significantly reducing inadequate gamma function. Recollect that when using SIC technology, the OP of the close Rx w relies on two decoding processes. First one is R_x w compiles the R_x v input. Secondly, R_x w decodes the signal of by its own. So, the close R_x w can be defined as [21]

$$\begin{aligned} OP_w = P&\left(\log_2\left(1+\frac{P_u|h_w|^2 d_w^{-\alpha}\alpha_v^2}{\alpha^2+P_u|h_w|^2 d_w^{-\alpha}\alpha_w^2}\right)<R_v\right)+P \\ &\left(\log_2\left(1+\frac{P_u|h_w|^2 d_w^{-\alpha}\alpha_v^2}{\alpha^2+P_u|h_w|^2 d_w^{-\alpha}\alpha_w^2}\right)>R_v, \log_2\right. \\ &\left.\left(1+\frac{P_u|h_w|^2 d_w^{-\alpha}\alpha_w^2}{\alpha^2}\right)<R_w\right. \end{aligned} \tag{4}$$

Although, $a_w^2 < a_v^2$ is followed by the channel assignment variables of the paired NOMA $R_x s$ and the paired NOMA $R_x s$ are chosen from two power areas, the encoding can therefore be consolidated.
In the case of RIS-NOMA, if CU is unable to identify the signal from FAU, the outage probability will be as [24].

$$OP_{CU} = P_r\left(\frac{|A|^2 a2}{|A|^2\left(k_t^2+k_r^2\right)+\frac{1}{P_s}}\leq\rho_{th1}\right) \tag{5}$$

Here, $A=\sum_{i=1}^{N}|h_i||g_{iN}|$, ρ_{th2} is the SINR threshold. The outage probability of FAU can be determined as [24]

$$OP_{FAU} = P_r\left(\frac{|B|^2 a2\rho_s}{|B|^2\left(k_t^2+k_r^2\right)+P_s+1+a2\rho_s}\leq\rho_{th2}\right) \tag{6}$$

Here, $B=\sum_{i=1}^{N}|h_i||g_{iF}|$, $\left(k_t^2+k_r^2\right)$ is the interference phrase.
For HS-UAV NOMA, the following three scenarios, the outage activity will occur in the far user D_f . Firstly, the switch would not be able to correctly decipher x_1 . Then, x_2 can not be correctly deciphered by the switch. The switch will effectively decipher x_1 and x_2, although its signal can not be decoded successfully by D_f . So, the OP of the Faraway user, D_f can be expressed as in [27]

$$OP_{D_f} = P_r\left(\min\left(\frac{\gamma_{x_1}^{SR}}{\gamma_{th,f}}, \frac{\gamma_{x_2}^{SR}}{\gamma_{th,n}}, \frac{\gamma_{x_1}^{RD_f}}{\gamma_{th,f}}\right)<1\right) \tag{7}$$

Where, respectively, the outage thresholds at D_f and D_n are th_f and th_n [27]. For close user D_n , outage events will not occur until x_1 and x_2 are effectively deciphered by both the relay and close user Dn. So, the OP of close user will be as [27]

$$OP_{D_n} = P_r\left(\min\left(\frac{\gamma_{x_1}^{SR}}{\gamma_{th,f}}, \frac{\gamma_{x_2}^{SR}}{\gamma_{th,n}}\right) < 1\right) \\ + P_r\left(\min\left(\frac{\gamma_{x_1}^{SR}}{\gamma_{th,f}}, \frac{\gamma_{x_2}^{SR}}{\gamma_{th,n}}\right) \geq 1, min\left(\frac{\gamma_{x_1}^{RD_n}}{\gamma_{th,f}}, \frac{\gamma_{x_2}^{RD_n}}{\gamma_{th,n}}\right)\right) \tag{8}$$

For the case of GFDM-NOMA, the situation with one BS and four users is planned in downlink of the proposed scheme. Increasing time-frequency resource block is shared by four users, separate from conventional GFDM systems. Considering a closer user with less power delegated and a much more power transmitted to far user. Due to the larger share of total power in NOMA networks, U_2 first deciphered by treating the U_1 signal as interference. Same case happened for U_3 and U_4 [29]. Nevertheless, whenever two users endure within the same fading channels and decoding processes, from the point of view of theoretical study, the signal power of U_2 is enhanced or diminished by that of U_1, with a 50 percent chance. So, the outage probability for $U_1 and U_2$ will be considered as in [29]

$$OP_{U_1} = Q(\sqrt{\frac{\alpha.\gamma}{\xi_1}}) \tag{9}$$

Here, ξ_1 is derived with $\theta = 1$ in (31) and the same power management technique is considered in [32].

$$OP_{U_2} = \frac{1}{2}Q(\sqrt{\frac{T_1.\gamma}{\xi_2}} + \sqrt{\frac{T_2.\gamma}{\xi_2}}) \tag{10}$$

Here, ξ_1 is derived with $\theta = 1$ in (31) and In addition, the factor of power enhancement T1 and the factor of power impedance T2 are represented as [29]

$$\{\begin{matrix} T_1 = (\sqrt{1-\alpha} + \sqrt{\alpha})^2 \\ T_2 = (\sqrt{1-\alpha} - \sqrt{\alpha})^2 \end{matrix} \tag{11}$$

here, $0.5 > \alpha > 0$.

$$OP_{U_{3,4}} = \frac{1}{2}Q(\sqrt{\frac{\gamma_3}{\xi}} + \sqrt{\frac{\gamma_4}{\xi}}) + \frac{1}{2}Q(\sqrt{\frac{\gamma_3}{\xi}} - \sqrt{\frac{\gamma_4}{\xi}}) \tag{12}$$

Here, $\gamma_3 = \frac{P_3}{\sigma_v^2}$ and $\gamma_4 = \frac{P_4}{\sigma_v^2}$ indicates U_3 and U_4 SNR on the receiver side.

B. Sum Rate Analysis

For CR-NOMA, with the symbol x_1,the sum rate capacity for S_1 to D_1 can be expressed as [19]

$$C_{G_1}^{FD} = \sum_{i=1}^{K}\sum_{j=1}^{K}(\begin{matrix}K \\ i\end{matrix} \quad)(\begin{matrix}K \\ j\end{matrix} \quad)(-1)^{i+j-2}\frac{1}{\ln 2} \\ \times \int_0^{\frac{a_3}{a_4}} \frac{q_1+q_2+q_3}{1+x} exp(\frac{-i\epsilon_3^{FD}x}{\lambda_{1,uk}} \\ - \frac{x}{(a_3\rho R - a_4\rho Rx)\lambda_{1,dk}} - \frac{x}{(a_3\rho R - a_4\rho Rx)\lambda_{2,dk}}) \tag{13}$$

$$q_1 = \frac{j\lambda_{1,uk}\lambda_{1,uk}}{(i\epsilon_1\lambda_{2,uk}x + j\lambda_{1,uk})(i\epsilon_2\lambda_{r,ps}x + \lambda_{1,uk})} \tag{14}$$

$$q_2 = \frac{(a_3\rho R - a_4\rho Rx)\lambda_{1,dk}}{\rho P\lambda_{1,ps}x + (a_3\rho R - a_4\rho Rx)\lambda_{1,dk}} \tag{15}$$

$$q_3 = \frac{(a_3\rho R - a_4\rho Rx)\lambda_{2,dk}}{\rho P\lambda_{2,ps}x + (a_3\rho R - a_4\rho Rx)\lambda_{2,dk}} \tag{16}$$

Using the symbol x_2, the capacity from S_2 to D_2 can be expressed as [19]

$$C_{G_2}^{FD} = E(\log_2(1+H_2)) = \frac{1}{\ln 2}\int_0^\infty \frac{1-FH_2(x)}{1+x}dx \tag{17}$$

Here, $H_2 = min(\gamma_{x2}^{S1-RK}, \gamma_{x2}^{RK-D2})$.

$$FH_2(x) = Pr(\min(\gamma_{x2}^{S1-RK}, \gamma_{x2}^{RK-D2}) < x) \\ = 1 - (1 - F_{\gamma_{x2}^{S1-RK}}(x))(1 - F_{\gamma_{x2}^{RK-D2}}(x)) \tag{18}$$

The sum rate in the U2X-NOMA networks is a prominent parameter that is necessary calculating for assessment process. Consequently, we have closed-form representation in the following consequence in terms of the ordinal ergodic rates of individual U2X R_xs. It is possible to express the attainable sum rate of the close R_x as follows [21].

$$C_w = \frac{3}{\alpha\ln(2)(R^3 - r_0^3)}\sum_{n=0}^{m-1}\frac{1}{n!}\sum_{k=0}^{\infty}\frac{\Gamma(\varphi_2 - 1)\Gamma(n+1+k)}{\Gamma(\varphi_2+k)} \\ \times C^{n+k}(R^{\varphi_1}exp(CR^\infty)\Gamma(k-n, CR^\infty) - r_0^{\varphi_1}exp(Cr_0^\infty) \\ \Gamma(-k-n, Cr_0^\infty)) \tag{19}$$

Here, $\Gamma(.)$ signifies higher unfinished gamma function. $C = \frac{m\sigma^2}{P_u\sigma w^2}$, $\varphi_1 = (\alpha n + 3 + \alpha k)$ and $\varphi_2 = n + \frac{3}{2} + 1$.

To express the attainable sum rate of the faraway R_x as follows [21].

$$C_v = \log_2(1 + Q_1 - Q_2)(2 + \frac{\alpha_v^2}{\alpha_w^2}) \tag{20}$$

Here, $Q_1 = \frac{3(D^{\alpha n+3} - R^{\alpha n+3})}{(D^3 - R^3)(\alpha n+3)}\sum_{n=0}^{m-1}\frac{1}{n!}(\frac{m\alpha^2}{P_u})^n$ and $Q_2 = \frac{3(D^{\alpha(n+1)+3} - R^{\alpha(n+1)+3})}{(D^3 - R^3)(\alpha(n+1)+3)}\sum_0^{m-1}\frac{1}{n!}(\frac{m\alpha^2}{P_u})^{(n+1)}$.
In this case, we consider the time limit communication sum rate for the NOMA system supported by RIS. The sum rate of the entire system corresponding to the set bit rate R1, R2 for various service specifications for CU and FAU users can be determined as a further output variable, and such sum rate can be measured by [24]

$$C_{CU,FAU} = R_1(1 - OP_{CU}) + R_2(1 - OP_{FAU}) \tag{21}$$

For HS-UAV NOMA, the services rely under a flat amount in the time limit transmitting mode, and the wireless faded streams specify the system throughput. So, the sum rate can be calculated as [27]

$$C_{HS-UAVNOMA} = \sum_{i=1}^{M}(1 - PD_m) \times R_{th,m} \tag{22}$$

Here, $R_{th,m}$ signifies D_m's rate of propagation period and can be described as $R_{th,m} = \frac{1}{2}\log_2(1+\gamma_{th,m})$. The constants $\frac{1}{2}$means that the communicating method is split into two time slots. The outage rate of user D_m is implied by $\gamma_{th,m}$.
The sum rate of SCMA-NOMA can be expressed as [30]

$$C_{SCMA-NOMA} = \sum_{i=1}^{F}\sum_{n=1}^{N}\sum_{k=1}^{K} s_{k,n}\log_2(1+\Gamma_{k,n}^{FUE,i}) \quad (23)$$

Here, the factor $s_{k,n}$ is a device management with $s_{k,n} = 1$ indicating that a relation occurs between user dataset $s_{k,n} = 0$ indicates that the it has not been connected to any data.
The achievable rate of GFDM-NOMA can be represented as [29]

$$C_{GFDM-NOMA} = \log_2(1+\frac{\alpha.\gamma}{\xi_1}) + \log_2(1+\frac{(1-\alpha).\gamma}{\alpha,\gamma+\xi_2}) \quad (24)$$

IV. Simulation Results

Simulations in this portion have been shown to verify the results generated in Matlab. In the simulations described in section III, we make sure that all users have the very same standard rate criteria and total transmission power specifications. Fig. 7 highlights the differences of outage reliability between the top sections addressed by the hybrid schemes of NOMA. It must be noticed that the study of the necessary outage matches well with the U2X-NOMA schemes at a high SNR rates. In order to highlight U2X-NOMA, we offer small-scale channel advantages for near and distant consumers in the contract. It is an evident from the outcome that the probability of outage for U2X-NOMA is dramatically lower than other discussed schemes and also theoretical outage performance. On the reverse, CR-NOMA places a very low SNR(5dB) in a better position than all other methods described here.

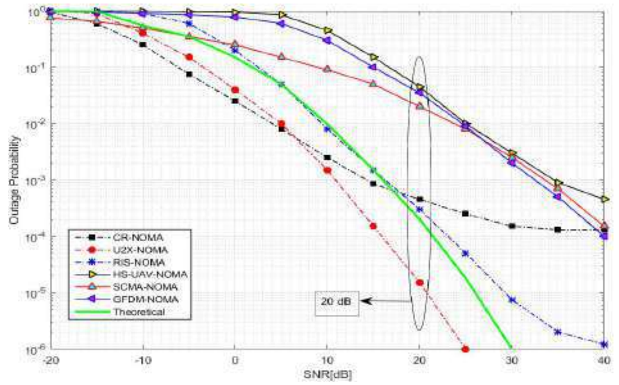


Fig. 7. Outage Probability performance among six hybrid NOMA schemes

In Fig 8 of the proposed hybrid NOMA schemes, the utility of the sum-rate is calculated in the next context. In this portion, due to its benefits of the small scale fading for all users, it is noticeable that U2X-NOMA perform better than the other schemes for high SNR.

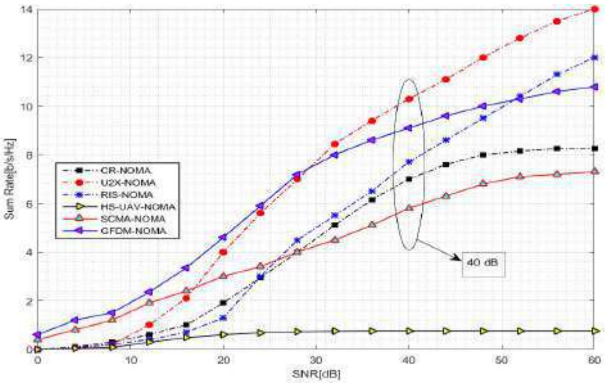


Fig. 8. Sum Rate Comparison among six hybrid NOMA schemes

V. Conclusion

For the hybrid NOMA structures discussed in this paper, we give a comparative theoretical framework as well as a measured structure. We use closed-form calculations for the sum rate and outage probability from the frameworks above. All derived phrases are checked through calculations. U2X-NOMA, considering all performance metrics, outperforms the other hybrid NOMA schemes underlying the configuration of outage and sum rate. The reason behind U2X-NOMA out forming the other schemes described here is U2X-NOMA uses the large scale fading. As a result, it can depict the path loss between transmitter and receiver side users. For future research, combination of UAV and STBC(Space time block coding) in terms of NOMA could be a better solution for next generation wireless communication.

References

[1] M. Z. Chowdhury, M. Shahjalal, S. Ahmed and Y. M. Jang, ”6G Wireless Communication Systems: Applications, Requirements, Technologies, Challenges, and Research Directions,” in IEEE Open Journal of the Communications Society, vol. 1, pp. 957-975, 2020, doi: 10.1109/OJCOMS.2020.3010270.

[2] N. Bhushan et al., ”Network densification: the dominant theme for wireless evolution into 5G,” in IEEE Communications Magazine, vol. 52, no. 2, pp. 82-89, February 2014, doi: 10.1109/MCOM.2014.6736747.

[3] K. N. R. S. V. Prasad, E. Hossain and V. K. Bhargava, ”Energy Efficiency in Massive MIMO-Based 5G Networks: Opportunities and Challenges,” in IEEE Wireless Communications, vol. 24, no. 3, pp. 86-94, June 2017, doi: 10.1109/MWC.2016.1500374WC.

[4] Z. Wei, L. Dai, D. W. K. Ng and J. Yuan, ”Performance Analysis of a Hybrid Downlink-Uplink Cooperative NOMA Scheme,” 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, Australia, 2017, pp. 1-7, doi: 10.1109/VTCSpring.2017.8108407.

[5] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada and T. Nakamura, ”Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access,” 2013 International Symposium on Intelligent Signal Processing and Communication Systems, Naha, Japan, 2013, pp. 770-774, doi: 10.1109/ISPACS.2013.6704653.

[6] S. Ali, E. Hossain and D. I. Kim, ”Non-Orthogonal Multiple Access (NOMA) for Downlink Multiuser MIMO Systems: User Clustering, Beamforming, and Power Allocation,” in IEEE Access, vol. 5, pp. 565-577, 2017, doi: 10.1109/ACCESS.2016.2646183.

[7] Y. Saito, A. Benjebbour, Y. Kishiyama and T. Nakamura, ”System-level performance evaluation of downlink non-orthogonal multiple access (NOMA),” 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, UK, 2013, pp. 611-615, doi: 10.1109/PIMRC.2013.6666209.

[8] B. Makki, K. Chitti, A. Behravan and M. -S. Alouini, ”A Survey of NOMA: Current Status and Open Research Challenges,” in IEEE Open Journal of the Communications Society, vol. 1, pp. 179-189, 2020, doi: 10.1109/OJCOMS.2020.2969899.

[9] L. Lei, D. Yuan and P. Värbrand, ”On Power Minimization for Non-orthogonal Multiple Access (NOMA),” in IEEE Communications Letters, vol. 20, no. 12, pp. 2458-2461, Dec. 2016, doi: 10.1109/LCOMM.2016.2606596.

[10] B. Zheng, Q. Wu and R. Zhang, ”Intelligent Reflecting Surface-Assisted Multiple Access With User Pairing: NOMA or OMA?,” in IEEE Communications Letters, vol. 24, no. 4, pp. 753-757, April 2020, doi: 10.1109/LCOMM.2020.2969870.

[11] X. Yue, Y. Liu, S. Kang, A. Nallanathan and Y. Chen, ”Modeling and Analysis of Two-Way Relay Non-Orthogonal Multiple Access Systems,” in IEEE Transactions on Communications, vol. 66, no. 9, pp. 3784-3796, Sept. 2018, doi: 10.1109/TCOMM.2018.2816063.

[12] G. Gui, H. Huang, Y. Song and H. Sari, ”Deep Learning for an Effective Nonorthogonal Multiple Access Scheme,” in IEEE Transactions on Vehicular Technology, vol. 67, no. 9, pp. 8440-8450, Sept. 2018, doi: 10.1109/TVT.2018.2848294.

[13] Y. Zhang, H. Wang, T. Zheng and Q. Yang, ”Energy-Efficient Transmission Design in Non-orthogonal Multiple Access,” in IEEE Transactions on Vehicular Technology, vol. 66, no. 3, pp. 2852-2857, March 2017, doi: 10.1109/TVT.2016.2578949.

[14] J. Choi, ”NOMA-Based Random Access With Multichannel ALOHA,” in IEEE Journal on Selected Areas in Communications, vol. 35, no. 12, pp. 2736-2743, Dec. 2017, doi: 10.1109/JSAC.2017.2766778.

[15] Y. Yin, Y. Peng, M. Liu, J. Yang and G. Gui, ”Dynamic User Grouping-Based NOMA Over Rayleigh Fading Channels,” in IEEE Access, vol. 7, pp. 110964-110971, 2019, doi: 10.1109/ACCESS.2019.2934111.

[16] B. Wang, K. Wang, Z. Lu, T. Xie and J. Quan, ”Comparison study of non-orthogonal multiple access schemes for 5G,” 2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Ghent, Belgium, 2015, pp. 1-5, doi: 10.1109/BMSB.2015.7177186.

[17] Q. Wu, W. Chen, D. W. K. Ng and R. Schober, ”Spectral and Energy-Efficient Wireless Powered IoT Networks: NOMA or TDMA?,” in IEEE Transactions on Vehicular Technology, vol. 67, no. 7, pp. 6663-6667, July 2018, doi: 10.1109/TVT.2018.2799947.

[18] L. Lv, J. Chen, Q. Ni and Z. Ding, ”Design of Cooperative Non-Orthogonal Multicast Cognitive Multiple Access for 5G Systems: User Scheduling and Performance Analysis,” in IEEE Transactions on Communications, vol. 65, no. 6, pp. 2641-2656, June 2017, doi: 10.1109/TCOMM.2017.2677942.

[19] D. -T. Do, M. -S. V. Nguyen, F. Jameel, R. Jäntti and I. S. Ansari, ”Performance Evaluation of Relay-Aided CR-NOMA for Beyond 5G Communications,” in IEEE Access, vol. 8, pp. 134838-134855, 2020, doi: 10.1109/ACCESS.2020.3010842.

[20] H. Tabassum, M. S. Ali, E. Hossain, M. J. Hossain and D. I. Kim, ”Uplink Vs. Downlink NOMA in Cellular Networks: Challenges and Research Directions,” 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, Australia, 2017, pp. 1-7, doi: 10.1109/VTCSpring.2017.8108691.

[21] T. Hou, Y. Liu, Z. Song, X. Sun and Y. Chen, ”UAV-to-Everything (U2X) Networks Relying on NOMA: A Stochastic Geometry Model,” in IEEE Transactions on Vehicular Technology, vol. 69, no. 7, pp. 7558-7568, July 2020, doi: 10.1109/TVT.2020.2994167.

[22] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao and L. Hanzo, ”Enhancing the Physical Layer Security of Non-Orthogonal Multiple Access in Large-Scale Networks,” in IEEE Transactions on Wireless Communications, vol. 16, no. 3, pp. 1656-1672, March 2017, doi: 10.1109/TWC.2017.2650987.

[23] Y. Liu, Z. Ding, M. Elkashlan and H. V. Poor, ”Cooperative Non-orthogonal Multiple Access With Simultaneous Wireless Information and Power Transfer,” in IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 938-953, April 2016, doi: 10.1109/JSAC.2016.2549378.

[24] A. Hemanth, K. Umamaheswari, A. C. Pogaku, D. -T. Do and B. M. Lee, ”Outage Performance Analysis of Reconfigurable Intelligent Surfaces-Aided NOMA Under Presence of Hardware Impairment,” in IEEE Access, vol. 8, pp. 212156-212165, 2020, doi: 10.1109/ACCESS.2020.3039966.

[25] Q. Wu and R. Zhang, ”Beamforming Optimization for Wireless Network Aided by Intelligent Reflecting Surface With Discrete Phase Shifts,” in IEEE Transactions on Communications, vol. 68, no. 3, pp. 1838-1851, March 2020, doi: 10.1109/TCOMM.2019.2958916.

[26] V. C. Thirumavalavan and T. S. Jayaraman, ”BER Analysis of Reconfigurable Intelligent Surface Assisted Downlink Power Domain NOMA System,” 2020 International Conference on COMmunication Systems NETworkS (COMSNETS), Bengaluru, India, 2020, pp. 519-522, doi: 10.1109/COMSNETS48256.2020.9027303.

[27] X. Li et al., ”A Unified Framework for HS-UAV NOMA Networks: Performance Analysis and Location Optimization,” in IEEE Access, vol. 8, pp. 13329-13340, 2020, doi: 10.1109/ACCESS.2020.2964730.

[28] S. Sharma, K. Deka, V. Bhatia and A. Gupta, ”Joint Power-Domain and SCMA-Based NOMA System for Downlink in 5G and Beyond,” in IEEE Communications Letters, vol. 23, no. 6, pp. 971-974, June 2019, doi: 10.1109/LCOMM.2019.2911082.

[29] X. Zhang, Z. Wang, X. Ning and H. Xie, ”On the Performance of GFDM Assisted NOMA Schemes,” in IEEE Access, vol. 8, pp. 88961-88968, 2020, doi: 10.1109/ACCESS.2020.2994083.

[30] T. Sefako and T. Walingo, ”Application of Biological Resource Allocation Techniques to SCMA NOMA Networks,” 2019 IEEE AFRICON, Accra, Ghana, 2019, pp. 1-7, doi: 10.1109/AFRICON46755.2019.9133767.

[31] P. Guan et al., ”5G Field Trials: OFDM-Based Waveforms and Mixed Numerologies,” in IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1234-1243, June 2017, doi: 10.1109/JSAC.2017.2687718.

[32] Y. Tao, L. Liu, S. Liu and Z. Zhang, ”A survey: Several technologies of non-orthogonal transmission for 5G,” in China Communications, vol. 12, no. 10, pp. 1-15, Oct. 2015, doi: 10.1109/CC.2015.7315054.

Software and Hardware Security of IoT

Ashwini Kumar Singh
M.Tech(ECE)
Department of Electronics and Communication
Indian Institute of Information Technology
Pune, India
ashwinisingh19@ece.iiitp.ac.in

Dr. Nagendra Kushwaha
Assistant Professor
Department of Electronics and Communication
Indian Institute of Information Technology
Pune, India
nagendra@iiitp.ac.in

***Abstract*—With the tremendous growth of IoT application,providing security to IoT systems has become more critical.In this paper, a technique is presented to ensure the safety of Internet of Things (IoT) devices. This technique ensures hardware and software security of IoT devices. Blockchain technology is used for software security and hardware logics are used for hardware security. For enabling a Blockchain, Ethereum Network is used for secure peer-to-peer transmission. A prototype model is also used using two IoT nodes to demonstrate the security logic.**

***Index Terms*—Internet of Things, Blockchain, Security.**

I. Introduction

In recent years the security issue of the IoT ecosystem has become subject of discussion which often troubles the end-users.The internet of things has become one common solution for better and scalable automation technologies. The deployment of the IoT technologies is now in limited applications but soon it is going to be included in every day-to-day activities. With the increase in the application areas of IoT technology, there is a high chance of misuse of these technologies. Therefore, there is an urgent need of ensuring the safety of these devices and make them such that they are effective to deal with all the threats in the future. Existing technology will be not enough to protect the devices.

With the increase of involvement of IoT system in human life there is exponential increase of cyber attacks on the IoT systems [1].Various IoT specific malware are being formulated to perform DOS attacks,eavesdropping attacks on the system [2].Due to it's open structure and involvement of third part, access,leaking the information out of the IoT nodes is not very much difficult.The IoT system is not only attacked at software level rather it is being compromised at the hardware level also.Since sometimes these IoT devices are installed remotely some foreign objects or a leakage path is being formed at the hardware of the IoT system by authorized persons [3].

In order to have a secure communication between IoT nodes,we have to provide security at every level of system i.e from hardware level to the software level.For providing hardware level security we can use dummy or fake logic to make the circuit more complex to understand by other and also we can install self-assessment test to check the proper working of the hardware before configuring it as IoT node. Now if we look for software level security(data transmission level) we can use blockchain technology in order to have secure peer to peer communication along with cryptographics method.The Blockchain has gained so much attention from the world since its launch. Fig 1 shows the basic components of the blockchain technology. Blockchain has the ability to optimize the global structure of the technologies connected with each other through internet [4]. It has mainly two fields that are going to be affected which are:

- It creates a decentralized system which removes the interference of the central server and provides peer to peer interaction.
- It creates a fully transparent and open to all database which brings transparency

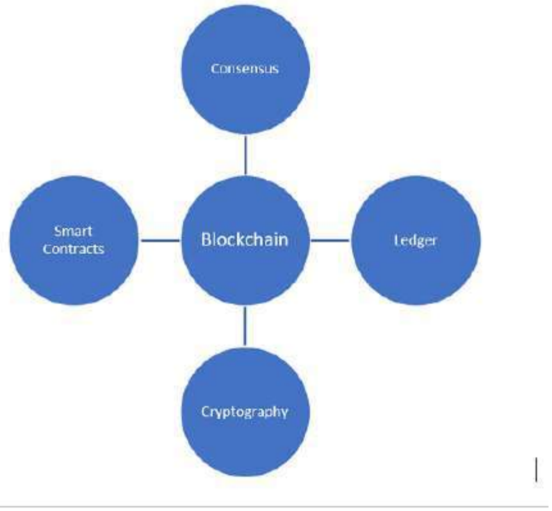


Fig. 1. Basic components of Blockchain

Block has 4 components, first of which is consensus which provides proof of work (PoW) and it verifies the actions occurring around the networks, second is ledger which provide the details of the transaction that occurred between the nodes. Third is cryptography, it make sure that the data on the network and the ledger is safely encrypted and only authorized user can have the access of it and the last one is smart contract which are termed as the verification and validation of the connecting nodes [5].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. IoT Security Concerns

IoT has tremendous amount of application for example in making smart city, improving health, autonomous vehicle and many more. But with kind of connection on which the foundation of IoT is built has many issues so if an unwanted person(hacker) gets access to the system by exploiting the issues can compromise the privacy of users and can also harm them [6].

Thus, while implementing IoT, the security of the systems should be strengthened. Keeping these systems secure is a difficult task because the actual sensor is sometimes placed remotely. Various cluster of the security [7] is shown in the fig 2. Now the security of these device needs a robust solution right from the hardware level to the application layer. Existing technology is not enough for the security of the vastly implemented technology.

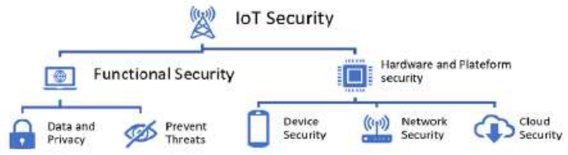


Fig. 2. IoT Security

Here the smart technology called Blockchain will give us a suitable solution to provide end-to-end security to the IoT system. Blockchain provides a decentralized system that stores the database of every transaction made in a network. Rather than having a traditional central database like that of the government, it has a ledger distributed over a network of nodes.

Now blockchain can provide the security for the application layer or we can say it will provide the way to share data between nodes securely, so we have to also look for the threats due to the hardware. Since the IoT specific hardware are generally third-party build so there is a chance that the third-party person may tamper with the circuit and provide a leakage path (Trojan). And also, since it is devices are remotely installed it is rather easy to install foreign materials on the IoT circuitry.

So, end to end security is a must for the IoT environment to provide a seamless experience.

III. Overview of Blockchain

Blockchain's decentralized cryptographic model allows users to trust each other and make peer-to-peer transactions, eliminating the need for intermediaries. This technology is not only affecting the way we use internet, but the global economy is also being revolutionized.

Following are the main parts of the blockchain which helps in completing the infrastructure:

1) Network Node: All the connected nodes maintain a database of transactions made on a blockchain network. The authentication process is also done by mutually decided the function so it eliminates the involvement of the third-party validation process. When a transaction is done, its records are added to the ledger of past transaction, this process is known as 'mining'. The proof of work has to be verified by the other nodes present on the network.
2) Distributed database: The database composed of blocks of information and each block has following data in it:
   - A list of transaction
   - TimeStamp
   - Information which links it to previous chain of the blocks
3) Shared Ledger: The ledger is updated when the transaction is made. It is available for all and it is incorruptible which brings the transparency.
4) Cryptography: The data is secured using crypto mechanism and unauthorized users cannot alter or tamper the data.

The blockchain network can be visualized as shown in fig.3

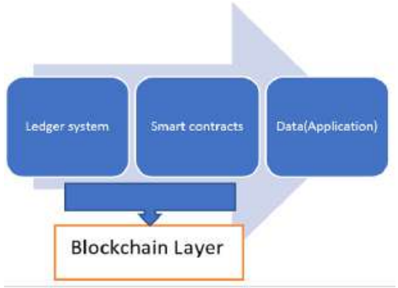


Fig. 3. Blockchain Network

The best example of implementation of blockchain is the Bitcoin and the transaction system in it.

Blockchain is implanted in 3 domains:

1) Public Domain: No-Permissioned area and each and every node send, read data without requiring any permission. Bitcoin and Ethereum are the examples for it.
2) Consortium Domain: Partial permissioned. Only defined nodes can take part in process and permission to read, write data may or may not be made public.
3) Private: Permissioned domain. Only the members whom the network belongs can write transaction. Reading can be made public or restricted to few nodes.This kind of private domain is applied in the industries.

IV. Architecture of IoT Based on Blockchain

The internet of things can have many nodes which will be communicating with each other, so we can build an architecture which would have a distributed connection.

Multiple models for communication:

1) Fundamental function of blockchain network can be used.
   - Peer to Peer messaging
   - Distributed data sharing
   - Differential coordination with the device

But there will be issues while implementing the model like Slow processing due to low end CPU because blockchain requires more memory and high CPU.

Also, the size of the ledger will increase day by day thus making the process slow.

In fig. 4 shows the basic architecture of the blockchain network implementation of technology i.e. mining, cryptography, ledger ,etc. Each IoT node is the member of the network participating in the transaction. Validation is done through mining and node can be a computer, cloud. These IoT devices are the client and theses clients will interact with each other with APIs. Devices create transaction and these transactions are sent to the blockchain nodes for storing and processing data into distributed ledger.

REST-APIs can be used to make communication with the IoT devices and the blockchain. Each of them will have different set of APIs.

Various attacks like "man-in-the-middle" attack can be avoided by establishing the trust between node and IoT devices.

Now while implementing this technology we have to take into the consideration of the delays caused due to blockchain [8] .Therefore we have to adjust our system in order to sustain the data transition delays.

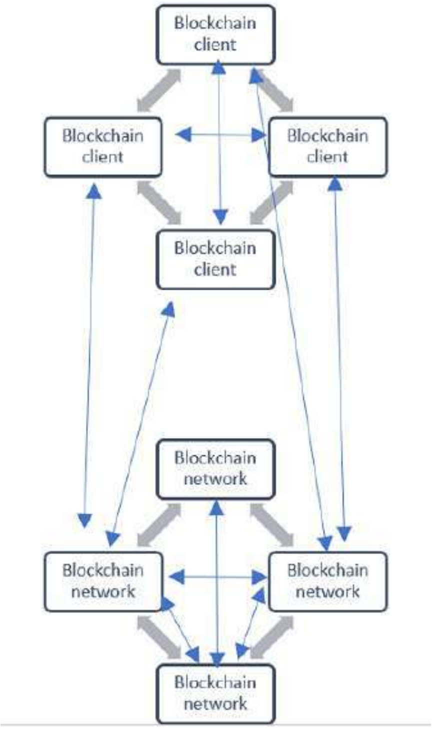


Fig. 4. Distributed Data Sharing

- Various recent research shows that future IoT environment will have multiple blockchains with having distinctive features and services.Also, the implementation of Artificial Intelligence will be there along with the blockchain. So, this kind of arrangements will be robust to outcome the vulnerably.

V. Enhancing the Security of IoT

A. Hardware Security

First line of security has to be implemented on the hardware level. Generally, the threats are known as Hardware trojan i.e. unwanted circuit which is leaking the data.

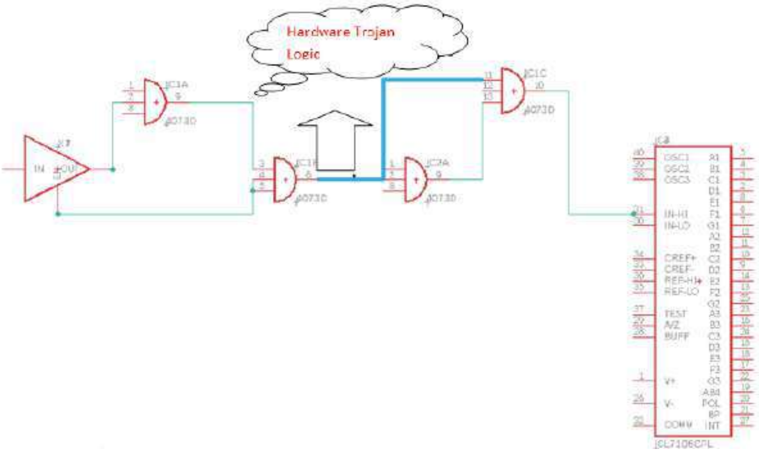


Fig. 5. Hardware Trojan

To avoid this kind of malpractice we can use following methods:

- Installing or fabricating a fake or dummy logic to create confusion to another person.
- Also adding multiple functioning mode of operation like normal mode and malfunction node and as the circuit designer we only know how to configure for normal mode operation.
- Also, often the attackers exploit the empty spaces on the PCB, so installing a Built-in-self-test circuit will help us to determine whether there has been any alteration to original circuit.

Now above will give us some extent of security to the devices because these devices are installed remotely.

B. Securing Data transmission

Now the next aspect and most important one is to provide secure data transmission. In this area there always maximum attack happens.

Following are the area through which we have to ensure for security:

1) Secure Transmission (Communication):Here the blockchain features will come into effect. The IoT devices will utilize a ledger to store the encryption key

in order to send encrypted data.
IoT devices sends the encrypted message using the Public Key of the receiver which is stored in the blockchain network after asking for it.
Then the legal or authenticate receiver can only decrypt the message using its Private Key. So, no third person can have the access of that data.

2) Authentication of the Legal IoT nodes: In this system the sender will digitally sign the message before sending to anyone. Then the receiver will obtain the public key of the sender to verify the signature of the received message.
The process will be as follows:
a) Sender calculates the hash of the message and encrypt it with it's Private key.:
b) This constitute a digital signature and it is transmitted along with the data.:
c) The receiver will obtain the Public Key of the sender through ledger and calculate the hash of the message.:

C. Configuring the legitimate IoT node:

Everyday large number of IoT devices are to be added so there should be ability to differentiate between legitimate and non-legitimate nodes.
As soon as the new device is switched on, this device register itself on the root server of that network. Then the exchange of information takes place and it sends its information to other devices and also receives the same from others.
Now an attack called "Spoofing" can happen at this stage. To avoid that we can use Domain name security extension to provide name resolution of root server.
In Private networks to make sure legitimate Node is connected, root server must authenticate the device before starting any bi-directional communication. This can be done in following ways:

- The credential can be generated by blockchain network and we can install these at the device setup.
- Some credential can be given to user and after initializing the process the device can be registered.

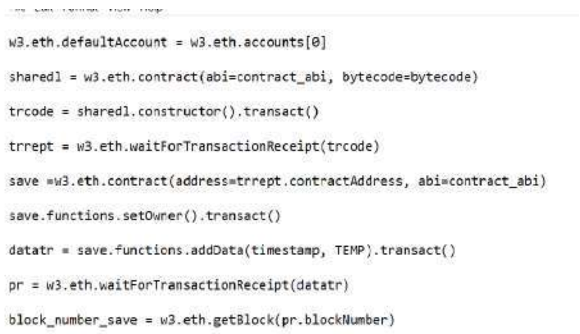
Thus, the blockchain features really help in securing and establishing the IoT environment. The properties of the devices, configuration details are securely stored on the ledger.The hash value feature also helps us to validate the received message. So, having a network of Blockchain connected device i.e. a decentralized system each and every node will have the copy of data. Whenever any device node user wants to access it, all the other nodes must validate it. After the validation then only a transaction can happen.

VI. Implementation

For implementation of the architecture with respect to above section, we have used a Raspberry pi and a laptop.We will use Ethereum's based system library Web3.py Python library and we will set a private Ethereum Blockchain network between Raspberry pi and a laptop.

A. Raspberry Pi configuration

We have to configure the pi in order to run the Blockhain network.Required operating system is "linux/Raspbian GNU" and Python3 must be installed.Also the Pi is connected to internet and a fixed IP address is assigned to it.A default account is created in the network using web3 module and also the IP of the connected node is stored in Pi address book.A part of code is referred as in fig.6
Also a pair of Public and Private key will be generated for each node and the public key of each node is shared along the shared ledger of nodes and private key is being kept safe.

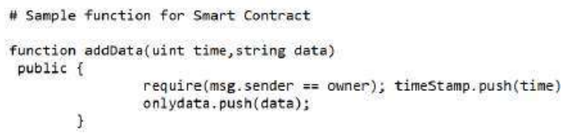


```
w3.eth.defaultAccount = w3.eth.accounts[0]

shared1 = w3.eth.contract(abi=contract_abi, bytecode=bytecode)

trcode = shared1.constructor().transact()

trrept = w3.eth.waitForTransactionReceipt(trcode)

save =w3.eth.contract(address=trrept.contractAddress, abi=contract_abi)

save.functions.setOwner().transact()

datatr = save.functions.addData(timestamp, TEMP).transact()

pr = w3.eth.waitForTransactionReceipt(datatr)

block_number_save = w3.eth.getBlock(pr.blockNumber)
```

Fig. 6. Web3(Ethereum)

B. Blockchain Smart Contract

Now we will use the laptop 2 node of the network. Now a smart contract a formed using senders information condition .The data part in the contract referred in fig. 7 which will be message to be transmitted is encrypted using the public key of node 2 and only the node 2 will able to decrypt the data using its private key.Both the laptop and the pi will use this smart contract in order to transfer data.



```
# Sample function for Smart Contract

function addData(uint time,string data)
 public {
                require(msg.sender == owner); timeStamp.push(time)
                onlydata.push(data);
        }
```

Fig. 7. Important Function of Contract

C. Integration using Python

In order to connect both we will use node id and IP address of the nodes and we will use "addPeer" command and mining will be done at a node.
Once the connection is established,the smart contract is being called and the message is stored at the node2.The message is

decrypted using the private key of the node 2. Each block of message is stored with a time stamp of reception along with the data.Alternatively the data can be appended into a .json file for better transparency.

VII. Conclusion

In this paper we have discussed the security vulnerability of IoT communication and growing need for a robust solution.A prototype of blockchain component is implemented using raspberry pi.The proposed hardware security methodology will help us to counter the hardware trojans and blockchain has been used as a solution to the application/software level security.In this way a decentralized communication system is being established without interference of third party for the validation and the data is transmitted securely between the nodes.This system demonstrates the important features of blockchain in terms of provide a secure environment for communicating between the nodes.

References

[1] Carlos Alberca, Sergio Pastrana, Guillermo Suarez-Tangil, and Paolo Palmieri.2016. Security analysis and exploitation of arduino devices in the internet of things.In Proceedings of the ACM International Conference on Computing Frontiers. Association for Computing Machinery, New York, NY,USA, 437–442. DOI:https://doi.org/10.1145/2903150.2911708

[2] H. Tao, M. Z. A. Bhuiyan, A. N. Abdalla, M. M. Hassan, J. M. Zain and T. Hayajneh, ”Secured Data Collection With Hardware-Based Ciphers for IoT-Based Healthcare,” in IEEE Internet of Things Journal, vol. 6, no. 1, pp. 410-420, Feb. 2019, doi: 10.1109/JIOT.2018.2854714.

[3] Sidhu S, Mohd BJ, Hayajneh T. Hardware Security in IoT Devices with Emphasis on Hardware Trojans. Journal of Sensor and Actuator Networks.2019; 8(3):42.

[4] M. Singh, A. Singh and S. Kim, ”Blockchain: A game changer for securing IoT data,” 2018 IEEE 4th World Forum on Internet of Things (WF-IoT),Singapore, 2018, pp. 51-55,doi: 10.1109/WF-IoT.2018.8355182.

[5] J.Xu,S.Wang,A.Zhou and F.Yang,”Edgence: A blockchain-enabled edge computing platform for intelligent IoT-based dApps,” in China Communications, vol. 17, no. 4, pp. 78-87, April 2020,DOI:10.23919/JCC.2020.04.008.

[6] N.Neshenko, E.Bou-Harb, J.Crichigno, G. Kaddoum and N.Ghani,”Demystifying IoT Security: An Exhaustive Survey on IoT Vulnerabilities and a First Empirical Look on Internet Scale IoT Exploitations”,IEEE Communications Surveys Tutorials, vol. 21,no. 3, pp. 2702-2733, 2019.

[7] A. Mosenia and N. K. Jha, ”A Comprehensive Study of Security of Internet-of-Things,” in IEEE Transactions on Emerging Topics in Computing, vol. 5, no. 4, pp. 586-602, 1 Oct.-Dec. 2017, doi: 10.1109/TETC.2016.2606384.

[8] P.Danzi,A.E.Kalør,Č.Stefanović and P.Popovski, ”Delay and Communication Tradeoffs for Blockchain Systems With Lightweight IoT Clients,” in IEEE Internet of Things Journal,vol. 6, no.2, pp.2354-2365,April 2019, doi:10.1109/JIOT.2019.2906615.

An IoT based System with Edge Intelligence for Rice Leaf Disease Detection using Machine Learning

S. M. Shahidur Harun Rumy
Department of Computer Science & Engineering
Southeast University
Dhaka, Bangladesh
rumywithy@gmail.com

Md. Ishan Arefin Hossain
Department of Computer Science & Engineering
Southeast University
Dhaka, Bangladesh
ishanarefin@gmail.com

Forji Jahan
Department of Computer Science & Engineering
Southeast University
Dhaka, Bangladesh
swarna.4g@gmail.com

Tanjina Tanvin
Department of Computer Science & Engineering
Southeast University
Dhaka, Bangladesh
tanvintonni06@gmail.com

***Abstract*—Bangladesh is one of the top five rice-producing and consuming countries in the world. Its economy dramatically depends on rice-producing. Rice leaf disease is the biggest problem in the agriculture sector. This is the main reason for the reduction of the quality and quantity of the crops. The spread of the disease can be avoided by continuous monitoring. However, manual monitoring of diseases will cost a large amount of time and labor. So, it is a good idea to have an automated system. This paper presents a rice leaf disease detection system using a lightweight Artificial Intelligent technique. We are applying the edge computing concept here. Our edge device is Raspberry Pi. We have processed all our data in Raspberry Pi. We consider three rice plant diseases, namely Brown Spot, Hispa, and Leaf Blast. They are the most common type of rice leaf disease in Bangladesh. We have used clear images of healthy and infected rice leaves with white background. After applying the necessary preprocessing, we have extracted the necessary features from the images. Then we have made an image classification model with various machine learning algorithms by feeding these features. We have learned that the Random Forest algorithm performed the best. By using our image classification model, we have achieved 97.50% accuracy on our edge device.**

Keywords*—*Edge Intelligence, IoT, Machine Learning, Rice Leaf Disease Detection, Image Classification

I. Introduction

As an agricultural country, about 135 million people of Bangladesh consider rice as their staple food. In this region, rice considers as low cost and most nutrient food. More or less, 48% of rural employment is the blessings of it. Besides, half of the agricultural GDP comes from the rice sector. The Rice sector contributes to one-sixth of the national income of Bangladesh. Around 80% of the total irrigated area is under this sector [1]. The livelihood of the people in rural areas depends mostly on rice cultivation. To ensure stable economic growth and maintain desired goals, disease-free rice cultivation can play an important role. So, it is evident that the proper cultivation of rice is a primary concern for Bangladesh.

Therefore, we have come up with an automated system, which will contribute to the development of Bangladesh's agriculture sector. Our proposed system will automatically detect if a leaf is healthy or infected from the leaf's image. Our system is based on the edge computing concept. Edge computing places networked computing resources as close as possible to where data is created. Edge computing is associated with the Internet of Things and the application of small computing devices like Raspberry Pi [2].

We are using Raspberry Pi as our edge device. We process all of our data in Raspberry Pi. Raspberry Pi is a credit-card-sized single-board computer. It is very low in price. It has some GPIO pins to communicate with external components. Its default operating system name is Raspberry Pi, which is based on Debian [3].

We are detecting three rice leaf diseases, namely Brown Spot, Hispa, and Leaf Blast, because these are the most common rice leaf disease in Bangladesh [4].

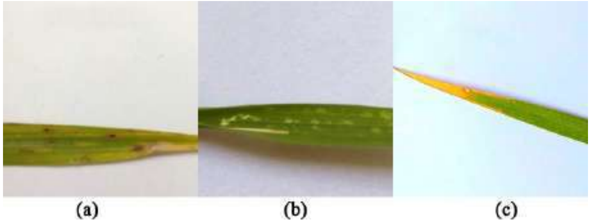


Fig. 1. Infected Leaves (a) Brown Spot, (b) Hispa, (c) Leaf Blast

Brown Spot is caused by Cochliobolus miyabeanus fungus. It causes a prominent spot on the leaves [5]. A species of leaf beetle named Dicladispa armigera is the reason for Hispa. It scrapes the upper surface of leaves and causes white patches [6]. Magnaporthe oryzae is a fungus that causes Leaf Blast disease in rice leaves. It causes grey-green or white spots with dark green or reddish-brown border [7].

We have taken images of healthy and infected leaves with white backgrounds [8]. After preprocessing the images, we have applied machine learning algorithms to make our image classification model. We have tried different machine learning algorithms, but random forest turns out to be the most efficient.

A. Internet of Things:

Internet of Things (IoT) refers to devices embedded with sensors and technologies that are connected to the internet

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

for exchanging information. It can be a self-driving car, an automatic door, a coffee maker, a wearable device, or a washing machine connected to the internet for collecting and sharing information. Most of the time, this information exchanging happens without the interference of humans. In some cases, these devices make use of artificial intelligence [10].

B. *Edge Computing:*

Edge computing is a distributed computing system that processes data on an edge device instead of a data center. It makes real-time data processing effective. This reduces problems with latency and connectivity. It ensures that only critical data will be sent over the network, and no-critical data will be processed on the edge device. This reduces latency for time-sensitive devices [11]. Since edge devices process data themselves, the latency is almost zero.

C. *Machine Learning:*

Machine learning is a method that enables systems to learn automatically from data and improve their accuracy over time from experience without being programmed to do so. It is a branch of artificial intelligence. In machine learning, training data is given to the learning algorithm. Then learning algorithm generates a set of rules from the pattern and feature of the data. Based on these rules, it makes a prediction on new data. The quality of the algorithm defines the accuracy of the prediction. [12].

D. *Random Forest Algorithm:*

Random forest algorithm is a supervised learning technique. This is a popular machine learning algorithm for classification. However, it can be used for regression problems also. It consists of a combination of multiple decision trees. It uses the bagging method for training. The prediction of random forest is based on the majority of all the decision trees. Random forest eliminates the risk of overfitting and it takes less time for training [9].

E. *Motivation:*

Bangladesh is the fourth-largest rice producer. Rice is grown on about 10.5 million hectares. Almost all of the 13 million farm families of Bangladesh grow rice [1]. So, it is essential to produce good quality and quantity of rice. But the main problem is the disease of rice crops. Detection of rice plant disease has always been challenging. It requires constant human observations. But it is much more time-consuming and requires more labor.

This research presents an automatic rice leaf disease detection system. It detects if any of the crops are infected by disease or not. This will make the life of farmers easier, reduce the cost of labor and save time. Our proposed system is cost-friendly also, as we are using Raspberry Pi, a cheap microcomputer.

F. *Research Paper Outline:*

Our research paper has six sections. The sections and their contents are as follows:

- Section 1: Introduction. This section gives the overview, objectives, and motivation of our research report.
- Section 2: Literature Review. This section presents a concise survey of different image processing and machine learning operations applied in rice leaf disease detection.
- Section 3: Methodology. This section gives an overview of the working process.
- Section 4: Implementation. This section presents our proposed work for rice leaf disease detection. It describes the detailed methodology of our proposed work.
- Section 5: Results and Discussion. This section presents the results we have achieved.
- Section 6: Conclusion. The summary and scope of our research report has been discussed here.

II. Literature Reviews

In [13], the image is treated as a matrix of M rows and N columns. Their work extracted color texture using chromatography concepts of CIE XYZ color space. They have extracted color features using CIE L*a*b* color space. They used those features to form a simple metric indicating the roundness of the spot. They have achieved 70% accuracy with 50 sample images.

Authors in [14] divided their work into two phases. First, they prepare the healthy leaf image using HS histogram, and then they prepare diseased leaf image by extracting significant colors. After that, they have used image segmentation to apply the outlier detection method.

Q. Yao et al. [15] used Support Vector Machine (SVM) for detecting rice diseases. The achieved accuracy is 97.2%. After segmenting the diseased spot from the image, shape and texture features were extracted from the spot. Bacterial leaf blight, rice sheath blight, and rice blast disease were classified using SVM.

Orillo et al. [16] used 55 images for Bacterial leaf Blight, 37 images for Brown Spot, and 42 images for Rice Blast. After reducing noise and contrast adjustment, image segmentation was done for feature extracting. Then the back-propagation neural network is applied for the classification.

In this [17], 300 images were used for classification. After converting the image to HSV color format, K-Means clustering is used for image segmentation. Statistical features such as Mean Value, Standard Deviation, and GLCM are calculated for feeding to artificial neuron network (ANN). 90% and 86% accuracy achieved on test dataset for the infected and healthy images, respectively.

Research work in [18] used two datasets. One has 48 images, and another has 23 images. They have used the Simple Linear Iterative Clustering (SLIC) algorithm for over-segmentation of images. They extracted three features, namely color, shape, and texture. They have extracted texture from Gray Level Co-occurrence matrix (GLCM). They have used Random Forest classifier for classification.

The authors of this [19] paper used K-means clustering for the segmentation of the diseased area of the rice leaves. For classification, they used the Support Vector Machine (SVM). They achieved 93.33% accuracy on the training data set but 73.33% accuracy on the test dataset.

The author in [20] used fermi energy-based segmentation to extract the infected parts from the images. The shape of the infected part was extracted using the DRLSE method. To reduce computation time, they have used rough set theory.

III. Methodology

The methodology of the research work can be divided into five steps. They are image preprocessing, image segmentation, feature extraction, training model, and detecting healthy and infected leaves in Raspberry Pi. Block diagram of the system is shown in Fig. 2.

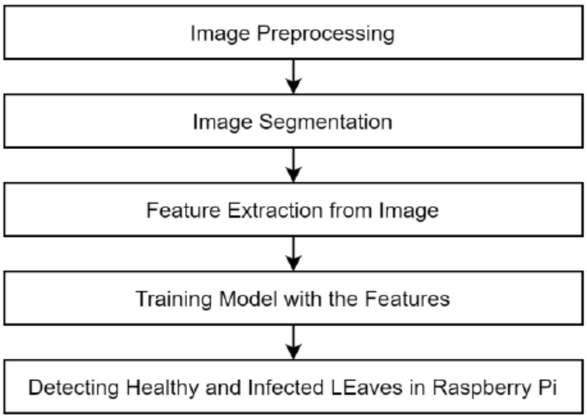


Image Preprocessing
Image Segmentation
Feature Extraction from Image
Training Model with the Features
Detecting Healthy and Infected LEaves in Raspberry Pi

Fig. 2. Block Diagram

The system architecture of our proposed system is shown in Fig. 3.

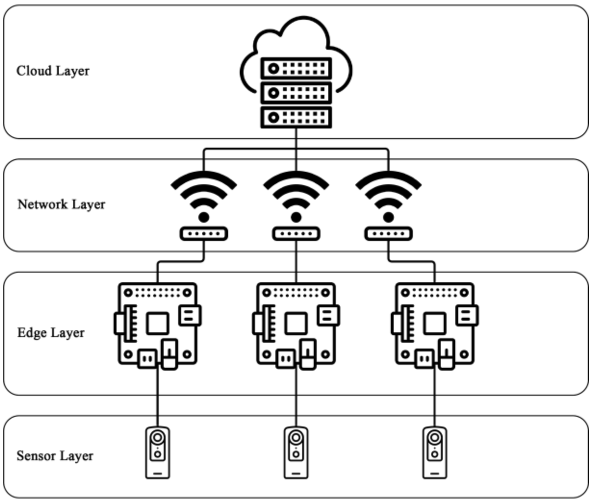


Fig. 3. System Architecture

IV. Implementation

We are working with an image dataset [8] from Kaggle. The dataset is not well sorted. In the healthy folder, there are some infected leaf images, and in the infected folder, there are some healthy leaf images. After separating them, we discovered a lower number of healthy leaf images than infected leaf images. Furthermore, the picture sizes are not consistent.

As we have to make an image classification model which can be run on Raspberry Pi, we are not classifying individual diseases. We will just find out if the leaf is infected or not. That is why we have labeled our dataset in two classes called infected and healthy. For the training dataset, we have taken 310 infected leaf images and 190 healthy leaf images. Furthermore, we have taken 50 infected leaf images and 30 healthy leaf images for the testing dataset. All of the images were selected at random.

A. Components

1) Raspberry Pi 3 Model B v1.2: We are using Raspberry Pi as our edge device. It is packed with Quad-Core 1.2GHz 64bit CPU. It comes with 1GB of RAM. The classification of the leaves is done on Raspberry Pi.

2) Raspberry Pi Camera Board v1.3: We will use a camera module to capture rice leaves images for detecting purposes. It is a 5MP resolution camera. The camera module can be connected to Raspberry Pi through the CSI Camera Port of Raspberry Pi.

B. Creating Model

We are creating an image classification model which can be run on Raspberry Pi. For that reason, we are not going to use complicating filters and image processing methods. Our goal is to get higher performance with a lesser computational load.

1) Image Preprocessing: For reducing computation time, we have resized the images to 256 × 256 pixels. Then we have changed the color format of images from RGB color format to HSV color format. Because it is easier to separate colors in HSV color format. HSV represents Hue, Saturation, and Value part of an image. The output is shown in Fig. 4.

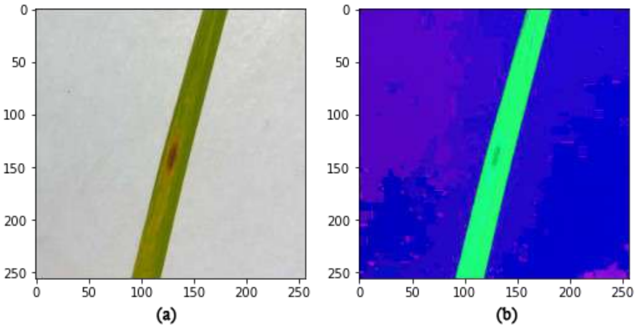


Fig. 4. (a) Image in RGB Format, (b) Image in HSV Format

2) Image Segmentation: After that, we have removed the background from the images. For that, we have performed image segmentation. In segmentation, we have first made a mask that will extract green color from the images, shown in Fig. 5.

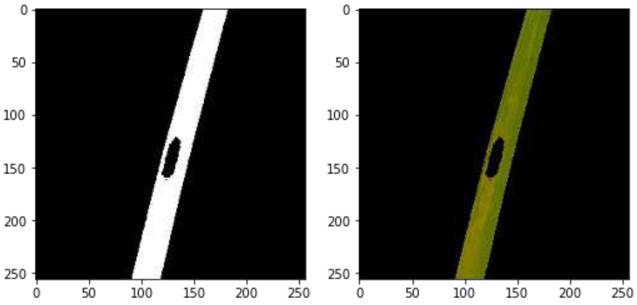


Fig. 5. The output of the green mask

Then we have made another mask that will extract brown color from the images, shown in Fig. 6.

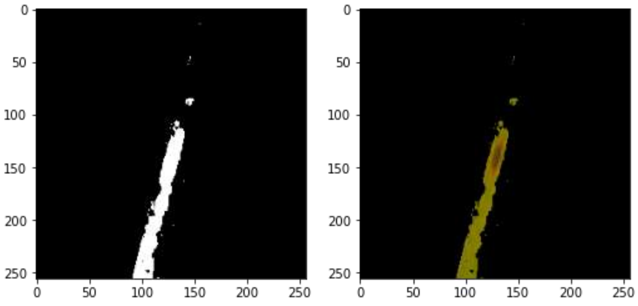


Fig. 6. The output of the brown mask

After that, we have combined these two masks, which will extract leaf from the images. After applying segmentation, the output image becomes like in Fig. 7.

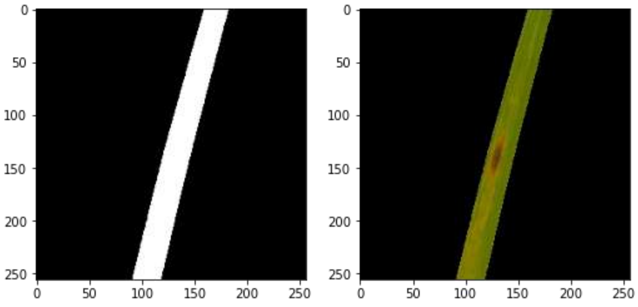


Fig. 7. The output of the final mask

3) Feature Extraction: We have used three feature descriptors to extract features from the images. For extracting color features, we have used Color Histogram. The color distribution in an image is represented by a histogram. Calculating an image's histogram is extremely useful because it offers insight into some of its properties, such as total range, contrast, and brightness. We have used Hu Moments for extracting shape features. A weighted average of the image pixel intensities is used to measure the image moment. Hu moments are moments that are not affected by translation, scaling, or rotation. Texture features were extracted using Haralick. A Gray Level Co-occurrence Matrix, or GLCM, is used to compute Haralick texture features. Haralick Texture is a texture-based approach for quantifying images.

4) Training Model: We have trained our model with six machine learning algorithms. All the models are validated using the 10-Fold Cross-Validation technique. The algorithms are,

a) Random Forest [RF]

b) Naïve Bayes [NB].

c) Decision Trees [DT].

d) Logistic Regression [LR]

e) K Nearest Neighbors [KNN].

f) Support Vector Machine [SVM].

A comparison of the machine learning algorithms is shown in Fig. 8. We can see that Random Forest has performed the best. That is why we have trained the whole dataset with the Random Forest algorithm.

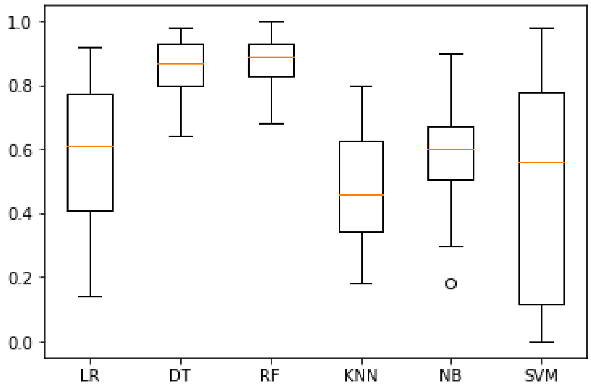
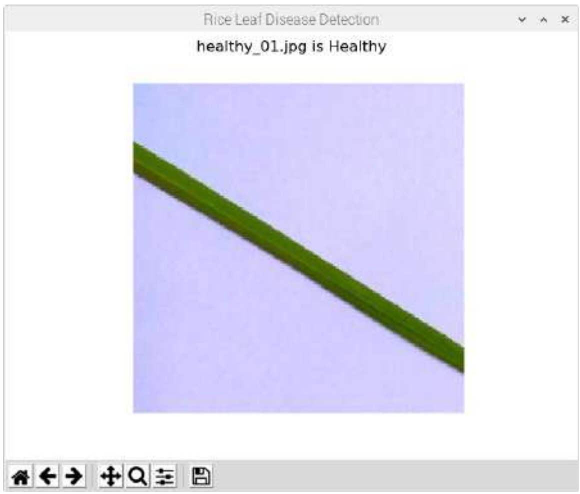


Fig. 8. Comparison of machine learning algorithms.

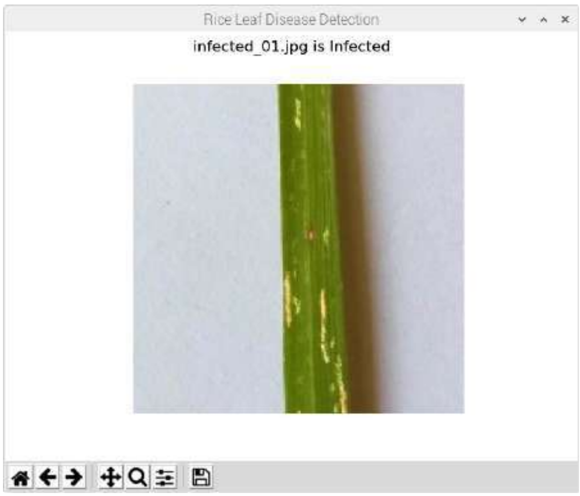
5) Detecting Healthy and Infected Leaves: Our main focus of this research is to implement edge computing concept in rice leaf disease detection. That is why after training our image classification model, we have exported the image classification model. Then we have transferred the image classification model to our Raspberry Pi. Then we have executed our program. We could detect healthy and infected leaves with our system without any hassle. We have shown in Fig. 9 and Fig. 10.



Rice Leaf Disease Detection

healthy_01.jpg is Healthy

Fig. 9. Classified as Healthy



Rice Leaf Disease Detection

infected_01.jpg is Infected

Fig. 10. Classified as Infected

V. Results and Discussion

A. Model Accuracy

After performing prediction on our test dataset, we achieved an impressive 97.50% accuracy. The classification report of our image classification model is given in Table I.

TABLE I. Classification Report

	precision	recall	f1-score	support
healthy	0.94	1.00	0.97	30
infected	1.00	0.96	0.98	50
micro avg	0.97	0.97	0.97	80
macro avg	0.97	0.98	0.97	80
weighted avg	0.98	0.97	0.98	80

The confusion matrix of our image classification model is shown in Fig. 11.

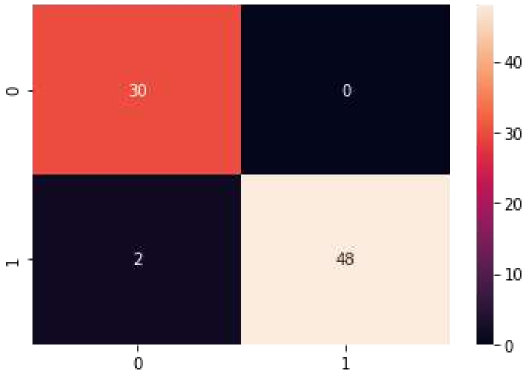


Fig. 11. Confusion Matrix

B. Raspberry Pi Performance

As a microcomputer, the performance we are getting from Raspberry Pi is pretty good. A total of 80 images were used to test the performance. We compared the performance of Raspberry Pi with a computer having a Hexa-Core 3.6GHz 64bit CPU and 8GB of RAM. Comparison of performance between Raspberry Pi and Computer is shown in Table II.

TABLE II. Performance Comparison

	Average time to classify one image
Raspberry Pi	0.20 sec
Computer	0.02 sec

C. Discussion

This paper has introduced the use of edge computing in agriculture. To the best of our knowledge, we are not aware of any other work in this field relating to edge computing. We are able to successfully identify healthy and infected leaves with the help of our system. We have got a good amount of accuracy from our image classification model. As an edge device, the performance of our system is promising. As we are classifying the images in our edge device, it will reduce the problems related to latency and connectivity.

We have randomly taken 30 images. Out of 30 images, only 4 images were the image of infected leaves. So, it can be said that our system can reduce approximately 86.66% of data transmission cost which will improve the latency.

VI. Conclusion and Future Scope

A. Conclusion

Rice leaf diseases can cause a significant amount of loss to the agriculture sector. This paper presents a system that will detect healthy and infected leaves automatically.

Many types of research have been done on detecting rice leaf diseases. But most of them use server-side data processing. The cost of a server is not cheap. That is why we have proposed a system that will classify the images on an edge device. We are using Raspberry Pi as our edge device, which is low in cost and power consumption. Our system can classify a leaf within 0.18 sec to 0.25 sec, and the accuracy of our image classification model is 97.50%. We are pleased with the performance of our system.

We hope our system will contribute a little to the advancement of the agriculture sector. The work has significant economic importance for Bangladesh.

B. Future Scope

Our current image classification model is built with machine learning. In the future, we are thinking of implementing deep learning for detecting leaf diseases. Currently, we are classifying all diseases as infected. Nevertheless, for future work, we plan to classify each disease by its name.

References

[1] "Rice in Bangladesh" Bangladesh Rice Knowledge Bank, http://www.knowledgebank-brri.org/riceinban.php. Accessed 02-01-2021.

[2] "Explaining Edge Computing" Youtube, uploaded by ExplainingComputers, 20-10-2019, https://www.youtube.com/watch?v=0idvaOCnF9E.

[3] "What is a Raspberry Pi?" https://www.raspberrypi.org/help/what-%20is-a-raspberry-pi. Accessed 04-01-2021.

[4] "Rice Disease and Its Managment" Bangladesh Rice Knowledge Bank, http://knowledgebank-brri.org/rice-disease-and-its-managment. Accessed 02-01-2021.

[5] "Brown spot" Bangladesh Rice Knowledge Bank, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/diseases/item/brown-spot. Accessed 02-01-2021.

[6] "Rice hispa" Bangladesh Rice Knowledge Bank, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/diseases/item/blast-leaf-collar. Accessed 02-01-2021.

[7] "Blast (leaf and collar)" Bangladesh Rice Knowledge Bank, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/insects/item/rice-hispa. Accessed 02-01-2021.

[8] Huy Minh Do. "Rice Diseases Image Dataset" Kaggle, 16-11-2019, https://www.kaggle.com/minhhuy2810/rice-diseases-image-dataset.

[9] "Random forest" Wikipedia, https://en.wikipedia.org/wiki/Random_forest. Accessed 15-01-2021.

[10] Alexander S. Gillis. "internet of things (IoT)" Tech Target, https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT. Accessed 19-01-2021.

[11] By Eric Hamilton. "What is Edge Computing: The Network Edge Explained", https://www.cloudwards.net/what-is-edge-computing. Accessed 15-01-2021.

[12] 2 IBM Cloud Education. "Machine Learning" IBM, 15-07-2020, https://www.ibm.com/cloud/learn/machine-learning.

[13] G. Anthonys and N. Wickramarachchi, "An image recognition system for crop disease identification of paddy fields in Sri Lanka," 2009 International Conference on Industrial and Information Systems (ICIIS), Sri Lanka, 2009, pp. 403-407, doi: 10.1109/ICIINFS.2009.5429828.

[14] Reinald Adrian D. L. Pugoy and Vladimir Y. Mariano "Automated rice leaf disease detection using color image analysis", Proc. SPIE 8009, Third International Conference on Digital Image Processing (ICDIP 2011), 80090F (8 July 2011); doi: doi.org/10.1117/12.896494

[15] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu and B. Yang, "Application of Support Vector Machine for Detecting Rice Diseases Using Shape and Color Texture Features," 2009 International Conference on Engineering Computation, Hong Kong, 2009, pp. 79-83, doi: 10.1109/ICEC.2009.73.

[16] J. W. Orillo, J. Dela Cruz, L. Agapito, P. J. Satimbre and I. Valenzuela, "Identification of diseases in rice plant (oryza sativa) using back propagation Artificial Neural Network," 2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Palawan, 2014, pp. 1-6, doi: 10.1109/HNICEM.2014.7016248.

[17] S. Ramesh and D. Vydeki, "Rice Blast Disease Detection and Classification Using Machine Learning Algorithm," 2018 2nd International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, India, 2018, pp. 255-259, doi: 10.1109/ICMETE.2018.00063.

[18] X. Mai and M. Q. -. Meng, "Automatic lesion segmentation from rice leaf blast field images based on random forest," 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR), Angkor Wat, 2016, pp. 255-259, doi: 10.1109/RCAR.2016.7784035.

[19] F. T. Pinki, N. Khatun and S. M. M. Islam, "Content based paddy leaf disease recognition and remedy prediction using support vector machine," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, 2017, pp. 1-5, doi: 10.1109/ICCITECHN.2017.8281764.

[20] H. B. Prajapati, J. P. Shah, and V. K. Dabhi, “Detection and classification of rice plant diseases,” Intelligent Decision Technologies, vol. 11, no. 3, pp. 357–373, 2017, doi: 10.3233/IDT-170301.

On the Event Reporting of Intra/Inter-Cluster Sensor Networks

Raed T. Al-Zubi
Department of Electrical Engineering
The University of Jordan
r.alzubi@ju.edu.jo

Abdulraheem A. Kreishan
Department of Electrical Engineering
The University of Jordan
Amman, 11942, Jordan
Abdulraheem.Kreishan97@gmail.com

Mohammad Q. Alawad
Department of Electrical Engineering
The University of Jordan
Amman, 11942, Jordan
m.qasem@bk.ru

Khalid A. Darabkh
Department of Computer Engineering
The University of Jordan
Amman, 11942, Jordan
k.darabkeh@ju.edu.jo

***Abstract*— Many protocols have been proposed to reduce energy consumption in wireless sensor networks and extend their lifetime. One of the most efficient class of these protocols is known as cluster-based routing protocols. In this paper, we consider a key issue that impacts the performance of this class of protocols. We call it Intra/Inter-Cluster Event-Reporting Problem (IICERP). We show that IICERP significantly degrades the performance of cluster-based routing protocols. Accordingly, we propose an efficient solution for IICERP (SIICERP). Thorough simulations have been conducted to show the performance of different cluster-based protocols with and without SIICERP. The simulations showed that SIICERP significantly improves the performance of cluster-based routing protocols.**

Keywords—Cluster-based networks; Routing protocols; Wireless sensor networks; SIICERP

I. Introduction

Wireless Sensor Networks (WSNs) consist of a large number of tiny and smart devices called sensor nodes [1-3]. Sensors are randomly deployed in remote areas for monitoring purposes [4]. In WSNs, the main task of deployed nodes is collecting data from the sensed region and transmitting it to the base station (BS) or the sink node through predefined channels [5-6]. Energy constraint is considered as the most disturbing issue to many researchers in the field of WSNs [7-8]. Sensor nodes consume a lot of energy due to sensing information, processing data, transmitting and aggregating data. Indeed, the main source of power in these nodes is the battery which is almost impossible to be changed or recharged after the nodes deployment in the network [9-10]. Therefore, well-designed protocols for data collection in WSN are needed to reduce the energy consumption in the network and so extending the lifetime of the network [11-13]. Different energy-aware data collection schemes (i.e., routing protocols) have been proposed for WSNs. These protocols can be classified into three main categories; flat-based routing protocols [14-16], tree-based routing protocols [17-18], cluster-based routing protocols [19-26]. In this paper, we mainly considered the cluster-based protocols which have proven to be the most energy efficient routing protocols [19-26]. In this type of protocols, there are some nodes that have a unique identification over the others. There are two types of nodes: normal nodes and Cluster Head (CH) nodes. The network is divided into many groups called clusters. Each cluster contains one CH node and a number of normal nodes. All of the normal nodes in any cluster transmit their sensed data to their CH which is responsible of aggregating the overall data and transmitting it to the BS. Clustering may be fixed or variable. In fixed clustering protocols, the clusters are formed after the deployment of the nodes and they still the same until the end of the network lifetime. The role of CH is rotated among all nodes in each cluster. In variable clustering protocols, the lifetime of the network is divided into multiple rounds. New clusters are formed in the beginning of each round. There are some protocols that use a third type of nodes called Relay Nodes (RNs) which are responsible of receiving data from CH nodes and then transmitting it to BS. Cluster-based routing protocols efficiently save the energy in WSNs.

All cluster-based routing protocols have a key issue which severely drains the batteries of the nodes. We call this problem as Intra/Inter-Cluster Event-Reporting Problem (IICERP). This problem appears on two levels; intra transmission and inter transmission levels. In first one, if an event occurs in a certain region in the network, then more than one node from the same cluster or different clusters may report on this event by sending similar data packets to their CHs. In the second one, if an event occurs in a certain region in the network, then more than one CH node may report on this event by transmitting similar data packets to the BS.

In this paper, we propose an efficient solution to IICERP problem (SIICERP) which improves the performance of cluster-based routing protocols. SIICERP aims at reporting on an event by no more than one normal node and no more than one CH node. SIICERP is compatible with cluster-based routing protocols. In order to evaluate the performance of SIICERP, we integrate into two cluster-based routing protocols. One of them uses fixed clustering scheme and the other applies variable clustering scheme. The simulation results show that SIICERP significantly saves energy and so extends the lifetime of the network. However, the rest of the paper is organized as follows. Section II presents some related work. Section III presents and explains SIICERP. In section

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

IV, we conduct simulations to evaluate the performance of SIICERP. Finally, concluding remarks are drawn in section V.

II. Related Work

Many cluster-based routing protocols have been proposed in literature [19-26]. However, in our work, to show the effect of the proposed solution, we consider only two cluster-based protocols proposed in [19] and [20]. We select these protocols since one of them is based on variable clustering scheme [19] and the other is based on fixed clustering scheme [20]. In [19], the authors proposed the first and the most popular cluster-based routing protocol called Low Energy Adaptive Clustering Hierarchy (LEACH). The overall network lifetime in this protocol is divided into time slots called rounds. Each round contains one short setup phase followed by one long steady state phase. In the setup phase, selecting CH nodes and forming clusters take place. Selecting CH nodes is done in randomness probabilistic form. Each node in the network generates random number between 0 and 1. Then by comparing this generated number with predefined threshold value, this node will be either a CH or normal node. Then, CH nodes send advertisement messages for all nodes in the network based on carrier sense multiple access (CSMA) protocol to inform them that they were elected as CHs. These messages contain the locations of elected CH nodes. Normal nodes will choose their CHs based on the strength of the received message signals. Each node sends join message to the nearest CH to inform it that it is a member of its cluster. Accordingly, the clusters are formed. In general, the clusters contain different numbers of member nodes. In order to organize data transmission of member nodes in each cluster, CH nodes create time division multiple access (TDMA) schedules. Then, they broadcast these schedules for their member nodes to tell them when each node must transmit its sensed data. Each TDMA schedule is divided into equal time slots. The number of time slots equals the number of member nodes in the cluster and the duration of each slot is the time needed by each node for transmitting its data packet. By this step, the setup phase ends and the steady state phase starts. The steady state phase is broken into time slots called frames. The number of these frames equals the largest number of member nodes exist in a cluster. Each member node in any cluster transmits its data packet to its CH node in its own frame. The CH then aggregates these packets with its own sensed data and transmits the aggregated data packet to the BS. CHs transmit data towards BS based on CSMA protocol.

In [20], an energy efficient routing protocol named Load Balancing Cluster Head (LBCH) was proposed. This protocol uses fixed clustering scheme where each cluster contains one RN, one CH node, and member nodes. The CH collects data packets from member nodes, aggregates them, and transmit these packets to the RN. In a multi-hop routing, RNs transmit to the BS the collected data in each cluster. The network lifetime in this protocol is broken into one short setup phase and one long steady state phase. In setup phase, the BS broadcasts a hello message for all nodes in the network. Then, each node sends its location and energy level to the BS. The BS then divides the network area into equal size subfields called clusters. The length and the width of each cluster are no longer than a predefined threshold distance. After that, the BS selects for each cluster one RN and one CH node (i.e., initial CHs and RNs). The role of RN requires consuming a lot of energy since it receives the overall data packets in the cluster and transmits them to the next RN. Hence, to minimize the energy consumption of these nodes, selecting RNs is done by calculating the magnitudes of nodes in each cluster. The magnitude of a node is linearly proportional to the node's energy and inversely proportional to its distance to the BS. The node that has the largest magnitude in each cluster will be selected as RN for that cluster. After selecting RNs, the BS defines the routing paths and sends to each RN the next hop RN. Choosing a CH for each cluster is done by the BS based on the weight of each node. For the following rounds, the role of RNs and CHs rotates among all nodes in the cluster which ensures load balancing between nodes. The node that was selected as RN in the previous rounds will not be selected as RN in the current round. Moreover, the node that was selected as CH in the previous rounds will not be selected as CH for the current round. After that all nodes in the cluster have played the role of RN and CH for once and consequently the same procedure for choosing RNs and CHs will be repeated. After that, the BS creates TDMA schedules for the nodes in each cluster in order to organize data transmission between nodes and CH. Moreover, to organize data transmission between each cluster and the BS, the BS assigns a unique CDMA code for each cluster. By this step, the setup phase ends and the steady state phase starts. The steady state phase is divided into equal frames called rounds. Each round is divided into number of time slots that is equal to the largest number of sensor nodes exist in a cluster plus an extra time slot that is reserved for controlling purposes of the following round. Each member node transmits its data packet in its time slot to the corresponding CH. The CH aggregates the overall data packets and transmit them to the corresponding RN in its cluster. In a multi-hop routing manner, RNs transmit the collected data from their clusters to the BS. In the extra time slot, the BS chooses a new CHs and RNs for the following round and broadcasts them for the nodes before the starting of the next round.

III. The proposed solution: SIICERP

A) Main Idea

The SIICERP is mainly based on the fact that if there is a group of sensor nodes that their sensing regions are overlapped, then they will unnecessarily send similar packets to report on an event occurs in the intersection region. The sensing region of a node is the region around the node where if an event occurs within it, the node will sense that event. The radius of the sensing region is called the sensing range Rs of a node [27]. Figure 1 shows a group of sensor nodes that will generate similar data packets while sensing the same event.

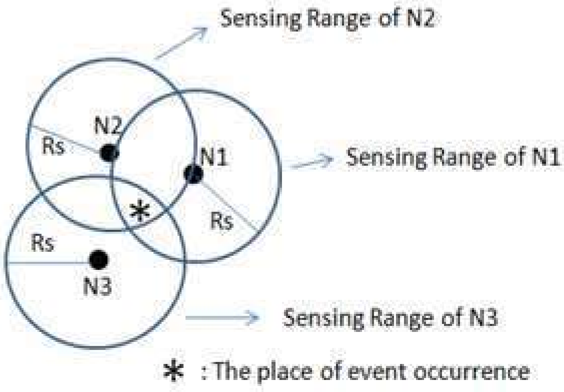


Figure 1. Group of sensor nodes generate similar data packets to report on the same event

Accordingly, SIICERP aims at reporting on any event by only one packet generated by one node such that saving energy in the network as well as achieving the load balancing in the network. To achieve the objective of SIICERP, firstly, it requires that if an event occurs in a certain region, then the node with the highest energy (among all the nodes from the same cluster or different clusters that sense the same event) will report on this event by sending a packet to its corresponding CH. Secondly, the transmission of this packet should be done over at least 2Rs range such that all the possible nodes that could sense this event will know that there is a node has higher energy than them and it reports this event instead of them. This means that a well-designed TDMA schedule is needed to satisfy the following requirements that are considered in the design of SIICERP:

- The sensor nodes (from the same cluster or different clusters) that may sense the same event should be assigned different time slots. This is required to allow these nodes to hear the transmission of each other and hence avoid sending similar packets and so reducing the energy consumption.
- The sensor node that has higher residual energy should be assigned the lower time slot index. This is required to ensure that the event is reported by the node that has the highest residual energy among its neighbors so load balancing is satisfied.

B) *Design of SIICERP*

SIICERP can be integrated into any cluster-based protocol as follows:

- In the setup phase of each round, the CHs are chosen and the clusters are formed based on the used cluster-based routing protocol. Each CH informs the BS the position and the residual energy of each member of its cluster.
- To prevent interference between transmissions from different clusters, the BS generates a unique CDMA code for each cluster. All the nodes of one cluster use the same code for transmission.
- The BS runs an efficient TDMA scheduling algorithm (discussed in the next section) that determines the transmission time for each node in the network.
- Finally, the BS announces the generated TDMA schedule to all nodes in the network using a predefined CDMA code. This TDMA schedule must be followed by all the nodes during the steady state phase. Based on this TDMA schedule, each node will know at which time slot (call it sending slot) and on which CDMA code it should send packets. Additionally, it will know at which time slots (call them listening slots) and on which CDMA codes it should listen. Now, if a node senses an event and the time of its sending slot becomes without receiving any announcement about the same event, then the node will report this event. Otherwise, at the moment of receiving any announcement about the event, the node ignores this event and turns off its transceiver electronics to save its energy.

In a nutshell, SIICERP is compatible with cluster-based routing protocols since it only requires a modification on the TDMA schedule of the routing protocol such that each node knows its sending slot and listening slots.

C) *TDMA Scheduling Algorithm (TSA)*

The TSA is illustrated in the following steps:

Step1: The BS creates for all nodes in the network an initial TDMA schedule based on the residual energy in each node whereas the first time slot is assigned for the node that has the highest energy level in the network and the second time-slot is assigned for the node that has the second energy level and so on. The length of the initial TDMA schedule is impractical; it equals to the number of the nodes in the network. Therefore, in the next steps, the length of the initial TDMA will be reduced by assigning one time-slot for different nodes.

Step2: Initially, each node is assigned a Flag Number (FN) equals to the index of its time-slot assigned in Step1.

Step3: Starting from the node with FN=1, the FNs of all its neighbors (within 2Rs range) are changed to 1. Similarly, for the neighbors of the node with FN=2, their FNs are changed to 2. This process is done for all nodes and the FN of a node should not be changed more than once in this step. In this step, for example, the first time slot is assigned for the node that has originally FN=1 and no one of its neighbors (that their FNs changed to 1) can use this slot for transmission.

Step4: Starting from the second time-slot, if the node in the second time-slot and the node in the first time-slot are not neighbors and not members of the same cluster, then the node in the second time-slot is moved to the first time-slot and its FN is changed to 1 as well as the FNs of all its neighbors are changed to 1 instead of 2. This process is repeated for all nodes. To generalize this step, there are two cases should be clarified:

- If the node has FN equals to the index of its time-slot, this means that its FN was not changed and no neighbor in the previous time-slots. Therefore, it checks if it is possible to move to one of the previous time-slots (starting from the first one); if there is no member from the same cluster in the checked time-slot, then it can be move to it. Otherwise, it cannot be moved.
- If the FN of the node does not equal to the index of its time-slot, this means that there is a neighbor in the time-slot of index equals to FN (that has higher energy). Therefore, the checking process should start from the time-slot of index equals to FN+1.

IV. Performance Evaluation

In this section, we analyze and evaluate the performance of our proposed solution, namely, SIICERP. We evaluate the performance of the protocols proposed in [19] and [20] with and without SIICERP. Thus, respectively, we use the names SIICERP-LEACH and SIICERP-LBCH to refer to the LEACH and LBCH protocols with SIICERP. Table I presents some common simulation parameters used in [19] and [20].

TABLE I. The common simulation parameters

Parameter	Value
Data packet size	1024 bits
Control packet size	176 bits
Initial energy	2.4 J
ETX (Energy for transferring one bit)	50 nJ/bit
ERX (Energy for receiving one bit)	50 nJ/bit
EDA (Energy for data aggregation)	5 nJ/bit
Transmit amplifier energy in free- space	10 PJ / bit / m^2
Transmit amplifier energy in multi-path	0.0013 PJ / bit / m^4
Simulation area	100 m x100 m
Distribution of nodes over the simulation area	Random
Base-station location	(50 m, 125 m)
Sensing range for each sensor	10 m

To evaluate and compare the performance of SIICERP-LEACH and LEACH protocols, we conduct several simulation scenarios. Figure 2 shows the result of one simulation scenario. In this scenario, the number of simulation rounds is 1000 and the location of one event is randomly generated per round (i.e., the total number of generated events during the simulation time is 1000 events). In this scenario, we want to focus on one part of IIER problem, which is sending similar packets from normal nodes to their CHs to report same events. Therefore, we change the number of all nodes in the network and keep the average number of CHs fixed. This is done by selecting the proper percentage of CHs (p) for each number of nodes; p=0.1 for 100 nodes, p=0.034 for 300 nodes, p=0.02 for 500 nodes, and p=0.0143 for 700 nodes. Figure 2 shows that, in SIICERP-LEACH protocol, the total number of sent packets to CHs equals to the total number of events (i.e., 1000 events). However, the figure shows that the effect of IIER problem (i.e., sending more similar packets to report one event) on the performance of LEACH protocol severely increases with increasing the number of nodes in the network. This is because increasing the total number of nodes in the network will increase the density of the nodes around each event which means more similar packets will be unnecessarily sent to CHs to report the same event.

Figure 3 shows the results of another simulation scenario. In this scenario, we try to show the effect of the second part of IIER problem (i.e., sending similar packets from different CHs to the BS in order to inform about the same events). Therefore, we fix the number of nodes to be 300 and change the average number of CHs (i.e., by changing the percentages p of CHs). In this scenario, the number of simulation rounds is 1000 and the location of one event is randomly generated per round. It can be shown from this figure that the performance of SIICERP-LEACH is fixed (i.e., total number of sent packets to BS equals to the total number of events). On the other hand, the performance of LEACH degrades with increasing the percentage of CHs. This is because increasing the number of CHs means more overlapped clusters and so more different CHs report the same event. To show the effect of SIICERP on the overall performance of LBCH protocol, we conduct a third simulation scenario. In this scenario, the percentage of clusters is 0.05 and the locations of events per round are distributed on a grid shape. The results of this simulation scenario are shown in Figure 4. From this figure, we can see that SIIER effectively improves the performance of LBCH protocol.

V. Conclusion

In this paper, we proposed an efficient solution for a problem in cluster-based routing protocols for WSN, namely, SIICERP. We called this problem intra/inter-cluster event-reporting problem. By integrating SIICERP solution into cluster-based protocols, only one packet will be sent by one node (with the highest energy) to report on an event instead of reporting on the event by many packets sent by different neighbor nodes. Accordingly, SIICERP results in saving energy and achieving load balancing in WSNs.

References

[1] S. A. Sibi and R. V. Prabhu, "Survey on Clustering and Depletion of Energy in Wireless Sensor network," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 1341-1345.

[2] T. Jagannadha Swamy, G. Ramamurthy and P. Nayak, "Optimal, Secure Cluster Head Placement Through Source Coding Techniques in Wireless Sensor Networks," *IEEE Communications Letters*, vol. 24, no. 2, pp. 443-446, Feb. 2020.

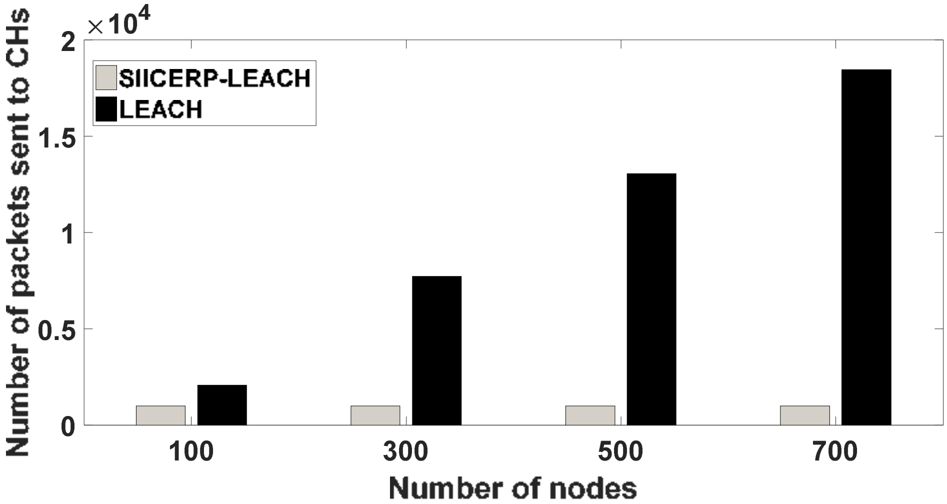


Figure 2. Total number of packets sent to CHs versus number of nodes. One event per round is generated in a random location on the area of the network

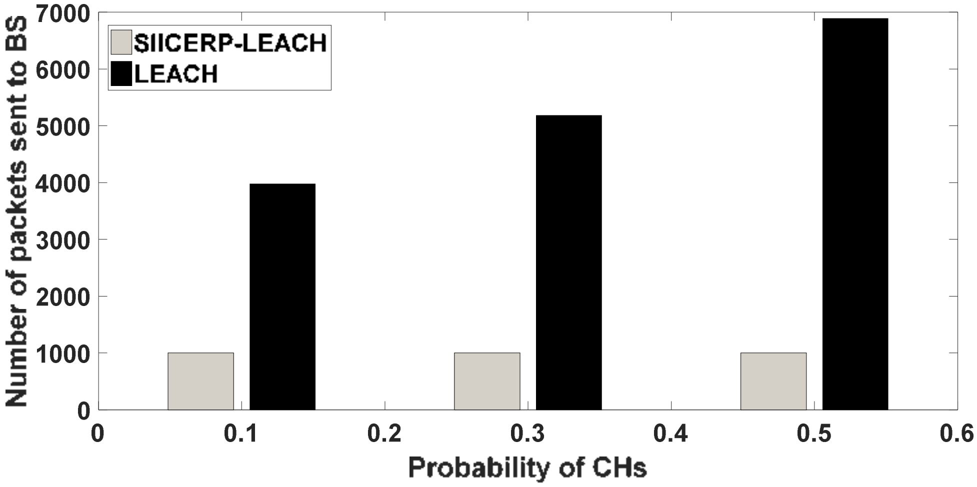


Figure 3. Total number of packets sent to BS versus probability of nodes. One event per round is generated in a random location on the area of the network

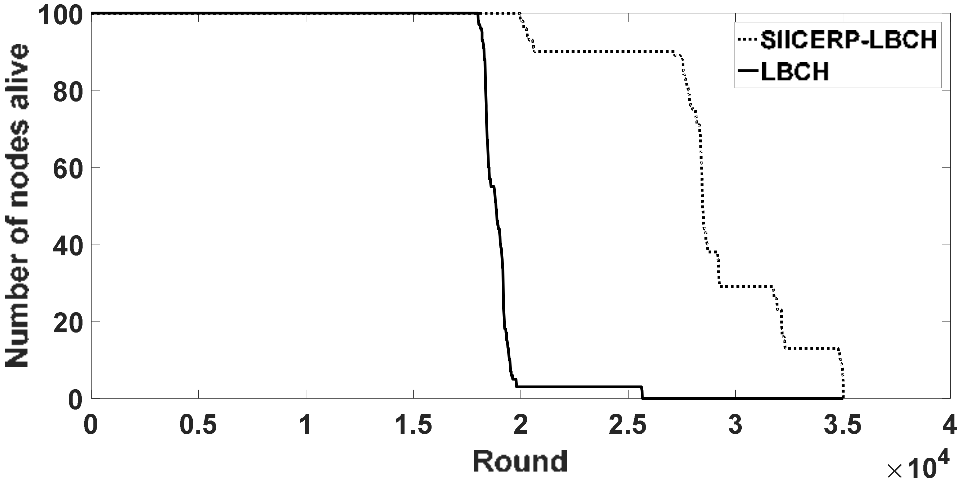


Figure 4. Number of nodes alive per simulation round. The locations of events per round are distributed on a grid shape; all nodes have data to send in each round

[3] L. Chen, W. Liu, D. Gong and Y. Chen, "Clustering and Routing Optimization Algorithm for Heterogeneous Wireless Sensor Networks," *2020 International Wireless Communications and Mobile Computing (IWCMC)*, Limassol, Cyprus, 2020, pp. 407-411.

[4] X. Liu, K. Mei and S. Yu, "Clustering Algorithm in Wireless Sensor Networks Based on Differential Evolution Algorithm," *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chongqing, China, 2020, pp. 478-482.

[5] K. A. Darabkh, J. N. Zomot, Z. Al-qudah and A. F. Khalifeh, "IEDB-CHS-BOF: Improved Energy and Distance Based CH Selection with Balanced Objective Function for Wireless Sensor Networks," in p*roc. 2020 6th International Workshop on Internet of Things: Networking Applications and Technologies (IoTNAT 2020) in conjunction with 5th IEEE International Conference on Fog and Mobile Edge Computing (FMEC 2020*), Paris, France, 2020, pp. 275-279.

[6] K. Darabkh, J. Zomot, and Z. Al-qudah, "EDB-CHS-BOF: Energy and Distance Based Cluster Head Selection with Balanced Objective Function Protocol," *IET Communications, Special Issue: Future of Intelligent Wireless LANs*, vol. 13, no. 19, pp. 3168 – 3180, 2019.

[7] S. Jain and N. Agrawal, "Development of Energy Efficient Modified LEACH Protocol for IoT Applications," *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, Bhimtal, India, 2020, pp. 160-164.

[8] W. Dargie and J. Wen, "A Simple Clustering Strategy for Wireless Sensor Networks," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1-4, June 2020, Art no. 7500804, doi: 10.1109/LSENS.2020.2991221.

[9] B. Kumar, U. K. Tiwari and S. Kumar, "Energy Efficient Quad Clustering based on K-means Algorithm for Wireless Sensor Network," *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, India, 2020, pp. 73-77.

[10] M. Rajkumar, A. Sureshkumar and J. Karthika, "An Enhanced Energy Efficient Clustering Approach for Wireless Sensor Networks," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2020, pp. 672-675.

[11] P. K. Mishra and S. K. Verma, "A survey on clustering in wireless sensor network," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-5.

[12] W. Neji, S. B. Othman and H. Sakli, "T-LEACH: Threshold sensitive Low Energy Adaptive Clustering Hierarchy for Wireless Sensor Networks," *2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, Monastir, Tunisia, 2020, pp. 336-342.

[13] Priyanka Rawat, Kamal Deep Singh, Jean-Marie Bonnin, Hakima Chaouchi. Wireless sensor networks: a survey on recent developments and potential synergies. Journal of Supercomputing, Springer Verlag, 2013, ⟨10.1007/s11227-013-1021-9⟩. ⟨hal-00955283⟩.

[14] H. Echoukairi, K. Bourgba, M. Ouzzif, "A Survey on Flat Routing Protocols in Wireless Sensor Networks". In: E. Sabir, H. Medromi, M. Sadik (eds) Advances in Ubiquitous Networking. UNet 2015. Lecture Notes in Electrical Engineering, vol 366. Springer, Singapore. https://doi.org/10.1007/978-981-287-990-5_25.

[15] M. Anisi, A. Abdullah, S. Razak, M. Ngadi, "An overview of data routing approaches for wireless sensor networks," *Sensors (Basel),* pp. 3964–3996, vol. 12, no. 4, Mar 2012.

[16] A. Kanavalli, D. Sserubiri, P. D. Shenoy, K. R. Venugopal and L. M. Patnaik, "A flat routing protocol for sensor networks," *2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS)*, Delhi, 2009, pp. 1-5. doi: 10.1109/ICM2CS.2009.5397948.

[17] B. Gong and T. Jiang, "A tree-based routing protocol in wireless sensor networks," *2011 International Conference on Electrical and Control Engineering*, Yichang, pp. 5729-5732, 2011. doi: 10.1109/ICECENG.2011.6057953.

[18] S. Hong and K. Han "Tree-based routing algorithms on wireless sensor networks: survey", *Journal of Systems and Information Technology*, vol. 16 no. 2, pp.113-121, 2014, https://doi.org/10.1108/JSIT-11-2013-0065.

[19] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, " Energy-efficient communication protocol for wireless microsensor networks," *Proceedings of the 33rd Hawaii International Conference on System Sciences*, USA, 2000.

[20] R. T. Al-Zubi, N. Abedsalam, A. Atieh and K. A. Darabkh, "LBCH: Load Balancing Cluster Head Protocol for Wireless Sensor Networks", *INFORMATICA*, Vol. 29, No. 4, pp. 633–650, 2018.

[21] R. T. Al-Zubi, N. Abedsalam, A. Atieh and K. A. Darabkh, "Lifetime-Improvement Routing Protocol for Wireless Sensor Networks," *Proceedings of 15th IEEE International Multi-Conference on Systems, Signals, and Devices (SSD)*, Hammamet, Tunisia 2018.

[22] Khalid A. Darabkh, Wijdan Y. Albtoush, and Iyad F. Jafar, "Improved Clustering Algorithms for Target Tracking in Wireless Sensor Networks," *Journal of Supercomputing*, vol. 73, no. 5, pp 1952–1977, May 2017.

[23] Khalid A. Darabkh, Noor J. Al-Maaitah, Iyad F. Jafar, and Ala' F. Khalifeh, "EA-CRP: A Novel Energy-aware Clustering and Routing Protocol in Wireless Sensor Networks, *Computers and Electrical Engineering*, vol. 72, pp. 702-718, November 2018.

[24] Khalid A. Darabkh, Wala'a S. Al-Rawashdeh, Raed T. Al-Zubi, and Sharhabeel H. Alnabelsi‚ "C-DTB-CHR: Centralized Density- and Threshold-based Cluster Head Replacement Protocols for Wireless Sensor Networks," *Journal of Supercomputing*, vol. 73, no. 12, pp. 5332-5353, December 2017.

[25] Khalid A. Darabkh, Wala'a S. Al-Rawashdeh, Mohammed Hawa, and Ramzi Saifan "MT-CHR: A Modified Threshold-based Cluster Head Replacement Protocol for Wireless Sensor Networks," *Computers and Electrical Engineering*, vol. 72, pp. 926-938, November 2018.

[26] Khalid A. Darabkh, Mohammad Z. El-Yabroudi, and Ali H. El-Mousa, "BPA-CRP: A Balanced Power-Aware Clustering and Routing Protocol for Wireless Sensor Networks," *Ad hoc Networks*, vol. 82, pp. 155-171, January 2019.

[27] Avinash More and Vijay Raisinghani, "A survey on energy efficient coverage protocols in wireless sensor networks ", Journal of King Saud University – Computer and Information Sciences, vol. 29, pp. 428–448, 2017.

A Yet Efficient Event Reporting Protocol for Wireless Sensor Networks

Raed T. Al-Zubi
Department of Electrical Engineering
The University of Jordan
Amman, 11942, Jordan
r.alzubi@ju.edu.jo

Mohammad Q. Alawad
Department of Electrical Engineering
The University of Jordan
Amman, 11942, Jordan
m.qasem@bk.ru

Abdulraheem A. Kreishan
Department of Electrical Engineering
The University of Jordan
Amman, 11942, Jordan
Abdulraheem.Kreishan97@gmail.com

Khalid A. Darabkh
Department of Computer Engineering
The University of Jordan
Amman, 11942, Jordan
k.darabkeh@ju.edu.jo

Abstract—**Event reporting in Wireless Sensor Networks (WSNs) has always been a major challenge in terms of energy consumption. Event reporting requires sensing the event and sending a reporting packet from the sensor to the centralized base station (BS). However, sensor's energy is limited and stored in a non-rechargeable battery. Therefore, several event-reporting (or data collection) protocols have been proposed to improve energy consumption in WSNs, and consequently, to extend their lifetime. In this paper, we propose an efficient event reporting protocol for WSNs which mainly aims at reporting any event by no more than one sensor node such that energy saving is satisfied in the whole network. To achieve this goal, our proposed protocol is mainly based on the following features: it is a cluster-based protocol, it is a multi-hop routing protocol, it applies distributed data aggregation, and it employs variable clustering and cluster head selection. Simulations show that our proposed protocol significantly extends the lifetime of WSNs compared to other related protocols.**

Keywords—Event Reporting; Cluster-based; Data Aggregation; Multi-hop Routing; WSNs; ERP-DDA

I. Introduction

To monitor a wide region via a WSN, the sensors are randomly distributed over this region [1-2]. The function of the sensors is summarized as follows: sensing the surrounding environment, converting the sensed information into data packets, and sending the data packets over a wireless channel to a base station (BS) [3-5]. Even though WSNs have different important and useful applications, they have many issues and challenges [6-7]. In this paper, we consider the issue of energy constraint in WSNs. In sensors, the main source of energy is a non-rechargeable battery which almost cannot be changed after the deployment of the sensors in the network [8-9]. Accordingly, several energy-aware data collection protocols (i.e., routing protocols) for WSNs have been proposed in the literature [7-12]. The cluster-based protocols have been proven to be efficient energy-aware routing protocols for WSNs [13-20]. Therefore, in this paper, we mainly consider this category of protocols. According to this type of protocols, the nodes are grouped into many clusters. For each cluster, there is one Cluster Head (CH) and different number of normal nodes. The normal nodes in each cluster send their data to their corresponding CH. After that, each CH aggregates the overall data (i.e., removes redundancy in the collected data) and sends it to the BS. The network lifetime is divided into time intervals called rounds. The formation of clusters may be done in a variable manner as in [13] (i.e., it is changed every round) or in a fixed manner as in [14] (i.e., the clustering is done after sensors deployment and it is not changed during the network lifetime).

After studying different event reporting and data collection protocols proposed for WSNs, we have determined some of their elegant features that could be employed to propose a new energy-aware event reporting protocol for WSNs. These features mainly are the cluster-based, the multi-hop, and the variable clustering. In addition to these features, we found that most cluster-based routing protocols (fixed or variable clustering) have a critical problem that severely degrades their performance in terms of energy saving. We call this problem as Intra/Inter-Cluster Event-Reporting Problem (IICERP). This problem appears on intra transmission and inter transmission levels. In the intra transmission level, if an event occurs in a certain area in the WSN, then different nearby nodes from the same cluster or different clusters will unnecessarily send similar packets to their CHs in order to report this event. In the inter transmission level, if an event occurs in a certain area in the network, then more than one nearby CH will unnecessarily send similar packets to inform the BS about the event.

To this extent, we present the works that are quite related to our proposed protocol. In [18], the Load Balancing Cluster Head (LBCH) protocol was proposed. LBCH applies fixed clustering technique. Each cluster has one CH node, one Relay Node (RN), and many member nodes. The member nodes, in each cluster, sense the events and then send reporting packets to their corresponding CH. After that, the CH aggregates the received packets and resends these packets to the RN. In a multi hop manner, RNs send to the BS the received packets. According to this protocol, the lifetime is divided into setup phase and steady state phase that is divided into time periods called rounds. In the setup phase, the BS sends a hello message to all nodes in the network. After that, the nodes send their locations and energy levels to the BS. Then, the BS divides the network area into subfields called clusters. These clusters are fixed over the lifetime of the network. After that,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

the BS, chooses one RN and one CH node for each cluster. The selection of the RN and CH is based on two calculated values called magnitude and weight, respectively. Thereafter, the BS determines for each RN the next hop RN. For the following rounds, the selection process of RNs and CHs is repeated in each round to ensure load balancing in the network. In [19], the Round-Robin Cluster Header (RRCH) protocol was proposed. The RRCH is a fixed clustering protocol which achieves load balancing and a high energy efficiency in WSNs. In the setup phase, clustering process is performed only once. In the steady state phase, there is no RNs selection. However, for each round, the selection of CH nodes for each cluster is done according to round robin method. In [20], the first fixed clustering routing protocol based on a well-known protocol Low Energy Adaptive Clustering Hierarchy (LEACH) was proposed. The new protocol is called LEACH-F. The clusters are constructed during the setup phase by applying a centralized cluster formation algorithm. In the steady state phase, the CHs selection is done in a round robin manner. In [21], the dynamic round time-based fixed LEACH scheme (we call it Adaptive LEACH-F) was proposed. Its main idea is to reduce the problem of the fixed round time in LEACH-F. The round time in the Adaptive LEACH is adaptively changed according to the current energy of the member nodes (not to their initial energy) and the total energy consumption in the cluster for that round. This leads certainly to reducing the possibility of early death of CHs in addition to extending the network lifetime.

In [22], the Adaptive Energy-aware Fixed Clustering Protocol (AEA-FCP) was proposed. This protocol is based on adaptive fixed clustering. In the setup phase, the network is divided into a number of clusters that are fixed during the network lifetime. In the steady state phase, the RNs route over a multi-hop path the aggregated data from the CHs to the BS. The selection of RNs and CHs is performed in each round based on the residual energy in the nodes. The CH and RN in a cluster switch their roles if their residual energy is less than certain percentages of the average residual energy of the cluster when they start working as CH and RN nodes. AEA-FCP is evaluated under two scenarios; continuous availability of data and event-based availability of data. In [23], the Self-incentive and Semi Re-clustering (SISR) protocol was proposed. It is a fixed clustering protocol. In the setup phase, each node in the network elects itself as a candidate CH with probability P and then broadcasts an *ADVERTISE_MESSAGE* with the initial radio range R_i. Then, the node gradually increases R_i until it receives at least one *ADVERTISE_MESSAGE* from other nodes. Accordingly, the node checks the probability (P) values of other nodes. If one of the nodes has higher value of P, it considers this node as its CH. Otherwise, the node gives up the competition. The selected CHs send an *INVITE_MESSAGE* and wait for a feedback from the normal nodes. When the normal node finds its CH, it sends a *JOIN_REQ_MESSAGE* to its CH to inform it about its decision. After the construction of clusters, each CH determines its CH sequence based on the signal strength of *JOIN_REQ_MESSAGE* from the normal nodes in its cluster. In the steady state phase, *HEARTBEAT_MESSAGEs* are broadcasted by the CHs to their member nodes. The member nodes that do not respond by *HEARTBEAT_ACK_MESSAGE* are considered as dead nodes. However, alive nodes send their *HEARTBEAT_ACK_MESSAGEs* that include their incentive values to be CHs.

In this paper, we propose an Event Reporting Protocol Based on Distributed Data Aggregation (ERP-DDA) for WSNs. The design of ERP-DDA is based on the good features of related protocols. In particular, it is a multi-hop and variable clustering protocol as well as it considers and solves the problem of IICERP via distributed data aggregation. For distributed data aggregation, we propose an algorithm which strictly implies that any node senses an event should, first, listen to its neighbors which have higher residual energy and could sense the same event. Then, if the node hears that one of its neighbors reported the same event, it will ignore this event and will not send any packet to report the event. Therefore, the redundancy in packet reporting is removed in a distributed manner (i.e., distributed data aggregation). The process is not as in most cluster-based protocols where data aggregation is centralized (i.e., it is only done by the CH in each cluster). Hence, distributed data aggregation ensures that any event is reported by no more than one sensor. This leads to reducing the energy consumption and extending the network lifetime. Simulation results show that ERP-DDA significantly saves energy and so extends the network lifetime compared to other related protocols. It is worthwhile to mention that the novelty of our work can be summarized as follows: to the best of our knowledge, no previous work considered the integration of multi-hop feature, variable clustering feature, and specifically the problem of IICERP into one protocol. However, the rest of the paper is organized as follows. Section II presents and explains ERP-DDA. In Section III, we conduct simulations to evaluate and compare the performance of ERP-DDA with other protocols. Finally, concluding remarks are drawn in Section IV.

II. ERP-DDA Protocol

The ERP-DDA mainly aims at reducing the energy consumed for reporting events in WSN in order to extend the network lifetime. The main features of ERP-DDA that help in achieving this goal are summarized as follows: it is a variable clustering and multi-hop protocol as well as it considers and solves IICERP problem. These features are explained by presenting the design of ERP-DDA.

A) *ERP-DDA Design*

According to ERP-DDA, the network lifetime is divided into multiple time intervals called rounds. Each round is divided into setup phase and steady state phase. In the setup phase of each round, the following steps are executed:

Step 1: All the alive nodes send to the BS its identification number, residual energy, and position (assuming the nodes have GPS or they apply a localization algorithm).

Step 2: The BS runs the algorithm proposed in LEACH protocol [20] for CHs selection. Other algorithms can be employed but in our work, we select the one proposed in LEACH protocol [20] because it is an energy efficient and satisfies load balancing.
Step 3: The BS runs Dijkstra's algorithm [24] to find the best route (in terms of minimum energy) between each CH and the BS through other CHs.
Step 4: The BS divides the nodes into clusters where each node is associated to the nearest CH. To prevent interference between transmissions from different clusters, the BS generates a unique CDMA code for each cluster. All the nodes of one cluster use the same code for transmission.
Step 5: To solve the problem of sending more than one packet in order to report the same event (we call this problem as IICERP), the BS runs our proposed algorithm called Listening and Sending Scheduling Algorithm (LSSA) that is explained in the next subsection. This algorithm generates a Time Division Multiple Access (TDMA) schedule that is followed by all nodes during the steady state phase. Based on this schedule, each node knows at which time slot (we call it sending slot) and on which CDMA code it will send its packets. Also, it knows at which time slots (we call them listening slots) and on which CDMA codes it will listen to other nodes. At this point, if a node senses an event and the time of its sending slot passes without receiving any announcement from its neighbors about the same event, then the node will report this event. Otherwise, at the moment of receiving any announcement about the event, the node ignores this event and turns off its transceiver electronics to save its energy.
Step 6: Finally, the BS announces to all nodes information about the result of LSSA, the corresponding CH of each node, and the next hop of each CH.

B) Listening and Sending Scheduling Algorithm

The LSSA mainly relies on the fact that if there is a group of sensor nodes where their sensing regions are overlapped, then they will unnecessarily send similar packets for reporting an event occurs in the intersection region. The sensing region of a node is the region around the node where if an event occurs within it, the node will sense that event. The radius of the sensing region is called the sensing range R_s of a node [25]. Figure 1 shows a group of sensor nodes that will generate similar data packets while sensing the same event. Accordingly, LSSA aims at reporting any event by only one packet generated by one node in order to save energy and achieve load balancing in the network. In other word, LSSA aggregates data in a distributed manner.

To achieve the objective of LSSA, firstly, it requires that if an event occurs in a certain region, then the node with the highest energy (among all the nodes from the same cluster or different clusters that sense the same event) will report this event by sending a packet to its corresponding CH. Secondly, the transmission of this packet should be done over at least $2R_s$ range such that all the possible nodes that could sense this event will know that there is a node that has higher energy than them reports this event instead of them. This means that a well-designed TDMA schedule is needed to satisfy the following requirements:

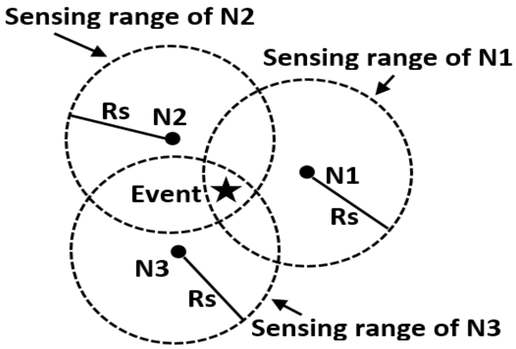


Figure 1. Group of sensor nodes generate similar data packets to report the same event

- The sensor nodes (from the same cluster or different clusters) that may sense the same event should be assigned different time slots. This is required to allow these nodes to hear the transmission of each other and hence avoid sending similar packets, and consequently, reducing the energy consumption.
- The sensor node that has higher residual energy should be assigned the lower time slot index. This is required to ensure that the event is reported by the node that has the highest residual energy among its neighbors so load balancing is satisfied.

These requirements are considered in the design of LSSA. Interestingly, the LSSA is illustrated in the following steps:

Step1: The BS creates for all nodes in the network an initial TDMA schedule based on the residual energy in each node where the first time slot is assigned for the node that has the highest energy level in the network and the second time-slot is assigned for the node that has the second highest energy level and so forth. The length of the initial TDMA schedule is impractical and equals to the number of the nodes in the network. Therefore, in the next steps, the length of the initial TDMA will be reduced by assigning one time-slot for different nodes.
Step2: Initially, each node is assigned a Flag Number (FN) equals to the index of its time-slot assigned in Step1.
Step3: Starting from the node with FN=1, the FNs of all its neighbors (within 2Rs range) are changed to 1. Also, for the neighbors of the node with FN=2, their FNs are changed to 2. This process is done for all nodes and the FN of a node should not be changed more than once in this step. In this step, for example, the first time slot is assigned for the node that has originally FN=1 and no one of its neighbors (that their FNs changed to 1) can use this slot for transmission.
Step4: Starting from the second time-slot, if the node in the second time-slot and the node in the first time-slot are not neighbors and not members of the same cluster, then the node in the second time-slot is moved to the first time-

slot and its FN is changed to 1 as well as the FNs of all its neighbors are changed to 1 instead of 2. This process is repeated for all nodes. To generalize this step, there are two cases should be clarified:

- If the node has FN equals to the index of its time-slot, then this means that its FN is not changed and no neighbor in the previous time-slots exists. Therefore, it checks if it is possible to move to one of the previous time-slots (starting from the first one). If there is no member from the same cluster in the checked time-slot, then it can move to it. Otherwise, it cannot move.
- If the FN of the node does not equal to the index of its time-slot, then this means that there is a neighbor in the time-slot of index equals to FN (that has the higher energy). Thus, the checking process should start from the time-slot of index equals to FN+1.

III. Simulation Results and Discussion

In this section, we conduct simulations to evaluate and compare the performance of ERP-DDA with different protocols presented in Section I. Table I summarizes the parameters used in the simulation. The same parameters were used in [13], [14], and [18].

TABLE I. THE COMMON SIMULATION PARAMETERS

Parameter	Value
Data packet size	1024 bits
Control packet size	176 bits
Initial energy	2 J
ETX (Energy for transmitting one bit)	50 nJ/bit
ERX (Energy for receiving one bit)	50 nJ/bit
EDA (Energy for data aggregation)	5 nJ/bit
Transmit amplifier energy in free-space	10 PJ / bit / m^2
Transmit amplifier energy in multi-path	0.0013 PJ / bit / m^4
Simulation Area	100 m x 100 m
Number of nodes	100
Distribution of nodes over the simulation area	Random
Base-station location	(50 m, 125 m)
Sensing range for each sensor	10 m
Percentage of CHs per round	5%

In our simulations, we consider continuous availability of data; the nodes in the network always have data to send. Therefore, the nodes will consume more energy to send their data packets and so short network lifetime is expected. To setup this scenario in simulation, the locations of events per round are distributed on a grid shape (i.e., the minimum distance between events is one sensing range). As a result, the total number of events per round is 50. This distribution of events is chosen such that each node, in each round, senses the closest event to it and has a packet to report this event. Since the nodes are randomly distributed over the network area, then it is possible that different nodes sense the same event. The performance of ERP-DDA is compared with other protocols that assume the same scenario. Figure 2 shows the total number of nodes alive per simulation round. Interestingly, it can be seen from Figure 2 that ERP-DDA shows a performance superiority over all compared protocols. In the case of applying ERP-DDA, the number of nodes alive in the network is slowly decreasing with time which results in extending the lifetime of the network. This is because ERP-DDA integrates the good features in different protocols in one protocol as well as it has an ability to overcome the weaknesses in all compared protocols. ERP-DDA is mainly based on four features that significantly reduce the energy consumed in reporting events in the network. First, ERP-DDA is a cluster-based protocol. This feature results in preventing the nodes in the same cluster to send the same data to the BS, and accordingly, reducing the consumed energy in the network. Moreover, different research works have shown that clustering is efficient in WSNs in terms of energy saving [13-20]. Second, ERP-DDA is a multi-hop routing protocol. Multi-hop routing between the CHs and the BS reduces the transmission distances, thereby reducing the transmission power and achieving load balancing in the network. It has been shown in the literature that multi-hop routing protocols are more efficient than single-hop protocols in terms of energy consumption [16] [26] [27] [28]. Third, ERP-DDA is based on variable clustering and variable cluster-head selection in each round which result in load balancing in the network, i.e., the task of reporting the events will be distributed over different nodes in the network. Fourth, as discussed before in Section II, ERP-DDA considers and solves IICERP problem such that one packet is sent by one node to report one event. Thus, the consumed energy in the network is significantly reduced.

IV. Conclusion

In this paper, we proposed an efficient energy-aware data collection protocol for improving the lifetime of wireless sensor network. This protocol is called Event Reporting Protocol based on Distributed Data Aggregation (ERP-DDA). This protocol aims at reducing the consumed energy in the network by reporting the event that occurs in a certain region by only one sensor node that has the highest residual energy in that region. Moreover, ERP-DDA applies clustering and multi-hop routing schemes in order to reduce the total consumed energy in the network. Simulation results showed that ERP-DDA protocol, compared with other protocols, achieves higher performance in terms of network lifetime.

References

[1] Darabkh, K. A. and Zomot, J. N., "An improved cluster head selection algorithm for wireless sensor networks," in *proc. 2018 14th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Limassol, Cyprus, Jun. 2018, pp. 65-70.

[2] Darabkh, K. A., Zomot, J. N., Al-qudah, Z. and Khalifeh, A. F., "IEDB-CHS-BOF: Improved Energy and Distance Based CH Selection with Balanced Objective Function for Wireless Sensor Networks," in *proc. 2020 6th International Workshop on Internet of Things: Networking Applications and Technologies (IoTNAT 2020) in conjunction with 5th IEEE International Conference on Fog and Mobile Edge Computing (FMEC 2020)*, Paris, France, 2020, pp. 275-279.

[3] Dargie, W. and Wen, J., "A Simple Clustering Strategy for Wireless Sensor Networks," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1-4, June 2020, Art no. 7500804, doi: 10.1109/LSENS.2020.2991221.

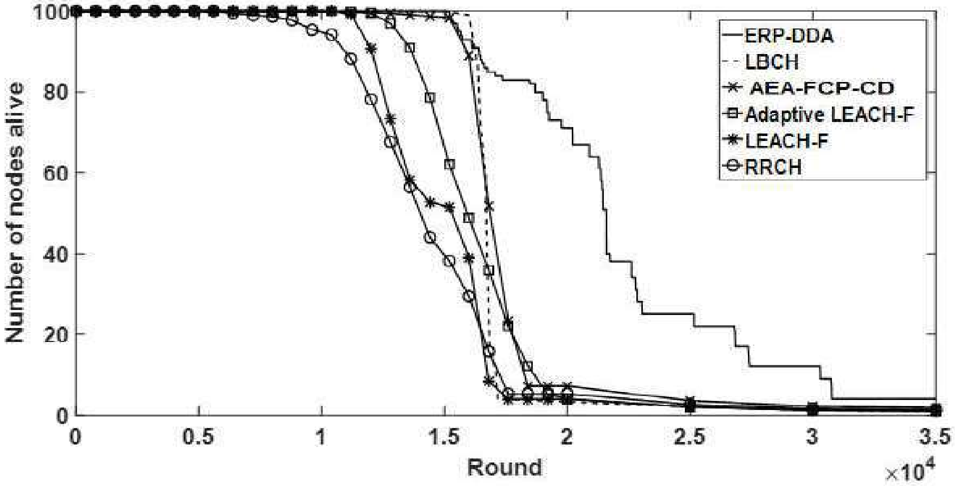


Figure 2. Number of nodes alive per simulation round. The locations of events per round are distributed on a grid shape; all the nodes have data to send in each round

[4] Kumar, B., Tiwari, U. K. and Kumar, S., "Energy Efficient Quad Clustering based on K-means Algorithm for Wireless Sensor Network," *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, India, 2020, pp. 73-77.

[5] Chen, L., Liu, W., Gong, D. and Chen, Y., "Clustering and Routing Optimization Algorithm for Heterogeneous Wireless Sensor Networks," *2020 International Wireless Communications and Mobile Computing (IWCMC)*, Limassol, Cyprus, 2020, pp. 407-411.

[6] Mishra, P. K. and Verma, S. K., "A survey on clustering in wireless sensor network," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-5.

[7] Neji, W., Othman, S. B. and Sakli, H., "T-LEACH: Threshold sensitive Low Energy Adaptive Clustering Hierarchy for Wireless Sensor Networks," *2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, Monastir, Tunisia, 2020, pp. 336-342.

[8] Swamy, T. Jagannadha, Ramamurthy, G. and Nayak, P., "Optimal, Secure Cluster Head Placement Through Source Coding Techniques in Wireless Sensor Networks," *IEEE Communications Letters*, vol. 24, no. 2, pp. 443-446, Feb. 2020.

[9] Liu, X., Mei, K. and Yu, S., "Clustering Algorithm in Wireless Sensor Networks Based on Differential Evolution Algorithm," *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chongqing, China, 2020, pp. 478-482.

[10] Sibi, S. A. and Prabhu, R. V., "Survey on Clustering and Depletion of Energy in Wireless Sensor network," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 1341-1345.

[11] Rajkumar, M., Sureshkumar, A. and Karthika, J., "An Enhanced Energy Efficient Clustering Approach for Wireless Sensor Networks," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2020, pp. 672-675.

[12] Jain, S. and Agrawal, N., "Development of Energy Efficient Modified LEACH Protocol for IoT Applications," *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, Bhimtal, India, 2020, pp. 160-164.

[13] Darabkh, K.A., Al-Rawashdeh, W.S., Hawa, M., Saifan, R. "MT-CHR: a modified threshold-based cluster head replacement protocol for wireless sensor networks," *Computers and Electrical Engineering*, vol. 72, pp. 926-938, November 2018.

[14] Darabkh, K.A., Al-Rawashdeh, W.S., Al-Zubi, R.T. Alnabelsi, S.H. "C-DTB-CHR: centralized density- and threshold-based cluster head replacement protocols for wireless sensor networks," Journal of Supercomputing, vol. 73, no. 12, pp. 5332–5353, 2017.

[15] Akhtar, M., Ali, A., Ali, Z., Hashmi, M., and Atif, M. " Cluster based Routing Protocols for Wireless Sensor Networks: An Overview," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 12, 2018.

[16] Rhim, Hana, Tamine, Karim, Abassi, Ryma, Sauveron, Damien, and Guemara, S. "A multi-hop graph-based approach for an energy-efficient routing protocol in wireless sensor networks," *Human-centric Computing and Information Sciences*, December 2018, https://doi.org/10.1186/s13673-018-0153-6.

[17] Gawade, R. D. and Nalbalwar, S. L., "A centralized energy efficient distance based routing protocol for wireless sensor networks," *Journal of Sensors*, vol. 2016, Article ID 8313986, 2016.

[18] Al-Zubi, R. T., Abedsalam, N., Atieh, A. and Darabkh, K. A., "LBCH: Load Balancing Cluster Head Protocol for Wireless Sensor Networks", *INFORMATICA*, Vol. 29, No. 4, pp. 633–650, 2018.

[19] Nam, D.-H., Min, H.-K. "An energy-efficient clustering using a round-robin method in a wireless sensor network," *In: Proceedings of the 5th ACIS International Conference on Software Engineering Research, Management & Applications (SERA 2007),* 2007.

[20] Heinzelman, W. R., Chandrakasan, A. and Balakrishnan, H., "Energy-efficient communication protocol for wireless microsensor networks," *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[21] Azim, A. and Islam, M. M., "A relay node based hybrid low energy adaptive clustering hierarchy for wireless sensor networks," *International journal of energy, information and communications*, vol. 3, no. 3, pp. 41-54, 2012.

[22] Darabkh, Khalid A. and Al-Jdayeh, Laith, "AEA-FCP: An Adaptive Energy-aware Fixed Clustering Protocol for Data Dissemination in Wireless Sensor Networks," *Personal and Ubiquitous Computing,* vol. 23, no. 5-6, pp. 819-837, Nov. 2019.

[23] Baek, J., An, S. K. and Fisher, P., "Dynamic cluster header selection and conditional re-clustering for wireless sensor networks," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2249-2257, November 2010, doi: 10.1109/TCE.2010.5681097.

[24] Dijkstra, E. W., "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[25] More, Avinash and Raisinghani, Vijay, "A survey on energy efficient coverage protocols in wireless sensor networks," *Journal of King Saud University – Computer and Information Sciences*, vol. 29, pp. 428–448, 2017.

[26] Darabkh, K., Zomot, J. and Al-qudah, Z., "EDB-CHS-BOF: Energy and Distance Based Cluster Head Selection with Balanced Objective Function Protocol," *IET Communications, Special Issue: Future of Intelligent Wireless LANs*, vol. 13, no. 19, pp. 3168 – 3180, 2019.

[27] Darabkh, K., Odetallah, S., Al-qudah, Khalifeh, Z., A. and Shurman, M., "Energy–Aware and Density-Based Clustering and Relaying Protocol (EA-DB-CRP) for Gathering Data in Wireless Sensor Networks," Applied Soft Computing, vol. 80, pp. 154-166, 2019.

[28] Darabkh, Khalid A., Kassab, Wafa'a K., and Khalifeh, Ala' F., "LiM-AHP-G-C: Life Time Maximizing Based on Analytical Hierarchal Process and Genetic Clustering Protocol for the Internet of Things Environment," *Computer Networks*, vol.176, p. 107253, July 2020.

Stability Analysis of Two-Stage OTA with Frequency Compensation

Misari K Shah
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science Pilani,
Dubai Campus, Dubai, UAE
misukhel@gmail.com

Vilas H Gaidhane
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science Pilani,
Dubai Campus, Dubai, UAE
Vilasgd612@gmail.com

Chippy Prasannan
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science Pilani,
Dubai Campus, Dubai, UAE
chippyprasannan23021991@gmail.com

Abstract— **In this paper, the design implementation of a two-stage operational transconductance amplifier (OTA) using a passive frequency compensation technique is presented. It is based on the Miller and passive frequency compensation techniques. The use of these techniques achieves the high gain-bandwidth product without an increase in the overall power consumption. It produced a -21.17dB of gain margin with a phase margin of 77° at a 1.8V power supply voltage. A resistor-capacitor pair in feed-forward compensation is connected in the first stage and a Miller compensation capacitor is introduced in the second stage of the circuit. The proposed op-amp design results in 1205 MHz of gain-bandwidth product. It is observed that the presented circuit performs better as compared to existing designed circuits**.

Keywords— ***stability analysis; frequency compensation; Gain-bandwidth product; OTA amplifier Miller compensation.***

I. Introduction

An operational transconductance amplifier (OTA) is the basic building block used in analog large-scale integration systems. However, the design of such a circuit has numerous design challenges to satisfy the performance parameters of circuits. These parameters are gain, bandwidth, and power consumption. In the past years, various researchers have suggested high-performance amplifiers [1]. The parameters such as high slew rates (SR), gain-bandwidth product, and low static power dissipation are required in a variety of applications especially in wireless and battery-powered systems [2]. The OTA can be used to fulfill these requirements. It can be implemented using a cascode topology in which one or two transistors are stack together in series. Using such cascode connections the overall output resistance and especially gain can be increased. However, it has some drawbacks that they suffer from voltage headroom constraints [3] which will limit the closed-loop gain. To overcome these issues, in multiple cascaded amplifiers, gain stages can be realized to attain the high voltage swing as well as high gain. Nevertheless, this design methodology is susceptible to instability. Each stage will introduce a lower frequency pole which causes a reduction in the phase margin thus leading to instability [4]. This issue is improved by adding a suitable compensation network to guarantee stability under the operating frequencies [5]. In closed-loop conditions, Miller compensation techniques are commonly used to provide stability [1]. The basic principle of the Miller compensation technique is splitting the dominant and non-dominant pole to acquire the desirable phase margin (PM). The drawback of this technique is the reduction of a dominant pole which causes the reduction of -3dB bandwidth and unity gain-bandwidth of the OTA [6].

The compensation network can be realized as a nulling resistor, current buffer or voltage buffer, or amplifier [7]. The best choice for moderate and higher capacitive loads are current amplifier and voltage buffer approaches, respectively [7, 8]. Moreover, there are some methods which do not use miller capacitor to compensated multi-stage amplifiers [9].

Classical topologies of two-stage amplifiers were introduced in 1967 [10]. Since then, many efforts were brought up about the stability issue of the two-stage amplifier and a lot of designs with improved performance are reported in the literature [11-14]. Moreover, a variety of high-performance amplifiers have been proposed by several researchers. The scaling of the supply voltage would eventually lead to a reduction in power consumption. Therefore, designers had come up with the idea of cascading, however, still limitations exist. Thus, topologies like inverter-based amplifiers, bulk-driven, and self cascode topologies help to solve the power consumption issue. Although the stability and gain in the above topologies are not considered, it has a variety of applications such as energy harvesting, wireless, IoT sensors, biomedical sensors, and so on. These applications are widely used in recent technologies and modifying different versions of OTA helps to make use of the above technologies effectively and efficiently. However, they did not consider the power consumption issue in the analysis. It is observed from the literature that the existing approaches lack stable output. Therefore, it is very much important to design a circuit with stable output and low power consumption [15]. Therefore, in this paper, a new circuit design is proposed. It consists of a mixture of frequency compensation [16] and miller compensation techniques.

The rest of the paper is divided into the following sections. An analysis of the existing circuits is conducted and

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

various combinations of the compensation networks are drawn and analyzed in Section II. The proposed circuit and analysis are presented in Section III. Moreover, the conclusions and future scope are drawn in Section IV.

II. Study of Existing Circuits

A simple circuit diagram of two-stage OTA is presented in Fig. 1. In this, transistors M_1 and M_2 are the NMOS transistors and M_3 and M_4 are PMOS transistors that form a current mirror circuit in the first stage of the amplifier. The second stage is a common source amplifier and is made up of NMOS and PMOS transistors, M_5 and M_6, respectively.

In Fig. 1, C_e is the capacitance at the mirror node, C_L is the load capacitance, and V_b is the bias voltage. The value of capacitor C_C and C_L should be very small [16]. By increasing C_L, the dominant pole, as well as unity gain frequency, can be shifted toward the origin [18]. Hence, phase margin (PM) can be increased, which improves stability [19-20]. The compensation network R_C and C_C generate a left-hand plane (LHP) zero, which can be represented as

$$Z_1 = \frac{1}{R_C C_C} \tag{1}$$

This zero helps to increase the gain-bandwidth product of the amplifier. Second LHP zero represented as

$$Z_2 = \frac{g_{m3}}{C_E} \tag{2}$$

where g_{m3} is the transconductance of M_3 transistor. Second stage has RHP zero represented as

$$Z_3 = \frac{g_{m6}}{C_{gd6}} \tag{3}$$

where g_{m6} is the transconductance of M_6 transistor. Above two zeros, Z_2 and Z_3 can be ignored as it is seen at high frequency.

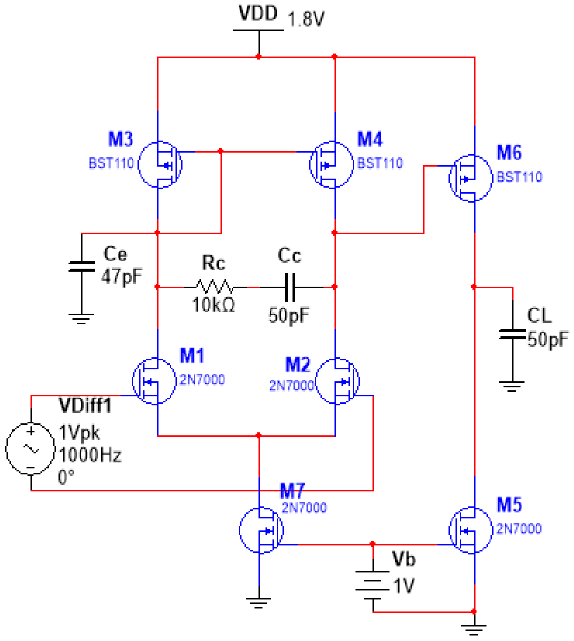


Fig. 1. Two-stage OTA circuit with compensation network.

The voltage gain A_V can be represented as

$$\frac{V_{out}}{V_{in}} = \frac{g_{m1} g_{m6} R_1 R_L \left(1 + sR_C C_C\right)}{\left(1 + sR_L C_L\right)\left[1 + sC_C\left(\alpha + R_C\right)\right]} \tag{4}$$

where V_{out} is voltage across the C_L capacitor and $V_{in} = V_{diff1}$ (1 Volts peak), g_{m1} is the transconductance of M_1 transistor; R_1 is the resistance at the output of the first stage, R_L is the equivalent resistance at the output node in parallel with the load. Moreover, α is a constant given as

$$\alpha = \frac{\left(1 + 2g_{m3} R_1\right)}{g_{m3}} \tag{5}$$

From the transfer function, dominant pole P_1 at the output of the first stage,

$$P_1 = \frac{1}{C_C\left(\alpha + R_C\right)} \tag{6}$$

The second pole at the load can be canceled by the zero with $R_L C_L = R_C C_C$, unless and until phase margin is more than 45° [17]. A two-stage OTA circuit without compensation and with feed-forward compensation is shown in Fig. 2.

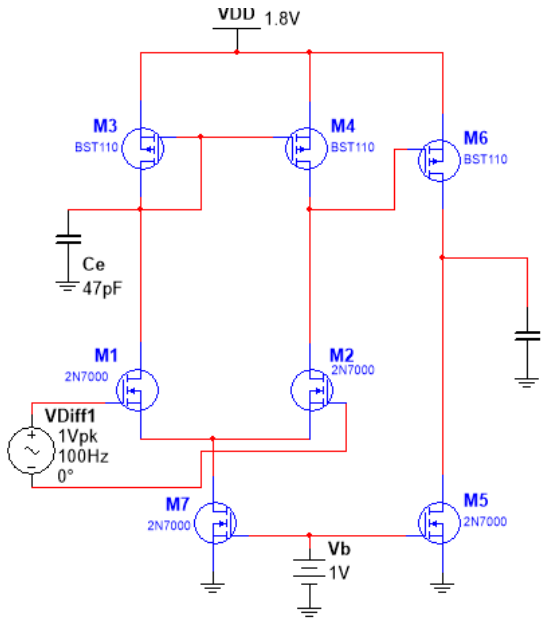


Fig. 2. Two-stage Non-compensated OTA circuit

Adding a capacitor in parallel to the resistor as a compensation network (CN), can further increase the gain-bandwidth product of the implemented circuit as shown in Fig. 3(a). Nodal equations for the circuit shown in Fig.3 are given as

$$-g_{m1} V_{in} = sC_{in} V_{in} + \left(sC_C + \frac{1}{R_C}\right) V_2 + G_1 V_2 \tag{7}$$

$$-g_{m6} V_2 = sC_L V_{out} + G_{11} V_{out} \tag{8}$$

$$\frac{V_{out}}{V_{in}} = \frac{\left(g_{m1} + sC_m\right) g_{m6} R_1 R_L R_C}{\left(1 + sR_L C_L\right)\left(R_1 + R_C + sR_1 R_C C_C\right)} \tag{9}$$

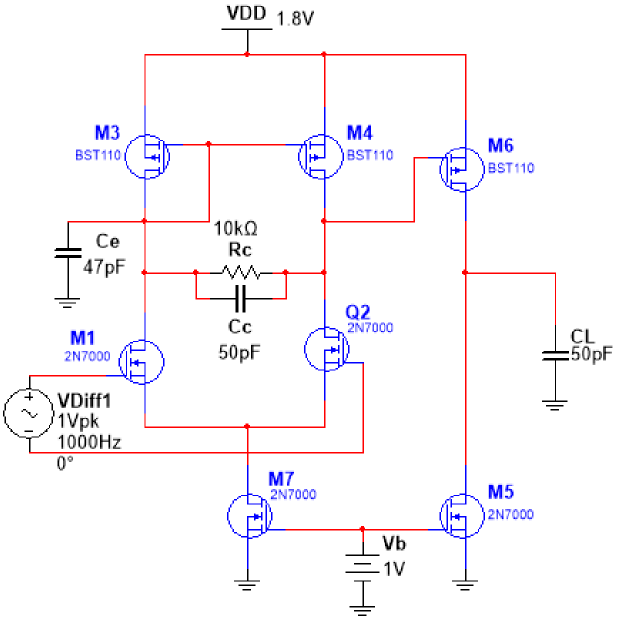


Fig. 3. Two-stage compensated OTA circuit

In Analog VLSI design researchers are continuously trying to reduce the area by using new MOS technology. In this paper, the compensation network components are replaced by MOS technology.

The performance of the OTA further can be improved by adding depletion type of MOSFET in parallel with a resistor as shown in Fig. 4.

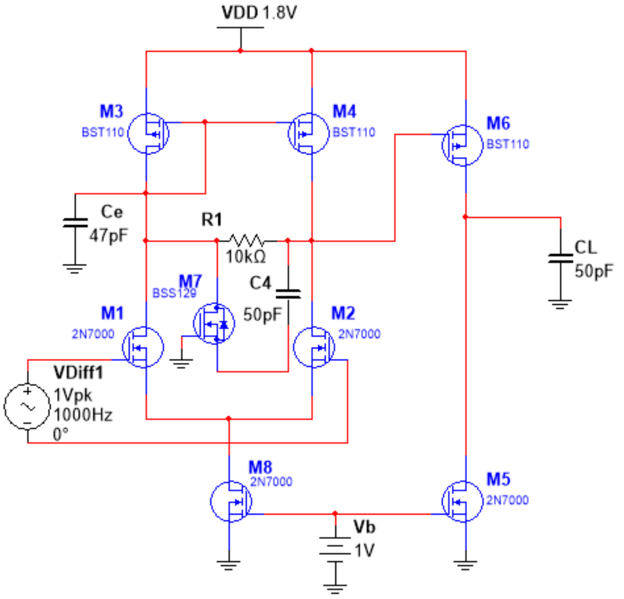


Fig. 4. CN with capacitor and resistor in series and parallel to resistor R_1

The transfer function in the s-domain for circuits can be calculated as

$$-g_{m1}V_{in} = sC_mV_{in} + [{}^{1}\!/_{R_C} + (R_9 + {}^{1}\!/_{sC_C}) - 1]V_2 + G_1 \quad (10)$$

$$-g_{m6}V_{in} = sC_LV_{out} + G_{11}V_{out} \quad (11)$$

$$T(s) = \frac{g_{m6}(g_{m1}+sC_m)R_1R_C(1+sC_CR_9)}{R_1[1+sC_C(R_9+R_C)]+R_C(1+sC_CR_9)}\left(\frac{R_{11}}{1+sC_LR_{11}}\right)$$

(12)

A depletion type transistor M_7 which is diode-connected will act as a resistor connected in series with a capacitor C_4 and this whole connection is in parallel with the resistor R_1.

However, its stability analysis can be improved by replacing the parallel capacitor and resistor with depletion-mode transistors and Miller compensation network. Based on this concept a new design is proposed. The proposed design consists of depletion type of MOSFET connected with capacitor either in series or parallel combination.

III. Modified OTA Circuit

The proposed design and implementation of two-stage OTA with passive frequency compensation in a 200um CMOS process technology, which draws a total current of $0.95mA$ from a $1.8V$ supply. The compensation network can be modified in which a depletion mode transistor, M_7 which is a diode-connected transistor that acts as a resistor. It is connected in parallel with a capacitor. Moreover, Fig.5 shows the performance of OTA can be improved by replacing the capacitor with a depletion mode transistor, M_8 by shorting source and drain terminals. The transfer function can be calculated as

$$-g_{m1}V_{in} = sC_mV_{in} + (sC_C + g_{m9})V_2 + G_1V_2 \quad (13)$$

$$-g_{m6}V_2 = sC_LV_{out} + G_{11}V_{out} \quad (14)$$

$$T(s) = \frac{g_{m6}(g_{m1}+sC_m)R_1R_{11}}{(1+sC_LR_{11})[1+R_1(sC_C+g_{m9})]} \quad (15)$$

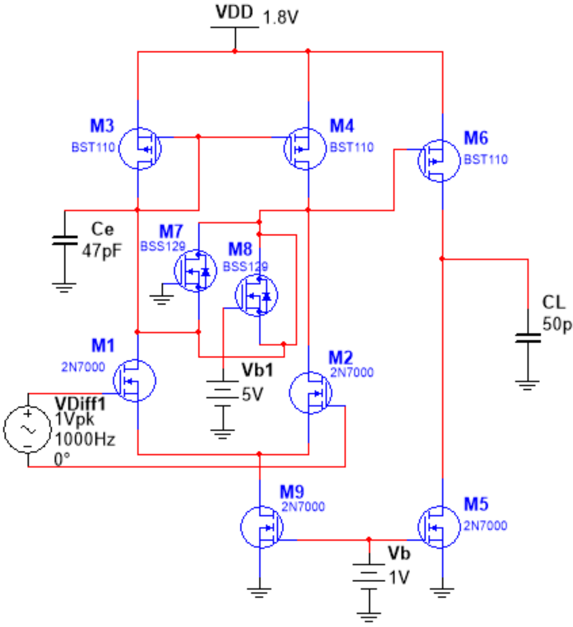


Fig. 5. Compensation network with capacitor and resistor in parallel, both of which is replaced by depletion-mode transistors

However, the proposed circuit can obtain a stable output and high gain. The design consists of a mixture of a feed-forward path technique along with the miller compensation technique. The circuit diagram of the proposed modified two-stage OTA is seen in Fig. 6.

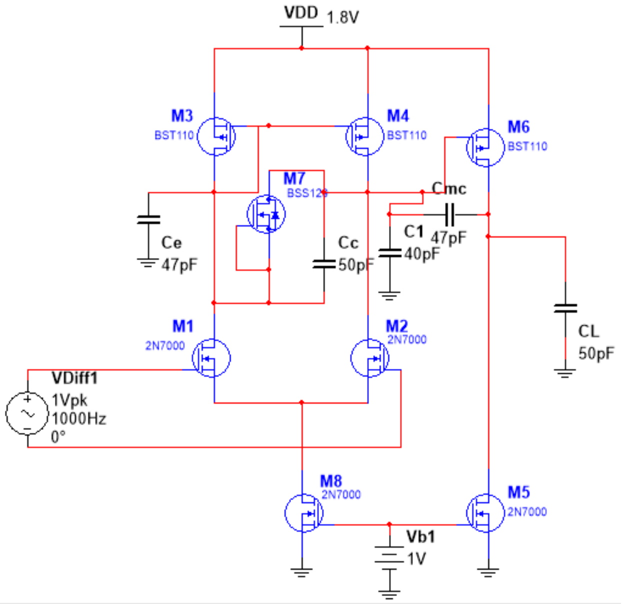


Fig. 6. Proposed two-stage OTA circuit

In this, M_1 and M_2 are the NMOS transistors and M_4 and M_3 (current mirror circuit) are the PMOS transistors in the first stage of the amplifier. These four transistors form the differential amplifier. C_E is the capacitance of the Mirror node. A depletion mode transistor, M_7 is connected instead of the resistor in the compensation network. A capacitor C_C is connected in parallel with M_7. The Miller compensation technique is added to the second stage of the circuit. A bias voltage, V_{b1} is applied at the gate of M_5 transistor, and capacitors C_1 and C_{mc} are the miller compensation capacitors along with C_L is the load capacitor.

The transfer function of the proposed circuit as shown in Fig. 6, is calculated as

$$T(s) = g_{m1}R_{p1}\left(\frac{1}{1+s(C_1+C_{mc})R_{p1}}\right)\left(\frac{1+\frac{sC_C}{g_{m7}}}{1+sC_L R_{11}}\right) \qquad (16)$$

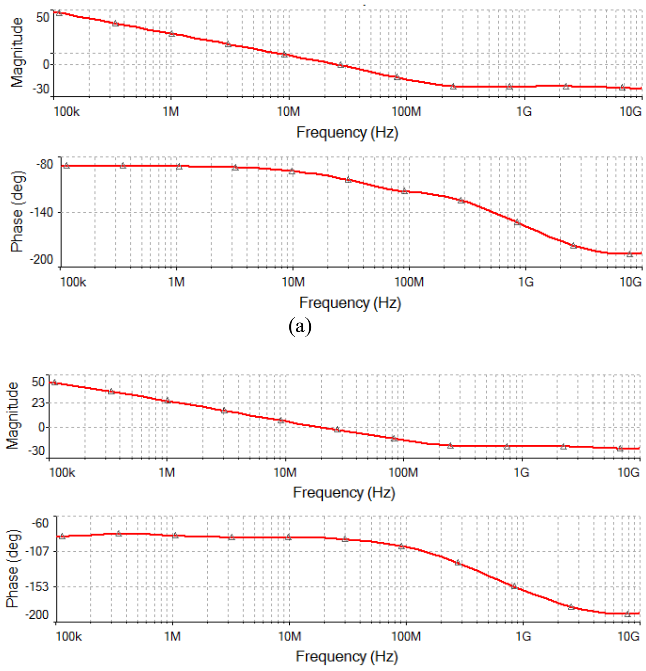
$$\times\left(\frac{-1}{g_{m7}+sC_C}\right)$$

Simulation results of all the circuits discussed will be further discussed in the upcoming section.

Experimentations and Results

A. Compensated Circuit and Non-Compensated Circuit

Fig.7(a) and Fig.7(b) show frequency response of a two-stage uncompensated OTA and a two-stage compensated OTA on simulating Fig.2. and Fig.1 on MULTI-SIM.



(a)

(b)

Fig.7. (a) Frequency response without compensation network, (b) Frequency response with R_C, C_C in series.

The frequency response without the use of a compensation network and with compensation network is shown in Fig.7.(a), and Fig.7.(b), respectively. The circuit simulated in Multi-Sim software and gain-bandwidth product, gain margin, and phase margin has been obtained. The gain-bandwidth product is 33275*MHz*. However, the stability of the amplifier is very less. The phase margin is 99° which is comparatively large. This will push the circuit into instability. Therefore, to improve the phase margin a compensation network should be introduced in the circuit. A compensation network shown in Fig.1, helps to overcome this issue. However, the gain-bandwidth product was reduced to 25360*MHz*. Thus, even though the gain-bandwidth product has decreased but the stability has improved as compared to the circuit without the compensation network. Further, the performance of the circuit can be improved by modifying the circuit shown in Fig.1.

B. Compensated Circuit with Capacitor and Resistor in Parallel

Furthermore, on simulating Fig.3, the below frequency response is observed.

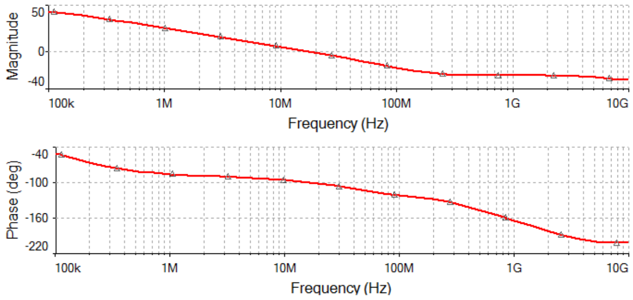


Fig. 8. Frequency response of CN with R_C and C_C in parallel.

Fig.8. shows the frequency response of CN with a resistor and capacitor connected in parallel. The components R_C and

C_C in parallel can further increase the gain margin to 39.3dB and phase margin 39° as shown in Fig. 8. However, the gain-bandwidth product was reduced. Thus, we move forward to another analysis to overcome this issue.

C. Resistor In Parallel With Capacitor And MOSFET In Series Which Are Connected In Series:

Furthermore, on simulating Fig.4, the below frequency response is observed.

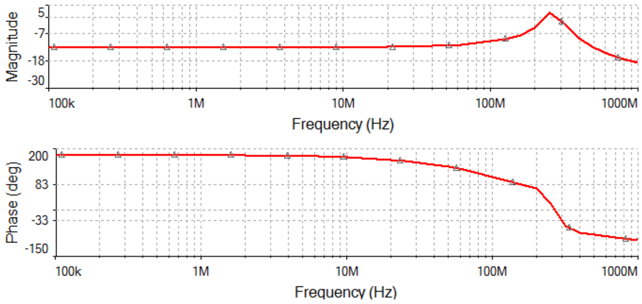


Fig. 9. Frequency response of CN with R_1 parallel with MOS and capacitor C_4

Using the depletion type NMOS and capacitor in series as shown in Fig. 4, provides a phase margin of 142^0. However, the GBW is very less and close to 581*MHz*. Thus, we move forward to another analysis to overcome this issue.

D. Capacitor and Resistor in Parallel, Replaced by Depletion Mode Transistors

Furthermore, on simulating Fig.5, the below frequency response is observed.

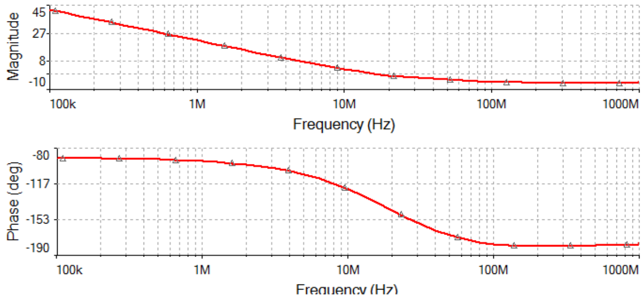


Fig. 10. Frequency response of CN with *Resistor and Capacitor in parallel, both replaced by depletion mode transistors.*

Fig.5 shows the proposed circuit and Fig.10 shows the frequency response of the proposed circuit that directs to its stability analysis. However, the gain-bandwidth product is large and nearly equal to 1181.18 *GHz* which is higher than the other circuits. However, the gain margin is -82.6dB and the phase margin is nearly 131°.

E. The Proposed Circuit

The gain-bandwidth product, phase margin, and gain margin as compared to the previous circuit can be improved by replacing the proposed design. In this capacitor C_1 and C_{mc} are connected as an extra circuitry to the OTA circuit as shown in Fig. 6.

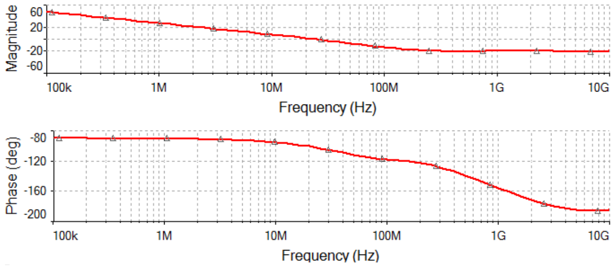


Fig.7. Frequency response of a proposed circuit

The proposed method includes a combination of the miller compensation technique along with the passive frequency compensation technique, which helps to increased stability and gain. This is achieved by introducing a new pole which increased the stability of the system. The Gain-Bandwidth (GBW) product of the amplifier is found to be 1205*MHz*. It is observed from the graph that the Phase Margin of the amplifier is 77°, which has been improved. Moreover, the gain margin is nearly -21.17dB which is good enough as compared to existing methods.

The various combinations with compensation networks connected to the first stage and their performance are summarized in Table 1. It is observed from Table 1 that the proposed design results in a higher gain-bandwidth product, gain margin as well phase margin. Although, the gain margin of the proposed approach is less but overlooked by the higher phase margin and thus the stability.

TABLE I. PERFORMANCE OF THE EXISTING MODELS AND PROPOSED DESIGN MODEL

Circuit Model	Parameters		
	Gain-Bandwidth Product (MHz)	Gain Margin (dB)	Phase Margin (deg)
Fig. 1	25360	-36.7	35
Fig. 2	33275	-30	90
Fig. 3	22560	-39.3	39
Fig. 4	581	-12.7	142
Fig. 6	1181180	-82.6	131
Fig. 7	1205	-21.7	77

IV. CONCLUSION AND FUTURE WORK

A new OTA amplifier with a compensation network is proposed to achieve a higher gain-bandwidth product as well as a stable output. The introduction of a compensation network helped to overcome the stability issue. The proposed circuit achieves a high gain-bandwidth product of 1205 *MHz* without an increase in the overall power consumption. It produced a -21.17dB of gain margin with a phase margin of 77°. The simulation results show that the proposed design is effective in improving the stability and gain-bandwidth product. As stability and gain are attained by using a combination of miller compensation technique and passive frequency compensation technique, it can be used in wireless and battery-powered systems. Moreover, it can be implemented in a variety of applications such as IoT sensors, biomedical applications, and so on. The results obtained from the proposed design can further be used to study and find out the sizes of the optimal transistors (length and width) in order to obtain operational amplifier performances for analog and mixed CMOS-based circuit applications using multi-objective genetic algorithms.

REFERENCES

[1] M. Tan and W.-H. Ki, "Current-mirror miller compensation: An improved frequency compensation technique for two-stage amplifiers," in Proc. Int. Sym. on VLSI Design, Automation, and Test (VLSI-DAT), pp.1–4, Apr. 2013.

[2] J. Ramirez-Angulo and M. Holmes, "Simple technique using local CMFB to enhance slew rate and bandwidth of one-Stage CMOS op-amps," Electron. Lett., vol. 38, no. 23, pp. 1409-1411, 2002.

[3] A. D. Grasso, G. Palumbo, and S. Pennisi, "High-performance four-stage CMOS OTA suitable for large capacitive loads," IEEE Trans. Circuits Sys. I: Regular Papers, vol. 62, no. 10, pp. 2476-2484, 2015.

[4] I. J. Nagrath and M. Gopal, "Control systems engineering," New Age International (p) Limited, 2015.

[5] S. Liu, Z. Zhu , J. Wang, L. Liu , and Y. Yang, "A 1.2-V 2.41-GHz Three-Stage CMOS OTA with efficient frequency compensation technique,", IEEE Trans. Circuits Sys. I: Reg. Papers, vol. 99, pp. 1-1, 2018.

[6] K. Lv, "Research on the high-sr and low-power enhancing transconductance folded cascade OTA", Boletín Técnico, vol. 55, no. 14, pp. 299-303, 2017.

[7] A. D. Grasso, G. Palumbo and S. Pennisi, "Comparison of the frequency compensation techniques for CMOS two-stage miller OTAs," IEEE Trans. Circuits Syst. II: Express Briefs, vol. 55, no. 11, pp. 1099-1103, 2008.

[8] M. Yavari, "Hybrid cascode compensation for two-stage CMOS opamps," IEICE Trans. Electron., vol. 88, no. 6, pp. 1161-1165, 2005.

[9] B. K. Thandri and J. S. Martínez, "A Robust feedforward compensation scheme for multistage operational transconductance amplifiers with no miller capacitors," IEEE J. Solid-State Circuits, vol. 38, no. 2, pp. 657-659, 2003. 10.

[10] R. J. Widlar, "Monolithic OP AMP with simplified frequency compensation(Monolithic operational amplifiers with simplified frequency-compensated network using minimum stages and integrated circuit components," EEE- Magazine Circuit Design Eng., vol. 15, pp. 58-63, 1967.

[11] P. R. Gray and R. G. Meyer, "MOS operation amplifier design-a tutorial overview," IEEE J. Solid-State Circuits, vol. 17, no. 6, pp. 969-982, 1982.

[12] B. Kamath, R. G. Meyer, and P. R. Gray, "Relationship between frequency response and settling time of operational amplifier," IEEE J. Solid-State Circuits, vol. 9, no. 6, pp. 347-352, 1974.

[13] J. E. Solomon, "The monolithic op amp:a tutotrial study," IEEE J. Solid-State Circuits, vol. 9, no. 6, pp. 314-332, Dec. 1974.

[14] M. Yang, and G. W. Roberts, "Synthesis of high gain operational transconductance amplifiers for closed-loop operation using a generalized controller-based compensation method," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 63, no. 11, pp. 1794-1806. 2016.

[15] Anna Richelli , Luigi Colalongo, Zsolt Kovacs-Vajna, Giacomo Calvetti, Davide Ferrari, Marco Finanzini, Simone Pinetti, Enrico Prevosti, Jacopo Savoldelli and Stefano Scarlassara, " A Survey of Low Voltage and Low Power Amplifier Topologies", Journal of Low Power Electronics and Applications, pp8030022, 2018.

[16] D. Marano, A. D. Grasso, G. Palumbo and S. Pennisi, "Optimized Active Single-Miller Capacitor Compensation With Inner Half-Feedforward Stage for Very High-Load Three-Stage OTAs," IEEE Trans. Circuits Syst. I, Reg. Papers, 2016.

[17] A. Mirvakili, and V. J. Koomson. "Passive frequency compensation for high gain-bandwidth and high slew-rate two-stage OTA," Electron. Lett., vol. 50, no. 9, pp. 657-659, 2015.

[18] A. S. Sedra and K. C. Smith, Microelectronics Circuits. Oxford University Press, 2015.

[19] V. H. Gaidhane, Y. V. Hote "An improved approach for stability analysis of discrete system," J. Control, Autom. Electr Syst., vol. 29, no. 5, pp. 535-540, 2018.

[20] V. H. Gaidhane, Y. V. Hote, "A new approach for stability analysis of discrete systems," IETE Technical Review, vol. 33, no. 5 466-471, 2016.

Design of Operational Amplifier using Artificial Neural Network

Ganta Yogesh
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science Pilani,
Dubai Campus, UAE
E-mail: kyogesh805@gmail.com

Vilas H Gaidhane
Department of Electrical and Electronics Engineering and APPCAIR
Birla Institute of Technology and Science Pilani,
Dubai Campus, UAE
E-mail:vilasgd612@gmail.com

Abstract—**It is well known that the mathematical modeling and manual computations of the transistor design parameters are sometimes impractical and remains a challenge for researchers. In submicron technologies, the basic component of a three-stage operational amplifier such as MOSFET is modeled by various complex nonlinear equations. The modeling equations include parameters such as channel length (L), channel width (W), node voltages, and branch currents. However, the design and analysis of such complex nonlinear equations are depending on the expertise of the designer. In this paper, a Neural network is used to design and implement a three-stage operational amplifier. The channel length (L) and width (W) which are most suitable for circuit characteristics are calculated using neural network training. The Cadence tool is used to simulate the various circuit diagrams. The Neural Network model is developed and trained using MATLAB 2019b software platform. The effectiveness of the proposed neural network model is tested using the various circuits.**

Keywords— ***MOSFET; Artificial neural network; Gain-bandwidth product; OTA amplifier Miller compensation.***

I. Introduction

In recent years, MOSFET design has become an integral part of almost all electronic devices. Therefore, an efficient and implementable design still remains a challenging area for the researchers. It is known that the MOSFET consists of Gate-Source, Drain-Source, and Bulk-Source potentials as inputs and current flow through the drain. We know that the MOSFET is a voltage-controlled device and it is because of the fact that the conduction is started by some minimum gate voltage. So, depending upon the input voltages provided such as V_{GS} and V_{DS} to MOSFET, the channel parameters results into a controlled drain current I_D. The channel length (L) and width (W) parameters provides a non-linear output current relationship with respect to the constant input and output voltages. However, the evaluation of such parameters $(L \text{ and } W)$ is very difficult since the MOSFETS are generally modeled using complex and non-linear equations along with the independent and dependent parameters. Therefore, to overcome these issues neural network-based approach can be used.

In past years, many researchers carried out numerous research work on the MOSFET modeling and design based on the neural network models. Kothapalli in 1995 [1] proposed an approach based on the neural network to design an operational transconductance amplifier (OTA). Moreover, a similar approach is presented by Langeheine *et al.* [2] and it is used for the transistor array design using a genetic algorithm. This model is developed only for the purpose of a single transistor and with constant drain-source voltage which cannot be used the general-purpose approach in analog integrated circuit design. Further, similar work has been carried out using the same process parameters and genetic algorithms in [3].

From literature, it is observed that the neural network model has been extensively used for MOSFET design as well as the various electronic circuits such as current mirror [4], differential amplifier [5], inverter, and logic gates [6]. In 2007, Avec *et al.* [7] has proposed a model which is a general design approach for analog integrated circuits. This model consists of Multi-Layer Perceptron (MLP) neural network with three inputs, two hidden layers, and two outputs. In this, training and testing data are prepared using the simulation of MOSFETs using Cadence software.

The channel length (L) and width (W) are important parameters for the modeling of MOSFET and Integrated circuits. It has been observed that the theoretical calculation of dimensions of MOSFET becomes impractical and complicated due to the other varying and nonlinear parameters. The ratio of channel length and width is called aspect ratio and it can be predicted by using a supervised neural network. Neural networks are mostly one which takes input data and train themselves to build a model and then predict the output for a new set of similar data. The artificial neural networks are made up of layers of a neuron. These neurons are the core processing units of the network. The input layer receives the input and the output layer predicts the final output. The hidden layer performs most of the computations required for the process. In a multi-layer perceptron (MLP) neural network, the modeling of the transistor can be done by providing input voltages and output current [8-10].

In this paper, an artificial neural network model is proposed to design the basic building blocks of the three-stage operational amplifier. The manual calculations of various design parameters of MOSFET are calculated using the Cadence software. These parameters are considered to train the proposed neural network model. In the first stage, basic MOSFETs circuits are simulated and their current-voltage characteristics are studied. Later, the current mirror,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

differential amplifier, two-stage, and three-stage operational amplifier were also simulated and characteristics have been studied in detail. The rest of the paper is organized as follows. The proposed approach is explained in Section II. Experimentations and results are shown in Section III and finally, conclusions are drawn in Section IV.

II. Proposed Neural Network Model

A neural network is generally termed as a system that consists of many neurons. It produces the best possible output though there is a change in the input. Generally, it consists of source nodes which include an input layer, one or more hidden layers, and an output layer. In general, in neural network architecture, the set of sensory nodes called as input neurons constitute the input layer, one more computational hidden layer, and summed output layer. The output of the neural network is decided by the activation function and threshold [11]. It has been observed from the literature that the sigmoidal function can perform better as compared to the binary and step function. The basic neural network with three inputs and one out is shown in Fig. 1. In Fig. 1, x_1, x_2 and x_3 are the input to the neurons and $w_1, w_2,$ and w_3 are the weights to the respective inputs. The activation function is used to calculate the output of the neural network using the threshold value.

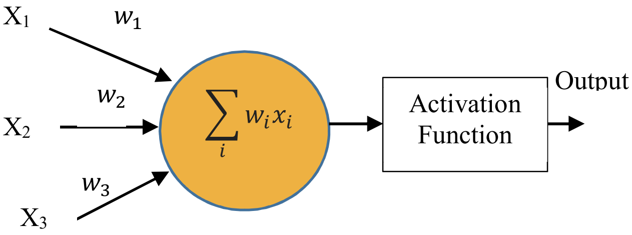


Fig.1. Basic block diagram neuron model.

In this work, the NMOS1 MOSFET model is used for the design and analysis. The input parameters such as gate-to-source (V_{GS}), drain-to-source (V_{DS}) and output drain-current (I_D) are considered as a input to the neural by which the design parameters channel length (L) and width (W) are estimated. The proposed neural network architecture is shown in Fig. 2.

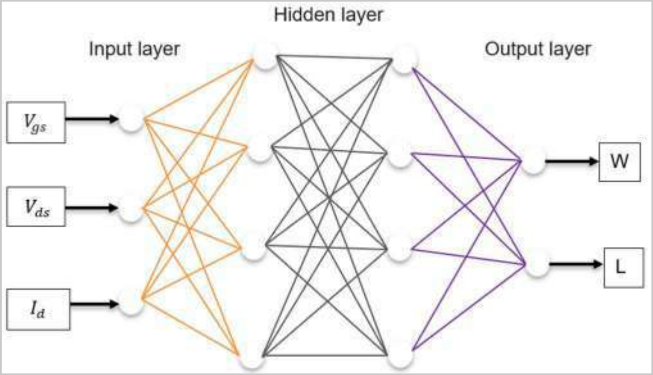


Fig. 2. Proposed Neural network model for circuit design and analysis.

Training and testing data are obtained by various simulations of MOSFET (n-channel and p-channel) in a Cadence software environment. The activation of the hidden neuron is the function of weighted inputs and bias. It can be calculated as

$$x_i = w_1 V_{GS} + w_2 V_{ds} + w_3 I_d + \theta_j \qquad (1)$$

III. Circuit Implementation

The circuit in Fig. 3 shows the design developed for the proposed neural network using Cadence software. The voltage values at different nodes and current are shown in Table II. The value of voltages and currents decides the design parameters length (L) and width (W) of the transistors. The channel length and width are calculated using the simulation of the circuit using Cadence and predicted using the proposed neural network.

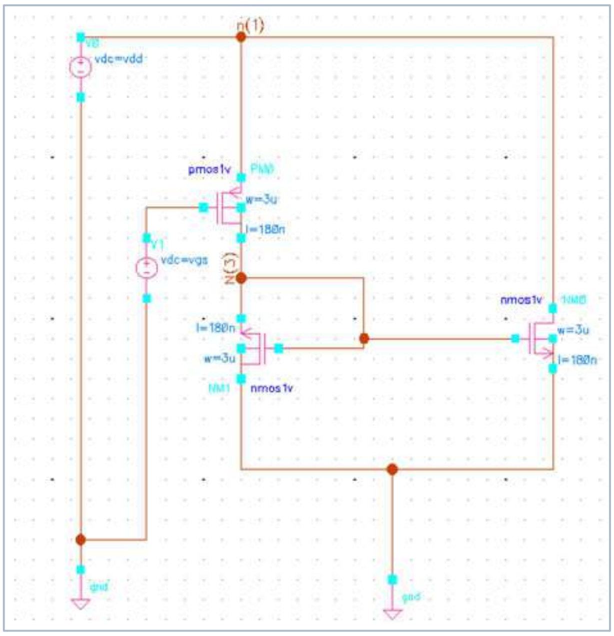


Fig. 3. A current mirror circuits.

TABLE 1. Simulated and Predicted Values of Width and Length for NMOS 1

S. No.	V_{GS} (V)	V_{DS} (V)	I_D (μA)	Simulated		Predicted	
				W	L	W	L
1	0.35	0.70	0.019	0.12	0.045	0.119	0.052
2	0.51	0.80	4.30	0.12	0.045	0.118	0.049
3	0.81	1.01	20.88	0.24	0.18	0.238	0.182
4	0.98	1.10	36.96	0.24	0.18	0.238	0.133
5	1.14	1.20	73.33	0.92	0.60	1.034	0.419
6	1.3	1.31	165.5	3.60	1.8	3.60	1.799

Table II shows the simulation and predicted results for NMOS 2 used in current mirror circuits shown in Fig. 3. Moreover, Table III shows the simulated voltages, current, and design parameters for PMOS 1.

TABLE II. SIMULATED AND PREDICTED VALUES OF WIDTH AND LENGTH FOR NMOS 2

S. No.	V_{GS} (V)	V_{DS} (V)	I_D (μA)	Simulated		Predicted	
				W	L	W	L
1	0.35	0.70	0.018	0.12	0.045	0.126	0.043
2	0.51	0.80	0.47	0.12	0.045	0.122	0.044
3	0.67	0.9	8.42	0.24	0.18	0.242	0.178
4	0.81	1.01	20.56	0.24	0.18	0.511	0.454
5	0.98	1.10	47.84	0.72	0.60	0.721	0.595
6	1.14	1.20	73.23	0.72	0.60	0.722	0.600
7	1.3	1.31	165.3	3.60	1.80	1.688	1.380

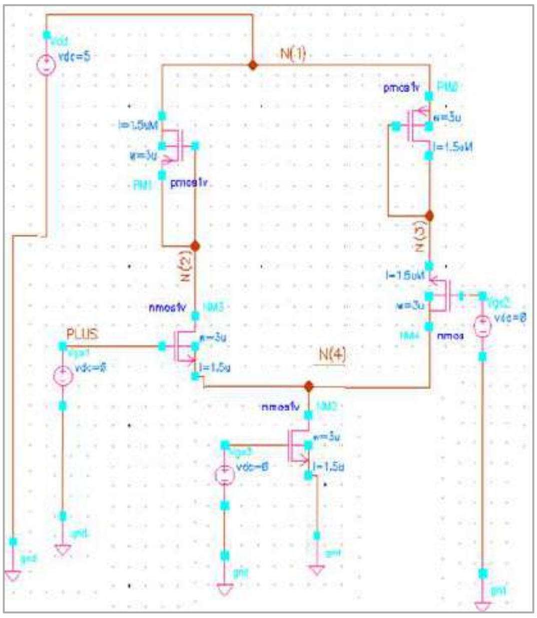
TABLE III. SIMULATED AND PREDICTED VALUES OF WIDTH AND LENGTH FOR NMOS 2

S. No.	V_{GS} (V)	V_{DS} (V)	I_D (μA)	Simulated		Predicted	
				W	L	W	L
1	0.35	0.70	0.018	0.36	0.045	0.360	0.046
2	0.51	0.80	0.477	0.36	0.045	0.389	0.054
3	0.67	0.9	1.640	0.72	0.18	0.773	0.156
4	0.81	1.01	8.421	0.72	0.18	2.271	0.438
5	0.98	1.10	47.35	2.16	0.60	2.162	0.598
6	1.14	1.20	72.39	2.16	0.60	2.132	0.602
7	1.3	1.31	163.5	8.00	1.80	8.024	1.786

It is observed from Table I, Table II, and Table III that the channel length (L) and width (W) parameters are fitted well with the simulated values.

Further, using the current mirror circuit, a differential amplifier circuit has been designed as shown in Fig. 4. The node voltages, currents, and design parameters of all MOSFETS are calculated using circuit simulation in Cadence software. In this design, two PMOS and three NMOS are used. All the MOSFETs are simulated and design parameters are predicted using the proposed neural network. The analysis of all MOSFETs are connected in differential amplifier is summarized in Table IV and Table V. It is observed that the simulated and predicted design parameters such as channel length (L) and width (W) are approximately similar which indicates the effectuality of the proposed neural network model.

The three-stage operational amplifier can be designed and implemented using the current mirror, differential amplifier, and two-stage operational amplifier. It is shown in Fig. 5.



(a)

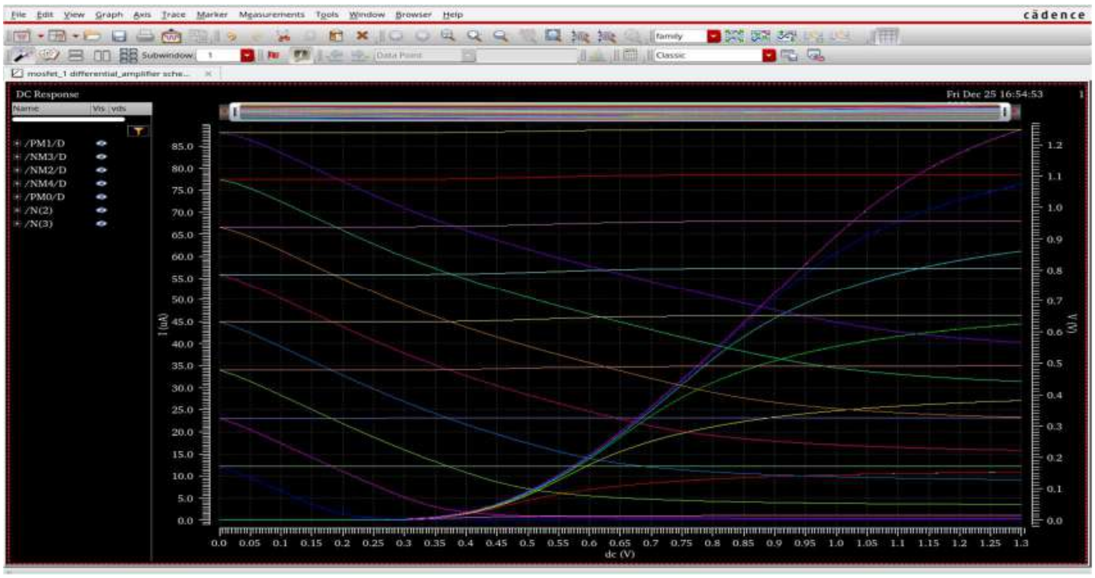


Fig. 4. Differential amplifier (a) Circuit diagram (b) simulated response of MOSFETS

TABLE IV. SIMULATED AND PREDICTED VALUES OF WIDTH AND LENGTH FOR PMOSFETS USED IN DIFFERENTIAL AMPLIFIER

V_{DS} (V)	V_{GS} (V)	PMOS N1					PMOS N2				
		I_D (pA)	Simulated		Predicted		I_D (μA)	Simulated		Predicted	
			W	L	W	L		W	L	W	L
0.20	0.16	0.001	0.36	0.04	0.237	0.019	0.004	0.36	0.04	0.400	0.038
0.35	0.31	0.002	0.36	0.04	0.356	0.041	0.048	0.36	0.04	0.677	0.109
0.51	0.46	0.83	0.72	0.18	0.708	0.183	3.601	0.72	0.18	0.788	0.253
0.67	0.61	0.94	0.72	0.18	0.716	0.181	11.32	0.72	0.18	1.193	0.306
0.82	0.77	9.93	2.16	0.60	2.525	0.642	25.53	2.16	0.60	2.874	0.402
0.98	0.92	11.7	2.16	0.60	2.160	0.599	34.66	2.16	0.60	1.988	0.977
1.14	1.07	43.3	8.00	1.80	8.000	1.799	76.32	8.00	1.80	7.797	1.785
1.30	1.22	61.1	8.00	1.80	8.000	1.800	88.80	8.00	1.80	7.703	1.763

TABLE V. SIMULATED AND PREDICTED VALUES OF WIDTH AND LENGTH FOR NMOSFETS USED IN DIFFERENTIAL AMPLIFIER

V_{DS} (V)	V_{GS} (V)	Simulated W and L		NMOS 1			NMOS 2			NMOS 3		
				I_D (μA)	Predicted		I_D (μA)	Predicted		I_D (pA)	Predicted	
		W	L		W	L		W	L		W	L
0.20	0.16	0.36	0.04	0.04	0.361	0.071	0.004	0.326	0.032	1.18	0.3600	0.0400
0.35	0.31	0.36	0.04	0.48	0.612	0.036	0.480	0.351	0.039	1.67	0.2535	0.0468
0.51	0.46	0.72	0.18	3.67	0.753	0.086	3.67	0.736	0.182	2.98	0.7200	0.1800
0.67	0.61	0.72	0.18	11.3	2.107	0.176	11.3	0.697	0.164	2.98	1.1366	0.2458
0.82	0.77	2.16	0.60	25.4	2.171	0.561	25.41	2.065	0.571	7.27	2.1600	0.6000
0.98	0.92	2.16	0.60	34.6	2.789	0.579	34.66	2.083	0.581	7.87	2.1600	0.6000
1.14	1.07	8.00	1.80	76.3	7.886	1.701	76.32	6.854	1.661	23.08	8.0000	1.8000
1.30	1.22	8.00	1.80	88.8	8.001	1.812	88.80	7.934	1.782	24.18	8.0000	1.8000

TABLE VI. SIMULATED AND PREDICTED VALUES OF WIDTH AND LENGTH FOR NMOSFETS USED IN DIFFERENTIAL AMPLIFIER

Simulated		PMOS 1		PMOS 2		PMOS 3		PMOS 4		PMOS 5		PMOS 6	
W	L	W	L	W	L	W	L	W	L	W	L	W	L
0.36	0.045	0.296	0.040	0.36	0.045	0.295	0.042	0.361	0.044	0.36	0.045	0.35	0.043
8.00	1.80	7.900	1.70	8.00	1.80	7.901	1.72	8.003	1.730	7.99	1.82	8.10	1.82

TABLE VII. SIMULATED AND PREDICTED VALUES OF WIDTH AND LENGTH FOR NMOSFETS USED IN DIFFERENTIAL AMPLIFIER

Simulated		I_D (μA)	NMOS 1		I_D (μA)	NMOS 2		I_D (μA)	NMOS 3		I_D (μA)	NMOS 4	
W	L		W	L		W	L		W	L		W	L
0.12	0.045	0.331	0.12	0.044	0.368	0.12	0.045	0.008	0.122	0.046	0.115	0.112	0.045
3.60	1.80	0.712	3.61	1.822	0.383	3.60	1.82	0.009	3.58	1.79	0.551	3.588	1.810
Simulated		I_D (μA)	NMOS 5		I_D (μA)	NMOS 6		I_D (μA)	NMOS7		I_D (μA)	NMOS 8	
W	L		W	L		W	L		W	L		W	L
0.12	0.045	0.045	0.13	0.042	0.381	0.11	0.044	0.68	0.13	0.045	0.122	0.121	0.045
3.60	1.80	0.181	3.60	1.78	0.463	3.60	1.80	0.38	3.61	1.82	0.57	3.603	1.807

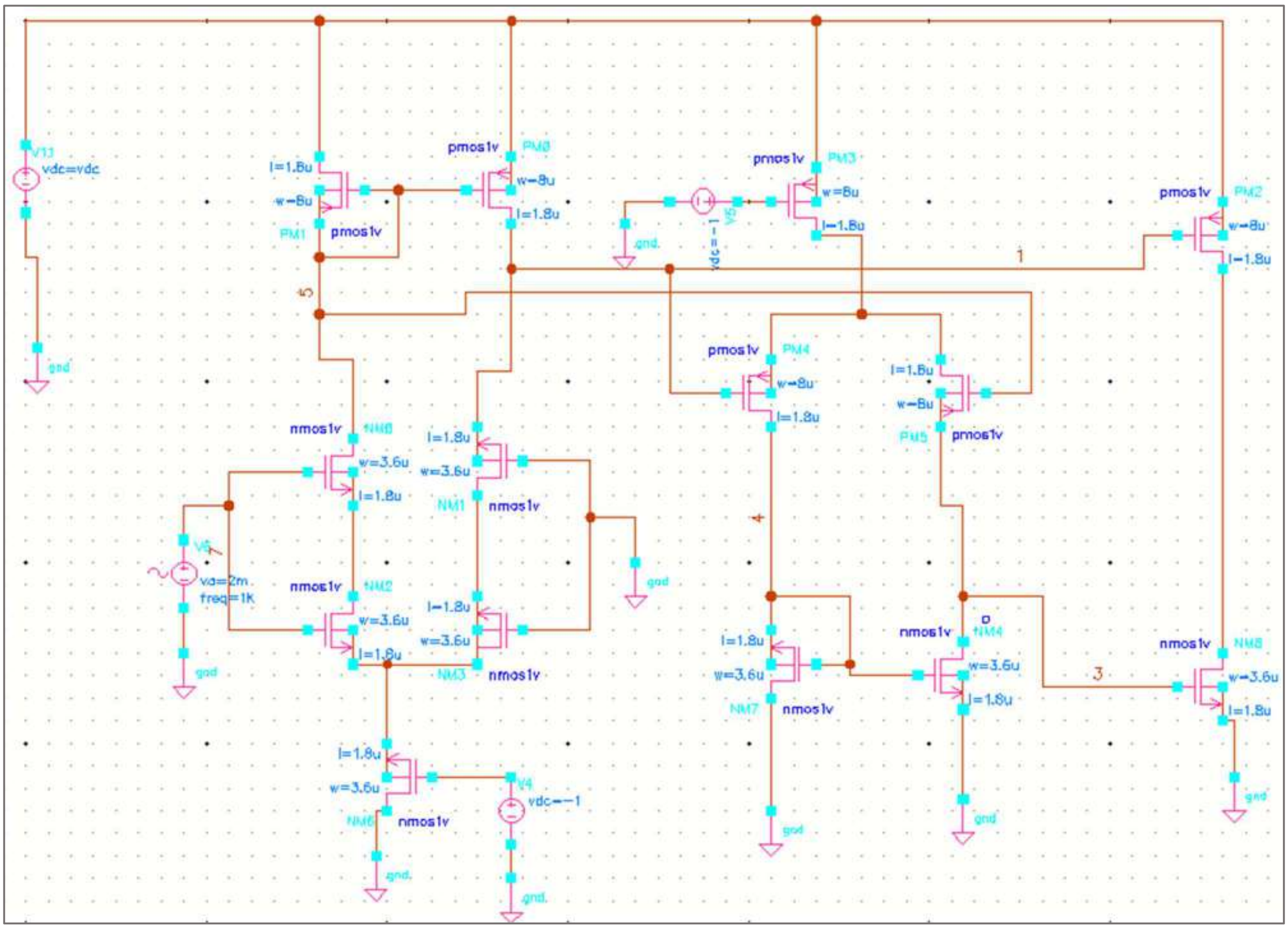


Fig. 5. Design and implementation of three-stage Operational Amplifier

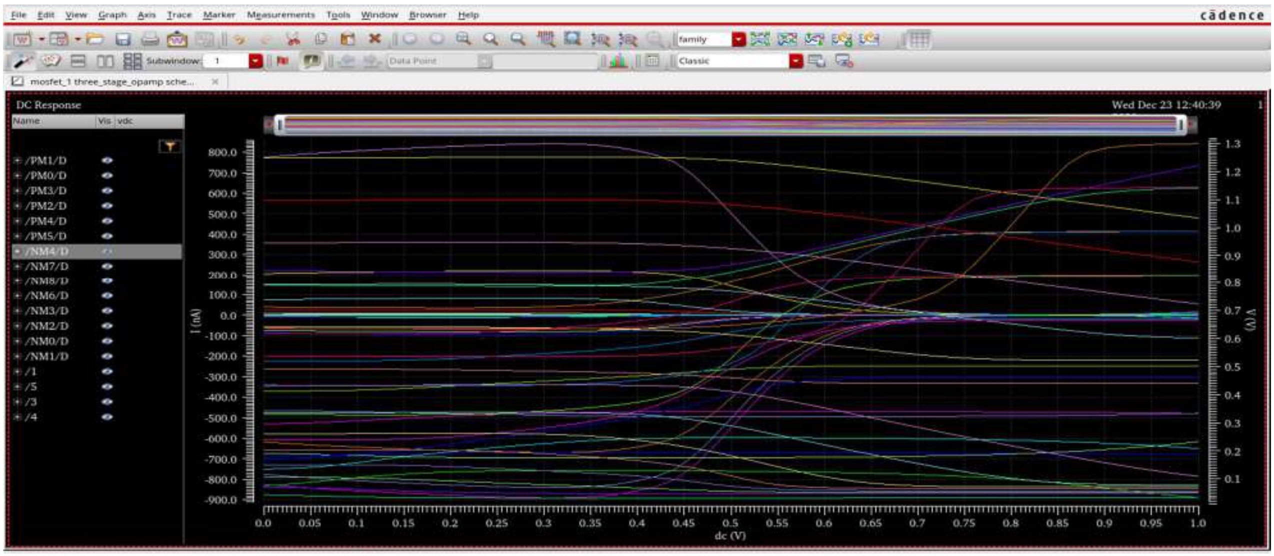


Fig. 6. The simulated output of three-stage Operational Amplifier

The three-stage operational amplifier is simulated using the CADENCE software and the circuit schematic is shown in Fig.5. In this circuit, six PMOS and eight NMOS are used. The analysis of all the MOSFETs has been carried out and design parameters such as channel length and the channel width are predicted using the neural network. However, the limited readings are summarized in Table VI and Table VII, since the test generates huge data which can not be recorded in the form of a table in the paper. Moreover, Fig. 6 shows the characteristics of the three-stage operational amplifier circuit.

IV. Conclusion

The basic building blocks of the operational amplifier are designed and simulated using the Cadence software. The aspect ratio of each MOSFET is calculated using the simulation. Moreover, the channel length and width are designed using the artificial neural network. The various experimentations are carried out using the current mirror, differential amplifier, and three-stage operational amplifier circuits. It is observed from the results that the presented neural network model can predict the design parameters such as channel length and width accurately and it is closed to simulated results. Thus, the proposed approach can be used to design analog and digital VLSI circuits.

References

[1] G. Kothapalli, "Artificial neural networks as aids in circuit design," Microelectron. J., vol. 26, no. 6, pp. 569-578, 1995.

[2] J. Langeheine, S. Fölling, K. Meier and J. Schemmel, "Towards a silicon primordial soup: A fast approach to hardware evolution with a VLSI transistor array," In Int. Conf. Evolvable Syst., 2000, pp. 123-132, Springer, Berlin, Heidelberg.

[3] M. Avcı, M. Y. Babaç and T. Yıldırım, "Neural network based transistor modeling and aspect ratio estimation for yital 1.5 micron process," In Third Int. Conf. Electr. Electron. Eng., ELECO, 2003, pp. 54-57.

[4] M. Avci and T. Yildirim, "Neural network based MOS transistor geometry decision for TSMC 0.18 µ process technology," In Int. Conf. Comput. Sci., 2006, pp. 615-622, Springer, Berlin, Heidelberg.

[5] T. H. Borgstrom, M. Ismail and S. B. Bibyk, "Programmable current-mode neural network for implementation in analogue MOS VLSI," IEE Proc. G (Circuits, Devices Syst.), vol. 137, no. 2, pp. 175-184, 1990.

[6] M. Hayati, A. Rezaei and M. Seifi, "CNT-MOSFET modeling based on artificial neural network: Application to simulation of nanoscale circuits," Solid-State Electron., vol. 54, no. 1, pp. 52-57, 2010.

[7] M. Avci, M. Y. Babac and T. Yildirim, "Neural network based MOSFET channel length and width decision method for analogue integrated circuits," Int. J. Electron., vol. 92, no. 5, pp. 281-293, 2005.

[8] N. Kumar, V. H. Gaidhane, and R. K. Mittal, "Cloud-based electricity consumption analysis using neural network," Int. J. Comput. Appl. Technol., vol. 62, no. 1, pp. 45-56, 2020.

[9] V. H. Gaidhane, N. Kumar, R. K. Mittal and J. Rajevenceltha, "An efficient approach for cement strength prediction," Int. Journal of Comput. Appl., pp. 1-11, 2019.

[10] V. H. Gaidhane, Y. V. Hote, and Vijander Singh, "Emotion recognition using eigenvalues and Levenberg–Marquardt algorithm-based classifier," Sādhanā, vol 41, no. 4, pp. 415-423, 2016.

[11] V. H. Gaidhane, and Y. V. Hote, "An efficient edge extraction approach for flame image analysis" Pattern Anal. Appl., vol. 21, no. 4 pp. 1139-1150, 2018

Constant Tera-ohm Pseudo-resistor Over Wide Dynamic Range

Israa Y. AbuShawish
Electrical Engineering Department
University of Sharjah
Sharjah, UAE
iabushawish@sharjah.ac.ae

Soliman A. Mahmoud
Electrical Engineering Department
University of Sharjah
Sharjah, UAE
solimanm@sharjah.ac.ae

Abstract—This work presents the design of a constant, programmable and extremely high (Tera ohm) MOS pseudo-resistor over a wide dynamic range of ± 0.6 V based on MOS transistors operating in the weak inversion region. The major challenge faced by the circuit designers in designing high resistance pseudo-resistor is the linearity problem and it is resolved in this work. The proposed pseudo-resistor is utilized as a feedback resistor in the second stage of the two-stages bio-medical amplifiers in the portable bio-detection system. By controlling the resistance of this pseudo-resistor, the lower cut-off frequency of the amplifier is automatically adjusted and thus allowing the detection of the small frequency range. Simulations in LT-spice using 130 nm CMOS technology under ± 0.6 V supply voltage are performed to validate the realization of the extremely high resistance and its application.

Keywords **— level-shifter, pseudo-resistor, bio-medical amplifier, bio-detection system, BMI.**

I. Introduction

The Pseudo-resistors are enormously essential devices in any bio-medical system. The relevant signals extracted in the biomedical applications having the frequency range of 0.1 Hz – 10 kHz, while the amplitude is ranging between 20µV – 10mV [1]–[4] Therefore, the need for the pseudo-resistors is vital to realize extremely high resistance values that can be reached to kilo-ohms resistance as in [5] and few giga-ohm resistance as in [6], [7], and few Tera-ohm resistance as in [8]–[11]. Besides, it will ensure a smaller chip area despite the usage of the high resistance value, with minimal designing cost and lower power consumption [7]. These resistors are preferred to be utilized over the conventional ones in many important circuit blocks such as the trans-impedance amplifiers, data converters, operational transconductance amplifiers, and multipliers as an integral part of the biomedical application, temperature sensing application and low current sensing application, etc. [12].

Recently, the neurobiologist and neuroscience researchers have devoted much of their research and studies to the field of bio-medical interfacing systems [13], [14]. The industrial revolution, the remarkable development in the Integrated Circuits, and the great achievements of the scientists up to date have been directed the attention towards the bio-medical interfacing systems field. The neural recording implants are the core element in the BMIs which consist of three main blocks, bio-medical amplifiers, ADCs and compressors [15], [16].

The main motivation behind this work is to provide an almost constant high resistance (reach to Tera ohm) pseudo-resistor, applicable for a wide dynamic range of ± 0.6 V. Thus, the performance of the application in the bio-medical analog circuits won't be vitiated and distorted. based on the proposed pseudo-resistor as a feedback element in the second stage amplifier. In this work, the proposed pseudo-resistor circuit will be investigated and designed in section II. The design of the bio-medical amplifier as an application of the proposed pseudo-resistor is thoroughly discussed, analyzed and simulated in section III. A conclusion is presented in section IV.

II. Proposed Pseudo-resistor

The proposed pseudo-resistor presented in Fig. 1. is designed using two series (NMOS- M_{a2}) and (PMOS- M_{b2}) transistors, each connected with a level shifter (PMOS- M_{c2}) and (NMOS- M_{d2}), respectively, which affords the capability to dynamically adjusting the gate to source voltages of the two transistors. This adjustment will maintain a constant V_{GS} on the pseudo-resistor over large output voltage swings. Meanwhile, by controlling the bias currents in the level shifters, the resistance value can be controlled. However, utilizing two different types of transistors (NMOS and PMOS) will be useful in implementing a large and constant R over the wide dynamic range, by ensuring at least one transistor has the proper operation and adjusted V_{GS}. Utilizing one NMOS transistor (M_{a2}) with a level shifter will guarantee that $V_{4,1} = V_{GSa2}$

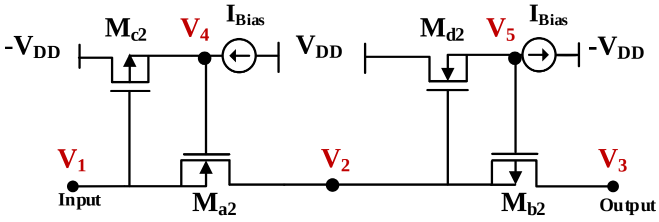


Fig. 1. The structure of the proposed pseudo-resistor

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

= V_{SGc2}, which will maintain a proper and constant V_{GSa2} value, meanwhile ensure that M_{a2} operating in the subthreshold region, thus a constant, linear and large resistance value over the positive voltage swing will be emulated. Knowing that, The I_{DS} expression for the NMOS transistor is given in (1), where $I_{Dn} = 2n\, \mu_n C_{ox} U_T^2\, W/L$, and U_T, V_{To}, n, μ_n, C_{ox}, W and L are defining the thermal voltage, MOS threshold voltage, subthreshold slope factor, mobility of electrons, gate oxide capacitance per unit, channel width, and length of the MOS transistor respectively. Hence, its equivalent resistance is given in (2), where I_{bias} is the biasing current source in the level shifter M_{c2}.

$$I_{DS} = I_{Dn}\left(e^{-\frac{V_{SB}}{U_T}} - e^{-\frac{V_{DB}}{U_T}}\right)e^{\frac{V_{GB}-V_{To}}{n\,U_T}} \qquad (1)$$

$$R_{2,1} = 1/\frac{\partial I_{2,1}}{\partial V_{2,1}} = \frac{U_T}{I_{bias}}.\frac{I_{Dpc2}}{I_{Dna2}}.\frac{1-e^{-\frac{V_{SDc2}}{U_T}}}{e^{-\frac{V_{2,1}}{U_T}}} \qquad (2)$$

On the contrary, utilizing one PMOS transistor (M_{b2}) with a level shifter will maintain a fixed value of $V_{2,5} = V_{SGb2}$ which will guarantee that M_{b2} is operating in the subthreshold region, accordingly preserve a constant, linear and large resistance value over the negative voltage swing. The equivalent R of this pseudo-resistor is given in (3), where I_{bias} is the biasing current source in the level shifter M_{d2}.

$$R_{3,2} = \frac{U_T}{I_{bias}}.\frac{I_{Dnd2}}{I_{Dpb2}}\left(\frac{1-e^{\frac{-V_{DSd2}}{U_T}}}{e^{\frac{V_{3,2}}{U_T}}}\right) \qquad (3)$$

By connecting the two cells mentioned in a series connection, as declared in Fig. 1 the stated drawbacks of using only one cell will be avoided, and thus validate that the cells emulating a constant, linear and high resistance value (1.08 TΩ) over the entire positive and negative voltage swing as demonstrated in Fig. 2. The size of the transistors M_{a2} and M_{b2} is (W= 0.26 μm, L= 65 μm) and M_{c2} and M_{d2} is (W= 200 μm, L= 1.5 μm).

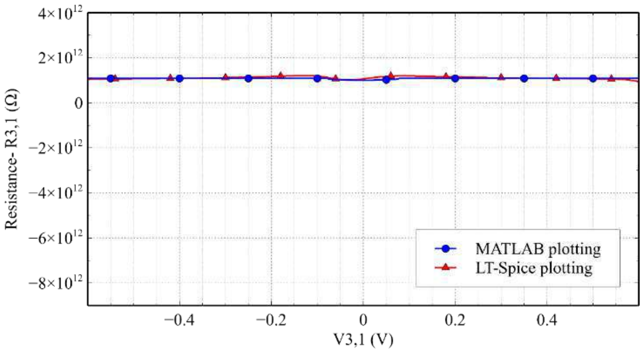


Fig. 2. The $R_{3,1}$ vs. $V_{3,1}$ curve of the proposed pseudo-resistor using both the LT-Spice and MATLAB simulators.

It's demonstrated also from Fig. 2 that $R_{3,1}$ curve plotted using LT-Spice is almost identical to the one plotted using MATLAB. Besides, this fact can be justified by the theoretical analysis, where the derived expression of $R_{3,1}$ under both conditions of $V_3 < V_1$ and $V_1 > V_3$ is expressed by the same equation, presented in (4). However, a small drop in the R reached 1.01 TΩ is observed near $V_{3,1}$= 0V; due to the gap between the realized resistance by NMOS and PMOS pseudo-resistors.

$$R_{3,1} = \frac{U_T}{I_{bias}}\left[\frac{I_{Dpc2}}{I_{Dna2}}\left(\frac{1-e^{\frac{-V_{SDc2}}{U_T}}}{e^{\frac{-V_{2,1}}{U_T}}}\right) + \frac{I_{Dnd2}}{I_{Dpb2}}\left(\frac{1-e^{\frac{-V_{DSd2}}{U_T}}}{e^{\frac{V_{3,2}}{U_T}}}\right)\right] \qquad (4)$$

III. Bio-medical Amplifier Design

As an application of the proposed pseudo-resistor, an op-amp-based bio-medical two-stage amplifier has been designed by employing the proposed pseudo-resistor as a feedback resistor in the second stage of the amplifier while using another non-linear pseudo-resistor as a feedback resistor in the 1st stage of the amplifier.

A. First Stage Amplifier Design and Simulation

This stage presented in Fig. 3 involves input capacitors utilized to block the DC offset voltages which are produced as a result of the electrochemical interaction at the electrode-tissue interface. This voltage considered a huge problem since it can alter and affect the mode of operation of the transistors and saturate the amplifier. The issue is arising since these offset voltages which range between (1 mV – 50 mV) have higher voltages values compared to the brain's signals, thus it is compulsory to get rid of these offset voltages [15].

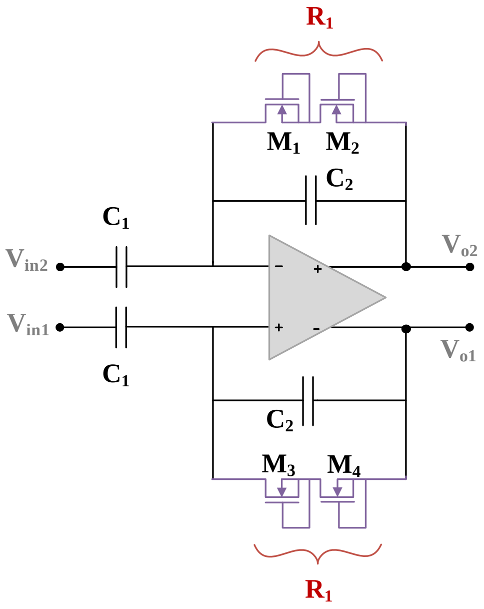


Fig. 3. The structure of the first stage bio-medical amplifier

This stage is also comprising of feedback capacitors along with a feedback resistor R_1 design using two series NMOS transistors at an equal aspect ratio of (1 µm /100 µm) to produce the pole frequency at 0.1 Hz. This pseudo-resistor is based on the symmetrical biasing for the body and V_G of the two NMOS transistors operating in the subthreshold region. This architecture of pseudo-resistor, is able to emulate an extremely high resistance value, reached to tens of Tera ohm but on the contrary, its value is not constant therefore it is not suitable to be used for the last stage of the amplifier due to the high output voltage swing and due to the linearity problem. Nevertheless, using this cell of pseudo-resistor is appropriate to emulate high resistivity in the first stage of the amplifier to achieve a very low cutoff frequency.

The gain of this stage is 31.8 dB which is the ratio between the input capacitor C_1 to the feedback capacitor C_2. Where the active block designed using a fully differential folded cascaded op-amp with A_v of 74.28 dB and ϕ_m of 86.5° [17]. The design structure of this op-amp is presented in Fig. 4.

The simulation in LT-spice has been conducted for verification purposes using 130 nm model, BSIM032, level = 4, ± 0.6 V supply. TABLE I. showed the design specifications and components values of the 1st stage amplifier and its active block. The magnitude and phase responses of this stage are presented in Fig. 5 (a) & (b) respectively.

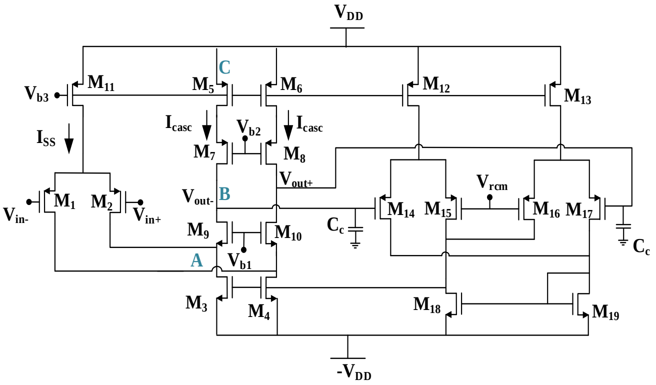
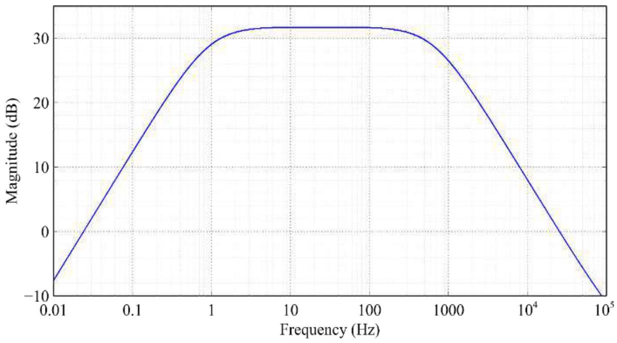
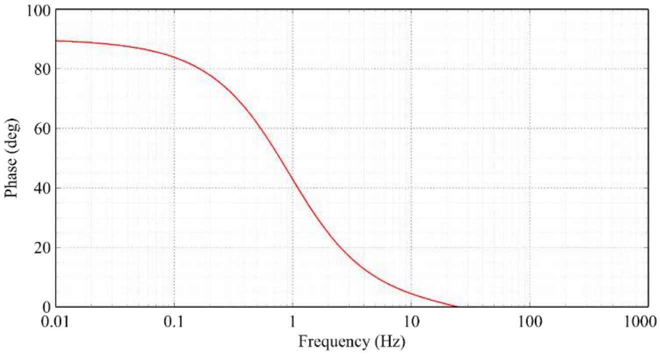


Fig. 4. The structure of the fully differential folded cascode Op-amp with the CMFB circuit



(a)



(b)

Fig. 5. The response of the first stage bio-medical amplifier: (a) The magnitude response and (b) The phase response

TABLE I. THE DESIGN AND COMPONENTS VALUES OF THE FIRST STAGE BIO-MEDICAL AMPLIFIER

	Components	**Value**
First stage amplifier	C_1	11.7 pF
	C_2	0.3 pF
	Design parameters	**Value**
	Gain	31.9 (dB)
	Cutoff frequency ω_o	0.629 (rad/sec)
	Transistor	**Aspect ratios (μm)**
Folded-cascode Op-amp	**Op-amp**	**W/L (µm)**
	M_1, M_2	50/1
	M_3-M_{10}	2/30
	M_{11}	5/1
	CMFB	**W/L (µm)**
	M_{12}, M_{13}	10/1
	M_{14}-M_{17}	71/1
	Design parameters	**Value**
	Supply Voltage (V)	± 0.6
	Gain (dB)	76.4
	Gain bandwidth (Hz)	44.8 k
	Phase margin (°)	86.6
	Power (Watt)	4.86 μ
	Slew rate (V/μS)	11.02 k
	Offset Voltage (V)	11.02 k

B. *Second Stage Amplifier Design and Simulation*

The structure of this stage shown in Fig. 6 comprises coupling input capacitors C_3, feedback capacitor C_4 and feedback resistor R_2 which designed using the proposed pseudo-resistor, since it required to design a pseudo-resistor that can emulate a high resistance value reached to T ohm resistance, provided that its linearity is high. Employing the proposed pseudo-resistor in this stage is the best solution to preserve an almost constant resistance value over a wide dynamic range of ±0.6 V, which leads to high linearity and almost constant pole frequency in the bio-

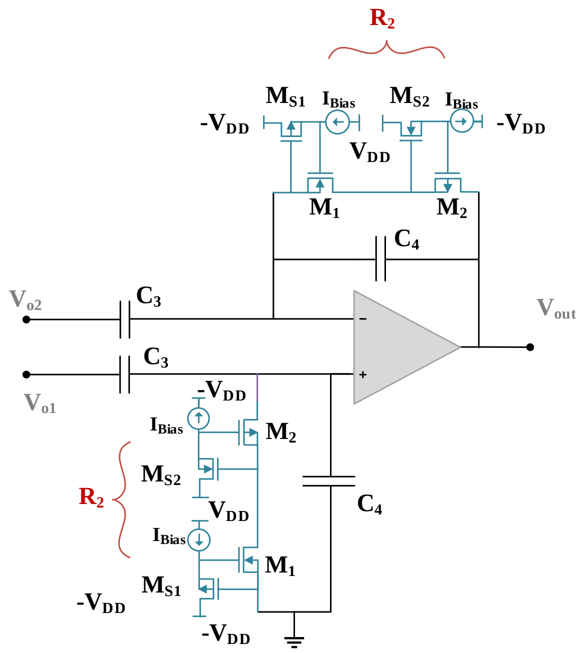


Fig. 6. The structure of the second stage bio-medical amplifier

medical amplifier design. The gain of this stage is 22.9 dB set by the ratio of the input capacitor C3 to the feedback capacitor C4. The active block was designed using a single-ended two-stage operational amplifier with a gain of 110.9 dB and 138.84° phase margin. The design structure of this op-amp is provided in Fig. 7. [17], [18].

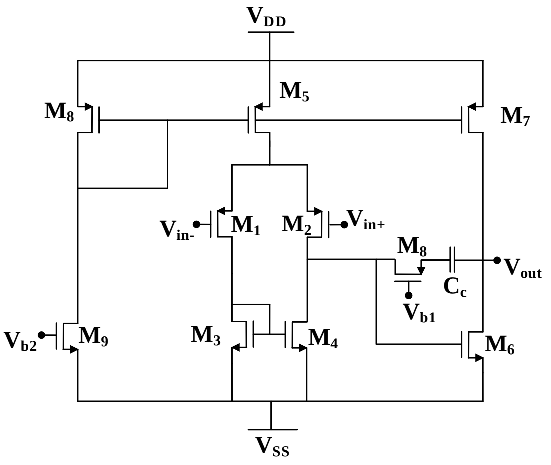
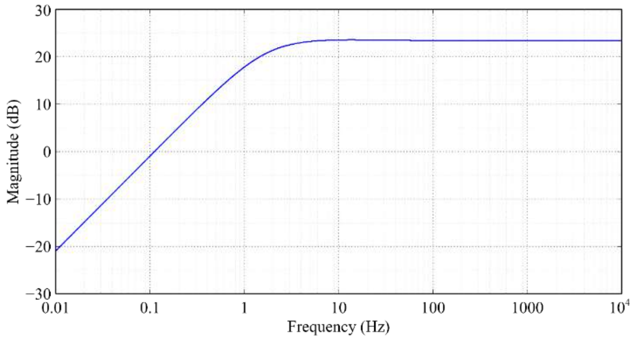
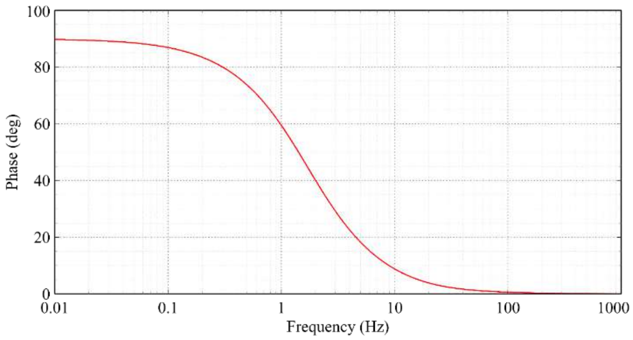


Fig. 7. The structure of the single-ended two-stages Op-amp



(a)



(b)

Fig. 8. The response of the second stage bio-medical amplifier: (a) The magnitude response and (b) The phase response

TABLE II. provides the design specifications and components value of the 2^{nd} stage bio-medical amplifier and its active block. The magnitude and phase of this stage are demonstrated in Fig. 8 (a) & (b) respectively.

TABLE II. THE DESIGN AND COMPONENTS VALUES OF THE SECOND STAGE BIO-MEDICAL AMPLIFIER

	Components	**Value**
Second stage amplifier	C_3	1.4 pF
	C_4	0.1 pF
	Design parameters	**Value**
	Gain	22.9 (dB)
	Cutoff frequency ω_o	0.628 (rad/sec)
	Transistor	**Aspect ratios (μm)**
Two-stage op-amp	M_1, M_2	2/5
	M_3, M_4	1/10
	M_5	1.6/3
	M_6	2/3.8
	M_7	4/3
	M_9	5/1
	M_8	1.6/3
	M_{10}	0.39/50.05
	M_{11}	5/1
	Design parameters	**Value**
	Supply Voltage (V)	± 0.6
	Gain (dB)	110.9
	Gain bandwidth (Hz)	54.9 k
	Phase margin (°)	138.84
	Power (Watt)	101.44 n
	Slew rate (V/μs)	857
	Offset Voltage (V)	0.65 μ

C. Overall Amplifier Design and Simulation

The structure of the two-stages bio-medical amplifier is presented in Fig. 9. It's worth noting that besides the linearity advantage provided by the usage of the proposed pseudo-resistor, it also facilitates controllability to the lower cutoff frequency of the amplifier. Accordingly, by decreasing the biasing current in the level shifter, the lower cutoff frequency will reduce thus, the bio-medical amplifier will be able

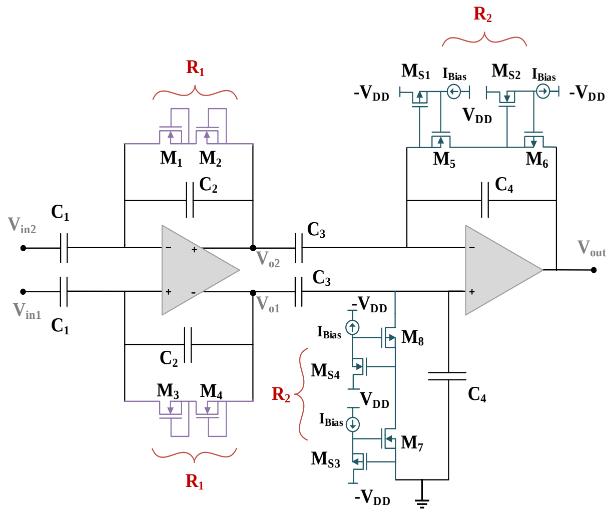
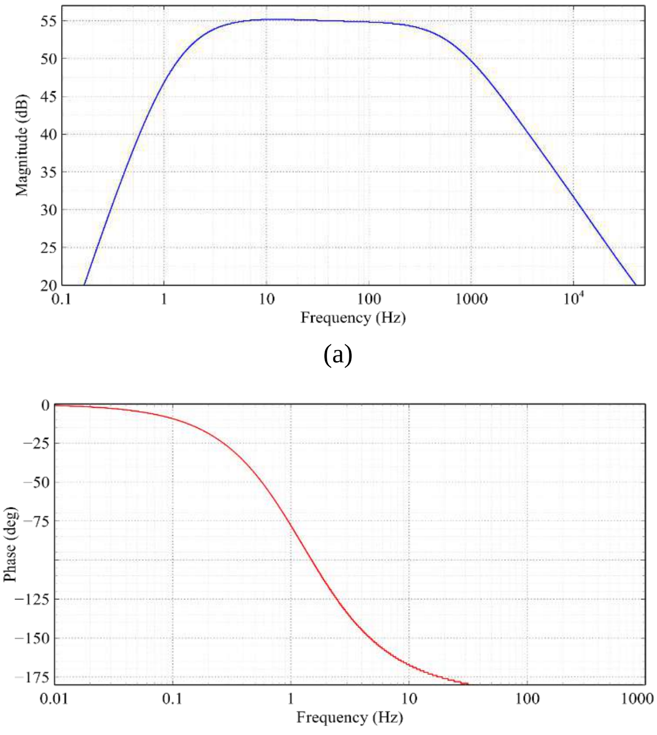


Fig. 9. The structure of the two-stages bio-medical amplifier

to detect the small frequency range. The magnitude and phase responses of the two-stages bio-medical amplifier are demonstrated in Fig. 10 (a) & (b) respectively. The response of the overall op-amp-based bio-medical amplifier is simulated over three different biasing currents; 1nA, 1.5nA, and 2nA as demonstrated in Fig. 11, where the corresponding lower cutoff frequencies are; 1.8 Hz, 2.3 Hz, and 2.8 Hz respectively. The simulated result of the input-referred noise spectral density of the two-stages bio-medical amplifier is shown in Fig. 12. The design parameters of the overall bio-medical amplifier are summarized in TABLE III.



(a)

(b)

Fig. 10. The response of the two-stages bio-medical amplifier: (a) The magnitude response and (b) The phase response

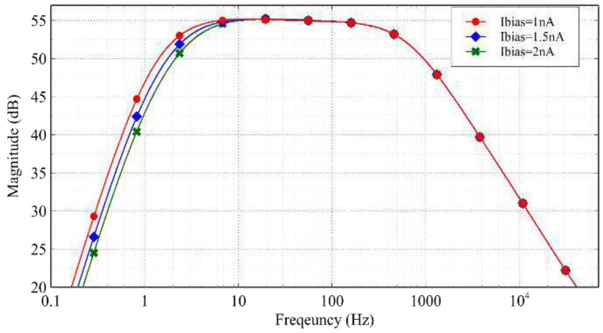


Fig. 11. The response of the two-stages bio-medical amplifier over three different biasing current values; 1nA, 1.5nA and 2nA

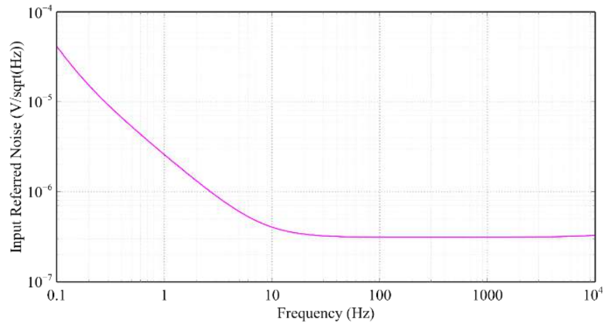


Fig. 12. The IRN spectral density of the two-stage bio-medical amplifier

TABLE III. THE DESIGN COMPONENTS AND PARAMETERS OF THE TWO-STAGE BIO-MEDICAL AMPLIFIER

Design parameters	Value
Gain	54.9 dB
Supply voltage	±0.6 V
Bandwidth	678.2 Hz
IRN (above 10 Hz)	$312.8\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$
IRN @0.1 Hz	$41.66\ \mu\mathrm{V}/\sqrt{Hz}$
Power consumption	4.69 µWatt
THD @100 Hz, 1mV$_{pk-pk}$	0.327% (-49.7 dB)
IM3 @50 and 60 Hz, 1mV$_{pk-pk}$	46.49 dB
IIP3	22.94 dBm
Gain	54.9 dB
Transistors of pseudo-resistors	**Value**
M$_1$ – M$_4$	1 µm /100 µm
M$_5$ – M$_8$	0.26 µm /65 µm
M$_{S1}$ – M$_{S4}$	200 µm /1.5 µm

IV. CONCLUSION

A novel CMOS pseudo-resistor using NMOS and PMOS transistors with level shifters is proposed. This cell affords an extremely high and almost constant resistance over a wide dynamic range of ±0.6 V, which leads to high linearity and almost constant pole frequency when it’s used in a bio-medical filter design application. The performance of the resistance is tested and applied in a two-stage bio-medical amplifier and simulation tests are provided and confirm the high linearity performance and the fine-tuning of the lower cut-off frequency.

REFERENCES

[1] A. A. Alhammadi, T. B. Nazzal, and S. A. Mahmoud, “A CMOS EEG detection system

with a configurable analog front-end architecture," *Analog Integr. Circuits Signal Process.*, vol. 89, no. 1, pp. 151–176, 2016, doi: 10.1007/s10470-016-0826-x.

[2] S. A. Mahmoud, H. A. Salem, and H. M. Albalooshi, "An 8-bit , 10 KS / s , 1 . 87 µ W Successive Approximation Analog to Digital Converter in 0 . 25 µ m CMOS Technology for ECG Detection Systems," *Circuits, Syst. Signal Process.*, vol. 34, pp. 2419–2439, 2015, doi: 10.1007/s00034-015-9973-z.

[3] S. I. Khan, M. S. Diab, and S. A. Mahmoud, "Design of low power Teager Energy Operator circuit for Sleep Spindle and K-Complex extraction," *Microelectronics J.*, vol. 100, no. March, p. 104785, 2020, doi: 10.1016/j.mejo.2020.104785.

[4] M. B. Elamien and S. A. Mahmoud, "On the design of highly linear CMOS digitally programmable operational transconductance amplifiers for low and high-frequency applications," *Analog Integr. Circuits Signal Process.*, vol. 97, no. Feb, pp. 225–241, 2018, doi: 10.1007/s10470-018-1128-2.

[5] R. R. Harrison *et al.*, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 123–133, 2007, doi: 10.1109/JSSC.2006.886567.

[6] J. N. Y. Aziz *et al.*, "256-channel neural recording and delta compression microsystem with 3D electrodes," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 995–1005, 2009, doi: 10.1109/JSSC.2008.2010997.

[7] R. H. Olsson, D. L. Buhl, A. M. Sirota, G. Buzsaki, and K. D. Wise, "Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 7, pp. 1303–1311, 2005, doi: 10.1109/TBME.2005.847540.

[8] K. Sharma, A. Pathania, R. Pandey, J. Madan, and R. Sharma, "MOS based pseudo-resistors exhibiting Tera Ohms of Incremental Resistance for biomedical applications : Analysis and proof of concept," *Integr. VLSI J.*, vol. 76, no. March 2020, pp. 25–39, 2021, doi: 10.1016/j.vlsi.2020.08.001.

[9] K. Sharma, L. Goyal, and L. Gupta, "Implementation of Tunable and Non-Tunable Pseudo-Resistors using 0.18µm Technology," *Int. J. Comput. Appl.*, no. December, 2016.

[10] R. R. Harrison, C. Charles, and S. Member, "A Low-Power Low-Noise CMOS Amplifier for Neural Recording Applications," vol. 38, no. 6, pp. 958–965, 2003.

[11] N. M. Laskar *et al.*, "Design of high gain, high bandwidth neural amplifier IC considering noise-power trade-off," *Microsyst. Technol.*, vol. 5, 2018, doi: 10.1007/s00542-018-4142-5.

[12] K. Abdelhalim, L. Kokarovtseva, J. L. Perez Velazquez, and R. Genov, "915-MHz FSK/OOK wireless neural recording soc with 64 mixed-signal fir filters," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2478–2493, 2013, doi: 10.1109/JSSC.2013.2272849.

[13] S. A. Mahmoud, A. Bamakhramah, and S. A. Al-tunaiji, "Six Order Cascaded Power Line Notch Filter for ECG Detection Systems with Noise Shaping," *Circuits, Syst. Signal Process.*, 2014, doi: 10.1007/s00034-014-9761-1.

[14] S. A. Mahmoud, A. Bamakhramah, and S. A. Al-tunaiji, "Low-Noise Low-Pass Filter for ECG Portable Detection Systems with Digitally Programmable Range," *Circuits, Syst. Signal Process.*, 2013, doi: 10.1007/s00034-013-9564-9.

[15] F. H. Noshahr, M. Nabavi, and M. Sawan, "Multi-Channel Neural Recording Implants : A Review," *sensors*, pp. 1–29, 2020, doi: 10.3390/s20030904.

[16] A. A. Alhammadi and S. A. Mahmoud, "Fully differential fifth-order dual-notch powerline interference filter oriented to EEG detection system with low pass feature," *Microelectronics J.*, vol. 56, pp. 122–133, 2016, doi: 10.1016/j.mejo.2016.08.014.

[17] H. Kassiri, K. Abdelhalim, and R. Genov, "Low-distortion super-GOhm subthreshold-MOS resistors for CMOS neural amplifiers," *2013 IEEE Biomed. Circuits Syst. Conf. BioCAS 2013*, pp. 270–273, 2013, doi: 10.1109/BioCAS.2013.6679691.

[18] I. B. Attili and S. A. Mahmoud, "Survey on Single Stage Amplifiers for Column Drivers in Active Matrix LCD Panels Leading to a Highly Linear Rail-to-Rail Robust Amplifier," *IEEE Access*, vol. 7, pp. 166629–166647, 2019, doi: 10.1109/ACCESS.2019.2954002.

Metadata Replication with Synchronous OpCodes Writing for Namenode Multiplexing in Hadoop

Taeha Kim
Dept. of Artificial Intelligence
Ajou University
Suwon, South Korea
kth@ajou.ac.kr

Sangyoon Oh †
Dept. of Artificial Intelligence
Ajou University
Suwon, South Korea
syoh@ajou.ac.kr

***Abstract*— A single Active Namenode (ANN) of Hadoop Distributed File System (HDFS) become a bottleneck when we require high-throughput read operations such as large-scale data analysis. Recently, various kinds of namenode schemes are proposed including asynchronous check pointing schemes to address the ANN bottleneck issue. Even if asynchronous schemes offers high throughput reading operations, they suffers in stale read problem where the latest data return is not guaranteed. In this paper, we propose a novel metadata replication scheme with synchronous OpCodes writing to achieve namenode multiplexing, where we can avoid the stale read problem. To reduce synchronization overhead, our proposed scheme conducts reduced replication only for metadata updates such as a write request, using quasi byte-level metadata operation codes. We conducted the empirical experiment to verify the effectiveness of our proposed schemes. The results show that our method reduces by 50.95% in the average required number of NNs when the number of NNs for read-only operation is 100.**

Keywords—HDFS, Metadata Replication, Staleness Elimination, Strong Consistency

I. Introduction

As the size of big data continues to grow, many distributed file systems have been used to store them. The Hadoop Distributed File System (HDFS) [1] became a popular choice for storage because of its high throughput and scalability. Thus, several hundred petabyte-scale data mining and analysis is feasible with HDFS. In data mining, a method to find specific patterns in the data, a mining application queries data to the HDFS. A transaction of those queries usually consist of high ratio of read operations [2] approximately 95% [3-4] of all requests. However, these high read operation ratio causes performance degradations in many cases. For example, Zhang et al. [3] reported a case of significant performance degradation from their experiments, where 500,000 queries to HDFS were produced and 90% of queries took more than 100 seconds. They analyze the experiments results and conclude that the limited read throughput was caused by the HDFS with A single Active Namenode (ANN), which is metadata.

Since the introduction of HDFS to the big data domain, there has been many studies that address the limitation of a single ANN especially for read requests. Shvachko et al. proposed the Observer Namenode (ObNN) in their study [5], which was designed only to process read requests. It is differentiated from ANN that processes both read and write requests. However, in order to use these ObNNs, Journal Nodes (JNs) and Quorum Journal Manager (QJM) are required as well as Standby Namenodes (SbNNs) to improve fault-tolerance of ANN. Thus, as the number of NN is required more and more, we can recognize the HDFS system complexity and cost is higher and higher. Therefore, HDFS needs asynchronous to improve performance. However, asynchronous approach can be occurred stale reads. Stale read problem where the latest data return is not guaranteed happens with high performance asynchronous techniques, and it may be occurring repeatedly that will yield serious performance degradation.

Stale read amplifies synchronizing overheads of the metadata changes. Stale read yields delay in read operations since to obtain the latest data, additional synchronization workload is required and it can be repeated. Additionally, stale read increases read latency and make it difficult to achieve real-time responsiveness. Furthermore, because it affects to other metadata processing operations [4], delay is amplified further. As a result, as the number of node that serves metadata grows, the throughput of metadata read operations is decreased. There are other problems of stale read. Even if a read request arrives after a write request is completed, the time at which the data are available is non-deterministic, which can affect the accuracy of data analysis [8]. Also, because data from stale read violates security requirements, vulnerabilities in data leak are detected [9] or restricted [10] to avoid general reads. Thus, stale reads are a critical problem to be resolved in a large-scale data analysis not only because it affects overall performances but also yields security vulnerabilities.

For better improvement in read throughput, the metadata multiplexing is better choice in many cases. The metadata is a descriptive information of data (e.g., file name, path, size, and created time). Among the all access, the number of accessing to metadata are majority. In the study [6], the metadata access occupies 50%–80% [6] of all accesses of the file system. Therefore, it significantly affects the overall performance.

To achieve metadata multiplexing, the replication technique is critical. Replication technique is well known for improving performance, availability and reliability. However, even if metadata occupies relatively very small (i.e., 0.1%–1% in [6]) the overhead resulting from synchronizing copies of metadata

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

whenever writing operation happens is a serious performance bottleneck. We may alleviate the synchronizing overhead of the metadata changes with a consistency protocol. However, still there will be stale reads problem to be resolved. The consistency protocol is used to maintain the metadata changes between copies of the metadata. Strong consistency [7] of protocols ensures that all metadata changes are applied uses a 'synchronous' synchronization technique. Therefore, there is no metadata mismatch but high overhead has to be endured. On the other hand, both weak consistency and eventual consistency [7] of protocol do not guarantee that all metadata changes are applied. In addition, these two consistency policies use 'asynchronous' synchronization techniques. Thus, the stale reads may occur (i.e., not the latest data is returned for the read request) even with relatively low overhead.

In this paper, we propose a novel read-only metadata multiplexing method to improve read throughput with reduced read latency as well as low metadata synchronizing overhead for changes. To achieve the reduced overhead, we introduce byte-level metadata and replicates only minimum metadata, i.e. we reduce size and the number of replications. We could achieve this by not replicating entire metadata. We also apply strong consistency policy to avoid stale reads.

To evaluate the effectiveness of our proposed method, we implement read-only namenode multiplexing on Hadoop. First, we implement the write operation codes (opcodes) of the metadata for replication in byte-level. Also, we replicate metadata only when they are changed (e.g., write requests occur) to reduce the number of replications. Second, for strong consistency, we use the eager primary-copy model [11]. Based on these two design/implementation, we achieve the increase of the number of ANN and the replication of metadata from the ANN to the ANN's replicas using the primary-copy model. The proposed ANN process both read and write requests. On the other hand, the replicas can process read-only requests.

We identify the main contributions of this paper are as follows:

1. Read-only metadata multiplexing method to maintain read throughput
2. Metadata replication method of byte-level opcodes for reduced synchronization
3. Synchronization technique of in-memory and on-disk metadata for read-after-write in a replica node

The remainder of this paper is organized as follows. Section II describe the related work. Then we describe the proposed scheme in Section III. The evaluation results are presented and analyzed in Section IV, and we conclude in Section V.

II. Related Work

A. Asynchronous checkpointing

Mohan et al. [14] investigated various checkpointing approaches such as asynchronous checkpointing in High Performance Computing. Among those approaches, synchronous checkpointing occurred large checkpoint stalls. The GPU is used for learning, and the CPU is used for checkpointing. In the past, GPU-based learning had to be stopped for checkpointing, but in the proposed technique, the CPU is minimized by pipelining through two-phase checkpointing called snapshot and persist so that the CPU performs as background regardless of learning. However, it may not be effective in CPU-based learning. Because checkpointing must also be performed while learning, the CPU load can be increased and learning may be delayed. Also, HDFS is not suitable because CPU is used, and such asynchronous checkpointing is known to cause stale. Because checkpointing is k-iteration, it must be performed periodically and repeatedly. However, it is necessary to ensure that it is performed only when metadata changes occur so that it cannot be always performed repeatedly.

In ObNN, Journal nodes (JNs) periodically back up the metadata of ANNs using a checkpoint; this is one of the reverse error recovery techniques involving a state-based backup method. However, the ObNN waits for the ANN's metadata to be transmitted to process a read request, and if a read request arrives before synchronization due to metadata inconsistency, then a stale read occurs. To mitigate this problem, the ObNN verifies the JNs for updated metadata before processing a read request. To process read requests that require the latest information, such as the quota usage API for the directory, the synchronization process is repeated until the latest metadata are obtained. ANN's metadata is asynchronously replicated and synchronized in several stages from the QJM's JNs to SbNN to the ObNN. The QJM periodically requests each JN to replicate the ANN's metadata to replicate and synchronize the metadata. When more than half of the odd-numbered JNs complete synchronization with the latest metadata, this cause write delay. Then the metadata are copied to the SbNN. When SbNNs download the metadata, they transmit the metadata to the ObNN. Thus, replicating asynchronously an entire single metadata over several nodes causes significant synchronization overhead.

B. Metadata distribution and replication for fault-tolerance

Niazi et al. [4] proposed a multiple namenode with single leader NN (e.g., ANN). They divided equally the entire metadata of single ANN into a distributed database. However, this causes relatively high network communication and synchronization overhead because this should be searched via network to get the metadata information. To alleviate the overhead, they placed the client side cache. This cause stale cache and the overhead that periodically needs to be updated. Moreover, the node number of distributed database that they used was 48 and recently it has increased to 64. This means that scalability is limited. Wang et al. [16] proposed multiple SbNNs that failover the ANN to remove single point of failure of the ANN. These all SbNNs cannot process read requests. SbNN of existing HDFS also still cannot process read requests.

C. Write delay of strong consistency

In CephFS [12], metadata was synchronized using strong consistency, such as primary-copy model. Therefore, metadata inconsistency due to writing did not occur, but high network traffic occurred between nodes due to metadata discovery and synchronization. Furthermore, owing to communication overhead, the low-latency requirement cannot be satisfied. In a

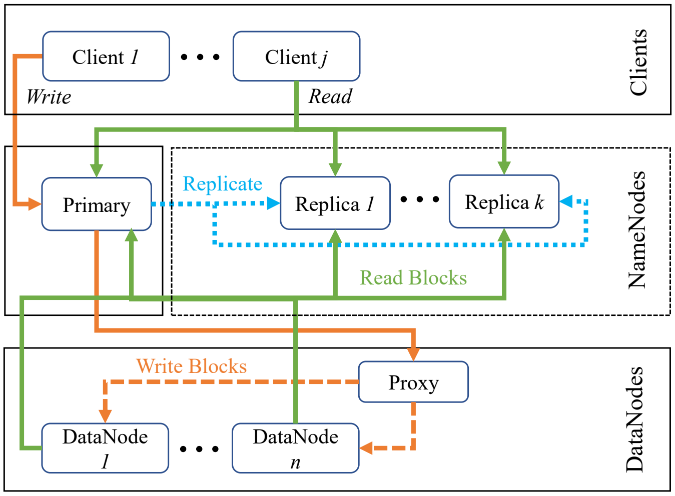


Fig. 1. Our proposed scheme for namenode replication

study by Jiayuan et al. [13], to lower the consistency level from strong to weak, the primary node that committed data asynchronously replicated metadata to the remaining replication nodes using a ring. This reduced the write latency but amplified stale reads. Thus, write overhead needs to be reduced or minimized.

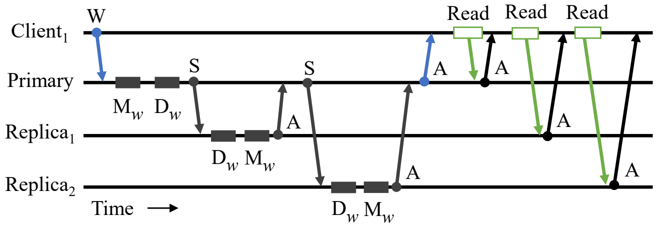
D. Stale reads

In CephFS, stale reads occurred owing to the size of the metadata cache; therefore, reads were repeated to obtain the latest data. To solve this problem, reading was repeated or metadata were reread when the lease expired such that the latest data were updated using read lease [15]. Ho et al. [8] proposed a stale synchronous parallel model that allows stale data in bulk synchronous parallel systems such as Hadoop as a parameter server system for distributed machine learning. However, regarding learning accuracy, even if specific data are not used for learning, accuracy loss occurs, and using invalid data affects accuracy. Therefore, if the data are old or stale, the corresponding parameter and variable values are not used through limiting the maximum age of the stale value. Snavely et al. [9] stated that stale data may violate security attributes, particularly confidentiality-based requirements, may contain sensitive information, and should not be read in general. Accordingly, the vulnerability of real programs, in which sensitive data were leaked due to stale reads, was detected. Aiyer et al. [10] used the quorum system, a malicious scheduler that maliciously delayed messages between the server and client guaranteed an upper bound to limit the insolvency of data read, but stale reads still occurred.

TABLE I
QUASI BYTE-LEVEL METADATA OPERATIONS

Identifier	Description	Size
OP_START_LOG_SEGMENT	Initialize edit log	12
OP_ADD	Calculate number of blocks	160[a]
OP_ALLOCATE_BLOCK_ID	Allocate block ID	20
OP_SET_GENSTAMP_V2	Generate timestamp	20
OP_ADD_BLOCK	Add block to DN	44
OP_UPDATE_BLOCKS	Update existing data blocks	58[a]
OP_RENAME_OLD	Rename file or directory	69
OP_MKDIR	Make directory	61
OP_DELETE	Delete file or directory	39
OP_CLOSE	Close the operation	101[a]

[a] The byte size can be increased depending on the lengths of the file name and path.



(W: write, M_w: commit to in-memory metadata, D_w: commit to on-disk metadata, S: synchronous synchronization, A: acknowledge)

Fig. 2. Synchronous synchronization to avoid stale reads in our proposed scheme

III. Proposed Scheme for Namenode Replication

In this section, we propose a metadata replication with synchronous opcodes writing for namenode multiplexing. For better performance of read throughput, our proposed scheme has a primary metadata and multiple replica metadata. In this paper, we call it an eager primary-copy model. The primary metadata processes both read and write requests. Any of both the primary and the replicas can process read-only requests from clients. This requires all of the metadata to be identical. To reduce synchronization overhead, we synchronize metadata operation codes (described in Table I) only when the metadata are changed (e.g., write requests occur).

For low metadata synchronizing overhead for changes, we replicate the write operation codes (opcodes) of the metadata in byte-level. A write opcode is an operation unit with transaction of storing metadata changes according to write requests. We synchronize the write opcodes are stored between the in-memory (e.g., RAM) and on-disk metadata (e.g., HDD or SSD). This enables read requests from clients to process immediately after processing a write request. Thus, this read-after-write can reduce read latency. Furthermore, the proposed scheme apply strong consistency policy to avoid stale reads. The remainder of this section exhibits the design and implementation of the proposed scheme.

A. Read-only Namenode multiplexing scheme

We introduce our proposed scheme to increase the number of ANNs which can process read requests. These primary and replicas are called the proposed ANNs in this paper. To increase the number of the ANN, we first prepare several nodes and then place identical ANNs on each node. As shown in Fig. 1, in the proposed scheme, each ANN should be able to process read requests from clients. To do that, all of the proposed ANNs should synchronize to keep identical metadata. However, the stale reads can be occurred when we use an asynchronous synchronization. To eliminate the stale reads, we use a synchronous synchronization approach (as shown in Fig. 2). In addition, synchronizing the metadata between the proposed ANNs causes the synchronization overhead. To reduce the overhead, we replicates byte-level metadata operations codes (i.e., opcodes as shown in Table I). We also replicates the opcodes from the primary to the replicas. This enable the

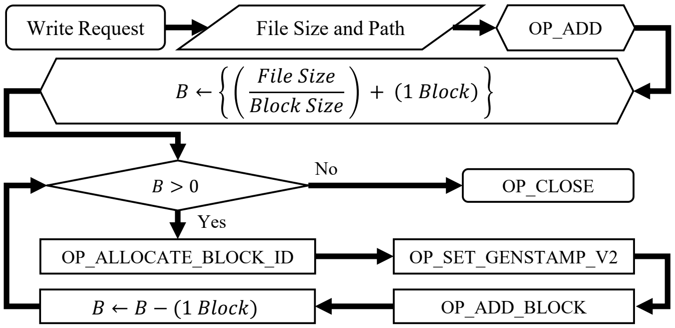

Write Request
File Size and Path
OP_ADD
$B \leftarrow \left\{ \left(\frac{File\ Size}{Block\ Size} \right) + (1\ Block) \right\}$
$B > 0$
No
OP_CLOSE
Yes
OP_ALLOCATE_BLOCK_ID
OP_SET_GENSTAMP_V2
$B \leftarrow B - (1\ Block)$
OP_ADD_BLOCK

Fig. 3. Flow chart of the write opcodes according to write request

proposed ANNs (i.e., primary and replicas in Fig. 1) and to be identical. Growing the number of the replicas enables better improvement for read throughput as in ObNN of existing HDFS. Thus, Fig. 1 shows the proposed scheme provides replicas to process read operations to achieve higher read throughput as the number of namenodes continues to grow as in ObNN of HDFS.

B. Write opcodes to reduce the synchronization overhead

The write opcodes are used to replicate only when changes are made to the metadata. We replicates only write opcodes that mutate the metadata among metadata opcodes for replicating the metadata in byte-level. This enable size of replications to be reduced. Also, we replicate metadata only when they are changed (e.g., write requests occur) to reduce the number of replications. Table I lists of the write opcodes and brief descriptions related to quasi byte-level changes in the metadata. They can be added according to the length of the file name; however, their size is in fact the size of bytes recorded in the metadata. The write opcodes do not have single-digit as size. Therefore, in this paper, we call it quasi byte-level opcodes. Fig. 3 shows the flow chart of the write opcodes according to write request. Algorithm 1 briefly shows an algorithm that replicates the write opcodes to synchronize metadata from the primary to the replicas.

```
Algorithm 1 Write opcodes replication with strong consistency
Inputs: $Req_{OPCODE}$ : set of metadata opcodes, $Write_{OPCODE}$ : set of write opcodes, $Num_{OPCODE}$ : number of $Req_{OPCODE}$, $Id_{TX}$ : identifier of transaction, $Id_{OPCODE}$ : identifier of $Req_{OPCODE}$, $Id_{NODE}$ : identifier of the primary and the replicas metadata, $Num_{NODE}$ : number of $Id_{NODE}$, $Bool_{\mathrm{ApNN}}$ : boolean whether our proposed scheme (i.e., ApNN) is activated
Output: $\mathrm{Ack}_{\mathrm{SYNC}}$ : result of syncronous synchronization
if ($Req_{OPCODE} \in Write_{OPCODE}$) then
    Set index of $Id_{NODE}$ to 1 as the primary
    for index $i \in \{1, 2, \ldots, Num_{OPCODE}\}$ do
        Get $Id_{OPCODE}$ at $i$ of $Req_{OPCODE}$
        // append $Id_{OPCODE}$ to metadata in the primary
        Append with $Id_{OPCODE}$ with $Id_{TX}$
        if ($Num_{NODE} > 1$ and $Bool_{\mathrm{ApNN}}$) then
            // append $Id_{OPCODE}$ to metadata in the replicas
            for index $j \in \{2, \ldots, Num_{NODE}\}$ do
                Get $Id_{NODE}$ at $j$ of the replicas
                $Ack_{\mathrm{SYNC}} \leftarrow$ Append with $Id_{OPCODE}$ with $Id_{TX}$
                if ($Ack_{\mathrm{SYNC}}$) then
                    // in-memory metadata for appended $Id_{OPCODE}$
                    Sync memory metadata with $Ack_{\mathrm{SYNC}}$
                Else
                    Retry to get $Ack_{\mathrm{SYNC}}$
            end if
        end for
    end if
end for
end if
Return $Ack_{\mathrm{SYNC}}$ from the replicas to the primary
```

TABLE II
EVALUATION ENVIRONMENTS

Host CPU, RAM	Intel Xeon E3-1270 v6 @ 3.80GHz x 8, 32GB
Host OS	Fedora 32
Guest CPU, RAM	1 vCore, 2GB RAM
Guest OS	CentOS 7.4
Java	OpenJDK 1.8.0_222
Hadoop	3.2.1

IV. Evaluation

To evaluate the effectiveness of our proposed method implemented on HDFS, we evaluate the synchronization overhead for the write opcodes by measuring the write performance of the multiplexed namenodes. To do that, we implemented read-only namenode (i.e., ANN replicas) multiplexing on HDFS. To implement our proposed scheme, we made the ANN replica node by modifying ANN to have functions of ObNNs and SbNNs. We also made ANN primary by modifying ANN to have functions JNs and QJM. To do that, we write 7,000 lines approximately and modify more than thirty files in Hadoop source codes. In this section, we describe the evaluation settings and analyze the results obtained from HDFS.

A. Evaluation Settings

We compare the synchronization overhead, as the number of nodes grow in our proposed scheme compared with the existing HDFS. To do that, we measure write time, write throughput, and average operations per second (ops). Specifically, in case of write time, to compare our proposed scheme with a single ANN of the existing HDFS, we compare the time it take to process write requests according to the file size. In addition, to compare write throughput and average ops, we measure them using

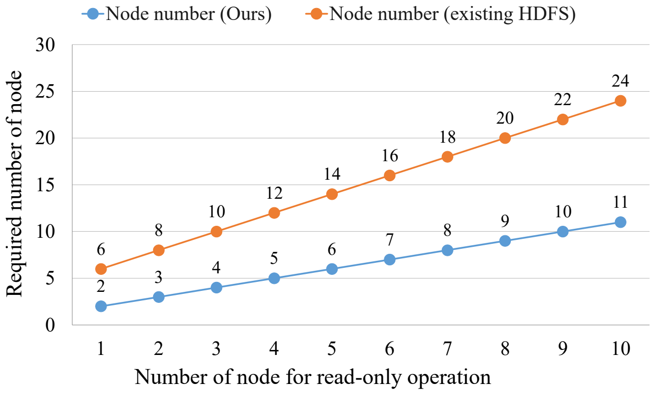


Fig. 4. Required number of namenode

TABLE III
AVERAGE WRITE OPERATION TIME

File size	Naive NN	Our scheme	Ratio
128MB (1 block)	3.786	3.324	113.8%
256MB (2 blocks)	3.412	3.954	86.3%
512MB (4 blocks)	4.502	5.267	85.5%
1024MB (8 blocks)	8.775	8.624	101.7%
2048MB (16 blocks)	12.988	16.178	80.3%
3072MB (24 blocks)	20.449	21.543	94.9%
4096MB (32 blocks)	26.164	29.198	89.6%

benchmark tools in Hadoop. Naive-NN is an existing HDFS with a single ANN. "Proposed-NN" is our proposed ANN and comprises 1 to 10 nodes. "1 Proposed-NN" is the primary namenode, and "2 through 10 Proposed-NN" is the replica namenode. The evaluation environment is presented in Table II. The existing HDFS method and the proposed method were applied separately to form a virtual machine, and each machine had one virtual core, 2 GB of memory, and 30 GB of disk space. An SSD was used in the host server for storage.

B. Evaluation Results

Fig. 4 shows the required number of namenode when the number of node for read operations (i.e., ObNNs) grows from one to ten. Equations (1) and (2) are a way to obtain the required number of node in existing HDFS and our proposed scheme, respectively. ANN and ANN primary is always 1. JNs is 2k+1 because of quorum and assume k to 1. Our proposed scheme does not need JNs and QJM because we do not use asynchronous synchronization. ANN replicas are almost equivalent to ObNNs and SbNNs because our proposed scheme has function of SbNN's ANN failover and ObNN's read operation processing capabilities.

$$\textbf{Number of node} = \textbf{ANN} + \textbf{JNs} + \textbf{ObNNs} + \textbf{SbNNs} \tag{1}$$

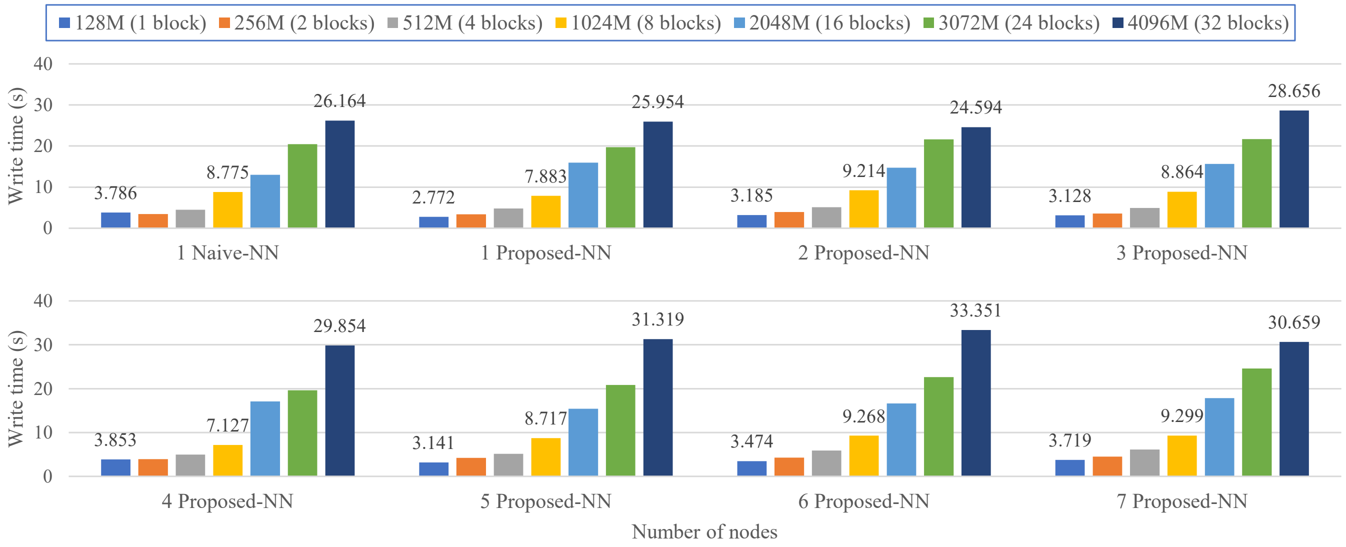
$$\textbf{Number of node} = \textbf{ANN primary} + \textbf{ANN replicas} \tag{2}$$

Therefore, we can obtain the difference between Equations (1) and (2). For instance, when the number of node for read operation is 1 to 10, our scheme needs only 2 to 11 nodes but existing HDFS should have 6 to 24 when JNs are three (as described in equation (1)). This means that our scheme can reduce the required number of node by 58.3%. This can be calculated by the following equations.

$$(\mathbf{2k} + \mathbf{1}) + \textbf{ANN replicas} \tag{3}$$

As shown in Fig. 5, the write time for each file size was measured to compare the synchronization overhead due to the increase in the number of nodes in the proposed scheme compared with the existing HDFS. The result shows that Naive-NN, the existing HDFS, had a write time of 26.164 s for a file size of 4096 MB (32 blocks), and the proposed scheme indicated an average of 29.198 s, differing by 3.034 s. The time required to complete a file upload to HDFS was measured using "-put" among the HDFS commands of Hadoop. The size of one block was 128 MB, the size of the file was the same as that of the block, and the number of DNs was set to one. Fig. 5 shows there was low overhead in given condition. We will analyze actual write time overhead in Table III.

We compared write time between our proposed scheme and existing ObNN. ObNN configuration applied as follows: ANN, SbNN, ObNN, DN is set to 1. JN is set to 3. When 1,000 files are written to existing HDFS, write and read time are 541,654 ms, 1,769 ms, respectively. In our proposed scheme with ANN primary and 1 replica, write and read time are 530,010 ms, 1,790 ms. To measure this, because we assigned one of unique number between 1 to 1,000 per text file, each file's size has from 1 to 3 bytes. Then we verified whether final summation is 500,500 (i.e., summation from 1 to 1,000) or not.

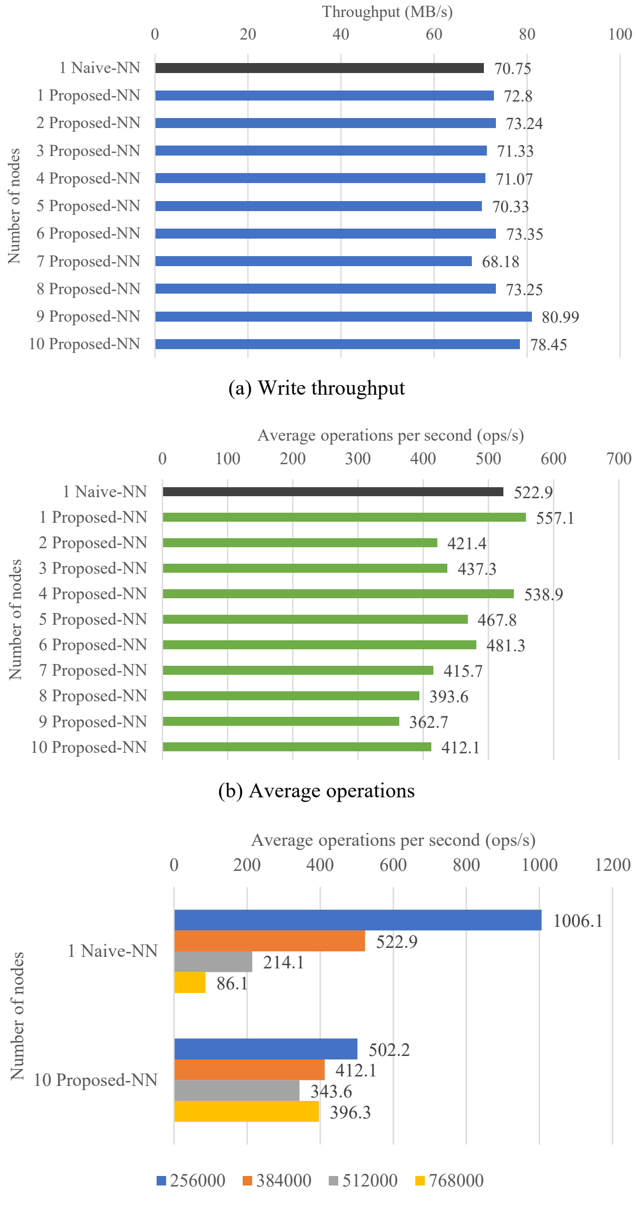


Write operation time for file with various blocks

Fig. 5. Synchronization overhead

Table III shows part of the result of Fig. 5. We can obtain 28 NNs through summation from 1 to 7 NNs. This results show that the write time per node in our proposed scheme took between 80%-113.8%. The average write operation time is 93.17% compared to the naive-NN. Thus, we verified that the overhead for write time is 6.83%.

As shown in Fig. 6(a), the write throughput was measured to compare the synchronization overhead due to the increase in the number of nodes in the existing HDFS and the proposed scheme. The results show that the throughput of naive-NN, i.e., the existing HDFS, was 70.75 MB/s, and the average throughput of the proposed scheme was 73.30 MB/s, which differed by 2.55 MB/s. It was measured using TestDFSIO, a benchmark tool of Hadoop, and the file size was set to 8 GB corresponding to 64 blocks, the buffer size to 10MB, the number of DNs to five, and the replication factor to five. Figs 5 and 6(a) show that the result values are slightly lower or higher than those of naive-NN. This means that the time required to add a block in the DN was slower or faster depending on the DN's disk input/output performance, which differed slightly temporarily.



(a) Write throughput

(b) Average operations

(c) Decrease in average operations

Fig. 6. Synchronization overhead

In Figs 6(b) and (c), the synchronization burden due to the increase in the number of nodes in the proposed scheme is compared with the existing HDFS. We used NNloadGenerator, a benchmark tool of Hadoop that generates workloads for a set time period and measures the average number of operations per second. The number of files was set to 1,000, the number of threads to 1,000, and the file size to 384,000 in Fig. 6(b), from 256,000 to 768,000 in Fig. 6(c), respectively. The file size is a basis value when creating a file for performance measurements and is used to create an arbitrary size (i.e., 10 times) rather than the actual size of the file. Both the number of DNs and the replication factor were set to five. Additionally, the ratio of reading and writing was set to 95% and 5%, respectively.

Fig. 6(b) shows that the existing HDFS naive-NN yielded 522.9 ops/s, whereas the proposed scheme yielded the average of 448.8 ops/s, indicating a difference of 74.1 ops/s. As shown in Fig. 6(d), naive-NN, an existing HDFS, decreased the average processing performance per second significantly as the file size increased. However, in the proposed scheme, even if the file size increased, the average processing performance per second remained constant compared with naive-NN. When the file size was 760,000, the computational performances of naive-NN and the proposed scheme were 86.1 and 396.3 ops/s, respectively. In addition, the lower limit of the file size of naive-NN was 86.1 ops/s, whereas that of the 10 Proposed-NN was 343.6 ops/s. Naive-NN does not intervene with synchronization but, in our proposed scheme, ANN intervenes with synchronization. Thus, Naive-NN can perform faster than ours when the file size is small. But, when the file size is bigger, naive-NN can look like performance degrade because naive-NN should wait for DNs to complete the job.

V. Conclusion

In this study, we proposed a metadata replication scheme with synchronous opcodes writing for namenode multiplexing. As the node number of metadata grows, the synchronization overhead can be high. To reduce the overhead, we replicated byte-level write opcodes. We synchronized in-memory and on-disk metadata so that all of the replicas can process read requests immediately for read-after-write in our proposed scheme. We applied our proposed scheme on HDFS and evaluated. As a result, our method reduces by 50.95% in the average required number of NNs when the number of NNs for read-only operation is 100. As the number of NNs is increased from one to seven, we also verified that the average write operation time for each file size (e.g., 128MB to 4096MB) was 6.83% higher than the existing HDFS with the given condition.

Acknowledgment

This work has been supported by the Future Combat System Network Technology Research Center program of Defense Acquisition Program Administration and Agency for Defense Development.(UD190033ED)

REFERENCES

[1] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The Hadoop Distributed File System,” in Mass Storage Systems and Technologies (MSST), IEEE 26th symposium, pp. 1–10, 2010.

[2] M. Wiesmann, F. Pedone, A. Schiper, B. Kemme and G. Alonso, "Database replication techniques: a three parameter classification," Proceedings 19th IEEE Symposium on Reliable Distributed Systems (SRDS), Nurnberg, Germany, pp. 206-215, 2000.

[3] Ang Zhang, Wei Yan, “Scaling Uber’s Apache Hadoop Distributed File System for Growth,” Uber Engineering, April 5, 2018, [Online]. Available: https://eng.uber.com/scaling-hdfs/

[4] Salman Niazi, Mahmoud Ismail, Seif Haridi, Jim Dowling, Steffen Grohsschmiedt, Mikael Ronström, "HopsFS: Scaling Hierarchical File System Metadata Using NewSQL Databases," in Proceedings of the 15th USENIX Conference on File and Storage Technologies (FAST), pp.89-103, 2017.

[5] Konstantin Shvachko, "Scaling HDFS with Consistent Reads from Standby Replicas," Linux Storage and Filesystems Conference (VAULT), USENIX Association, Santa Clara, CA, 2020.

[6] Q. Xu, R. V. Arumugam, K. L. Yong and S. Mahadevan, "Efficient and Scalable Metadata Management in EB-Scale File Systems," in IEEE Transactions on Parallel and Distributed Systems (TPDS), vol. 25, no. 11, pp. 2840-2850, Nov. 2014.

[7] Cheng Li, Daniel Porto, Allen Clement, Johannes Gehrke, Nuno Preguiça, Rodrigo Rodrigues, “Making geo-replicated systems fast as possible, consistent when necessary,” in Proceedings of the 10th USENIX conference on Operating Systems Design and Implementation (OSDI). USENIX Association, pp.265–278, 2012.

[8] Q. Ho, J. Cipar, H. Cui, J.-K Kim, S.-H Lee, P. Gibbons, G. Gibson, G. Ganger, and E. Xing, “More effective distributed ML via a Stale Synchronous Parallel parameter server,” in Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS), vol.1, pp.1223–1231. 2013.

[9] W. Snavely, W. Klieber, R. Steele, D. Svoboda and A. Kotov, "Detecting Leaks of Sensitive Data Due to Stale Reads," 2018 IEEE Cybersecurity Development (SecDev), Cambridge, MA, pp. 37-44, 2018.

[10] Aiyer A., Alvisi L., Bazzi R.A. “On the Availability of Non-strict Quorum Systems,” In: Fraigniaud P. (eds) Distributed Computing (DISC), Lecture Notes in Computer Science, vol. 3724. Springer. 2005.

[11] M. Wiesmann, F. Pedone, A. Schiper, B. Kemme and G. Alonso, "Understanding replication in databases and distributed systems," in the Proceedings 20th IEEE International Conference on Distributed Computing Systems (ICDCD), pp.464-474, 2000.

[12] Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, Carlos Maltzahn, "Ceph: A Scalable, High-Performance Distributed File System," 7th Symposium on Operating Systems Design and Implementation (OSDI), pp.307-320, USENIX Association, 2006.

[13] Jiayuan Zhang, Yongwei Wu, Yeh-Ching Chung, "PROAR: A Weak Consistency Model for Ceph," in IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS), 2016.

[14] Jayashree Mohan, Amar Phanishayee, Vijay Chidambaram, “CheckFreq: Frequent, Fine-Grained DNN Checkpointing,” in the Proceedings of the 19th USENIX Conference on File and Storage Technologies (FAST '21), pp.203-216. February 23-25, 2021.

[15] RedHat CephFS, “Preventing Stale Reads,” [Online]. Available: https://docs.ceph.com/docs/master/dev/osd_internals/stale_read/

[16] Feng Wang, Jie Qiu, Jie Yang, Bo Dong, Xinhui Li, Ying Li, "Hadoop High Availability through Metadata Replication," in Proceedings of the First International Workshop on Cloud Data Management (CloudDB), pp.37-44, 2009.

Design and development of Wearable multi-sensory smart device for human safety

Neamul Hossain
Dept. of Electrical & Electronics Engineering
American International University-Bangladesh
Dhaka, Bangladesh
Email: neamulhossain603@gmail.com

Johurul Haque Ovi
Dept. of Electrical & Electronics Engineering
American International University-Bangladesh
Dhaka, Bangladesh
Email: johurulovi@gmail.com

Shahriar Tasnim
Dept. of Electrical & Electronics Engineering
American International University-Bangladesh
Dhaka, Bangladesh
Email: stchowdhury11@gmail.com

Nuheen Islam
Dept. of Electrical & Electronics Engineering
American International University-Bangladesh
Dhaka, Bangladesh
Email: kni.nirob@gmail.com

Saniat Rahman Zishan
Dept. of Electrical & Electronics Engineering
American International University-Bangladesh
Dhaka, Bangladesh
Email: saniat@aiub.edu

***Abstract*— In the modern era, safety is one of the most concerning issues for people because of the increased rate of harassment, rape, hijacking, etc. For the sake of people's security, a device has been proposed in this paper. This device will help to prevent rape, hijacking, unwanted harassment, etc. The proposed system can be applied to make it smarter, safer, and automated. Flex Force pressure sensor, VR3 module will be used for activating the device automatically as well as artificial intelligence (AI) will be used to assume the attacker's wary behavior. In an unexpected situation, when the device will get activate it will send the location of that place to a pre-defined number through a message. This system will ensure the well-being of people of all gender and ages.**

***Keywords*— YOLOv3 algorithm with TensorFlow framework, VR3 Module, Flex Force pressure sensor, GSM, GPS, Human security**

I. Introduction

Around the world today, human safety has become a vital issue. Not only women, every human being is affected by physical/sexual abuse, hijacking, harassment, and a fear of violence. An organization working to protect women's rights known as Bangladesh Mahila Parishad revealed a statistic that approximately 2083 women and girl child were subjected to sexual violence from the first of 2019 to June 30, 2019, and out of that 731 were raped. From 2014 through 2018, a total of 5,274 women, and children faced different kinds of violence including 3,980 rapes [1]. It is assessed that around 35% of women everywhere in the world have encountered some sort of lewd behavior in the course of their life. Ladies who are between 18 -24 years of age are multiple times bound to be casualties of assault or rape. In Bangladesh assault rate 9.82% out of 11682 incidents [2]. Not only sexual assault or rape human trafficking, but kidnapping also has become a burden for people. Sometimes police can retrieve people from kidnappers sometimes they couldn't. As per figures detailed by the "Daily Star" newspaper, between February 2012 and June 2018, 4,152 instances of kidnappings because of human trafficking were recorded, yet just 25 individuals were indicted [3]. So, considering the safety of the people a model designed which will help to make sure the safety of people. These types of safety devices will lessen the crime rate. Moreover, it helps people to stay safe when needed and enables them to live more independently with greater confidence, peace of mind, and dignity when they are outside of the house .

II. Literature Review

Human security is one of the most significant issues as crime increments. There is some wellbeing device to lessen the crime. Here, some related distributed works are going to be discussed.

A security solution for women [4] developed by authors consists of a smart band and smartphone and both are connected with Bluetooth Low Energy (BLE). The Smart band consists of three sensors which are a temperature sensor, motion sensor, and pulse rate sensor. The app is installed in a smart phone to monitor the data directed by the smart band and to send the location of the victim using the GPS and the GSM system of smart phone. A smart watch [5] is designed by authors through which a woman can seek help by pressing the button of the smart watch. The smart watch module consists of three sub-module which are the Sensing, Control, and Transmission module. These three modules together help to activate the device and send the location of the victim, also the device is capable of giving severe shock to the culprit and generating an alarm. The paper [6] clarifies a gadget, which will work in three different ways, for example, voice, switch, and shock. At the point when the gadget got triggered by the sensor, it will begin working and then it will send location to police and message to the enlisted number through a GSM module and the same will be accomplished by a voice order. In ref. [7] the authors proposed a design which is implemented in the form of a smart ring, smart band, smart belt, etc. for ladies' wellbeing based on the Internet of Things (IoT). It is comprised of Raspberry Pi Zero, Raspberry Pi camera, buzzer, and button to activate the services. It is actuated by the victim by tapping the button. After clicking, the current location via GPS of the victim's and the camera catches the picture of the culprit which is then sent to police or pre-defined contact numbers using the victim's cell phone. This technique further uses the Uniform Resource Locator (URL) of the image and alert message to inform the family and police personnel. In the work [8] authors proposed a design which is a combination of several types of equipment such as IoT module, GPS & GSM, Neuro stimulator, Vibration sensor, Buzzer. In the device, GPS & GSM is used to detect & send the location of the

978-1-6654-4067-7/21/$31.00 ©2021 IEEEIEEE

victim, and the buzzer is used to generate an alarm to aware nearby people, whereas the IoT module is used to track the location continuously and update it on the webpage. For own safety using a Neuro Stimulator victim would be able to give a shock to the attacker. Moreover, a Vibration sensor is used to send the victim's last location if somehow the device gets damaged. In ref. [9] the authors proposed a design capable of self-defense, recording evidence, and tracing the location of a victim. In a critical situation, when the button is pressed nerve stimulator and buzzer is activated for self-defense and the camera is activated for recording evidence simultaneously also victim can send the current location of the victim and contact with a pre-define number through an Application. A wearable device [10] for women's safety and security designed by authors is comprised of an Arduino controller, temperature LM35 sensor, flex sensor, pulse rate sensor, GSM, and GPS. All those sensors help to activate the device. The hidden camera was installed in the victim's dress, when the device gets activated, the camera transmits the live scenario to the registered contacts so that they can be able to see what's happening there. And in the meantime, it will track and send the location of the victim to registered phone number using the GPS and GSM module. The principal objective of the work [11] is to make a wearable IOT gadget for the security and protection of human. This is executed by the measurement of physiological signs in simultaneousness with body motions. The signs are dissected and the internal heat level is estimated by galvanic skin obstruction. This work manages internal heat levels and stresses and skin obstruction and the connection between them. The gadget investigates skin opposition and internal heat level to break down the circumstance of the individual. A lot of research has been done and also running in the improvement of a human security system to decrease our social issues. Though all the device is made for human safety the majority of them are not perfect for human safety. Hence, a human safety system is proposed where some extra features are added to bring the perfection of the safety device.

III. Proposed Model And Simulation

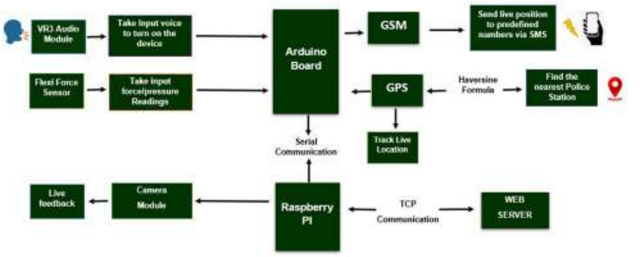


Fig.1. Block diagram of the design.

This paper is based on both Arduino and raspberry-pi. For the real-time recognition and detection process the YOLOv3 algorithm is used hence, raspberry-pi is used as microprocessor. For VR3, FlexiForce pressure sensor, GPS, GSM, Arduino is used as microcontroller. And to connect Arduino and Raspberry pi serial communication is used between them.

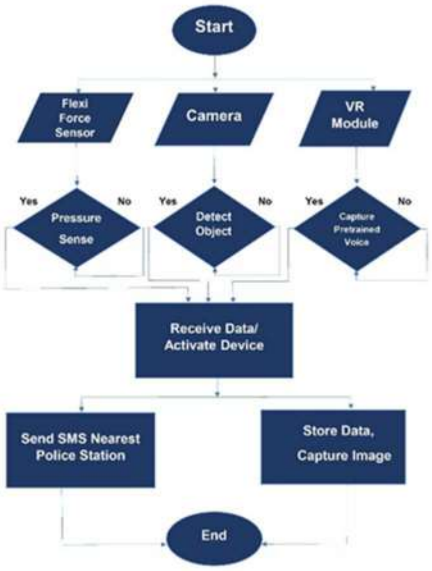


Fig.2. Flow chart of the design.

The device can be activated through the FlexiForce pressure sensor or VR3 module or from camera input. If victims face any critical situation, the device can be activated by saying pre-defined words using the VR3 module. Moreover, if somehow the FlexiForce pressure sensor get unexpected pressure then the device will also be activated. A camera is used to record real-time image and if the image match with pre-trained data set then also the device will be activated. So, when the device will activate, in the meantime it will send the location of the victim to the nearest police station and family members with the help of GPS & GSM.

Simulation of GSM and GPS Module:

Fig.3 shows the connection of both GPS and GSM together. GPS is used for tracing the victim's exact location. When the device will get activate the GPS will track that location. GSM is used for sending the location of victims through SMS to a pre-defined number.

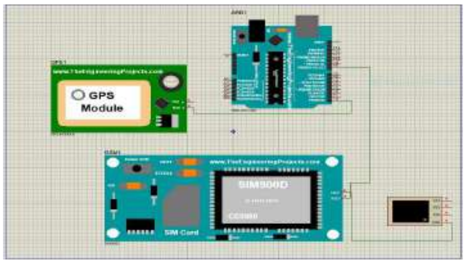
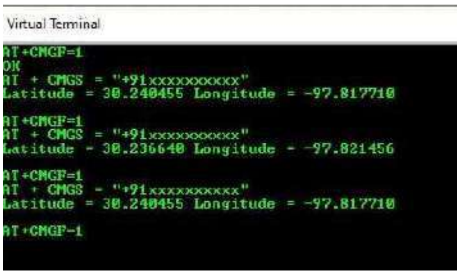


Fig.3. Combined simulation of GSM and GPS


Virtual Terminal

```
AT+CMGF=1
OK
AT + CMGS = "+91xxxxxxxxxx"
Latitude = 30.240455 Longitude = -97.817710

AT+CMGF=1
AT + CMGS = "+91xxxxxxxxxx"
Latitude - 30.236640 Longitude - -97.821456

AT+CMGF=1
AT + CMGS - "+91xxxxxxxxxx"
Latitude = 30.240455 Longitude = -97.817710

AT+CMGF-1
```

Fig.4. Coordinates of GPS

Simulation of FlexiForce Pressure sensor:

The simulation of the FlexiForce Pressure sensor is shown in fig.5. This sensor ranges from 0 to 25lbs of pressure. When the sensor will sense pressure more than the threshold value the sensor will be turned on. Where the threshold value has been set to 0.4lb (around 200gram).

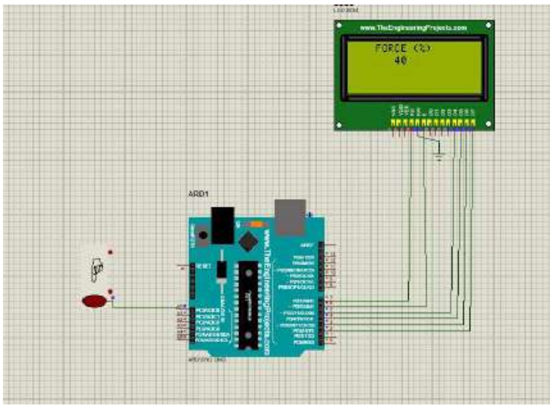


Fig.5. Simulation of Flexi Force pressure sensor.

Simulation of Object Detection:

The YOLOv3 algorithm is used in object detection using the TensorFlow framework. Harmful weapons such as Handguns, knives, and weapons are trained in Google Colab which is shown in fig 6. Around 2400 images were trained with 300 epochs.

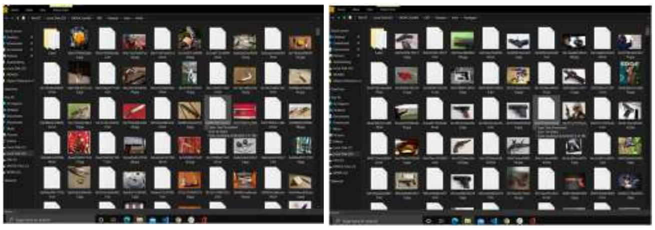


Fig.6. Trained Images of weapons and handguns

IV. Implementation Of The Design

Implementation of the design means the execution of design plans applying compatible processes to reach the utmost goal. It brings an idea to a real-life prototype. The device is made of several types of components such as Voice Recognition Module, Flexi Force pressure sensor, GPS, and GSM where Arduino is used as a microcontroller. The YOLOv3 algorithm is used for detecting the abnormal behavior where Raspberry Pi is used as a main processor. The device can be activated through VR3 module or FlexiForce pressure sensor or by analyzing dubious behavior from camera input.

A. VR3 Module

In this module, any sound could be trained as a command and users need to train the module first before let it recognizing any voice command. From fig.7, a mic is shown that is used to capture the voice and the voice process in the module. If the captured voice match with the pre-trained voice then the device will be activated. Several test has been carried out to check the accuracy of voice command from different distance. It has been found that the module worked accurately when the sound source is within 0-3 meter from VR3 module.

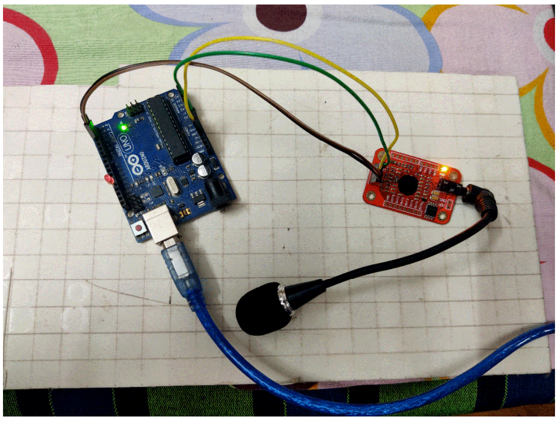


Fig.7. Voice Recognition Module

B. FlexiForce pressure sensor

The FlexiForce Pressure sensor acts as a force-sensing resistor. When there is no force in the sensor then its resistance is very high. When a force is applied to the sensor, the resistance proportionally decreases. If the sensor senses force more than threshold value which has been set to 0.4lb (around 200gram), it will send a signal to Arduino to activate the device.

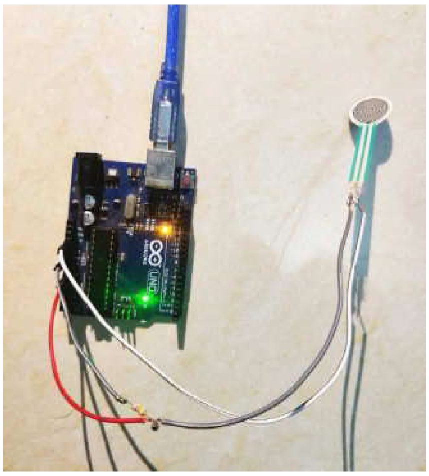


Fig.8. FlexiForce Pressure sensor

C. GPS

It is used for tracing the victim's exact location. A buck converter is used to convert the voltage into 5 voltage. An antenna is added for capturing the signal accurately. Here, "Haversine formula" is used to calculate the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface. This will find the nearest police station, additionally tracks live location.

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2)$$
$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$$
$$d = R \cdot c$$

Where, φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km);

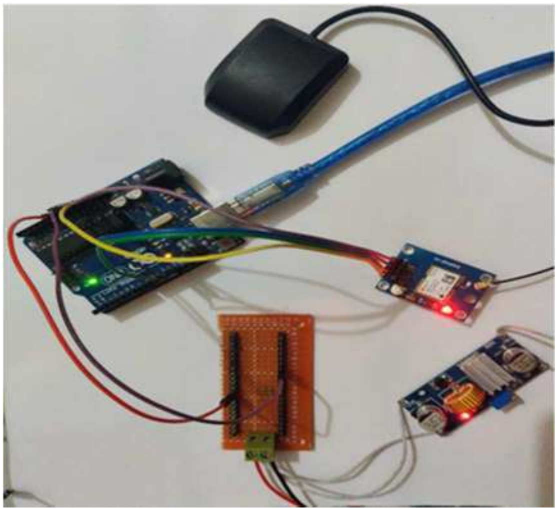


Fig.9. GPS Module

D. GSM

When the device is activated by getting the emergency signal, it sends the location information via SMS to predefined relatives or friend's phone numbers and nearest police station.

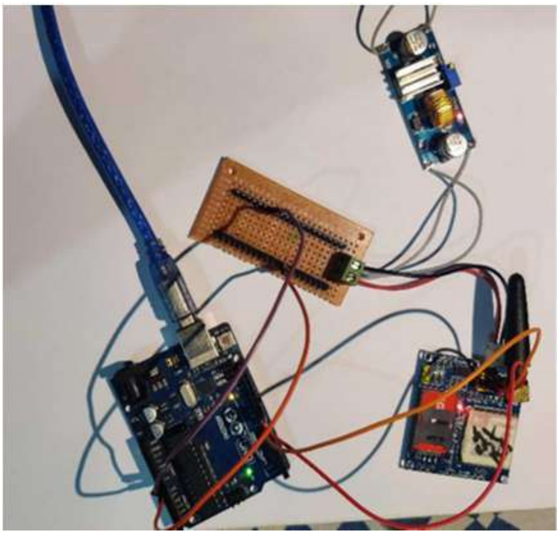


Fig.10. GSM Module

E. Object detection

The camera shown in fig.11 is used for the real-time recognition process. If the camera detects any weapons it will send a signal to the Raspberry pi then the device will be activated.

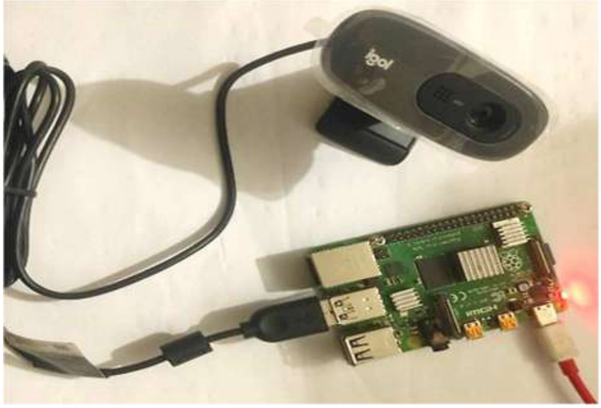


Fig11.Object detection system

F. Implementation of the complete system

A wearable multi-sensory device for human safety is shown in fig.12, where the device can be activated using either Flexi-force pressure sensor or VR3 module or by detecting suspicious behavior of the attacker. As soon as the device will be activated it will send the location of the victim to the predefined number, as well as it will capture the image of the attacker to find out the attacker later.

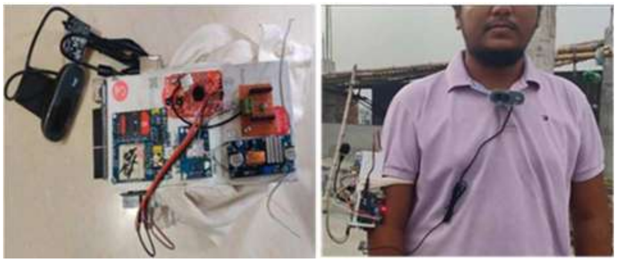


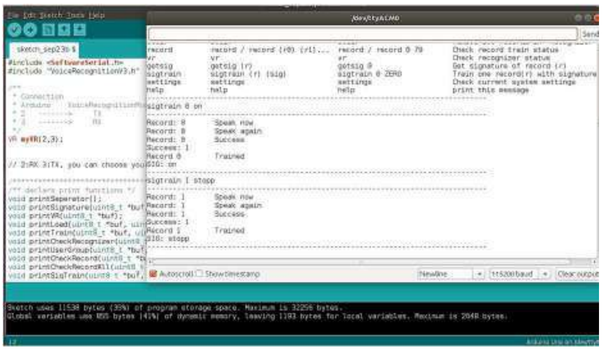
Fig.12. Implementation of Wearable multi-sensory device for human safety.

V. Result And Analysis

Before the implementation of the device, simulation of the component was done. In case of the VR3 module and FlexiForce pressure sensor, the result was quite perfect but for weapon detection through AI, the result wasn't 100% perfect because enough images were not trained. Moreover, sending the victim's location to the predefined number through GPS & GSM was done successfully.

A. VR3 Module

At first, the module is trained with a specific voice command as shown in fig.13 then with that word the VR3 module is checked whether the module worked or not.



```
#include <SoftwareSerial.h>
#include "VoiceRecognitionV3.h"

/**
 * Connection
 * Arduino    VoiceRecognitionMo
 * 2   ------->     TX
 * 3   ------->     RX
 */
VR myVR(2,3);

// 2:RX 3:TX, you can choose you

/***************************
/** declare print functions */
void printSeperator();
void printSignature(uint8_t *buf
void printVR(uint8_t *buf);
void printLoad(uint8_t *buf, uin
void printTrain(uint8_t *buf, ui
void printCheckRecognizer(uint8_
void printUserGroup(uint8_t *buf
void printCheckRecord(uint8_t *b
void printCheckRecordAll(uint8_t
void printSigTrain(uint8_t *buf,
```

```
record      record / record (r0) (r1)...   record / record 0 79     Check record train status
vr          vr                             vr                       Check recognizer status
getsig      getsig (r)                     getsig 0                 Get signature of record (r)
sigtrain    sigtrain (r) (sig)             sigtrain 0 ZERO          Train one record(r) with signature
settings    settings                       settings                 Check current system settings
help        help                           help                     print this message
------------------------------------------------------------------
sigtrain 0 on
------------------------------------------------------------------
Record: 0       Speak now
Record: 0       Speak again
Record: 0       Success
Success: 1
Record 0        Trained
SIG: on
------------------------------------------------------------------
sigtrain 1 stopp
------------------------------------------------------------------
Record: 1       Speak now
Record: 1       Speak again
Record: 1       Success
Success: 1
Record 1        Trained
SIG: stopp
------------------------------------------------------------------
```

```
Sketch uses 11538 bytes (35%) of program storage space. Maximum is 32256 bytes.
Global variables use 855 bytes (41%) of dynamic memory, leaving 1193 bytes for local variables. Maximum is 2048 bytes.
```

Fig.13.Taking Voice command of VR3 module.

As shown in fig.14 when the Module got the voice command same as the trained one then the device got activated which is understood by the red LED cause when the module isn't activated the LED is off but when the device gets activated the LED is on.

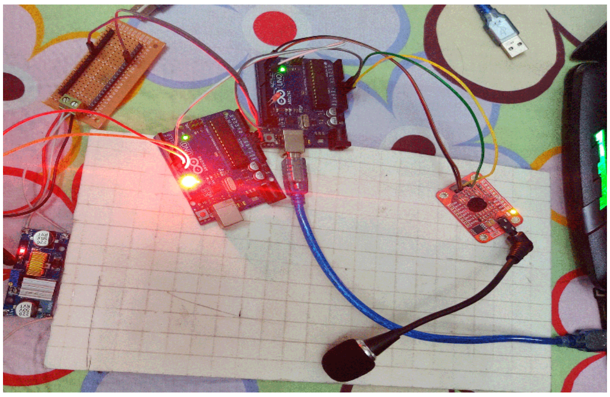
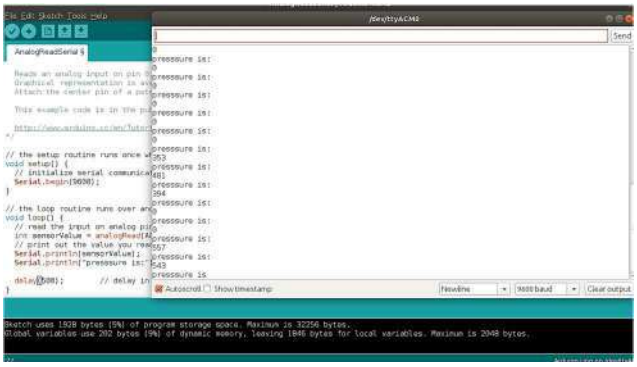


Fig.14. Result of Voice Recognition Module.

B. *Flexi Force Pressure Sensor*

The working of the FlexiForce pressure sensor depends on the pressure sensing on the sensor. From fig. 15, it is shown that pressure is 0 that means it doesn't sense any pressure. Which means device isn't activated.



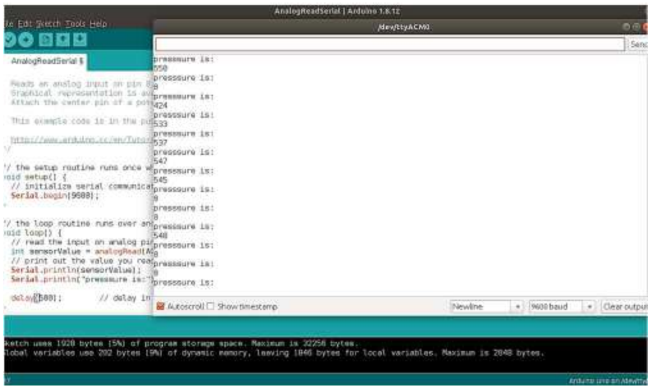
```
  Reads an analog input on pin 0
  Graphical representation is av
  Attach the center pin of a pot

  This example code is in the pu

  http://www.arduino.cc/en/Tutori
*/

// the setup routine runs once w
void setup() {
  // initialize serial communicat
  Serial.begin(9600);
}

// the loop routine runs over an
void loop() {
  // read the input on analog pi
  int sensorValue = analogRead(A
  // print out the value you rea
  Serial.println(sensorValue);
  Serial.println("pressure is:"

  delay(500);        // delay in
}
```

```
0
pressure is:
0
pressure is:
0
pressure is:
0
pressure is:
0
pressure is:
0
pressure is:
353
pressure is:
481
pressure is:
394
pressure is:
0
pressure is:
0
pressure is:
557
pressure is:
543
pressure is
```

```
Sketch uses 1928 bytes (5%) of program storage space. Maximum is 32256 bytes.
Global variables use 202 bytes (9%) of dynamic memory, leaving 1846 bytes for local variables. Maximum is 2048 bytes.
```

Fig.15. Value of Flexi Force pressure sensor

As shown in fig.16, there has some pressure more than threshold value that means after sensing pressure the device gets activated. And fig. 17, shows the turned-on state of FlexiForce Pressure sensor.



```
  Reads an analog input on pin 0
  Graphical representation is av
  Attach the center pin of a pot

  This example code is in the pu

  http://www.arduino.cc/en/Tutori
*/

// the setup routine runs once w
void setup() {
  // initialize serial communicat
  Serial.begin(9600);
}

// the loop routine runs over an
void loop() {
  // read the input on analog pi
  int sensorValue = analogRead(A
  // print out the value you rea
  Serial.println(sensorValue);
  Serial.println("pressure is:"

  delay(500);        // delay in
}
```

```
pressure is:
550
pressure is:
0
pressure is:
424
pressure is:
533
pressure is:
537
pressure is:
547
pressure is:
545
pressure is:
0
pressure is:
0
pressure is:
548
pressure is:
0
pressure is:
0
pressure is:
```

```
Sketch uses 1928 bytes (5%) of program storage space. Maximum is 32256 bytes.
Global variables use 202 bytes (9%) of dynamic memory, leaving 1846 bytes for local variables. Maximum is 2048 bytes.
```

Fig.16. Value of FlexiForce pressure sensor

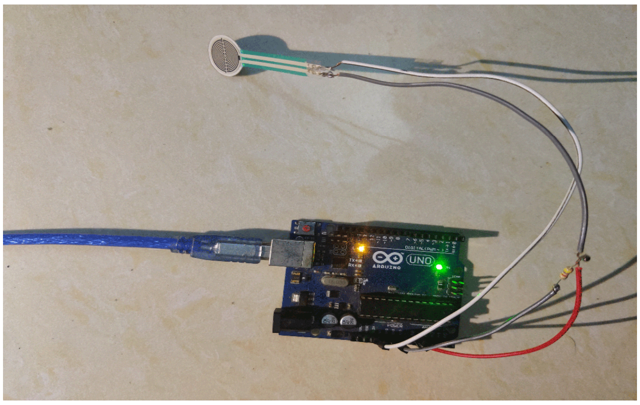
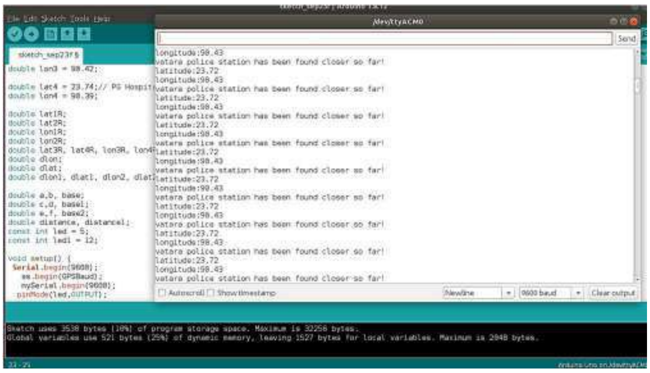


Fig.17. Result of FlexiForce Pressure Sensor

C. *GPS*



```
double lon3 = 90.42;

double lat4 = 23.74;// PS Hospit
double lon4 = 90.39;

double lat1R;
double lat2R;
double lon1R;
double lon2R;
double lat3R, lat4R, lon3R, lon4
double dlon;
double dlat;
double dlon1, dlat1, dlon2, dlat

double a,b, base;
double c,d, base1;
double e,f, base2;
double distance, distance1;
const int led = 5;
const int led1 = 12;

void setup() {
 Serial.begin(9600);
   ss.begin(GPSBaud);
   mySerial.begin(9600);
  pinMode(led,OUTPUT);
```

```
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
latitude:23.72
longitude:90.43
vatara police station has been found closer so far!
```

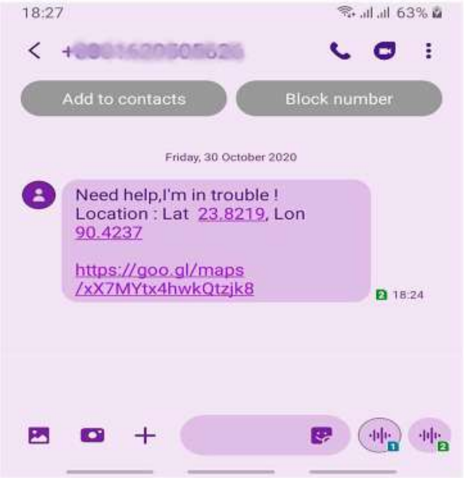
```
Sketch uses 3538 bytes (10%) of program storage space. Maximum is 32256 bytes.
Global variables use 521 bytes (25%) of dynamic memory, leaving 1527 bytes for local variables. Maximum is 2048 bytes.
```

Fig.18. Vatara police station by utilizing GPS module

From fig.18, the GPS module is initiated and it is showing the Vatara police station. That means when the device got activated at that moment the victim was near Vatara Thana. And Vatara Police station will get a message of the victim's location so that they can easily find out the victims and the attacker from that location.

D. *GSM*

When the system will activate it will send the victim's location to a pre-defined number through the GSM module. Fig.19 shows that, a message including the location of a place received by a person.



+[redacted]

Friday, 30 October 2020

Need help,I'm in trouble !
Location : Lat 23.8219, Lon 90.4237

https://goo.gl/maps/xX7MYtx4hwkQtzjk8

18:24

Fig.19. Message sent through GSM

E. *Result of object detection method*

From fig.20 shows that a person showing different kinds of weapons and the system detects the weapon along with the name of the weapon. As soon as it detects a weapon it will send a command and the device will be activated. For the detection process thousands of weapons images such as Hand Gun, Knife was trained.

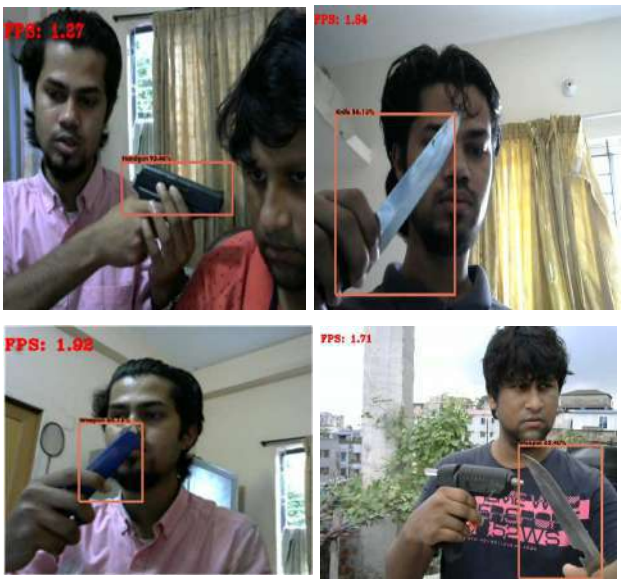


Fig.20. Weapon detection and recognition

At the moment of system activation, it will capture the image of the attacker. Here, fig.21 shows the captured image of the attacker at the time of device activation which is stored in the webserver.

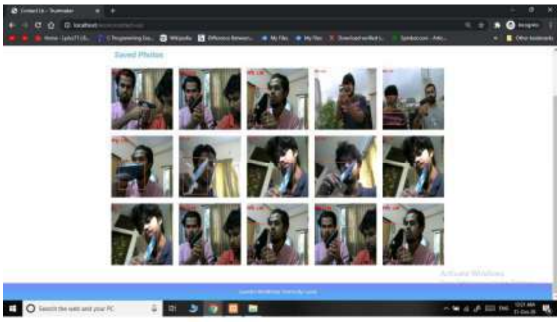


Fig.21. Stored image in Webserver

Fig.22 shows different confidence values in several detection periods. For handgun detection event the frame's confidence value is achieved more than 50%.

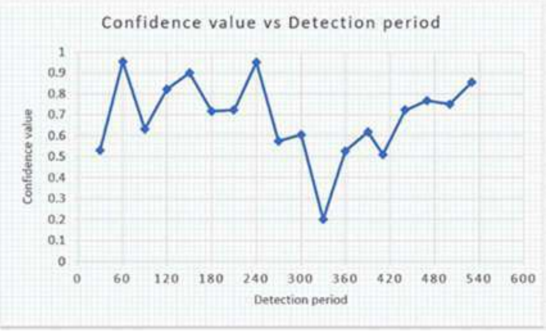


Fig.22: Values of Confidence ratio over total detection period

Fig.23. shows different confidence values in several detection periods in a graph. For knife detection event the frame's confidence value is achieved more than 60%.

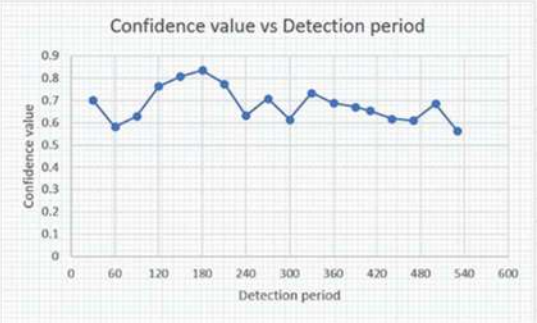


Fig.23: Values of Confidence ratio over total detection period

Fig.24. shows loss vs. epoch graph. In the event of weapon detection highest loss at epoch 10 is achieved 0.53. When the frames are ended at 300 epochs the loss came near 0.41.

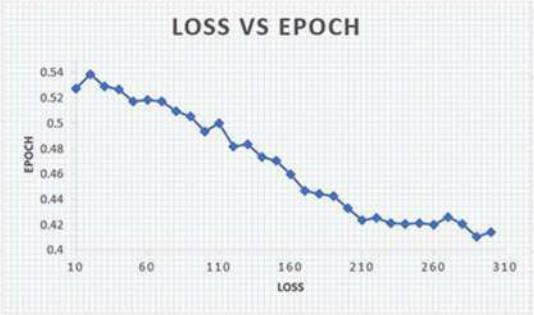


Fig.24: Loss Vs. epoch ratio for weapon detection event

VI. Conclution And Future Work

Statistics say that rape, kidnapping, hijacking, harassment, molesting, etc. is increasing day by day but very little progress has been made to find a noteworthy solution to the problem. Hence, a prototype is designed to make people safer in such situation and the purposes of the design have been achieved. Because using this smart device people can secure them, protect them, can take help in emergencies. Moreover, this device will capture the photo of the attacker which will help the authority to find the attacker easily. There is some scope for further improvement of this device. In the future, a 360-degree viewing camera can be used instead of a one-sided camera so that it can analyze all sides of the surrounding. If a huge dataset of the weapon can be trained then the efficiency of the weapon detection will be increased. Moreover, in future besides weapons, the suspicious behavior of a person can be analyzed to detect the culprit.

References

[1] A shocking 731 rapes reported in first six months of 2019. (2019,July 8). Dhaka Tribune. https://www.dhakatribune.com/bangladesh/nation/2019/07/08/a-shocking-731-rapes-reported-in-first-six-months-of-2019

[2]. Rape statistics by country 2021. (n.d.). Retrieved April 04,2021, from https://worldpopulationreview.com/countries/rape-statistics-by-country/

[3]. Page, F. and Correspondent, S., 2020. Culprit A Serial Rapist: Rab. [online] The Daily Star. Available at: https://www.thedailystar.net/frontpage/dhaka-university-student-rape-serial-rapist-he-is-1851484

[4]. Harikiran, G. C., Menasinkai, K., & Shirol, S. (2016, March). Smart security solution for women based on Internet of Things (IOT). In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 3551-3554). IEEE.

[5]. Shreyas, R., Varun, B., Kumar, H. S., Kumar, B. P., & Kalpavi, C. (2016). Design and development of women self defence smart watch prototype. International Journal of Advanced Research in Electronics and Communication Engineering, 5(4).

[6]. "Bhardwaj, N., & Aggarwal, N. (2014). Design and Development of "Suraksha"-A Women Safety Device.
International Journal of Information & Computational Technology, 4(8), 787-792.

[7]. Sogi, N. R., Chatterjee, P., Nethra, U., & Suma, V. (2018, July). SMARISA: a raspberry pi based smart ring for women safety using IoT. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 451-454). IEEE.

[8] Sathyasri, B., Vidhya, U. J., Sree, G. J., Pratheeba, T., & Ragapriya, K. (2019). Design and Implementation of Women Safety System Based on Iot Technology. International Journal of Recent Technology and Engineering (IJRTE) ISSN, 2277-3878.

[9] Sen, T., Dutta, A., Singh, S., & Kumar, V. N. (2019, June). ProTecht–Implementation of an IoT based 3–Way Women Safety Device. In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1377-1384). IEEE.

[10]. Nandhini, A., & Moorthi, K. Journal Homepage: -www. journalijar. Com

[11]. M. Pramod, C. V. (2018, January). IOT WEARABLE DEVICE FOR THE SAFETY AND SECURITY OF WOMEN AND GIRL CHILD. International Journal of Mechanical Engineering and Technology (IJMET), Volume 9(Issue 1), pp. 83–88. Retrieved from http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=9&ITyp

`MiniCon`: Automatic Enforcement of a Minimal Capability Set for Security-Enhanced Containers

Haney Kang, Jinwoo Kim, and Seungwon Shin
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology (KAIST)
Yuseong, Daejeon, South Korea
{haney1357, jinwoo.kim, claude}@kaist.ac.kr

***Abstract*—Nowadays, containers have been widely adopted not only for clouds but also for individual users. On the one hand, containers provide much more light-weight virtualized infrastructure, but on the other hand, it is unavoidable to handle security issues since the isolation of containers is relatively weak, compared to the legacy VMs. Under the container architecture, adversaries are able to exploit kernel vulnerabilities to escalate privilege to gain root privilege, and leak system, privacy critical information. Although previous solutions provide strong security protection, unfortunately, none of them do provide a way to apply policies. Therefore, in this paper, we present an eBPF-based capability enforcement system, `MiniCon`, that automatically generates and enforces a minimal capability set by using Seccomp Filter. It monitors capability requests from containers, merges those requests to create a minimal capability set, and enforces capability policies through Seccomp Filter.**

***Index Terms*—Cloud, Virtualization, Capability, Container**

I. Introduction

Recently, cloud has been emerging technology because of its flexibility and ease of management. Service providers (or tenants) are now shifting their service environment from their local servers to cloud to benefit by flexibility and management of cloud. Thanks to the on-demand resource allocation policy of the cloud, service providers now can request exact amount of hardware resources they required, and also they are able to request reallocation of resources in any time they need. Moreover, hardware resources are already physically installed at data centers, the only thing service providers have to do is just to follow simple configuration steps so as to set up their environment. Therefore, service providers can effectively reduce huge amount of cost for managing and purchasing resources.

However, to allocate resources to different users, resources should be properly isolated as like they are in separate hardware. This is important because it may cause security issues if users are available to access others tenants' resources. To address this issue, virtualization techniques have been adopted to isolate resources.

Virtualization can be implemented by two different architecture. Virtual Machine (VM) is the traditional way to virtualize entire software stack including kernel, which has been employed as the most popular approach over the last decade. Under the VM environment, each user has their individual kernel, and their applications are executed on top of the isolated kernel stack. Hypervisor, a middle layer for virtualization which locates under a VM kernel layer, restricts unauthorized actions such as resource access from other VMs, and it is responsible for orchestrating resource allocation among VMs. With VM, resources can be strongly isolated because hypervisor fundamentally limits resource view under kernel. However, performance degradation cannot be ignored due to the duplicated kernels.

To address the performance issues of VM, containers have been proposed. In contrast to VM, each container shares the kernel, so the user space is the only one isolated among container instances. This greatly reduces overhead for kernel-related operations, such as system calls. Due to its performance advantages, many IT vendors adopt container as a virtualization for their cloud [1]–[4]. Despite those benefits, some critical security issues inhere within containers. Unlike VMs, containers share a kernel stack, which is probably vulnerable, and can be a penetration channel to attacks. For example, adversaries are able to invoke system calls to trigger kernel vulnerabilities which affects not only host but also other containers.

Indeed, many vulnerabilities that stem from the (insecure) container architecture have been disclosed: CVE-2016-0728 showed that adversaries can cause memory leak by abusing system call with designed arguments [5]. CVE-2016-5195 showed that adversaries can also intentionally bring about race condition by simultaneously invoke `write` system call [6]. Moreover, privilege escalation can be done by `waitid` system call [7]. Besides, globally accessible information have to also be concerned. Luo et al. [8] proposed a covert channel attack using global information which is not limited by Linux security functions, total memory usage, and kernel logs. ContainerLeaks [9], likewise, showed system-wide host information can be leaked by using globally accessible files. Meanwhile, internal attacks from a compromised container to other benign containers have been reported. For instance, PWN Docker Redis Attack compromise vulnerable server which accept Docker API and it compromise other containers in same network [10]. Therefore, despite of its performance efficiency, it is important to handle security issues on containers.

978-1-6654-4067-7/21/$31.00 © 2021 IEEE

For this, several methods were recently introduced to address the security concerns about container. SCONE [11] utilizes Intel SGX to secure information inside of container. It can securely prevent information leakage from outside even from host, but it contains significant performance issue. SPEAKER [12] separates running phase of container to minimize available capability by using dynamic analysis. Unfortunately, it does not take account of the limitation of dynamic analysis which cannot cover entire logic of container. Cimplifier [13], however, choose debloating method which separate single large container to a number of container. Although separating container reduces attack surface such as vulnerability of container, these containers still threatened by vulnerability. Security Namespace [14] suggests to use Linux Security Module (LSM) with fine-grained policy. Most of container products, such as Docker [15] and Kubernetes [16] are compatible with LSM, it does not automatically provide policy guideline. Confine [17] reduces system call using static analysis. By using static analysis, this work can produce policy which covers a entire set of system calls in applications. However, some system calls might be never invoked while the static analysis can consider it to be called. In case of unoptimized application, Confine might provide a super set of system calls. On the other hand, LSMs such as SELinux, AppArmor, Seccomp [18]–[20] are used to apply security policy to containers, still they do not provide ideal policies to prevent containers from kernel vulnerabilities.

To tackle these challenges, in this paper, we propose a security enforcement framework, `MiniCon`, that confines minimal container capabilities for building a secure container environment. To find out minimal capability set of given container, we first implement eBPF [21] capability probe which collects set of capabilities required by containers. We then build an environment generator that creates realistic environment of running container. Finally, we create Seccomp [20] policy based on a minimal capability set to enforce it.

II. Background

A. Container

To satisfy requirements of cloud systems, containers aim to isolate system resources such as CPU usage, and network interfaces from other users. Not only for physical hardware resources, containers enable to isolate logical resources such as UID, PID, and even file systems. Isolating UID, for example, users in different containers do not have visibility to each others. By doing this, container environments abstract user's view such that a user is the only one who uses this machine.

Containers provide isolation of these resources by implementing a thin layer which uses Linux *namespaces* and *cgroups*. Thin layer resides in host user space as a daemon process and they create, initialize, stop, and delete according to administrators command. Once administrators create and initiate container, Linux namespaces inside of thin layer restrict a view of file system including devices and system file to isolate logical resources. Simultaneously, it applies available capacity of hardware resource by applying cgroups policy.

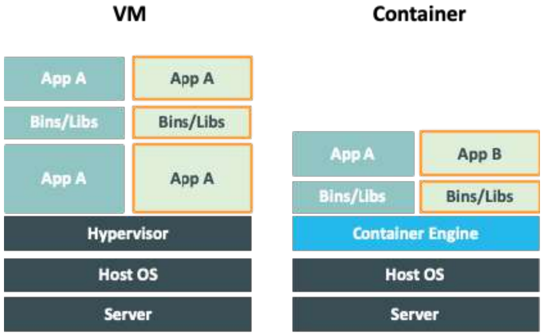


Fig. 1. The conceptual architecture of containers compared to VMs.

Therefore, containers are basically user spaces isolated by Linux namespaces which is allowed to use limited amount of hardware resource enforced by cgroups Policy. Fig. 1 shows conceptual architecture of container by comparing it to Virtual Machine.

B. Linux Capabilities

Linux kernel supports security sensitive, but compulsory operations such as system calls. Although these operations should be able to be invoked by users, triggering them must be controlled very strictly. For this purpose, Linux kernel introduces capabilities which define fine granularity as a key of root permissions. For example, containers which executed without CAP_CHOWN, are not able to change the UIDs and GIDs of files. There are 38 capabilities Linux kernel 4.18, and 15 are set in containers by default as described in Table I [22].

C. Motivating Examples

As a default, containers only allowed 15 capabilities, but adversaries are still able to abuse system calls to accomplish their goal. For example, CVE-2016-0728 [5] is a vulnerability which can cause memory leakage. In this example, adversaries invoke `keyctl` system call which only requires CAP_SETUID and CAP_SYS_ADMIN. From the result of the memory leakage vulnerability, adversaries can escalate privileges and achieve root permission even they have never been granted. Another example about CVE-2016-5195 [6] is a vulnerability of race condition. In the case that adversaries already have privilege to CAP_SYS_ADMIN and CAP_SYS_PTRACE, they can routinely invoke `madvise` or `ptrace` to provoke race condition. Then adversaries can write their arbitrary code on the memory where a race condition occurred even if memory area is read-only.

III. MiniCon Design

A. Overview

As mentioned in Section II-C, assigning capabilities without any verification may critically affect security of container, even though these capabilities are granted by default. Thus, it is important to find out unused capabilities and drop these capabilities. In this paper, we suggest minimal capability enforcement by using eBPF [21] and Seccomp [20] to mitigate chance

TABLE I
CAPABILITIES SUPPORTED BY LINUX KERNEL 4.18

Cap. No.	Capability	Set by default
0	CAP_CHOWN	Y
1	CAP_DAC_OVERRIDE	Y
2	CAP_DAC_READ_SEARCH	
3	CAP_FOWNER	Y
4	CAP_FSETID	Y
5	CAP_KILL	Y
6	CAP_SETGID	Y
7	CAP_SETUID	Y
8	CAP_SETPCAP	Y
9	CAP_LINUX_IMMUTABLE	
10	CAP_NET_BIND_SERVICE	Y
11	CAP_NET_BROADCAST	
12	CAP_NET_ADMIN	
13	CAP_NET_RAW	Y
14	CAP_IPC_LOCK	
15	CAP_IPC_OWNER	
16	CAP_SYS_MODULE	
17	CAP_SYS_RAWIO	
18	CAP_SYS_CHROOT	Y
19	CAP_SYS_PTRACE	
20	CAP_SYS_PACCT	
21	CAP_SYS_ADMIN	
22	CAP_SYS_BOOT	
23	CAP_SYS_NICE	
24	CAP_SYS_RESOURCE	
25	CAP_SYS_TIME	
26	CAP_SYS_TTY_CONFIG	
27	CAP_MKNOD	Y
28	CAP_LEASE	
29	CAP_AUDIT_WRITE	Y
30	CAP_AUDIT_CONTROL	
31	CAP_SETFCAP	Y
32	CAP_MAC_OVERRIDE	
33	CAP_MAC_ADMIN	
34	CAP_SYSLOG	
35	CAP_WAKE_ALARM	Y
36	CAP_BLOCK_SUSPEND	
37	CAP_AUDIT_READ	

of exploit by eliminating unnecessary capabilities. Fig. 2 shows the overall architecture of the work. The architecture mainly categorized to three steps: presenting realistic container execution environment, identifying capabilities required by container, applying security policy.

B. Presenting Container Execution Environment

Container can be executed by various command, configuration network environments. Various environments trigger container to be executed with different program logic path. Network environment, for example, server-client model application continuously invokes `epoll` and `wait` system calls in idle state. When it receives network packets from client, an application moves its state to handler to perform function related to incoming request. This is obviously essential, otherwise policies produced by system may disable functionality of application. To cover enough program logic path, we created execution environments by three methods.

Via Container Execution Command. Container execution commands such as `docker run` provide various arguments. Although these arguments supports users to build their unique and optimized environment, most of them are rarely used. The main reason of it is that these arguments are constructed for user convenience which do not largely affect environment. From this reason, we found that most of execution command guided from container image manual pages only contain a few frequently used options. Therefore, to find out these options out of arguments, we crawled execution command pattern from image hub of container vendor. As a result, we found that networking, volume mounting, attaching and detaching standard in/output, allocating TTY were mostly used option.

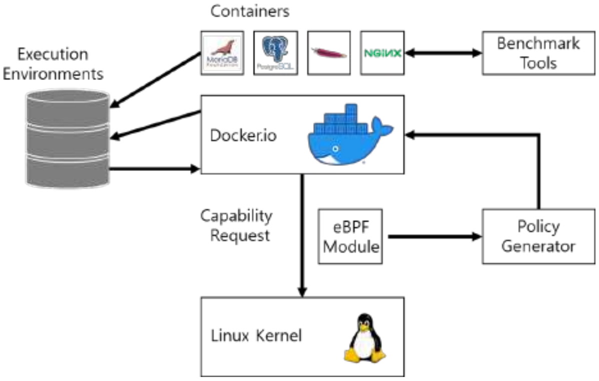


Fig. 2. `MiniCon` architecture overview

Via Application Execution Command. Applications reside in containers takes various arguments same as execution of container. We collect command manual by providing option `--help` to application execution command and formalize it to apply proper arguments to options. For instance, it is possible to describe existence of options by formalizing it to boolean value. Some options takes only integer or string or even a few defined string as a value of option, otherwise it returns error. By formalizing these arguments, we were able to run same containerized application with various environment. This can be similarly applied configuration, since configuration of applications are mostly given by dictionary format such as json.

Via Benchmark Tools. To generate realistic network interaction of container, we utilize benchmark tools for known applications. For example, we use `ab`, an apache benchmark tool, for `httpd` container to create network packets. Since benchmark tools are developed to verify functionality of target application, benchmark tools are appropriate to generate realistic network interaction.

C. Capability Probing

From the number of cases to execute container, we extract capability requirements for each cases and intersect them. This is because we do not know exact argument and configuration that user run, thus we have to produce a capability set which contains every capability possibly required by containers. To identify capability requirements, extended berkeley packet filter (eBPF) is used for fast and reliable monitoring. Our implementation was done by BPF Compiler Collection

(BCC) [23] which is toolkit for creating kernel tracing while using of eBPF. Since capability checking is done by kernel function `int cap_capable(...)`, we implement probing function for `int cap_capable(...)` to identify which process request and what capability was requested. By referring argument `int cap`, we were able to get capability number. For the final step, we filter out unrelated logs such as comparing process tree.

D. Policy Enforcer

From the minimal capability set given by Section III-C, Seccomp [20] policies were created. Our policy generation code separates capabilities to required set which is minimal capability set and unnecessary set, then it convert to Seccomp policy format. We apply policy by using option such as *–security-opt="seccomp=path/to/file"*. There can be randomness issue which cause failure of collecting few required capabilities, it is important to keep checking whether containers fail. If there is no failure regarding operation permission (`EPERM`) error, we finally conclude that `MiniCon` finds out a minimal capability set to run containers. Otherwise, we add capabilities which causes `EPERM` error and repeat capability checking procedure.

IV. Preliminary Evaluation

For the sake of showing our analysis result, we used 15 container images from Docker Hub [24].

Generally Required Capabilities: We classified capabilities into three categories by their functionality. Table II shows the capabilities that were used in namespace category are used to create isolated namespaces by host. When isolating namespaces, CAP_SETUID and CAP_SETGID were requested.

Another capability which always required for most containers was the file access. This is because applications in containers generally access shared files from a host. For example, some applications require configuration files, and they are mostly located in shared mounted volumes. In this case, containers are required to have CAP_FOWNER to read file inside of container.

On the other hand, we found that networking related capabilities are also generally used by containers. The reason is that majority of the applications are aimed to developed to act as a server. To do so, they have to be assigned well known port which is port less than 1024. In this case, containers are required to have CAP_NET_BIND_SERVICE. Moreover, these containerized applications have to send non-TCP/UDP packets such as ARP and ICMP, and they are required to have CAP_NET_RAW capability.

Required Linux Capabilities Fig. 3 shows number of required capabilities for each container images. In this result, we are able to find most of containers are likely to use similar capabilities, and they rarely use other capabilities given by default.

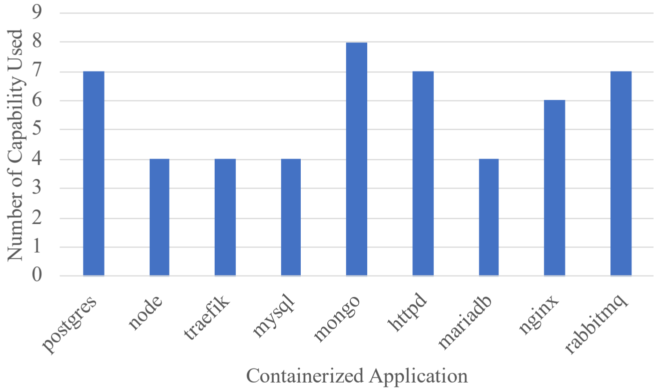


Fig. 3. Number of Linux Capability Usage

V. Conclusion and Future work

Containers have been adopted not only to a number of major cloud vendors, but also for developers. Design of containers dramatically increases performance compared to virtual machines, still security issues remain unsolved. Adversaries are able to abuse system calls to bypass isolation, and may grant root privilege or leak critical information. To handle it, we introduce eBPF base capability enforcement system by using Seccomp Filter. With this system, operators are able to give a minimal capability set which prevents unintentional operation of containers, even adversaries who tries to exploit host and other containers by abusing unnecessary capabilities.

Our preliminary result shows that majority of containers does not required entire default set of capabilities. In other words, default capabilities provides excessive privileges which may chance of exploit. With our work, it is possible to identify a proper capability set to be assigned for containers.

Our first target of future work is increasing coverage. Although we made an effort to cover as many case as we can, still unfounded capability can be remained. To handle this issue, we will extend our work by applying static analysis.

Acknowledgment

This work has been supported by the Future Combat System Network Technology Research Center program of Defense Acquisition Program Administration and Agency for Defense Development (UD190033ED).

References

[1] "Containers at Google" Accessed on: Feb. 28, 2021. [Online]. Available: https://cloud.google.com/containers

[2] "Container services" Accessed on: Feb. 28, 2021. [Online]. Available: https://azure.microsoft.com/en-us/product-categories/containers/

[3] "Containers on AWS" Accessed on: Feb. 28, 2021. [Online]. Available: https://aws.amazon.com/containers/

[4] "Containers on IBM Cloud" Accessed on: Feb. 28, 2021. [Online]. Available: https://www.ibm.com/cloud/containers

[5] "CVE-2016-0728" Dec. 16, 2015. Accessed on: Feb. 28, 2021. [Online]. Available: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2016-0728

[6] "CVE-2016-5195" May. 31, 2016. Accessed on: Feb. 28, 2021. [Online]. Available: https://cve.mitre.org/cgi-bin/cvename.cgi?name=cve-2016-5195

[7] "Linux Kernel 4.14.0-rc4+ - 'waitid()' Local Privilege Escalation" Oct. 22, 2017. Accessed on: Feb. 28, 2021. [Online]. Available: https://www.exploit-db.com/exploits/43029

TABLE II
CAPABILITIES SUPPORTED BY LINUX KERNEL 4.18

Category	Capabilities	Description
Namespace	CAP_SETGID	Manipulation of process GIDs
Namespace	CAP_SETUID	Manipulation of process UIDs
File	CAP_FOWNER	Bypass permission check on operations
Network	CAP_NET_BIND_SERVICE	Bind a socket to internet domain privileged ports
Network	CAP_NET_RAW	Open a non-TCP/UDP socket

[8] Luo, Yang, et al. ”Whispers between the containers: High-capacity covert channel attacks in docker.” 2016 IEEE Trustcom/BigDataSE/ISPA. IEEE, 2016.

[9] Gao, Xing, et al. ”ContainerLeaks: Emerging security threats of information leakages in container clouds.” 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, 2017.

[10] ”Threat Alert: TeamTNT is Back and Attacking Vulnerable Redis Servers” Sep. 30, 2020. Accessed on: Feb. 28, 2021. [Online]. Available: https://blog.aquasec.com/container-attacks-on-redis-servers

[11] Arnautov, Sergei, et al. ”SCONE: Secure linux containers with intel SGX.” 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). 2016.

[12] Lei, Lingguang, et al. ”SPEAKER: Split-phase execution of application containers.” International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, Cham, 2017.

[13] Rastogi, Vaibhav, et al. ”Cimplifier: automatically debloating containers.” Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. 2017.

[14] Sun, Yuqiong, et al. ”Security namespace: making linux security frameworks available to containers.” 27th USENIX Security Symposium (USENIX Security 18). 2018.

[15] ”Docker” Accessed on: Feb. 28, 2021. [Online]. Available: https://www.docker.com/

[16] ”Kubernetes” Accessed on: Feb. 28, 2021. [Online]. Available: https://kubernetes.io/

[17] Ghavamnia, Seyedhamed, et al. ”Confine: Automated system call policy generation for container attack surface reduction.” 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020). 2020.

[18] ”SELinux Project” Accessed on: Feb. 28, 2021. [Online]. Available: http://selinuxproject.org/page/Main_Page

[19] ”AppArmor” Accessed on: Feb. 28, 2021. [Online]. Available: https://gitlab.com/apparmor

[20] seccompsandbox - overview.wiki Accessed on: Feb. 28, 2021. [Online]. Available: https://code.google.com/archive/p/seccompsandbox/wikis/overview.wiki

[21] ”eBPF” Accessed on: Feb. 28, 2021. [Online]. Available: https://ebpf.io/

[22] ”Docker run reference” Accessed on: Feb. 28, 2021. [Online]. Available: https://docs.docker.com/engine/reference/run/

[23] ”github bcc” Accessed on: Feb. 28, 2021. [Online]. Available: https://github.com/iovisor/bcc

[24] ”Docker Hub” Accessed on: Feb. 28, 2021. [Online]. Available: https://hub.docker.com

Heterogenous Breast Phantom with Carcinoma for Ionizing Machines

Budour AlFares
Biomedical Engineering Department
Imam Abdulrahman Bin Faisal
Univeristy, P.O. Box 1981, Dammam,
31441, Saudi Arabia
2170001049@iau.edu.sa

AlJohara AlJabr
Biomedical Engineering Department
Imam Abdulrahman Bin Faisal
Univeristy, P.O. Box 1981, Dammam,
31441, Saudi Arabia
2170001319@iau.edu.sa

Maryam Zainalabedin
Biomedical Engineering Department
Imam Abdulrahman Bin Faisal
Univeristy, P.O. Box 1981, Dammam,
31441, Saudi Arabia
2170000424@iau.edu.sa

Munthreen AlMuzain
Biomedical Engineering Department
Imam Abdulrahman Bin Faisal
Univeristy, P.O. Box 1981, Dammam,
31441, Saudi Arabia
2150003968@iau.edu.sa

Gameel Saleh
Biomedical Engineering Department
Imam Abdulrahman Bin Faisal
Univeristy, P.O. Box 1981, Dammam,
31441, Saudi Arabia
gsmohammed@iau.edu.sa

Maryam AlHashim
Medical Imaging Physics Department
King Fahd Specialist Hospital
Dammam, Saudi Arabia
maryam.hashim@kfsh.med.sa

Abstract—**Breast cancer is one of the main reasons that lead to mortality, particularly in females. One of the factors that help in breast cancer detection is the existence of breast phantom with high similarity to the real breast associated with the utilized modality. Currently, the existing breast phantoms are limited in terms of anatomy and breast ionization reactions properties. Thus, the aim of this paper is to design a heterogenous patient-based breast phantom that mimics the real breast tissues properties concerning ionizing imaging modalities. The phantom includes skin tissue, fibroglandular tissue, adipose tissue, pectoral muscles, and insertion of malignant lesion. The process started with real breast MR images from King Fahd Specialist Hospital with BI-RADS I tissue segmentation for molds creation. 3D Slicer software was used for segmentation, and the Meshmixer was used to refine the segmented tissue. The molds are printed using PRUSA 3D printers. The tissue-mimicking materials were tailored in terms of their elemental composition weight fractions and three ionization properties. These are the mass attenuation coefficient (MAC), electron density (n_e) and effective atomic number (Z_{eff}). These characteristics were compared with the corresponding properties of the real breast ICRU reported elemental compositions. The achieved results showed an excellent agreement between the MAC of the tissue-mimicking materials and the ICRU-based breast tissues. The percentage of error in n_e and Z_{eff} amounts to only 2.93% and 5.76%, respectively. That means the phantom can optimize the function of breast ionizing imaging modalities and lead to higher breast cancer detection sensitivity.**

Keywords—breast, phantom, ionization, electron density, mass attenuation coefficient, effective atomic number, tissue-mimicking material.

I. Introduction

Breast cancer is the second most occurring cancer in females and considered as the second reason of mortality [1]. Based on the national breast cancer foundation, an estimation of 276 thousands new cases were diagnosed with breast cancer in 2020 and more than 42 thousands women are expected to die in the United States (US) [2]. However, early-stage cancer detection would increase treatment possibilities and the survival rate [3]. Breast phantoms are significant devices that help in early breast cancer detection. They must mimic the properties of the breast tissues with tissue-equivalent materials having similar response of the breast when it is screened by each imaging modality. Anthropomorphic breast phantoms are utilized to produce images that simulate aspects of clinical breast screening. They are beneficial in characterization and optimization of breast imaging modalities. In the development of imaging screening technologies, technical assessment for modality optimization through phantoms is essential before the modality can be utilized for clinical use. Nowadays, the existing medical imaging phantoms have a lot of limitations including unrealistic uniform background structure that does not mimic real organs and tissues. They are mostly designed with homogenous solutions for the whole aimed organ with no simulation of tumors, masses, or other lesions.

Phantoms are devices that are being used in the field of medical imaging physics and health science [4]. They are considered as artificial models that represent the human body to evaluate, examine and tune the performance of several imaging modalities. Phantoms are designed to help in assessing the optimal radiation that is subjected to the patient especially in new emerging imaging techniques and for quality assurance (QA) purposes. Such advanced phantoms that are patient-based developed can further benefit in hands-on operator training and image-guided interventional procedures. Based on the purpose that the phantom is developed for, the composition and the design process would be established.

A. Mammography Breast Phantoms

In 2015, a group at Duke University developed a breast phantom by matching virtual breast phantoms for mammography projections. The virtual phantoms were made into physical phantoms using additive manufacturing multi-material 3D printing. One design with printed with single additive materials then filled with oil. Second design with double additive printed materials for whole breast. The first phantom design offer a better result in term of breast contrast compared to the first design, but the second presented an undesirable air bubbles [5].

Another physical 3D anthropomorphic breast phantom was developed at the University of Pennsylvania for image quality assessment of 2D and 3D breast x-ray imaging systems. The fabricated phantom consists of 45% dense tissue and

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

already 5cm deformed in thickness for mammography scan. Digital mammographic images with W/Rh at 30 kVp and 104 mAs showed a less than 1% coefficient of variation of the relative attenuation between the two simulated tissues with acceptable appearance with presence of air bubbles [6].

Moreover, another method for fabricating breast phantoms was developed by a research group at the Committee for the defense of human right (CDHR). The aim was to model breast x-ray attenuation properties. The phantom was designed only to match full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT). After projections at 35kVp it was found that both glandular and adipose tissues were acceptable with limitation of masses and lesions insertion to the phantom [7].

B. Computed Tomography Breast Phantoms

A study was published with an aim to construct breast phantom made from polyethylene by segmenting the patient's breast image into adipose and fibro-glandular tissues. To mimic the different tissues, thermoplastic mold was placed as the outer layer of the skin, fibro-glandular tissue represents by filling the air gaps by water, while polyethylene and paraffin wax were used to mimic the adipose tissue. Moreover, calcium carbonate particles were used to represent the macrocalcification. The designed phantom model uncompressible breast with an ability to do measurements [8].

C. Multimodality Breast Phantoms

In 2016, a paper published with an aim to develop a multipurpose gel-based breast phantom consisting of a simulated tumor to function with ultrasound, CT, and MRI. The TMMs was ballistic gelatin powder and Metamusil^{TM} for breast background. Barium sulfate, copper sulfate, water and ballistic gelatin were used for the simulated tumor. The resulted Hounsfield units (HU) of the simulated breast background was 24 HU which was far from the reference value of –100 HU for adipose and 40 HU for fibro-glandular. On MRI, the tumor showed a signal-difference to noise ratio of 3.7. The results indicate a further development should be done [9].

In 2019, a study was conducted to design a 3D-printed breast phantom for multimodal imaging with TMMs based on a 3D printing. The developed phantom was based on polyvinyl chloride (PVC) including a structure of different lesions, adipose, and fibro-glandular tissues. CT and MRI were used to determine the tissue mimicking properties considering the HU and MRI relaxation times. The results showed that the temperature difference between the PVC softener mixture and the breast mold could present bubbles that affects the image quality and cannot be eliminated. Also, lack in heterogeneity present in the tissues reduces the similarity to the real breast tissues [10].

Recently in 2020, a group at Sapienza University published a phantom for multimodality use. The TMMs were based on different properties such as dielectric properties, acoustic properties, and attenuation coefficient. The phantom used TMMs to simulate the skin, fat tissue, glandular tissue, tumor, and muscle. The phantom resulted a good match between the reference and the physical measured values with ±10% for the majority of the TMMs. However, the phantom solid parts were compression irreversible because of the fat layer composition. Also, further investigation should be done in order to have more contrast between the tumor and surrounding tissues [11].

This paper aims to design and develop a heterogenous patient-based breast phantom that mimics the real breast tissues heterogeneous. The anatomy and tissues properties of the breast when screened with ionizing imaging modalities are considered. Three main tissue characteristics are investigated and analyzed to validate the performance of the proposed design. These are the mass attenuation coefficient, electron density and effective atomic number.

II. Materials and Methods

A. Patient data acquistion

Institutional Review Board (IRB) approval was obtained from King Fahd Specialist Hospital (KFSH) in Saudi Arabia, Dammam (RAD0319), in order to use patient data. A Digital Imaging and Communications in Medicine (DICOM) magnetic resonance imaging (MRI) breast images with dynamic contrast-enhancement is used. The criteria of patient data selection were built upon the BI-RADS which was established by American College of Radiology (ACR). The score of the chosen image was I (normal breast) with dense fibro-glandular tissues to be inserted with a malignant lesion.

B. Image segmentation and molds creation

For breast tissue segmentation, the acquired MR breast images were imported into segmentation software. This is to have a realistic separated geometry for the external shape, skin, fibro-glandular tissue and tumor. The patient DICOM images were imported into 3D slicer software with 128 slices for internal tissues segmentation to segment out the skin and fibro-glandular tissue from the surrounding adipose. The purpose was to ensure a complete elimination of all the surrounding structures. Each slice was segmented in different orientations to improve segmentation reliability using the threshold function, which was adjusted manually. Afterwards, the segmented fibro-glandular tissue was converted into 3D model and saved as Standard Triangle Language (STL) file for further processing. Negative and positive molds were created from the segmented skin layer to create a single flask for the skin and adipose tissue. Segmented fibro-glandular was modified to create a base for handling purposes. Segmented fibro-glandular was modified to create a base for handling purposes. Also, a tumor mold of 2 cm diameter was created. The molds of the external breast shape, skin, and fibro-glandular were printed using acrylonitrile butadiene styrene (ABS) plastic to have a high realistic distribution of the interior structure, specifically for the fibro-glandular mold [12].

C. Tissue-mimicking materials (TMMs)

The proposed components for the TMMs in Table 1 was based on choosing specific materials [11], [13]-[18] that have been modified and validated using the ionizing characteristic parameters. This is to provide realism to mimics the representation of the real breast tissues with an improved anthropomorphic nature as well as meeting the response of the breast ionization reactions. Deionized water and safflower were introduced to all TMMs to simulate the water and oil concentrations in the tissues, respectively. Additionally, Sodium chloride, Aluminum oxide and Potassium chloride were used as scattering particles when the tissue is introduced to the X-rays and to affect the attenuation with a linear dependence as a function of frequency.

TABLE 1: FORMULA FOR THE SKIN, ADIPOSE, FIBRO-GLANDULAR, CARCINOMA AND PECTORAL MUSCLES TISSUES.

Layer	Component	Weight (g)
Skin	NaCl	15
	Deionized water	620
	X-100 surfactant	30
	PVA	50
	Agar	5
	Benzalkonium chloride	4
	Sugar	300
Adipose Tissue	NaCl	2
	Deionized water	110
	X-100 surfactant	100
	Safflower oil	315
	Beeswax	400
	Agar	7
	SiC	4.9
	KCl	6.5
Fibro-glandular Tissue	SiC	4
	Deionized water	663.8
	X-100 surfactant	40
	Safflower oil	170
	Glycerol	130
	Agar	27
	Aluminum oxide	15
	KCl	5
	Benzalkonium chloride	5
Carcinoma	NaCl	7
	Agar	35
	Sugar	220
	Deionized water	700
	Benzalkonium chloride	3.5
	KCl	1.9
Pectoral muscles Tissue	Agar	20
	Sugar	360
	Deionized water	900
	Benzalkonium chloride	4.14
	X-100 surfactant	20
	White egg	50
	Safflower oil	40

D. *Tissue-mimicking materials characterization*

The tissue-mimicking materials would be characterized for modalities utilize ionizing radiation such as computed tomography and mammography. The TMMs of the proposed phantoms are characterized analytically by using three important parameters which are mass attenuation coefficient, electron density (n_e) and effective atomic number (Z_{eff}). Mass attenuation coefficient describes how simply it can be penetrated by a beam of light, sound, particles, or other energy or matter [19]. While the electron density measures the electron probability of being exist in a unit volume of an element [20], and for the Z_{eff} it is an element atomic number in which photon interact similar to a given composite material. It would be found numerically by using National Institute of Standards and Technology (NIST) XCOM. The energies will be specified according to the range used in each imaging modality after having the weight fractions for each tissue compounds.

From the mass density (ρ_m) and atomic composition ($\frac{Z}{A}$), the electron density of a material can be calculated according to the formula:

$$n_e = \rho_m . N_A . \left(\frac{Z}{A}\right) \quad (1)$$

$$where: \ \frac{Z}{A} = \sum_i a_i . \left(\frac{Z_i}{Ai}\right) \quad (2)$$

N_A is Avogadro's number and a_i is the fraction by weight of the ith element of atomic number Z_i and atomic weight A_i ranging from 0 to 8 depending on the elemental composition.

Z_{eff} can be obtained from:

$$Z_{eff} = \Big(\sum_{\substack{all\ tissue \\ components\ (n)}} a_n Z_n^{2.94}\Big)^{\frac{1}{2.94}} \quad (3)$$

Where a_n represent the fractional contribution of each element to the total number of electrons in mixture.

Based on the International Commission on Radiation Units (ICRU), reference values elemental compositions weight fractions of the real tissues would be compared and validated with the above calculations.

III. RESULTS AND DISCUSSION

A. *Image segmentation*

For the patient DICOM images segmentation to extract the region of interest (ROI) through detection of boundaries, breast outer shape and fibro-glandular tissue was separated from other surrounding tissues using 3D Slicer software to have a realistic separated geometry for the external shape, skin and fibro-glandular tissue. The resulted 3D fibro-glandular volume model comparable to the axial and coronal images slices with high anatomical precision shown in Figure 1 from different perspectives. The segmented model presented a need for further smoothing and processing to be able for 3D printing without defects.

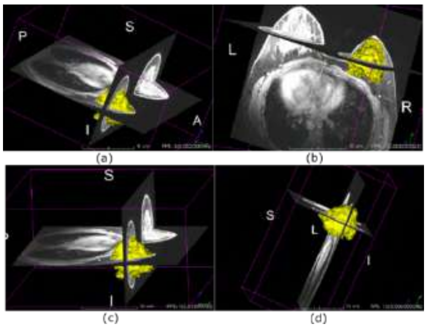


Figure 1: 3D segmented model comparable to the MRI image slices from different perspectives. 3D Slicer is used.

B. *3D mold fabrication*

Free-floating and deformities were removed, and model hollows were edited during this process. The skin was converted into a negative mold to represent the outer shape of the phantom. In this outer shape mold, a 3 mm thickness was added at the rounded corners to create the skin tissue thickness. The fibro-glandular STL file was edited to have an applicable adipose to fibro-glandular interface surface for printing. Figure 2 presents all processed molds.

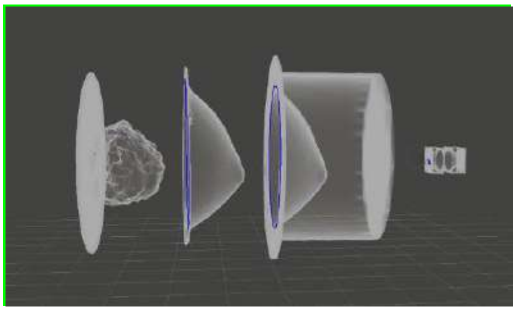


Figure 2: 3D phantom molds design.

The 3D models of the phantom were printed in Namthaja 3D printing manufacturing solutions, Saudi Arabia, Dammam. The printed external shape, skin, fibro-glandular, tumor molds are shown in Figure 3.

C. *TMMs characterization*

The weight fractions of the proposed TMMs elements were calculated for each breast tissue of skin, fibro-glandular, adipose, pectoral muscle, and malignant carcinoma. Table 2 presents the resulted weight fractions of TMMs comparable to the real breast tissue elemental compositions weight fractions found from ICRU reports. The values where close to the real especially when focusing on the main tissue elements which are C, H and O for all tissues. Table 3 was develop based on Equations (1) and (2) which present the specific tissue electron density and effective atomic number. Low errors were observed with maximum of 5.76% for fibroglandular Zeff and minimum of 0.00415% for pectoral muscle ne. That means the proposed materials and quantities can produce similar real breast tissue interactions when exposed to ionizing radiations with energies in the range 10-150 KeV.

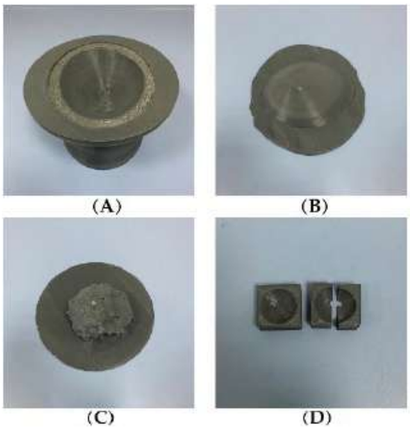


Figure 3: Breast phantom molds: (A) Outer skin, (B) skin, (C) fibro-glandular and (D) carcinoma mold.

Additionally, the mass attenuation coefficients in cm^2/g of the proposed phantom tissues were found and plotted against the real breast tissues mass attenuation coefficients in the energy range of 10 KeV to 150 KeV, as clinically used in mammography and CT machines.

Figures 4, 5, 6, 7 and 8 show the mass attenuation coefficient against the applied photonic energy for the skin, adipose tissue, fibroglandular tissue, malignant carcinoma, and pectoral muscles tissues, respectively for both real ICRU elemental compositions and the phantom calculated elemental compositions. All the introduced tissue-mimicking materials of the proposed phantom revealed a good agreement in all the energy range of energy with insignificant differences in the range of small applied photonic energy. The mass attenuation coefficient for the skin and adipose tissue real and calculated results showed a great overlapped graph for all points with high similarity.

TABLE 2: BREAST PHANTOM ELEMENTAL COMPISITIONS WEIGHT FRACTIONS.

		Na	Cl	C	H	O	Mg	K	N	S	P	Si	Al
Skin	Phantom	0.004821033	0.007692516	0.275323873	0.098487258	0.613573165	-	-	0.000102156	-	-	-	-
	ICRU	0.001	0.003	0.204	0.1	0.645	-	0.001	0.042	0.002	0.001	-	-
Adipose	Phantom	0.000832225	0.004552743	0.689825033	0.12133871	0.17621494	-	0.003605935	-	-	-	0.003630414	-
	ICRU	0.001	0.001	0.598	0.114	0.278	-	-	0.007	0.001	-	-	-
Fibro-glandular	phantom	-	0.002619926	0.218820618	0.105858707	0.659929689	-	0.002477372	0.000147657	-	-	0.002646044	0.007499987
	ICRU	0.001	0.001	0.332	0.106	0.527	-	-	0.03	0.002	0.001	-	-
Pectoral Muscles	Phantom	-	0.000234935	0.16631	0.096804967	0.73072	-	-	0.005929387	-	-	-	-
	ICRU	0.0008	-	0.123	0.101997	0.720993	0.002	0.0002	0.035	0.005	0.002	-	-
Carcinoma	Phantom	0.002849483	0.005615315	0.116584436	0.098753539	0.776084016	-	0.001031131	0.000113212	-	-	-	-
	ICRU	-	-	0.187626775	0.101419878	0.668356998	-	-	0.042596349	-	-	-	-

TABLE 3: SKIN, FIBROGLANDULAR, ADIPOSE, PECTORAL MUSCLES AND CARCINOMA ELECTRON DENSITY AND Z_{EFF}.

Tissue		Electron density	Error %	Z_{eff}	Error %
Skin	Phantom	3.59776E+23	0.163	7.22	0.558
	ICRU	3.60362E+23		7.2630298	
Fibro-glandular Tissue	Phantom	3.39E+23	0.0315	7.33E+00	5.76
	ICRU	3.17967E+23		6.93E+00	
Adipose Tissue	Phantom	3.19937E+23	0.619	6.391046106	0.956
	ICRU	3.17967E+23		6.330516684	
Pectoral muscles	phantom	3.43E+23	0.00415	8.29E-01	0.195
	ICRU	3.44E+23		8.184042E-1	
Carcinoma	Phantom	3.57E+23	2.93	7.46E+00	4.46
	ICRU	3.31E+23		7.11E+00	

IV. CONCLUSION

The purpose of this study was to develop a patient-based heterogenous breast phantom that mimics various breast tissues when screened using ionizing imaging modalities. The chosen phantom TMMs were validated using three main radiation characteristics which are the mass attenuation coefficient, electron density, and effective atomic number. The process of developing the phantom started with IRB that was requested and approved from KFSH, for data acquisition. Then it was followed by segmentation using 3D Slicer and refinement process using Solid work software. The designed molds are then printed using PRUSA 3D printers. For each tissue mimicking process, a combination of different materials with different weights were used. The weight fractions of the elements used in these materials are calculated numerically and compared with the real values from the ICRU reports. The mass attenuation coefficient, electron density and the effective atomic number for each TMMs are each the results showed a good agreement between them.

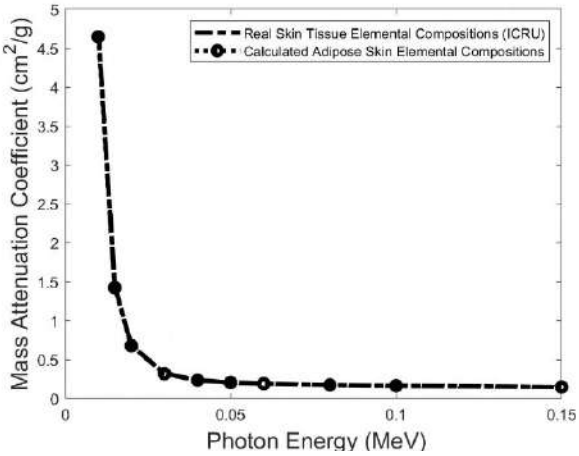


Figure 4: Real and calculated skin elemental compositions photon energy vs. mass attenuation coefficient graph.

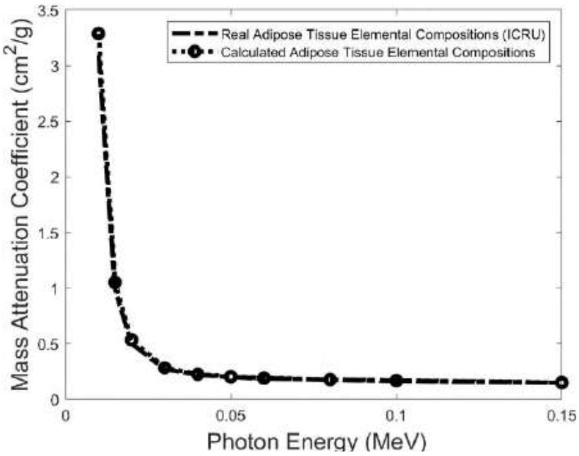


Figure 5: Real and calculated Adipose elemental compositions photon energy vs. mass attenuation coefficient graph.

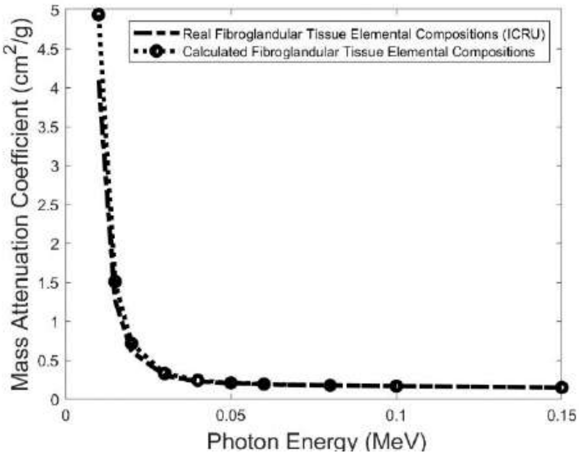


Figure 6: Real and calculated fibroglandular elemental compositions photon energy vs. mass attenuation coefficient graph.

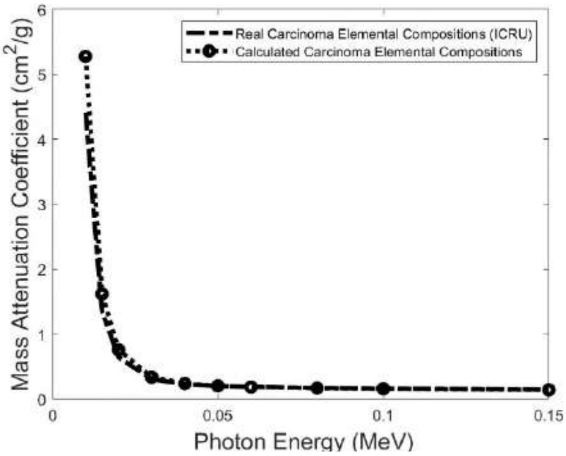


Figure 7: Real and calculated carcinoma elemental compositions photon energy vs. mass attenuation coefficient graph.

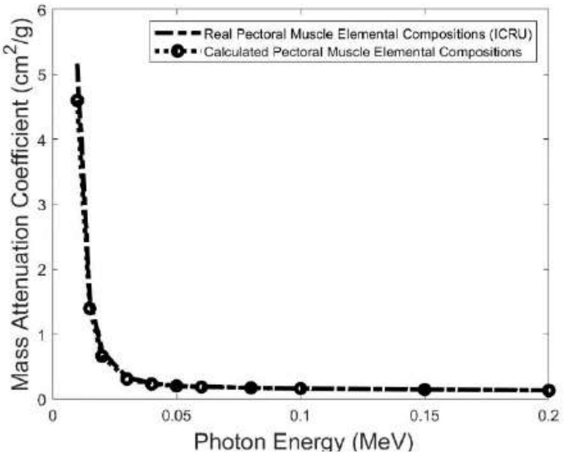


Figure 8: Real and calculated pectoral muscle elemental compositions photon energy vs. mass attenuation coefficient graph.

ACKNOWLEDGMENT

We thank Dr. Jawaher AlGhamdi for her dedicated support in the Chemistry Lab. We also thank Namthaja company, eastern state, for their help in the 3D printing. Finally, this research was funded by the Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University Dammam, Saudi Arabia Grant (2020-053-Eng).

REFERENCES

[1] CDCBreastCancer, “Basic Information About Breast Cancer,” *Centers for Disease Control and Prevention*, Sep. 14, 2020. https://www.cdc.gov/cancer/breast/basic_info/index.htm (accessed Feb. 27, 2021).

[2] “2020-Breast-Cancer-Stats.pdf.” Accessed: Nov. 26, 2020. [Online]. Available: https://www.nationalbreastcancer.org/wp-content/uploads/2020-Breast-Cancer-Stats.pdf.

[3] “Promoting Cancer Early Diagnosis.” https://www.who.int/activities/promoting-cancer-early-diagnosis (accessed Nov. 26, 2020).

[4] L. A. DeWerd and M. Lawless, “Introduction to Phantoms of Medical and Health Physics,” in *The Phantoms of Medical and Health Physics: Devices for Research and Development*, L. A. DeWerd and M. Kissick, Eds. New York, NY: Springer, 2014, pp. 1–14.

[5] N. Kiarashi, A. Nolte, G. Sturgeon, W. Segars, S. Ghate, L. Nolte, E. Samei, J. Lo, "Development of realistic physical breast phantoms matched to virtual breast phantoms based on human subject data", Medical Physics, vol. 42, no. 7, pp. 4116-4126, 2015. Available: 10.1118/1.4919771.

[6] A.-K. Carton, P. Bakic, C. Ullberg, H. Derand, and A. D. A. Maidment, “Development of a physical 3D anthropomorphic breast phantom,” *Med. Phys.*, vol. 38, no. 2, pp. 891–896, 2011, doi: https://doi.org/10.1118/1.3533896.

[7] A. H. Rossman, M. Catenacci, C. Zhao, D. Sikaria, J. Knudsen, D. Dawes, M. Gehm, E. Samei, B. Wiley, J. Lo, “Three-dimensionally-printed anthropomorphic physical phantom for mammography and digital breast tomosynthesis with custom materials, lesions, and uniform quality control region,” *J. Med. Imaging Bellingham Wash*, vol. 6, no. 2, p. 021604, Apr. 2019, doi: 10.1117/1.JMI.6.2.021604.

[8] N. D. Prionas, G. W. Burkett, S. E. McKenney, L. Chen, R. L. Stern, and J. M. Boone, “Development of a Patient-Specific Two-Compartment Anthropomorphic Breast Phantom,” *Phys. Med. Biol.*, vol. 57, no. 13, pp. 4293–4307, Jul. 2012, doi: 10.1088/0031-9155/57/13/4293.

[9] M. Ruschin, S. Davidson, W. Phounsy, T. Yoo , L. Chin, J. Pignol, A. Ravi, C. McCann., “Technical Note: Multipurpose CT, ultrasound, and MRI breast phantom for use in radiotherapy and minimally invasive interventions,” *Med. Phys.*, vol. 43, no. 5, p. 2508, May 2016, doi: 10.1118/1.4947124.

[10] Y. He, Y. Liu, B. A. Dyer, J. M. Boone, S. Liu, T. Chen, F. Zheng, Y. Zhu, Y. Sun, Y. Rong, J. Qiu., “3D-printed breast phantom for multi-purpose and multi-modality imaging,” *Quant. Imaging Med. Surg.*, vol. 9, no. 1, pp. 63–74, Jan. 2019, doi: 10.21037/qims.2019.01.05.

[11] G. Ruvio, R. Solimene, A. Cuccaro, G. Fiaschetti, A. J. Fagan, S. Cournane, J. Cooke, M. Ammann, J. Tobon, J. E. Browne., “Multimodal Breast Phantoms for Microwave, Ultrasound, Mammography, Magnetic Resonance and Computed Tomography Imaging,” *Sensors*, vol. 20, no. 8, Art. no. 8, Jan. 2020, doi: 10.3390/s20082400.

[12] M. J. Burfeindt, T. J. Colgan, R. Owen Mays, J. D. Shea, N. Behdad, B. D. Veen, S. C. Hagness, “MRI-Derived 3-D-Printed Breast Phantom for Microwave Breast Imaging Validation,” *IEEE Antennas Wirel. Propag. Lett.*, vol. 11, pp. 1610–1613, 2012, doi: 10.1109/LAWP.2012.2236293.

[13] Q. Duan, J. Duyn, N. Gudino, J. A. de Zwart, P. Gelderen, D. Sodickson, R. Brown , “Characterization of a dielectric phantom for high-field magnetic resonance imaging applications,” *Med. Phys.*, vol. 41, no. 10, Oct. 2014, doi: 10.1118/1.4895823.

[14] A. Oglat, M. Z. Matjafri, N. Suardi, M.A. Oqlat, M. A. Abdelrahman, A. A. Oqlat, O. F. Farhat, B. N. Alkhateb, R. Abdalrheem, M. S. Ahmad, M. Abujazar, “Chemical Items Used for Preparing Tissue-Mimicking Material of Wall-Less Flow Phantom for Doppler Ultrasound Imaging,” *J. Med. Ultrasound*, vol. 26, no. 3, pp. 123–127, 2018, doi: 10.4103/JMU.JMU_13_17.

[15] Lafon, C., Sapozhnikov, O., Kaczkowski, P., Vaezy, S., Noble, M. and Crum, L., 2001. An innovative synthetic tissue - mimicking material for high - intensity focused ultrasound. The Journal of the Acoustical Society of America, 110(5), pp.2613-2613.

[16] K. Manickam, M. R. Reddy, S. Seshadri, and B. Raghavan, “Development of a training phantom for compression breast elastography—comparison of various elastography systems and numerical simulations,” *J. Med. Imaging*, vol. 2, no. 4, Oct. 2015, doi: 10.1117/1.JMI.2.4.047002.

[17] M. Miyakawa, S. Takata and K. Inotsume, "Development of non-uniform breast phantom and its microwave imaging for tumor detection by CP-MCT," 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 2009, pp. 2723-2726, doi: 10.1109/IEMBS.2009.5333383.

[18] J. Browne, A. Fagan, and A. Doyle, “Tissue Mimicking Materials,” 20200170612, 04-Jun-2020.

[19] J. H. Hubbell and S. M. Seltzer, “X-Ray Mass Attenuation Coefficients,” NIST. [Online]. Available: https://www.nist.gov/pml/x-ray-mass-attenuation-coefficients.

[20] F. M. Khan and J. P. Gibbons, *Khan’s the physics of radiation therapy*, Fifth edition. Philadelphia, PA: Lippincott Williams & Wilkins/Wolters Kluwer, 2014.

A Causal Bayesian Network Model for Resolving Complex Wicked Problems

Daniel T. Semwayo
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
2292902@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—Wicked problems are a specific class of complex problems that emerge from complex adaptive systems (CAS) and stakeholder disagreements on the definition and character of these problems and their possible resolution. Attempts at resolving wicked problems through integration and use of formal methods such as ontologies, Bayesian networks (BN), and complex systems dynamic (CSD) models have been attempted recently but wicked problems continue to defy resolution. This paper argues that this is the result of a lack of ontologically precise causal Bayesian models that adequately represent the hierarchical, dynamic, emergent characteristics and multiple perceptions of CAS and their emergent wicked problems. This paper's contribution is the incorporation of complexity systems theory concepts, namely: perspective, granularity and context, as explicit ontological constructs in a high precision ontological causal BN model, the Granular Contextual Perspectives (GCP) causal Bayesian Network model, using Hidden Markov Model (HMM) formalism to address this shortcoming. Using an illustrative example this conceptual paper shows that the (GCP) causal Bayesian Network model performs better than baseline Bayesian Network models at the visual representation, compact and retractable inference, and machine learning of CAS and their emergent wicked problems. The model is useful at supporting the exploration of possible effects of proposed alternative interventions or prototypical design strategies for resolving a given wicked problem.**

***Index Terms*—Hidden Markov Models, Causal Hierarchical Dynamic Bayesian Networks , Ontology engineering, Wicked problems, Complex Adaptive Systems, Design Science strategies**

I. Introduction

Wicked problems, which have been described as specific class of ill-defined complex problems that emerge from complex adaptive systems (CAS) and stakeholder disagreements on the definition and character of these problems have, and continue to be difficult to resolve [1]–[4].

Wicked problems include ill-defined problems like pandemics, climate change effects, traffic jams, rapidly changing business environments, and financial market crashes. While these problems have not always been recognised or defined explicitly as wicked problems, attempts at resolving this class of complex problems has been receiving a lot of attention recently [2], [3], [5]–[7].

Resolving wicked problems has proved to be elusive primarily because of their complex and dynamic nature, the difficulty faced in defining them and disagreements on how to resolve them, especially where multiple stakeholders with divergent perspectives of a given problem are involved. The lack of recognition of these complex problems as "wicked", is in itself part of the problem as inappropriate tools suitable for resolving "tame", well defined static problems as puzzles are then applied incorrectly to resolve these non linear wicked problems [8].

Complex Systems Dynamic (CSD) models [9], Ontologies [10], [11], Hierarchical Bayesian Networks (HBN) [12], Hidden Markov Model (HMM) [13] and combinations of these have increasingly been used to try and solve wicked problems. The promise of these various efforts lies in the integration of the advantages provided for by each of these approaches into an integrated modeling framework.

While some effort at such integration has indeed been attempted in the recent past, [14]–[17], challenges still remain. This paper argues that this is the result of a lack of ontologically precise causal Bayesian Network (BN) models that adequately represent the hierarchical, dynamic, emergent characteristics, and multiple perceptions of CAS and their emergent wicked problems. This paper proposes a novel ontologically precise casual Bayesian model to address this shortcoming.

The paper's contribution is the incorporation of complexity systems theory concepts, namely: perspective; granularity; and context, as explicit BN model ontological constructs to develop a high precision ontological causal BN model, the Granular Contextual Perspectives Bayesian (GCP) causal Bayesian model. The model uses Hidden Markov Model (HMM) formalism to: i) enhance the visualisation of the character of CAS and their emergent wicked problems; ii) extend the scope of Bayesian inference to answer wicked problems' specific queries; iii) refine Bayesian machine learning of the structure and parameters of given wicked problems.

The rest of the paper is structured as follows. Section II gives an account of related work. Section III provides a broad overview of the character of wicked problems. In section IV the key complexity theory kernel concepts (granularity, context, and perspectives) incorporated in the GCP causal

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Bayesian model as modeling constructs are introduced and defined. In section V, using the COVID19 pandemic as an illustrative example, a further exploration of the character of wicked problems is carried out through the lens of the systems theory kernel concepts identified as useful in modeling the character of wicked problems.

Section VI provides the formal definition of the GCP causal BN model and details of the model architecture. In section VII the superiority of the (GCP) causal BN model to baseline Bayesian models in addressing wicked problems is demonstrated using the COVID19 pandemic as an illustrative example. In section VIII a discussion of the implications of the (GCP) causal BN model and its utility claims on modeling wicked problems is carried out. Section IX provides a conclusion and the way forward with respect to empirical experimentation using the GCP causal BN model.

II. Related work

The character of wicked problems as viewed through a systems paradigm lens appears in [1], [7], [18], [19]. Work on modeling CAS using CSD models from which wicked problems emerge is detailed in [9]. Systems thinking and design thinking disciplines having contributed to a better understanding of the systemic character of wicked problems and ways to address such problems [19]–[22].

[23] define an ontology as "a formal, explicit specification of a shared conceptualization". The application of ontology engineering to model the structure of complex phenomena to generate formal human and machine readable artifacts called domain ontologies is found in [11], [24]–[26]. The utility of ontologies as representing the structure of domain knowledge is well documented in [24], [26]–[28]. The application of ontological theory based on logical and philosophical principles to enhance clarity and precision in abstracting complex reality has been proposed in [11], [24], [26].

Bayesian networks, as graphical probabilistic models have been developed to compactly represent and reason over linked complex phenomena using computation techniques [29], [30]. Various attempts to integrate dynamic systems models, ontology engineering methods and Bayesian networks to model CAS are detailed in [16], [31], [32].

While these various disciplines have contributed to a better understanding of complexity and how to handle complexity through visualization and computation, attention to their integration for the purpose of addressing wicked problems emerging from CAS and multiple stakeholder perspectives of such problems has been fragmented and inadequate.

As highlighted in [33] Causal Bayesian Models CBMs need sound knowledge of causal big data generating processes to effectively support useful prediction, inference, and structure, parameter learning applications. CSD models provide insights into CAS dynamic and delayed feedback processes, their causes and effects. CBMs and their formalism in particular, as articulated in [34] hold much promise as the basis for integrating CAS' ontological constructs and processes to adequately represent and provide the computational mechanism to reason about, and resolve wicked problems.

Recent accounts in literature point to the inadequacy of ontological constructs in existing conceptual models to support the conceptual modeling of CAS to resolve wicked problems [35], [36]. Incorporation of highly precise ontological structures as Causal Bayesian Models (CBM) constructs representing CAS and emergent wicked problems' knowledge are thus necessary for such purposes. The GCP causal BN model provides a novel way to addresses these inadequacies.

III. The character of Wicked problems

The term "wicked problem" attributed to Churchman [37], was popularised by [38], and has been used to describe ill-defined, very difficult problems to solve [39]. A summary of the original ten characteristics of wicked problems as outlined in [38] are summarised in [40] as follows: i) non-solubility, i.e., inability to break the problem into parts; ii) non-definitiveness in problem resolution i.e., that there is "no single or definite computational formulation or a set of valid solutions or right answers, but only answers that are better or worse from different angles" [41]; iii) indeterminacy, i.e., conflicting perspectives of the problem and possible solutions emanating from different experiences, knowledge, goals; iv) irreversible consequentiality, i.e., trial and error strategies do not work [40]. Wicked problems include pandemics, inequality, poverty, traffic jams, rapidly changing business environments, and financial market crashes. Recent literature refers to super wicked problems which not only exhibit the characteristics of wicked problems but an added dimension of urgency in resolving them because time to resolve their irreversible effects is running out [42]. Global climate change effects fit into this category [6].

Wicked problems have been closely linked to CAS that are characterised by: non-linearity; positive and negative feedback loops between multiple interacting entities and across multiple dimensions [43]. These characteristics need to be grasped to fully appreciate the complexity of the interacting variables and the pathological emergent structures and behavioural patterns which are refereed to as wicked problems [3], [5], [43], [44].

Wicked problems can thus be viewed as a class of pathological effects that emerge from complex interactions between agents in bounded, albeit open, natural or artificial CAS manifest at varying granular levels of observation or engagement, whose definition for a given context / sub-context is contested by stakeholders with multi-dimensional perspectives of the problem [43]. Emergence is central to the characterisation of wicked problems. [45] provide a useful formal framework for defining emergence:

> Let $\{S_i\}_{i\in I}$ be a family of general systems or "agents". Let Obs^1 be "observation" mechanisms and Int^1 be interactions between agents.
> The observation mechanisms measure the properties of the agents to be used in the interactions. The interactions then generate a new kind of structure $S^2 = R\left(S_i^1, \mathrm{Obs}^1, \mathrm{Int}^1\right)$ which is the result of

> the interactions. This could be a stable pattern or a dynamically interacting system. We call S^2 an emergent structure which may be subject to new observational mechanisms Obs^2 [45].

A wicked problem is thus where S^2 is a pathological emergent structure with related behavioural patterns arising from the numerous dynamic interactions of related entities within natural or artificially defined bounded CAS, and observed at Obs^2.

Wicked problems manifest at two levels. The first level, which in this paper shall be referred to as level I, is about the ontological aspects of the problem, that is, what the problem is, what relationships characterize the problem, and where and how it manifests. The second level, level II, is about the epistemological aspects of wicked problems, that is, how and at which granular level wicked problems are perceived, explored, known, and understood, at say Obs^2, and how candidate solutions are conceptualised. At level II multiple divergent stakeholder contextual perspectives from varying granular levels of observation or abstraction of a level I wicked problem are explored.

[46] identify dynamic complexity, finitude, and normativity as the summary key factors that cause wicked problems. Complexity, they define as a feature that arises from interactions between system variables and feedback loops which make natural and engineered systems unpredictable. Finitude refers to cognitive ability limitations, experience and knowledge. Normativity is about different norms and values held dear by different stakeholders, the major source of conflict that makes consensus difficult to reach [46]. Complexity factors belong to level I, while finitude and normativity belong to level II.

The numerous interactions among variables, (the causes), and the unpredictable patterns of change, (the effects), make it difficult to make effective interventions to resolve wicked problems. Interventions can easily become the source of other problems. As a result these problems are notoriously difficult to abstract, represent in conceptual models and solve using existing computational models in information systems.

Systems theory and systems thinking as a practice have been applied to identify prototypical pathological patterns that emerge from such interactions and unintended consequences that arise from the inappropriate of use of deterministic [20], [47], [48].

IV. Granularity, Context, and Perspective

For any given context or sub-context complex phenomena is abstracted cognitively and described at different levels of granularity and from a specific perspective. [35] captured this phenomena thus:

> "There is only one Herbert (the frog) that we and the molecular biologist apprehend, but, depending upon our interests and our focus, we may each apprehend him from different granular perspectives" [35].

Granularity is a concept relating to the cognitive, spatial, or temporal level of abstraction of a phenomena from an observer's point of view [36], [49], [50]. It is also used to define the coarseness of an observation or an investigation [51].

The important factor about granularity with respect to complex phenomena is that the observer's interest and hence level of abstraction is a key determinant in what is defined, 'seen' and investigated [26]. [52] argues that to study and understand complex phenomena fully "we need to be able to keep an eye both the tree and the forest", that is, atomic elements, micro sub-systems that emerge from their interactions, and the macro system that result from the linkages between sub-systems.

Context in modeling systems is important. Natural and artificial systems are said to have a translucent boundary which defines what is inside of a system and what is outside, determined by the focus of interest [53]. [34] demonstrates the importance of context in deciding what is considered endogenous and exogenous to the system under review, and the impact this has on the modeling process.

Perspective refers to enduring beliefs by a person or a group of people with similar mental models of the world, a product of our knowledge, past experience, uncertainty, incomplete information and societal norms. Perspectives influence the definition, and proposed solution spaces for wicked problems and are a source of bounded rationality, that is, restricted understanding and explanation of the state of the world we live in and its problems [54].

Different perspectives of a given problem are the source of disagreements between stakeholders about the definition of a given common problem. The divergence in belief systems is typically driven by one or more of the following: local context; various scientific philosophical standpoints about complex phenomena; the related methods of studying them; experiences; culture; divergent stakeholder interests and biases [3], [6], [40].

Veridical partitioning is the essence of representation of some aspect of reality in a model for a given purpose [35]. [35] advocate embracing realist perspectivalism, a view that knowledge can be obtained by means of veridical granular partition integration. The GCP casual BN model is about exploring veridical partitions and promoting veridical integration as a way of promoting ontological commitment to a common view of a given wicked problem and its solution. It is thus important, given this objective, to explicitly represent multiple perspectives in ontologies and BNs to support the sharing of knowledge and gain wider understanding of a given wicked problem.

V. The COVID19 Pandemic as a Wicked Problem

In this section the Coronavirus Disease 2019 (COVID19) pandemic is used as an illustrative prototypical wicked problem example to further explore the character of wicked problems applying the key identified key systems theory modeling concepts: granularity; context; and perspective.

At the micro granular contextual sub-system level interactions between the Severe Acute Respiratory Syndrome

Coronavirus 2 (SARS-CoV-2), the lung cells and the human immune system take place. At the individual human granular contextual level, the COVID19 disease manifests (observed) as a complex clinical ailment that emerges from micro granular level interactions.

The problem is wicked for the following reasons. The problem is non-soluble, that is, it cannot be easily broken into parts that can be analysed and dealt with separately. The ailment is an emergent phenomenon from a wide range of interacting variables at the micro granular contextual sub-system level, i.e., the SARS-CoV-2, the lung cells, the human immune system, and pre-existing health conditions, which interact in both known (deterministic) and uncertain (unknown) ways. The solution to the ailment is non definitive. While development of specific vaccines provide a solution, such a solution is likely to be temporary. The coronavirus is subject to mutations, a result of the virus adapting to its dynamic ever changing environment, thus making it difficult to find a singular clinical solution to its eradication. At coarser granular contextual level, say the community level, large gatherings, the so called "super spreading events" lead to exponential spread of the coronavirus through close contact between those infected and those not yet infected, leading to surges or "waves" of community level infections. This phenomenon, as systemic pattern, is typical of pandemics, the 1918 H1N1 influenza 'Spanish flu' pandemic being a case in point. Here, consequences of inappropriate actions or behaviours are irreversible.

While the pandemic is broadly understood to be a health problem of level I 'wickedness' more complexity, level II complexity, is brought on by conflicting perspectives on the nature of the broader economic and social impacts of the level I health problem, and how these should be defined and resolved, a case of indeterminacy of the wicked problem. Non clinical solutions implemented in different countries like 'lock-downs' to limit the movement of people as a measure to stop the spread of the coronavirus have been contested by stakeholders with varying perspectives of the problem and with different specific interests.

Stakeholders with an interest in preserving economic activity argue that lock-down impositions threaten jobs and livelihoods. Travel restrictions have resulted in the collapse of businesses in the airline, hospitality and tourism sectors across the world, highlighting another characteristic of wicked problem, irreversible consequentiality, where decisions and their consequences cannot be reversed. Other interest groups argue that restrictions on movement impinge on their rights of freedom of movement, and this has led to riots in some countries. This is a classic character of wicked problems, where a solution to a given problem becomes the cause of another problem [38]. The divergent perspectives represent level II characteristics to do with finitude and normativity [46].

The foregoing represents a wide range of complex interactions between CAS at different levels that make the problem non-soluble, non-definitive, non-determinate, and subject to irreversible consequentiality [40]. CAS exhibit dynamism, emergence, counter-intuitiveness, inter-connectedness between interacting agents within systems at various granular contextual levels, which lead to non-linear, unexpected outcomes. These outcomes or impacts are emergent at at various granular levels of observation for given contexts and perspectives.

It has been argued in [55] that the hierarchical dynamic complexity of complex adaptive systems means that wicked problems cannot be solved in a deterministic kind of way, that is, by analysing the character and behaviour of the problem, and developing interventions at the same level of the problem observation. Interacting variables are different for each granularity, context, and perspective delineated sub-system, and thus have to be represented in models as such. Further, the inter-connectedness of these sub-systems, driven by the fact that they represent multiple perspectives of the same wicked problem have to be represented as such in ontologies to support the sharing of knowledge and perspectives.

It has been suggested in [43] that there is "an engaging symmetry between complexity and wicked problems, where complexity is both the source of intractable wicked problems, and a way to trace the pathway out". This is the approach taken in this paper in building a high ontologically precise casual hierarchical and dynamic Bayesian Network model to represent and support the resolution of wicked problems through Bayesian inference, structure, and parameter learning.

VI. The Granular Contextual Perspectives Causal Bayesian network model

Figure 1 shows a high level architecture of the GCP causal BN model depicted as a 2 time step unrolled dynamic HMM representation of two separate Granular Contextual Perspectives (GCP) systems and sub-systems A and B and an integrating GCP system C.

Each GCP system is modeled as a template, where the key ontological notions of granularity, context, and perspective are incorporated as template variables, see [56] for a detailed definition of templates as representational model frameworks.

Within each GCP sub-system entities and their attributes are modelled as observable, semi-observable and latent random variables and values, after [57]. Observable variables are colored grey. The relationships between the variables represent causal links between world entities concepts and their conditional probabilities.

The major difference between typical dynamic HMMs, and the GCP causal BN model, wherein lies the novelty of the GCP causal BN model, is that while the variables propagated and observed over time are the same for dynamic HMMs, represented by a single GCP sub-system, say GCP A in the GCP causal BN model model, each GCP template contains a set of interacting random variables which are either different for each GCP or interact in a unique way as their existence and behaviour is determined or influenced by the parent sub-system meta-properties: granularity, context, and perspective. Each variable in a GCP subsystem is essentially an entity playing a role defined by the GCP sub-system.

Further, intra - GCP sub-system variable interactions produce emergent patterns and properties at the sub-system level

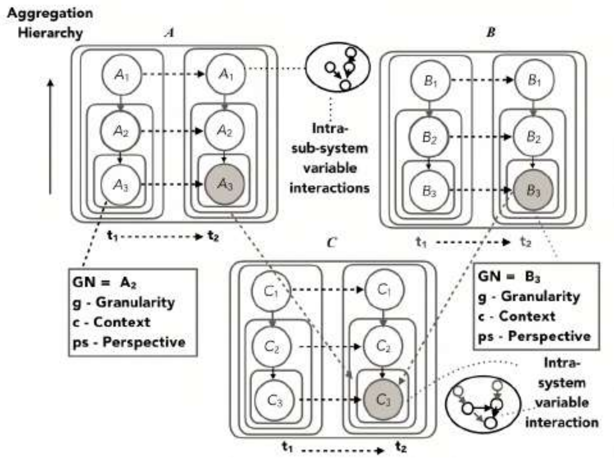


Fig. 1. Granular Contextual Perspectives Architecture

which are greater or smaller than the sum of their parts. In the COVID19 example provided this could be the disease, an emergent outcome of the interactions of the coronavirus and the lung cells and immune system of the individual.

The formal definition of the GCP system ontological architecture, which adapts and extends the definition of granular niches in [58] follows.

Definition 1:

A GCP is defined as an 9-tuple (S,T,M,N,Gl,C,Ps,A,E) where:

- (S) is a spatial or virtual location occupied by this GCP;
- (T) is a time interval granularity;
- (M) is a non-empty set of member entities, present at location S for part of time interval *T*, each representing a role;
- (N) is a non-empty set of interactions between entity roles, the normal behaviour in that gcp;
- (Gl) is a hierarchical structuring of the entities, *M* based on relative granularity of abstraction /
- observation/ manifestation i.e. *Gl* is a function mapping every *m* in *M* onto a granular level*gl* i.e. *Gl(m)* = *gl*;
- (C) is a description of contextual character for a given granular niche;
- (Ps) is a description of the world view that holds for a given GCP, a product of stakeholder belief system, knowledge, experience and culture;
- (A) is a set of emergent attributes of the GCP sub-system which are not direct attributes of its members *M*;
- (E) is a possibly empty set of environmental parameters that hold at location *S* during time *T*

The following constraints are applicable for GCP sub-systems:

- If *m* ∈ *M* then $Gl(m)$ *is unique i.e. every entity playing a specific role has exactly one granular level in a GCP.*
- *If* $m_1 \in M$ *and* $m_2 \in M$ *and* m_1 *is-part-of* m_2 *then* $Gl(m_1) \leq Gl(m_2)$
- *If* $m_1 \in$ *M,* $m_2 \in$ *M,Gl* ∈ *N and Gl* (m_1, m_2)*) then* ∃*t* ∈ *T s.t. In*(s, t, m_1) *and In*(s, t, m_2) *i.e. for two entities to interact in a GCP sub-system they must exist in that sub-system at the same time.*
- *For all m* ∈ *M,* ∃ *t* ∈ *T s.t. In(s,t,m) = false i.e. an entity does not have to remain in a GCP sub-system throughout its existence.*

The GCP causal BN model is defined formally as as a plate model, after [56] as follows:

Definition 2: A Plate model M_Plate defines, for each template attribute $A \in \aleph$ with argument signature $U_1, ..., U_k$:

- *A set of* template parents $Pa_A = \{(B_1(U_1)..., B_l(U_l)\ \}$, such that for each $B_i(U_i)$ we have that $U_i \subseteq \{U_1.., U_k\ \}$. The variables U_i are the argument signature of the parent B_i.
- A template CPD P$(A|Pa_A)$

The set of template parents B_i, $A_{1\cdot\cdot n}, B_{1\cdot\cdot n}, C_{1\cdot\cdot n}$, the template parents, in GCP causal BN model, and $\{granularity, context, perspective\}$, being instances of the argument signature, U_i.

The proposed GCP causal BN model artifacts are intended, on the one hand to specifically abstract an 'objective' ontological structure and stochastic processes character of a complex world, while on the other hand, surface, and lay bare, the various granular contextual perspectives and assumptions held by stakeholders about a given wicked problem. The GCP causal BN model is expected to support the design of model artifacts that more precisely represent complex reality and provide richer probabilistic reasoning and learning capabilities with respect to CAS and their emergent wicked problems than baseline Bayesian modeling frameworks.

A, for instance, using our illustrative example could represent the health services' perspective within the context of the COVID pandemic at different granular levels. A_3 then represents the micro granular contextual sub-system level. Here the variables of interest include the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the lung cells, the human immune system, and pre-existing health condition present. A_2 then represents the individual person granular contextual perspective sub-system level. Variables of interest at this level include COVID19 status, access to quality health care, severity of symptoms if present, and levels of anxiety.

The GCP system B could represent the economic perspective within the context of the Covid19 pandemic at different granular levels. C represents an integrative architecture to visually explore the linked sub-systems and Bayesian graphical pathways to enable the exploration of the likely effects of an intervention, within a GCP system (e.g., A), at a GCP subsystem level (e.g., A_3), across GCP systems and sub-systems (e.g., B_3). GCP C represents multi-stakeholder ontological commitment to a wicked problem definition and a candidate design solution/s, arrived at through consensus using the model.

VII. Utility of the GCP causal BN model

Figure 2 shows the application of causal diagrams, after [33] to inspect the precision of causal representation and the effects of interventions using the GCP causal BN model.

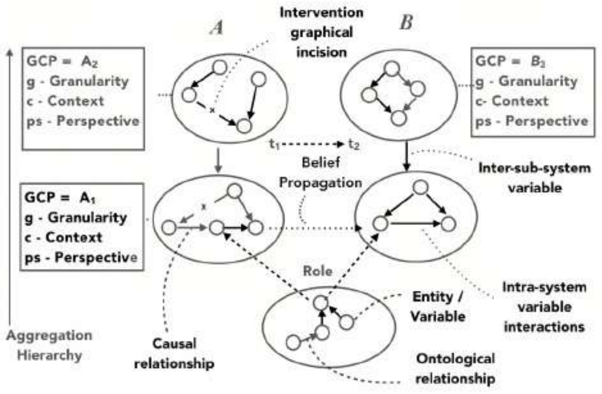


Fig. 2. Granular Contextual Perspectives causal graph showing an intervention

The X in the edges between the variables indicate an intervention, an incision of an edge to control the pathways between interacting variables. For example, wearing a mask "deletes" the pathway through which the coronavirus is transferred from person to person.

A comprehensive account of how to identify causal inconsistencies, such as spurious non-causal confounding variables, and the problems brought about by conditioning on collider nodes, i.e., nodes in a causal graph representing a variable causally influenced by two or more variables is provided in [33]. [33] also outlines techniques applied to handle such issues in causal graphs, such as the back door criterion.

Causal diagrams provide the means to visually inspect cause and effects to: i) handle any conceptual inconsistencies; and ii) illuminate likely effects of proposed interventions [34]. The *do* calculus, after [34] is applied to determine causal inference of the form:

$$P(y|do(x), z) \tag{1}$$

which describes the conditional probability of y if an intervention on x, given knowledge of z, where z represents existing knowledge of the causal influence of the triple factors: granularity; perspective; and context.

By explicitly incorporating granularity, perspective, and context as top level ontological knowledge structuring constructs in a plate model representing GCP causal BN subsystems, the Markov conditioning assumption [56] can be utilised to extend the range of queries answerable by the model to address wicked problems.

The Markov conditioning assumption links graph and probability functions where each variable is probabilistically independent of its non-descendants, conditional on its parents, and is defined more formally after [56] as:

$$P(x_j|pa_j) = P(x_1, ...x_j - 1) \tag{2}$$

Conditional probability x_j is sensitive only to a small subset of predecessors PA_j., the ontological meta-properties, the argument signature U_i of the plate model representing a GCP sub-system, which considerably simplifies and reduces input information required [34]. Thus the application of the GCP causal BN model leads to reduction in computational complexity, given that Bayesian inference is NP Hard. NP hardness refers to computational complexity that increases exponentially with the size of the network making a problem unsolvable in non-deterministic polynomial time [59].

The possible effects of an intervention on inter-linked variables, given pre-existing knowledge of granular contextual perspectives encoded as conditional probability distributions, can be computationally simulated. Using the Covid19 pandemic example, this facilitates decision making on important questions such as, which granular hierarchical level to target with an intervention for a given Covid19 context (e.g., health, economic, and human rights contexts), to obtain the most desirable outcomes, with the least negative effects.

Queries of the following kind are enabled: "What is the likely effect of a statutory requirement for everyone to stay at home for 3 weeks on hospital *x* , located at location *l*, and the likely effect on employee *y*, an airline pilot who works for airline *z*, given the different perspectives and belief systems of the affected stakeholders?"

The GCP causal BN model also supports counterfactual queries by applying the following calculus, after [33]

$$P(y_x|x', y') \tag{3}$$

which describes a hypothetical situation that says, "was it *x* that caused *y*, and, what if I acted differently". The calculus extends the query space to address complex scenario planning queries of the kind:

"Imagine today is date *d*, 30 September 2022, *(representing temporal granularity in the GCP causal BN model)*, the pandemic has been brought under control through wide spread vaccination, and people are free to travel to *l* locations *(representing sub-context in our model)*. Airline company *x* finds out that its most lucrative market segment *b* of business travellers *(representing a granular sub-context in the model)* has adapted to doing business online *o*, and business people do not travel as much as during *bc*, the pre-COVID19 era, *(representing a sub-context, and temporal granularity in the model)*.

What could company *x* have done to prepare for such an eventuality *e a*ffected by multiple perspectives *ps* in the model, given evolving knowledge *k* of changes in business cultural practices, such knowledge being fragmented and subject to divergent beliefs *db* of company executives of how the future could pan out?".

The foregoing inferential and counterfactual queries are essential for the exploration of possible effects of proposed alternative interventions or prototypical design strategies for resolving a given wicked problem, and it is argued in this paper, that such queries cannot be handled efficiently and effectively through existing baseline Bayesian models as they do not have the systemic meta-properties as modeling constructs.

VIII. Discussion

Wicked problems are emergent features of CAS and while modeling of emergence in CAS is provided for in some research work such as [60] the specific modeling of the epistemological elements of CAS and wicked problems from an observers view point, level II of wicked problems in this paper, is not represented in such models. By incorporating granularity, context, and perspective as constructs delineating CAS subsystems the GCP causal BN model, using HHM conditioning, the model provides the mechanism to compactly represent both levels, I and II of wicked problems to support tractable inference computation. While some baseline Bayesian models do utilise HHMs to compactly model hierarchy and systems nestedness as elaborated in [12] such Bayesian models are not able to simulate the effects of interventions across granular contextual perspective sub-systems through Bayesian belief propagation.

The use of a common formal ontological modeling language promotes semantic inter-operation between collaborating stakeholders and helps eliminate bounded rationality between divergent perspectives about a given wicked problem. The model can thus be used to support multi-stakeholder group learning and understanding of the ontological and stochastic character of a given wicked problem, paving the way for improved collaborative exploration, and solution design.

The GCP casual BN model further refines machine learning capabilities with respect to learning the structure of wicked problems. The argument is that the incorporating of granular contextual perspectives as ontological knowledge constructs in the model provides a causal model that supports an estimand that refines statistical classification of data to more precisely recover the structure, and learn about parameters of a given wicked problem than baseline Bayesian models. The need for such high precision estimands is well articulated in [33]. The constructs provide the means to model data generating structures and processes missing from models without the representational complex systems constructs.

IX. Conclusion and the way forward

The GCP causal BN model outlined in this paper provides a novel high precision ontological representation of wicked problems in dynamic, hierarchical, Bayesian networks, problems that emerge from complex adaptive systems, and stakeholder context, perspectives and granular level of observation.

The incorporation of granularity, context, and perspective of kernel complexity theory notions as Hidden Markov Model modeling constructs in the GCP causal BN model architecture extends the scope of queries answerable using baseline Bayesian models to specifically handle inference and counterfactual queries that support the understanding and resolution of wicked problems. The constructs further serve to refine machine learning of a given wicked problem structure and its parameters. To the best of our knowledge no such model has yet been developed to enhance understanding of wicked problems and their resolution through Bayesian reasoning and causal graphical logic.

By providing a modeling architecture that links sub-systems as multi-domain ontology / perspectives of bounded linked sub-systems the proposed GCP causal BN model architecture reduces computational complexity. The model also simplifies representation of the complex relational interactions for a given wicked problem for visual inspection and discussion by stakeholders with different perspectives and beliefs about the status of the wicked problem, and to arrive at informed consensus based solutions.

While a single prototypical example was used to illustrate the utility and novelty of the model, the model is equally applicable in resolving other wicked problems exhibiting similar challenges of dynamic complexity, finitude and normativity, which include market crashes, urban development, and negative climate change effects.

The intention, as a way forward, is to carry out empirical experiments to benchmark the GCP causal BN model against baseline Bayesian models to further validate the veracity of the model for Bayesian inference reasoning and machine learning of wicked problems. This is to be carried out using synthetic ground truth data, applying the Kullback–Leibler KL divergence method as applied in [13] for structure and parameter learning in dynamic environments.

Acknowledgment

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835).

References

[1] R. Costanza, L. Wainger, C. Folke, and K.-G. Mäler, "Modeling complex ecological economic systems: toward an evolutionary, dynamic understanding of people and nature," *BioScience*, vol. 43, no. 8, pp. 545–555, 1993.

[2] V. A. Brown, J. A. Harris, and J. Y. Russell, *Tackling wicked problems through the transdisciplinary imagination*. Earthscan, 2010.

[3] K. Crowley and B. W. Head, "The enduring challenge of 'wicked problems': revisiting rittel and webber," *Policy Sciences*, vol. 50, no. 4, pp. 539–547, 2017.

[4] H. A. Simon, *The sciences of the artificial*. MIT press, 2019.

[5] C. Burman, M. Aphane, and N. Mollel, "The taming wicked problems framework: Reflections in the making," *Journal for New Generation Sciences*, vol. 15, no. 1, pp. 51–73, 2017.

[6] K. Levin, B. Cashore, S. Bernstein, and G. Auld, "Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change," *Policy sciences*, vol. 45, no. 2, pp. 123–152, 2012.

[7] O. Mascarenhas, "Innovation as defining and resolving wicked problems," *Unpublished manuscript, ENT*, vol. 470, p. 570, 2009.

[8] M. King and J. Kay, *Radical Uncertainty: Decision-making for an unknowable future*. Hachette UK, 2020.

[9] S. Galea, M. Riddle, and G. A. Kaplan, "Causal thinking and complex system approaches in epidemiology," *International journal of epidemiology*, vol. 39, no. 1, pp. 97–106, 2010.

[10] N. Guarino and C. Welty, "Ontological analysis of taxonomic relationships," in *International Conference on Conceptual Modeling*. Springer, 2000, pp. 210–224.

[11] G. Guizzardi, G. Wagner, J. P. A. Almeida, and R. S. Guizzardi, "Towards ontological foundations for conceptual modeling: The unified foundational ontology (ufo) story," *Applied ontology*, vol. 10, no. 3-4, pp. 259–271, 2015.

[12] E. Gyftodimos and P. A. Flach, "Hierarchical bayesian networks: A probabilistic reasoning model for structured domains," in *Proceedings of the ICML-2002 Workshop on Development of Representations*. Citeseer, 2002, pp. 23–30.

[13] R. Ajoodha and B. Rosman, "Learning the influence structure between partially observed stochastic processes using iot sensor data," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] N. Benjamin-Fink and B. K. Reilly, "A road map for developing and applying object-oriented bayesian networks to "wicked" problems," *Ecological Modelling*, vol. 360, pp. 27–44, 2017.

[15] H. Al Harbi *et al.*, "Semantically aware hierarchical bayesian network model for knowledge discovery in data: an ontology-based framework," Ph.D. dissertation, University of Salford, 2017.

[16] S. Fenz, A. M. Tjoa, and M. Hudec, "Ontology-based generation of bayesian networks," in *2009 International Conference on Complex, Intelligent and Software Intensive Systems*. IEEE, 2009, pp. 712–717.

[17] B. G. Marcot and T. D. Penman, "Advances in bayesian network modelling: Integration of modelling technologies," *Environmental modelling & software*, vol. 111, pp. 386–393, 2019.

[18] W. C. Wimsatt, "The ontology of complex systems: levels of organization, perspectives, and causal thickets," *Canadian Journal of Philosophy*, vol. 24, no. sup1, pp. 207–274, 1994.

[19] J. Gharajedaghi, *Systems thinking: Managing chaos and complexity: A platform for designing business architecture*. Elsevier, 2011.

[20] P. M. Senge, "The fifth discipline: The art and practice of the learning organization (rev. ed.)," *New York, NY: Currency Doubleday*, 2006.

[21] D. Rousseau, J. Billingham, and J. Calvo-Amodio, "Systemic semantics: A systems approach to building ontologies and concept maps," *Systems*, vol. 6, no. 3, p. 32, 2018.

[22] A. Bousquet and S. Curtis, "Beyond models and metaphors: complexity theory, systems thinking and international relations," *Cambridge review of international affairs*, vol. 24, no. 01, pp. 43–62, 2011.

[23] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data & knowledge engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[24] N. Guarino, *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*. IOS press, 1998, vol. 46.

[25] E. P. B. Simperl and C. Tempich, "Ontology engineering: a reality check," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2006, pp. 836–854.

[26] J. Gignoux, G. Chérel, I. D. Davies, S. R. Flint, and E. Lateltin, "Emergence and complex systems: The contribution of dynamic graph theory," *Ecological Complexity*, vol. 31, pp. 34–49, 2017.

[27] F.-P. Pai, L.-J. Yang, and Y.-C. Chung, "Multi-layer ontology based information fusion for situation awareness," *Applied Intelligence*, vol. 46, no. 2, pp. 285–307, 2017.

[28] A. A. I. Abuazab, H. B. Selamat, R. B. C. M. Yusoff, and A. N. Abdalla, "Ontology-driven bayesian network model for semantic expression," *DEStech Transactions on Computer Science and Engineering*, no. cmee, 2017.

[29] D. Koller and A. Pfeffer, "Representations and solutions for game-theoretic problems," *Artificial intelligence*, vol. 94, no. 1-2, pp. 167–215, 1997.

[30] Q. Liu, F. Pérès, and A. Tchangani, "Object oriented bayesian network for complex system risk assessment," *IFAC-PapersOnLine*, vol. 49, no. 28, pp. 31–36, 2016.

[31] P. C. Costa, K. B. Laskey, and G. AlGhamdi, "Bayesian ontologies in ai systems," 2006.

[32] T. Love and R. Ajoodha, "Building undirected influence ontologies using pairwise similarity functions," in *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 2020, pp. 1–6.

[33] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[34] J. Pearl, *Causality*. Cambridge university press, 2009.

[35] K. Munn and B. Smith, *Applied ontology: An introduction*. Walter de Gruyter, 2013, vol. 9.

[36] B. Smith and D. M. Mark, "Ontology and geographic kinds," 1998.

[37] C. Churchman, "West (1967). wicked problems," *Management Science*, vol. 14, no. 4, pp. 141–2, 1967.

[38] H. W. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," *Policy sciences*, vol. 4, no. 2, pp. 155–169, 1973.

[39] H. Simon, "The architecture of complexity. reprinted in his the sciences of the artificial," 1996.

[40] W.-N. Xiang, "Working with wicked problems in socio-ecological systems: Awareness, acceptance, and adaptation," *Landscape and Urban Planning*, no. 110, pp. 1–4, 2013.

[41] G. Elia and A. Margherita, "Can we solve wicked problems? a conceptual framework and a collective intelligence system to support problem analysis and solution design for complex social issues," *Technological Forecasting and Social Change*, vol. 133, pp. 279–286, 2018.

[42] B. G. Peters, "What is so wicked about wicked problems? a conceptual analysis and a research program," *Policy and Society*, vol. 36, no. 3, pp. 385–396, 2017.

[43] M. Zellner and S. D. Campbell, "Planning for deep-rooted problems: What can we learn from aligning complex systems and wicked problems?" *Planning Theory & Practice*, vol. 16, no. 4, pp. 457–478, 2015.

[44] J. H. Holland, *Hidden order: How adaptation builds complexity*. Addison Wesley Longman Publishing Co., Inc., 1996.

[45] N. A. Baas and C. Emmeche, "On emergence and explanation," *Intellectica*, vol. 25, no. 2, pp. 67–83, 1997.

[46] R. Farrell and C. Hooker, "Design, science and wicked problems," *Design studies*, vol. 34, no. 6, pp. 681–705, 2013.

[47] D. H. Kim and C. Lannon, *Applying systems archetypes*. Pegasus Communications Waltham, 1997.

[48] W. Braun, "The system archetypes-the systems modeling workbook," *available at: wwwu. uniklu. ac. at/gossimit/pap/sd/wb_sysarch. pdf*, 2002.

[49] G. C. Santos, "Ontological emergence: How is that possible? towards a new relational ontology," *Foundations of Science*, vol. 20, no. 4, pp. 429–446, 2015.

[50] M. Calabrese, A. Amato, V. Di Lecce, and V. Piuri, "Hierarchical-granularity holonic modelling," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, no. 3, pp. 199–209, 2010.

[51] F. T. Fonseca and M. J. Egenhofer, "Ontology-driven geographic information systems," in *Proceedings of the 7th ACM international symposium on Advances in geographic information systems*, 1999, pp. 14–19.

[52] A. Jensen and T. Aven, "A new definition of complexity in a risk analysis setting," *Reliability Engineering & System Safety*, vol. 171, pp. 169–173, 2018.

[53] R. Hummelbrunner, "Systems thinking and evaluation," *Evaluation*, vol. 17, no. 4, pp. 395–403, 2011.

[54] V. Robert, G. Yoguel, and O. Lerena, "The ontology of complexity and the neo-schumpeterian evolutionary theory of economic change," *Journal of Evolutionary Economics*, vol. 27, no. 4, pp. 761–793, 2017.

[55] J. H. Miller and S. E. Page, *Complex adaptive systems: An introduction to computational models of social life*. Princeton university press, 2009.

[56] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[57] R. Ajoodha and B. Rosman, "Tracking influence between naïve bayes models using score-based structure learning," in *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*. IEEE, 2017, pp. 122–127.

[58] S. Berman and T. D. Semwayo, "A conceptual modeling methodology based on niches and granularity," in *International Conference on Conceptual Modeling*. Springer, 2007, pp. 338–358.

[59] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the third annual ACM symposium on Theory of computing*, 1971, pp. 151–158.

[60] A. Poitreger and J. Bishop, "The engineering of emergence in complex adaptive systems," 2002.

Detection of Oil Spill Pollution in Seawater Using Drones: Simulation & Lab-based Experimental Study

Ashraf Saleem
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
asaleem@squ.edu.om

Ahmed Al Maashri
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
amaashari@squ.edu.om

Omer Eldirdiry
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
o.eldirdiry@squ.edu.om

Jawhar Ghommam
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
jawher@squ.edu.om

Hadj Bourdoucen
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
hadj@squ.edu.om

Amran Al-Kamzari
Pollution Operations Monitoring Centre
Environment Authority
Muscat, Oman
amran.alkamzari@meca.gov.om

Ghazi Al Rawas
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
ghazi@squ.edu.om

Ahmed Ammari
Electrical and Computer Engineering Dept.
Sultan Qaboos University
Muscat, Oman
chiheb@squ.edu.om

***Abstract*— This paper presents a simulation and lab-based experimental study of detecting oil spills in the oceans using drones. The proposed simulation process is the first phase of long-term research that is targeting to limit the oil spill contaminations in Omani coastline. This study presents some methods and tools to detect an oil spill in a prearranged scenario. The methods and tools are tested in simulation and verified in lab-based experiments. Footage of oil spill cases is used to illustrate the process of oil spill detection. In addition, image processing techniques are applied to provide an accurate outline of the spill's perimeter. The results exhibit the effectiveness of using the proposed methods and tools in detecting oil spills.**

Keywords—simulation, lab-based experiment, oil spill, contamination, detection, image processing

I. Introduction

The last few decades have witnessed major oil spill incidents around the globe. It is a challenge to overcome these ecological problems and the pollutions caused by the accidental leaks of oil [1-3]. It has been recorded that around 750,000 m^3 of crude oil were spilled into the Gulf of Mexico in 2010 [4]. Such a leak in the oilfield can cause pollution on the coastline thousands of miles away. The most obvious sign of the leak is a film of oil spreading on the surface of the ocean, which can easily be noticed. Then, the nearby beaches and shorelines are polluted by oil stains. However, the most critical impact of the oil spill occurs long after the initial spill. Oil consists of at least 300 types of chemicals, most of which can negatively harm the living marine organisms. Once these hazardous chemicals are introduced to the food chain, they can cause serious health issues to sea life and humans alike.

Very few leaks are naturally caused by the infiltration of the oil deposits underneath the ocean floor. However, most of the oil leaks are caused by human errors due to a damaged tanker or a leaking oil rig. An example of the largest oil spills, caused by humans, is the leak near Alaska in 1989 when the oil tanker Exxon Valdes accidentally released millions of liters into the sea. Years of efforts have been spent to clean up the polluted area.

Containing oil spills and cleaning the seawater and shores are enormous tasks, and therefore several studies have investigated effective oil spill remediation techniques [5-8]. These techniques aim to either remove or separate the oil from the seawater. Such techniques include:

- Chemical cleanup by using chemical dispersants and emulsifiers [9, 10],
- Mechanical tools that showed high efficiency in recovering the oil such as booms, sorbents and skimmers [11- 14], and
- Bioremediation by using oil-eating bacteria, which can break down the oil into sub composites that can be consumed by marine life [15].

However, several challenges face the use of these techniques, which can be applied only under certain conditions. For example, adding chemical dispersants into the seawater can be harmful [11]. On the other hand, cleaning the oil spills using mechanical tools is expensive and requires a large number of tools and equipment [16]. For instance, Exxon has spent over 2 million dollars on the Valdez cleanup operations [17].

The cleaning process of the oil spill becomes more challenging when the oil spreads widely in the ocean. Therefore, it is important to detect oil spills immediately. Various studies have been focused on achieving rapid and accurate detection of oil spills in the sea and the ocean. During the last 15 years, the operational use of satellites in detecting oil spills has bloomed, especially when satellite images became available for researchers. Alternatively, Synthetic-Aperture Radar (SAR) imagery is used by many organizations (e.g. European Maritime Safety Agency) to monitor the oil spills in the ocean. SAR imagery has the capability of providing usable images under all weather conditions day and night [18-21]. On

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

the other hand, various studies have been performed to examine the capabilities of thermal infrared cameras in detecting oil spills in water [22].

Robotics, artificial intelligence, and computer vision are fields that can be integrated to provide solutions that would automate many processes. One of which is the detection and reporting of oil leakage. In recent years, remote sensing technologies were widely applied to investigate and monitor various environmental issues related to oil spills and red tides. Common techniques use cameras in the visible spectra. However, using standard cameras to inspect particular spectral regions in the visible spectrum cannot easily discriminate such ecological issues in water environments [23]. Many studies proposed modifications to standard cameras to unlock the major obstacles that interfere with the use of visible light [24-26]. Unmanned Aerial Vehicle (UAV) equipped with remote sensing is a very promising approach for oil spills and red tide investigations. It enables for higher accuracy, lower costs, and revisit times determined by the operator as opposed to fixed satellite revisit times [27]. All of these remote sensing methods are regularly applied. For example, a pollution control aircraft is operated by the German Navy is mainly used for oil spill detection [28]. UAVs equipped with multispectral or hyperspectral imagers are used in many studies to accurately map multiple types of water pollution [29].

II. Problem Statement

Oil spills in the oceans pose a threat to coastal countries such as the Sultanate of Oman. The Sultanate is blessed with a unique geographical location, with a long shoreline filled with diversified marine life forms. In particular, the peninsula of Musandam overlooks the Strait of Hormuz; a strategic location for maritime trade and a passage to 35% of the world's seaborne oil shipments. Also, the Arabian Gulf is a busy route for cargo ships and marine trading. Therefore, incidents of oil spills leaking from ships and tankers could happen. The government’s mandate is to watch for any spillage around the clock to protect the environment. In the occurrence of spillage, current, winds, and waves carry the spilled oil toward the coast. Therefore, before reaching the coast, the oil is subject to physical and chemical transformations, primarily due to the prevailing weather conditions [30]. As it reaches the shoreline, the oil is most likely to penetrate deeply in the sandy beaches or to be retained by rocky coastal tracts. Given the dangerous chemical composition of the crude oil (a mixture of thousands of hydrocarbons [31]), its primary environmental impact is on the atmosphere, on the surface of the sea, in the water column, onshore, as well as on the humans and marine life [32, 33]. Fig. 1 shows some examples of such leakage in the Musandam shoreline that would threaten the environment and marine life.

The current approach for detecting oil spills is both costly and time-consuming as it involves coordination between several parties. Drones’ applications have been spread widely in many fields [34] due to the drones’ flexibility. Therefore, as an alternative solution, we propose a UAV-based solution capable of detecting oil spills using remote sensing and machine vision techniques. The proposed alternative employs drones that are performing real-time, low-end imaging analysis on the target areas. Additionally, the system is capable of accurately locate the spill and measure weather parameters such as wind speed and direction. The latter is needed to predict the direction and speed at which the spill will extend.

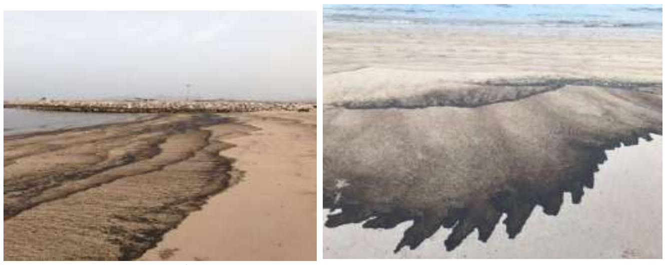


Fig. 1. Examples of oil leakage in Musandam shoreline threatening the environment and marine life [Courtesy: Environment Authority, Sultanate of Oman]

The proposed solution is a comprehensive one that entails many subsystems. As such the research project is conducted over several phases. This paper discusses the work done in the first phase, where both simulation and emulation environments were developed to experiment with the proposed solution. This provides the authors with a lab-based experimental study to test several remote sensing and machine vision techniques for oil detection before conducting field tests in the Gulf of Oman. The authors believe that these lab-based experiments should cut the development cost and time. Furthermore, the experiments will facilitate diversifying the circumstances in which the oil spill has occurred.

III. Development of Simulation Environment

Robot Operating System (ROS) and Gazebo are used to simulate UAV flight workspace, with the dimensions of 5×4×2.8 meters. These dimensions are based on the EIVS laboratory [35] as shown in Fig. 2 (a).

The simulation environment models a UAV that hovers over the seawater. The water includes an oil spill as shown in Fig. 2 (b). This simulation environment mimics the incident at the Gulf of Mexico, which took place in April 2010 [4].

During the simulation run, the modeled UAV flies over the oil spill a shown in Fig. 2 (b). The UAV surveys the region while capturing images simultaneously. The route taken by the UAV during the survey is determined by several factors such as mission time. It is worth mentioning that the UAV plans its missions based on two different modes. The first mode is when the oil spill location is known. In this mode, the task will be to survey the area and give exact locations for the oil spill area without the need to search for the ship that caused this spill. In contrast, the second mode entails patrolling the shores looking for potential oil spills. if a spill is detected, the UAV will undergo the search mode, looking for nearby ships that might have caused the spill.

A connection has been made between the simulated drone and ArduPilot. Using this software it is possible to locate the simulated drone at any point in Google map and give the drones point-locations or routes to follow. As an example, the starting location of the UAV is at Ras Al-Hadd in Oman, where the UAV was given a specific area to survey. This tool can be useful to simulate the task, which will be given to the actual UAVs; providing an estimate of the time and the battery life required to accomplish the task.

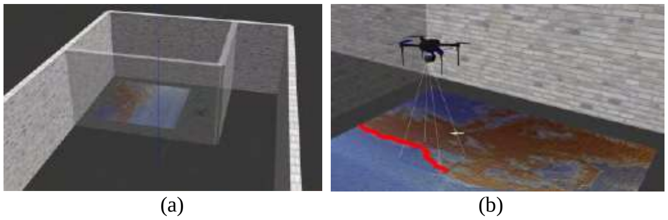

(a) (b)

Fig. 2. (a) Gazebo simulation of the lab room, and (b) a drone detecting the area of the oil spill in simulation

As mentioned earlier, the UAV is equipped with a thermal camera. The captured thermal images provide a means to segregate water from oil given the disparity in the measured temperatures. The thermal images were processed using several image processing techniques to automatically determine the outline of an oil spill by analyzing the temperature profile of the images.

IV. Emulation Environment

An indoor UAV emulation environment is used to experiment with the proposed oil spill detection system. Experimenting with the system in an indoor and controlled environment leads to a significant reduction in cost and time. Additionally, the emulation environment offers us the ability to test the system with several cases of oil spills under various conditions. It should be noted that the occurrence of an oil spill in oceans infrequent, which makes it very challenging to test the system. Therefore, the indoor emulation environment allows the team to experiment with different techniques and strategies to detect oil spills and be prepared for any unexpected scenarios. Fig. 3 shows the emulation environment that was constructed at the EIVS lab. The figure also shows the equipment used in the experiments.

The floor of the area is furnished with a printed poster that shows seawater contaminated with an oil spill (see Fig. 4). The poster size is 3×4 m. Indoor drones from Quanser are used in these experiments. These drones are equipped with downward-facing cameras, which capture images in grayscale color space. This camera is selected to resemble the output of thermal images similar to the ones captured by the FLIR Zenmume X2 camera that are mounted on Matrice 200 V2 (to be used in the 2nd phase of the proposed system development).

Given that indoor drones are unable to receive GPS signals, it is not possible to rely on GPS service for positing and localization. Instead, the emulation environment uses the OptiTrack camera system to localize the drones.

The indoor drones are equipped with fixed focal-length cameras. Therefore, the drones are maintained at a fixed altitude. Capturing images at the predetermined altitude mimics that the same images that would have been captured in actual field tests.

Then, a suitable flying trajectory in the x- and y- plane (the horizontal plane) is set in such a way that the field of view (FoV) of the cameras is adequate with the arena. This is depicted in Fig. 4. For this reason, the desired x- and y-trajectory in Fig. 5 was designed so that the drone flies in the boundary of ±0.8 m in the y-axis, while the x-axis location is incrementing by 0.45 m every time the drone makes a turn.

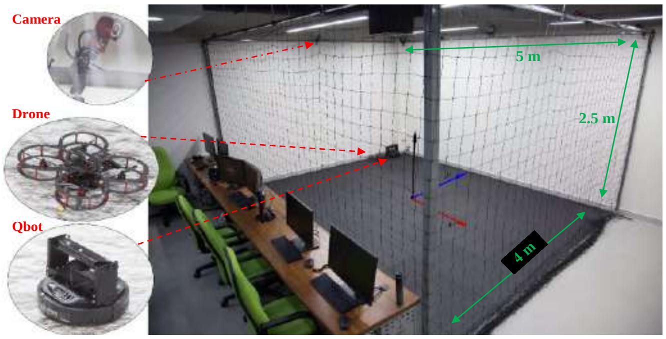


Fig. 3. The emulation environment constructed at the EIVS lab

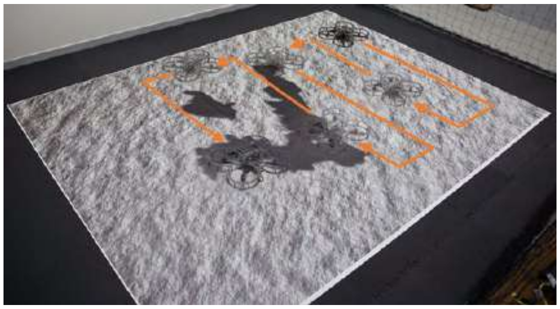


Fig. 4. The drone is flying above an image of an oil spill case in the ocean

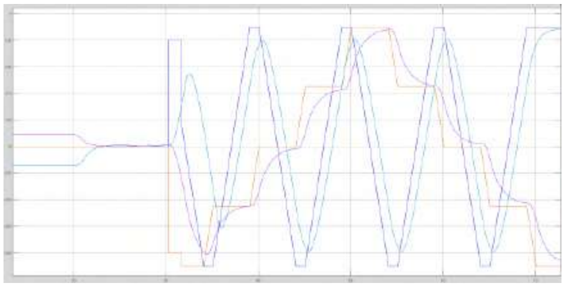


Fig. 5. Monitoring of the drone's location during the experiment

V. Image Processing Techniques

To perform the image processing method, the following steps are followed:

A. Image acquisition: During the flight of the drone with the designed trajectory, the drone was capturing an image every 0.3 seconds. This small period is enough for the drone to capture all 49 mini-images that represent the whole workspace area, as exhibited in Fig. 6.

B. Image stitching: The images that were captured in the previous step are stitched to one another. This yields an image that encompasses that the whole poster. This process was automated using a MATLAB script that was developed for this purpose. An example of stitching three images is shown in Fig. 7.

It is worth mentioning that in actual field tests, the image stitching process is not required. Image stitching was performed in this experiment to provide coordinates for all points in the stitched image since the location of the starting point is known by the OptiTracking cameras (the coordinates of the first mini-image taken by the drone).

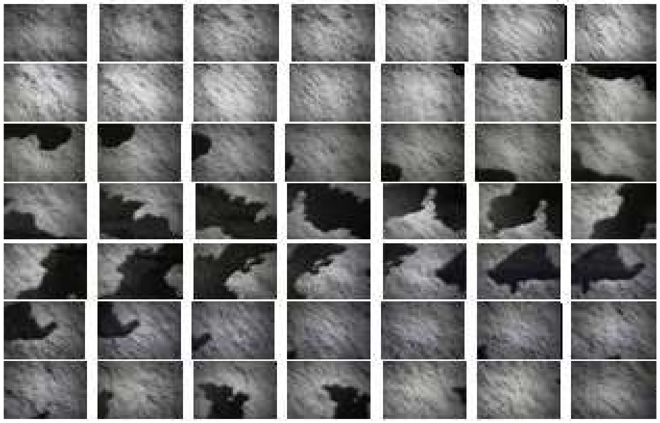


Fig. 6. 49 images captured by the drone while flying above the workspace area

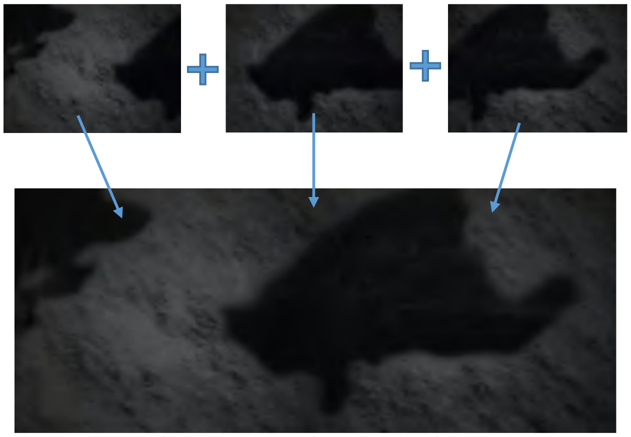


Fig. 7. An example of image stitching for a portion of the study case image

C. Image Analysis: the next step is to identify the region of the oil spill. Image processing techniques are used to detect the outline of a spill (contour). The input images are in grayscale color space. The images are first converted to HSV color space, as it gives better segregation between the seawater and oil spill regions. Fig. 8 shows the oil spill as represented in the HSV color space. It is noted that the spill is not distinguishable from the background when in the Hue and Saturation channels. This is because they both describe the color (for Hue) and the intensity of the color (for Saturation). On the other hand, the oil spill is much more visible in the Value channel. The Value channel represents the brightness of the area. Some adjustments in the HSV threshold for three channels can affect the detection of the oil spill. This overall process is used to identify the location of the oil spill area with respect to the room reference frame of the room.

Based on the outcome of the previous steps, the system outline the region where the oil spill is located (see Fig. 9). The spotted difference between oil and water in the V channel of the HSV color space was noticeable, where the black color was indicating the water and the white color was indicating the oil. However, in the current stage of this research the accurate calculation of the region was not implemented. Therefore, the detected region was presented as a rectangular shape that is bounded by the maximum and the minimum distance in x- and y- directions. Then, the system reports location of the contour around the oil spill.

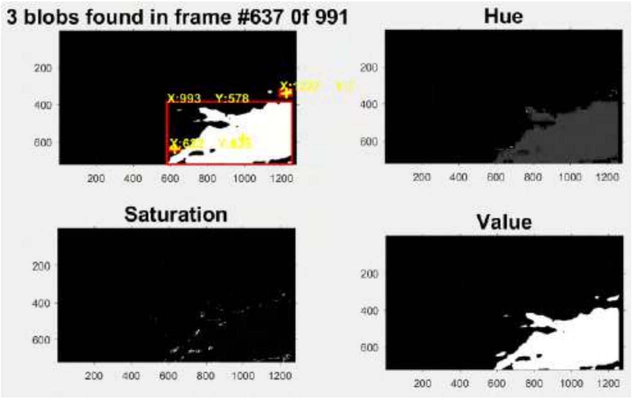


Fig. 8. Captured images are represented in HSV color space. Notice how the oil spill is more visible in the V channel.

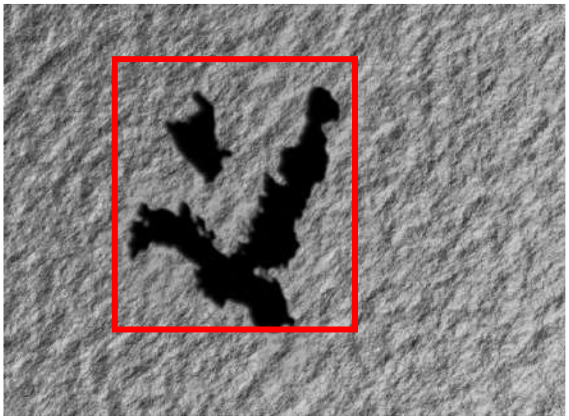


Fig. 9. The detected area of the total oil spill and contour location is reported

VI. Conclusion

This paper described a simulation and lab-based experimental study of a proposed UAV-based system for detecting oil spills in the oceans. Oil spill detection has always been an issue since it is infrequent to detect oil spills in real life and monitor the wide ocean all the time. Therefore, this study is conducted, using a cost-effective solution than the existing ones, to provide a training environment to deal with oil spill cases. The drone-based system can provide the required information for the detection process. This research has gone through two stages to implement the oil detection process. The first stage of the project was to simulate the environment, for which oil detection can be made possible for study and analysis. The second stage was a laboratory-based, emulation test to demonstrate the effectiveness of the proposed algorithms. Since the proposed system was successfully tested in simulation and emulation, the real drone will be used to implement the proposed solution and will be tested on real oil spill scenarios.

In the future, the proposed system presented in this paper as simulated and lab-based procedures will be carried out for real-life cases, where drones that are more sophisticated will be used. Moreover, the contour around the detected oil will be improved to detect exactly the area of the oil spill.

Acknowledgment

This work was funded in parts by research grants from BP Oman (EG/ENG/ECED/18/01), Smart City Platform (TRC/SRG/SC/IMCO/01/2019), His Majesty Trust Fund

(SR/ENG/ECED/19/01), and the Ministry of Higher Education, Research and Innovation's Block Fund (RC/RG-ENG/ECED/18/02). The authors would like to thanks Sultan Qaboos University for providing a conducive research environment. Additionally, the authors thank Mr. Abdul Hakim Alharthy for his support to the research efforts. Special thanks to Environment Authority for providing data and information related to oil leak incidents.

References

[1] C.H. Peterson, S.D. Rice, J.W. Short, D. Esler, J.L. Bodkin, B.E. Ballachey, and D.B. Irons, "Long-term ecosystem response to the Exxon Valdez oil spill," *Science,* vol. 80, no. 302, pp.2082–2086, 2003. https://doi.org/10.1126/science.1084282.

[2] S. Naz, M. F. Iqbal, I. Mahmood, & M. Allam, "Marine oil spill detection using Synthetic Aperture Radar over Indian Ocean," *Marine Pollution Bulletin*, vol. 162, pp. 111921, 2021.

[3] D. C. Zacharias, C. M. Gama, & A. Fornaro, "Mysterious oil spill on Brazilian coast: Analysis and estimates," *Marine Pollution Bulletin*, vol. 165, pp. 112125, 2021.

[4] S.E. Allan, B.W. Smith, and K.A. Anderson, "Impact of the deepwater horizon oil spill on bioavailable polycyclic aromatic hydrocarbons in Gulf of Mexico Coastal waters," *Environ. Sci. Technol.*, vol. 46, pp. 2033–2039, 2012. https://doi.org/10.1021/es202942q.

[5] H. Bidgoli, A. A. Khodadadi, and Y. Mortazavi, "A hydrophobic/oleophilic chitosan-based sorbent: toward an effective oil spill remediation technology," *Journal of Environmental Chemical Engineering*, pp. 103340, 2019.

[6] H. Bidgoli, Y. Mortazavi, and A. A. Khodadadi, "A functionalized nano-structured cellulosic sorbent aerogel for oil spill cleanup: Synthesis and characterization," *Journal of hazardous materials*, vol. 366, pp. 229-239, 2019.

[7] L. Huang, X. Peng, F. Zhang, Y. Wen, C. Xiao, & Y. Zhang, "A cloud-based platform for marine oil spill emergency response capability visualization," *IEEE 2019 5th International Conference on Transportation Information and Safety (ICTIS)*, pp. 810-815, July, 2019.

[8] J. Yang, J. Wan, Y. Ma, & Y. Hu, "Research on object-oriented decision fusion for oil spill detection on sea surface," *In IEEE IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 9772-9775, July, 2019.

[9] E.B. Kujawinski, M.C. Kido Soule, D.L. Valentine, A.K. Boysen, K. Longnecker, and M.C. Redmond, "Fate of dispersants associated with the deepwater horizon oil spill," *Environ. Sci. Technol.*, vol. 45, pp. 1298–1306, 2011. https://doi.org/10.1021/es103838p.

[10] H. Chapman, K. Purnell, R.J. Law, and M.F. Kirby, "The use of chemical dispersants to combat oil spills at sea: a review of practice and research needs in Europe," *Mar. Pollut. Bull.*, vol. 54, pp. 827–838, 2007. https://doi.org/10.1016/j.marpolbul.2007.03. 012.

[11] V. Broje, and A.A. Keller, "Improved mechanical oil spill recovery using an optimized geometry for the skimmer surface," *Environ. Sci. Technol.*, vol. 40, pp. 7914–7918, 2006. https://doi.org/10.1021/es061842m.

[12] D. Nowak, & M. Wąż, "Analysis Of Radar Early Warning Systems For Oil Spills In The Offshore Sector," *IEEE 2019 European Navigation Conference (ENC)*, pp. 1-4, April, 2019.

[13] J. Jiménez, & J. M. Girón-Sierra, "Modelling the automatic deployment of oil-spill booms: a simulation scenario for sea cleaning," *In IEEE 2018 Winter Simulation Conference (WSC)*, pp. 1192-1203, December, 2018.

[14] D.P. Prendergast, and P.M. Gschwend, "Assessing the performance and cost of oil spill remediation technologies," *J. Clean Prod.*, vol. 78, pp. 233–242, 2014. https://doi.org/10. 1016/j.jclepro.2014.04.054.

[15] B. M. Macaulay, and D. Rees, "Bioremediation of oil spills: a review of challenges for research advancement," *Annals of Environmental Science*, vol. 8, pp. 9-37, 2014.

[16] V. Broje, and A.A. Keller, "Effect of operational parameters on the recovery rate of an oleophilic drum skimmer," *J. Hazard. Mater.*, vol. 148, pp. 136–143, 2007. https://doi.org/ 10.1016/j.jhazmat.2007.02.017.

[17] S. Skinner, and W. K. Reilly, "The EXXON VALDEZ oil spill," *US Environmental Protection Agency*, 1989.

[18] F. Ronci, C. Avolio, M. Di Donna, M. Zavagli, V. Piccialli, & M. Costantini, "Oil Spill Detection from SAR Images by Deep Learning," *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2225-2228, 2020.

[19] C. Alexandrov, N. Kolev, Y. Sivkov, A. Hristov, & M. Tsvetkov, "Oil Spills Detection on Sea Surface by using Sentinel–1 SAR Images," *In IEEE 21st International Symposium on Electrical Apparatus & Technologies (SIELA)*, pp. 1-4, June, 2020.

[20] J. Yin, & J. Yang, "SYMMETRIC SCATTERING MODEL BASED FEATURE EXTRACTION FROM GENERAL COMPACT POLARIMETRIC SAR IMAGERY," *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1703-1706, 2020.

[21] F. Yu, W. Sun, J. Li, Y. Zhao, Y. Zhang, and G. Chen, "An improved Otsu method for oil spill detection from SAR images," *Oceanologia*, vol. 59, no.3, pp. 311-317, 2017.

[22] K. Pilžis, and V. Vaišis, "Detection of oil product on the water surface with thermal infrared camer/Naftos aptikimas vandens paviršiuje naudojant infraraudonųjų spindulių kamerą," *Mokslas–Lietuvos ateitis/Science–Future of Lithuania*, vol. 9, no. 4, pp. 357-362, 2017.

[23] I. Leifer, B. Lehr, D. Simecek-Beatty, E. Bradley, R. Clark, P. Dennison, Y. Hu, S. Matheson, C. Jones, and B. Holt, "State of the art satellite and airborne oil spill remote sensing: Application to the BP DeepWater Horizon oil spill," *Remote Sensing Environment*, vol. 124, pp. 185–209, 2012.

[24] D. Wang, F. Gong, D. Pan, Z. Hao, and Q. Zhu, "Introduction to the airborne marine surveillance platform and its application to water quality monitoring in China," Acta Oceanol. Sin., vol. 29, pp. 33–39, 2010.

[25] A.L. Iler, and P.D. Hamilton, "Detecting oil on water using polarimetric imaging," *Proc. SPIE Int. Soc. Opt. Eng.*, pp. 9459, 2015.

[26] M. Fingas, and C. E. Brown, "A Review of Oil Spill Remote Sensing," *Sensors*, 2018. Doi:10.3390/s18010091.

[27] A.M. Lechner, A. Fletcher, K. Johansen, and P. Erskine, "Characterizing upland swamps using object-based classification methods and hyper-spatial resolution imagery derived from an unmanned aerial vehicle," *Proceedings of the XXII ISPRS Congress Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, (Melbourne, Australia, ISPRS), vol. I–4, pp. 101–106, 2012.

[28] K. Gruener, "The Three-frequency microwave radiometer of a 2nd generation airborne surveillance system for remote sensing of maritime oil pollution," *In: Proceedings of IEEE Workshop RF and Microwave Noise*, Ilmenau, pp. 66–69, 1996.

[29] V. V. Klemas, "Coastal and Environmental Remote Sensing from Unmanned Aerial Vehicles: An Overview," Journal of Coastal Research, vol. 31, no. 5, pp. 1260-1267, 2015.

[30] D. Mackay, S. Paterson, K. Trudel, "A Mathematical Model of Oil Spill Behaviour, *Tech. rep., Environmental Impact Control Directorate. Environment*, Canada, 1980.

[31] M. Mahjoubi, S. Cappello, Y. Souissi, A. Jaouani, and A. Cherif, "Microbial Bioremediation of Petroleum Hydrocarbon–Contaminated Marine Environments," *chap. 16, In Recent Insights in Petroleum Science and Engineering*, 2018. DOI: 10.5772/intechopen.72207.

[32] G. Guidi, F. Cumo, F. Gugliermetti, "Best available techniques for oil spill containment and clean-up in the Mediterranean Sea," *WIT Transactions on Ecology and the Environment*, vol. 103, pp. 527-535, 2007.

[33] M.R. Swift, J. Belanger, B. Celikkol, R.R. Steen and D. Michelin, "Observations of Conventional Oil Boom Failure*", Proceedings of the Twenty- Third Arctic and Marine Oilspill Program (AMOP) Technical Seminar*, Environment Canada, Ottawa, pp. 481-492, 2000.

[34] M. Hassanalian, & A. Abdelkefi, "Classifications, applications, and design challenges of drones: A review," *Progress in Aerospace Sciences*, vol. 91, pp. 99-131, 2017.

[35] Embedded & Interconnected Vision Systems Lab, Twitter Account, [online]: https://twitter.com/eivs_oman

Tracking Control of Vertical Tail Damaged Aircraft with dissimilar Actuator Configuration

1st Muhammad Ammar Ashraf
School of Automation and Electrical Engineering
University of Science and Technology Beijing
Beijing, China
ammarashraf@xs.ustb.edu.cn

2nd Salman Ijaz
Department of Electrical and Electronics Engineering
University of Nottingham
Ningbo, China
salman.ijaz@nottingham.edu.cn

3rd Yao Zou
School of Automation and Electrical Engineering
University of Science and Technology Beijing
Beijing, China
zouyao20@126.com

4thHamdoon Ijaz
School of Aeronautics Engineering
Beijing University of Aeronautics and Astronautics
Beijing, China
hamdoonijaz@outlook.com

***Abstract*—This article presents a new control scheme for an aircraft that has suffered from actuator failure due to severe damage to its vertical tail. The damaged degree of the vertical tail is treated as uncertainty. Firstly an output feedback controller is designed for the nominal lateral dynamic model of aircraft. However, when the actuator's fault/failure occurs, the idea of generalized virtual actuators is adopted. In this way, the same baseline controller is used even in the presence of fault/failure. Linear matrix inequality is used to compute the controller and virtual actuator's gains. The simulation results show good tracking performance and fault-tolerant competency.**

***Index Terms*—linear matrix inequality, control reconfiguration, virtual actuator, linear quadratic regulator**

I. Introduction

An escalation in the safety and reliability requirements in civil aviation requires a tremendous amount of toil in aircraft protection and reducing the possibility of a significant failure. In an aircraft, most of the structural damage occurred due to control surface malfunctioning or vertical-tail loss that often causes a significant chore for the pilot to control the craft [1], [2]. The misfortune disaster in 1985 with Boeing 747 is one of the examples of vertical-tail damage, which lead to a plane crash. One more similar site surfaced in which airbus A-300 in 2001 caused the death of 265 passengers. Reconfiguring the control input immediately after the fault appears in the system can avoid such failures. However, in customary hybrid actuation systems (HAS), adequate measures in dealing with this problem are not present. Due to this reason, the trend of more electrically powered aircraft is increasing nowadays. Recently, the dual redundant actuation systems (DRAS) [3], [4] in modern aircraft has replaced the traditional similar actuation system to make the system more reliable and to solve the actuator failure issues due to common cause. Nowadays, commercial aeroplanes A-350, A-380, and A-400 M use 2H/2E type DRAS. The use of DRAS indicates the potential fault-tolerant control (FTC) capability to meet some extreme situations [5]. For example, when the vertical-tail total loss occurs, hydraulic lines pull apart, resulting in the system's total failure of hydraulic actuators. Therefore, to handle this situation, aircraft with DRAS is more effective than Electro-hydrostatic actuators (EHA) to drive significant control surfaces.

The overall effects of vertical tail loss on directional characteristics are comparable to the pitch axis from stabilizer damage. Much work has been performed studying the damaged aircraft modelling and the fault-tolerant strategy. The calculation of the vertical tail loss of Boeing-747 100/200 data is done by [6], [7]. However, the aircraft can only lose a centralized HAS input channel when exposed to vertical tail loss. Damage-induced aerodynamics in these studies does not characterize the nonlinear association and control derivatives. Passive fault-tolerant control (PFTC) for vertical tail loss of an aircraft is studied in [8], but after the degree of damage increases, the PFTC techniques are no longer valid. An active fault-tolerant control (AFTC) [9]–[11] is a more reasonable response for more extreme losses.

In a recent survey, several researchers have proposed solutions to deal with structural damage scenarios, such as model predictive control (MPC) in [12], model reference adaptive control (MRAC) in [8], [13]–[15], sliding mode control (SMC) in [16], [17] and adaptive control [18]. The authors in [19] proposed adaptive control law to compensate for the vertical-tail loss by utilizing differential thrust because all the actuators would lose their efficiency. In contrast, the vertical tail is significantly damaged with hydraulic loss. Because of the error management ability of more electric aircraft (MEA) fortified through DRAS, [20], [21] proposed a reconfiguration control law in which the nominal controller

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

is designed using a linear quadratic regulator (LQR) and MRAC is used to design reconfiguration controller in case of actuator failure.

This paper presents a different control scheme that utilized virtual actuator configuration to deal with actuator failure due to damage in the vertical-tail of aeroplanes driven by DRAS. In a fault-free situation, attaining tracking control is done by an output feedback controller. Fault-tolerant is achieved by placing a virtual-actuator circuit between the nominal controller and faulty plant for fault or failure conditions within specific actuators. The proposed scheme is dominant over the existing strategies in the following ways.

- The presented method is advantageous because both the nominal and the faulty cases use the same baseline controller. Therefore, the designer can choose the nominal control law according to the tracking performance requirement.
- Since the state feedback controller is not practical; therefore, an output feedback controller is designed based on the nominal system. The controller gains are obtained using an iterative LMI procedure that gives more optimal controller parameters than the simple LMI procedure.

The organization of this paper is: the plane model with a vertical-tail is specified in section II. The reconfiguration problem is discussed in section III, containing the overview of a virtual actuator, problem formulation, and reconfiguration using a virtual actuator. The nominal output feedback controller design is proposed in section IV. Section V contains the design of a virtual actuator. The simulation result is given in section VI. Finally, the conclusion in section VII proposed the effectiveness of the proposed approach.

II. MODELING OF DAMAGED AIRCRAFT

A. Configuration of DRAS

Fig. 1 shows the aircraft layout with the DRAS configuration. The complete elevator, inboard aileron, lower rudder, and spoilers are operational using a hydraulic actuator (HA) and electro-hydraulic actuator (EHA) system.

B. Nominal aircraft modelling

Two dissimilar actuators EHA and HA, derive the aircraft's vertical-tail in nominal condition as vertical-tail fault influences the lateral model constraints, taken into account for the paper. The nominal model of aircraft is taken from [13] and is represented as follows.

$$\begin{aligned} \dot{x}_p &= A_p x_p + B_p u_p \\ y_p &= C_p x_p(t) \end{aligned} \tag{1}$$

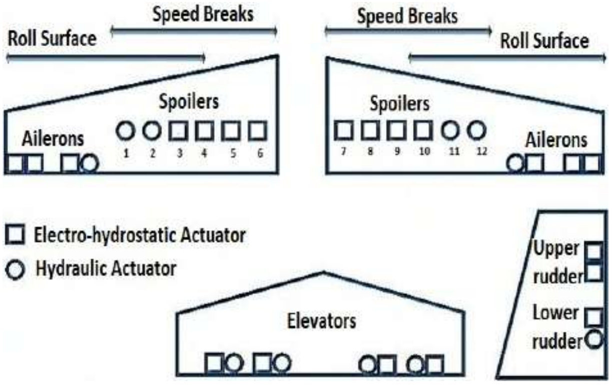


Fig. 1. The aircraft actuation system with DRAS

where the aircraft state vector $x_p = [\beta, p, r, \phi]^T$, and the state variables β denotes the angle of side-slip, r shows yaw angular rate, ϕ and p represents aircraft roll angle and its rate. The state-space matrices of the nominal aircraft model are given as

$$A_p = \begin{bmatrix} \frac{R_\beta}{m} & \frac{R_p}{m} & \frac{R_r}{m} - (u_o) & g_o cos\theta_o \\ \frac{K_\beta}{J_x^{'}} + J_{zx}^{'} Q_\beta & \frac{K_p}{J_x^{'}} + J_{zx}^{'} Q_p & \frac{K_r}{J_x^{'}} + J_{zx}^{'} Q_r & 0 \\ \frac{Q_\beta}{J_z^{'}} + J_{zx}^{'} K_\beta & \frac{Q_p}{J_z^{'}} + J_{zx}^{'} K_p & \frac{Q_r}{J_z^{'}} + J_{zx}^{'} K_r & 0 \\ 0 & 1 & tan\theta_o & 0 \end{bmatrix}$$

$$B_p = \begin{bmatrix} \frac{R_{\delta a}}{m} & \frac{R_{\delta r}}{m} \\ \frac{K_{\delta a}}{J_x^{'}} + J_{zx}^{'} N_{\delta a} & \frac{K_{\delta r}}{J_x^{'}} + J_{zx}^{'} N_{\delta r} \\ \frac{Q_{\delta a}}{J_x^{'}} + J_{zx}^{'} K_{\delta a} & \frac{Q_{\delta r}}{J_x^{'}} + J_{zx}^{'} K_{\delta r} \\ 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

where $J_x^{'} = \frac{J_x J_z - J_{zx}^2}{J_z}$, $J_z^{'} = \frac{J_x J_z - J_{zx}^2}{J_x}$, $J_{zx}^{'} = \frac{J_{zx}}{J_x J_z - J_{zx}^2}$,J_x and J_z denote moments along x and z axes; J_{zx} represents coupling moments along (z, x) axes. R_p, R_β, and R_r shows aerodynamics in 3-axes; K_β, K_p and K_r defines rolling moments; Q_β, Q_p and Q_r represent the yawing moments; $R_{\delta a}$ and $R_{\delta r}$ represent the side aerodynamics; $Q_{\delta r}$ and $Q_{\delta a}$ are the rolling moments provide by rudder and aileron; m is the total airplane mass; u_o denotes reference flight speed and θ_o represents reference angle of climb. Furthermore, $u_p = [\delta_a \quad \delta_r]^T$ represents the control input vector where δ_a and δ_r are the control surfaces inputs delivered via aileron and rudder. The outputs are β and ϕ.

III. THE RECONFIGURATION PROBLEM

A. An overview of the virtual actuator

The basic concept here is to avoid the whole system's failure after the fault is induced in one component. Without the proper control reconfiguration, the actuator fault deviates the system from its normal behaviour and even causes physical damage. The approach adopted in this paper aims for output to be free from fault. Possible means to deal with such faults are placing a reconfiguration block in the middle of all the actuator vector and controller output. The purpose is to reconfigure the control input signal to the faulty plant such that it produces the same effect as nominal. That block is named the virtual actuator [22].

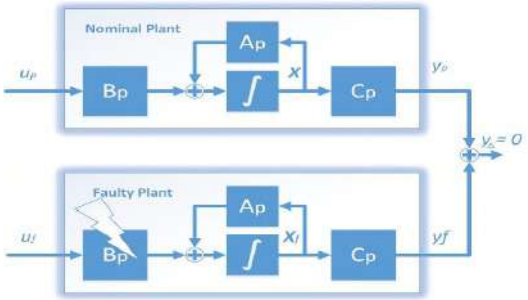


Fig. 2. The reconfiguration goal

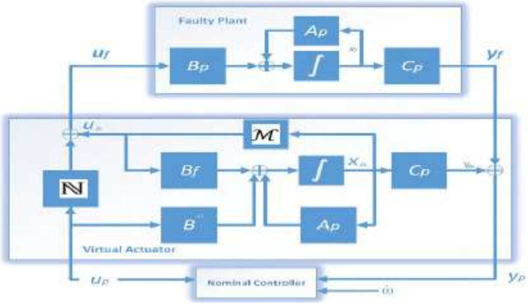


Fig. 3. The closed-loop configuration of a generalized virtual actuator

With the proper virtual actuator, a faulty plant and the virtual actuator behaviour match the nominal. Hence the nominal controller can be used for the defective plant, which is a significant benefit for the reforming controller because of its prolonged procedure and several closed-loop stability tests.

B. Nominal plant and faulty plant Modeling

The nominal plant $\Sigma_p(A_p, B_p, C_p, x_{p_0})$ is modeled in state-space form

$$\Sigma_p \begin{cases} \dot{x}_p = & A_p x_p + B_p u_p \\ & y_p = C_p x_p \qquad : x_p(0) = x_{p_0} \end{cases} \tag{2}$$

where $x_p \in \Re^{n_x}$, $y_p \in \Re^{n_y}$ and $u_p \in \Re^{n_u}$ the system's states, output and input vectors. $A_p \in \Re^{n_x \times n_x}$, $B_p \in \Re^{n_x \times n_p}$, $C_p \in \Re^{n_y \times n_y}$ is the system matrices. The nominal controller is an output feedback controller with an input y_p and reference input $\omega(t)$ is represented as

$$u_p(t) = -\mathcal{F} y_p(t) + \mathbb{N}\omega(t) \tag{3}$$

Where $\mathcal{F} \in \Re^{n_u \times n_y}$is the output feedback gain and $\mathbb{N}$ is the outer loop term to facilitate tracking. Since the actuator failure is considered in this paper, so the difference between the faulty and the nominal plant is in its input matrix B_p that is $B_p \neq B_f$. The model of the faulty actuator under the actuator failure is obtained as

$$\Sigma_f \begin{cases} \dot{x}_f = A_p x_f + B_f u_f \\ y_f = C_p x_f \qquad : x_f(o) = x_{f_o} \end{cases} \tag{4}$$

The initial state x_{f_o} is assumed to be the same as the plant's initial state. We assumed that actuators' failure information is known in prior; therefore, the model of faulty plant $\Sigma_f(A_p, B_f, C, x_{f_o})$ is known.

C. Reconfiguration Goal

A plant is configured successfully if the system contain a virtual actuator and the faulty plant, as shown in Fig. 2. That is, the reconfigured and nominal plant has the same input-output behaviour $(y_f(t) = y_p(t))$ and also the reconfigured plant is stabilizable.

D. Reconfiguration using the virtual actuator

Considering (4), the generalized virtual actuator is clear as

$$\begin{aligned} \dot{x}_\Delta &= A_\Delta x_\Delta + B_\Delta u_p \qquad : x_\Delta(o) = x_{\Delta_o} \\ u_f &= C_\Delta x_\Delta + D_\Delta u_c \\ y_p &= C_p x_\Delta + y_f \end{aligned} \tag{5}$$

where the state $x_\Delta \in \Re^{n_x}$ and the matrices

$$\begin{aligned} A_\Delta &= A_p - B_f \mathcal{M} \\ B_\Delta &= B - B_f \mathbb{N} \\ C_\Delta &= \mathcal{M} \\ D_\Delta &= \mathbb{N} \end{aligned} \tag{6}$$

where the matrices $\mathcal{M}$ and $\mathbb{N}$ are selected at liberty.

E. Analysis of Reconfigured Closed Loop System

For analyzing the closed-loop system, the combination of the model of plant (2) and virtual actuator (5)-(6) along with the nominal controller (3), gives the model of reconfiguration plant,

$$\begin{pmatrix} \dot{x}_p \\ \dot{x}_\Delta \end{pmatrix} = \begin{pmatrix} A_p - B_p \mathcal{F} C_p & 0 \\ -B_\Delta \mathcal{F} C_p & A_p - b_p \mathcal{M} \end{pmatrix} \begin{pmatrix} x_p \\ x_\Delta \end{pmatrix} + \begin{pmatrix} B_p \mathbb{N} \\ B_\Delta \mathbb{N} \end{pmatrix} \omega$$
$$\begin{pmatrix} x(0) \\ x_\Delta(0) \end{pmatrix} = \begin{pmatrix} x_0 - x_{\Delta_0} \\ x_{\Delta_0} \end{pmatrix} \tag{7}$$

$$y_c = \begin{pmatrix} C_p & 0 \end{pmatrix} \begin{pmatrix} x_p \\ x_\Delta \end{pmatrix} \tag{8}$$

$$y_p = \begin{pmatrix} C_p & -C_p \end{pmatrix} \begin{pmatrix} x_p \\ x_\Delta \end{pmatrix} \tag{9}$$

It can be seen from (7-9), the closed-loop system contain the eigenvalue of the nominal plant (2) with output feedback controller (3) and set of virtual actuator eigenvalues (5-6).

IV. DESIGN OF NOMINAL OUTPUT FEEDBACK CONTROLLER

This section will design the output feedback control to stabilize the plant (2) with the state feedback control law (3). The methodology followed in this paper is adopted from [23] and extended to the given plant. ILMI algorithm helps to calculate the controller gain. For a plant (2) and output feedback controller (3), the control task is to find output feedback gain, the closed-loop system Σ_c as

$$\Sigma_c : \dot{x}_p = (A_p - B_p \mathcal{F} C_p) x_p \tag{10}$$

is stable.

Theorem 1: The system Σ_c is stabilizable through output feedback control if $\mathcal{P} > 0, \mathcal{X} > 0$ and $\mathcal{F}$ satisfying

$$\begin{aligned} A_p^T \mathcal{P} + \mathcal{P} A_p - \mathcal{X} B_p B_p^T \mathcal{P} - \mathcal{P} B_p B_p^T \mathcal{X} + \mathcal{X} B B^T \mathcal{X} + \\ (B_p^T \mathcal{P} + \mathcal{F} C_p)^T (B_p^T \mathcal{P} + \mathcal{F} C_p) < 0 \end{aligned} \tag{11}$$

Apply Schur complement to the inequality (11), the following quadratic matrix inequality (QMI) is formed as:

$$\begin{bmatrix} \begin{matrix} (A_p^T \mathcal{P} + \mathcal{P} A_p - \mathcal{X} B_p B_p^T \mathcal{P} - \\ \mathcal{P} B_p B_p^T \mathcal{X} + \mathcal{X} B B^T \mathcal{X}) \end{matrix} & (B_p^T \mathcal{P} + \mathcal{F} C_p)^T \\ (B_p^T \mathcal{P} + \mathcal{F} C_p) & -I \end{bmatrix} < 0 \tag{12}$$

For a fixed value of X, (7) is reduced to the LMI problem with the unknown value of $\mathcal{F}$ and $\mathcal{P}$. To obtain the necessary condition of the feasibility of (11) perturb A_p to $A_p - (\alpha/2)I$ for some $\alpha \geq 0$ such that

$$\begin{aligned} A_p^T \mathcal{P} + \mathcal{P} A_p - \alpha \mathcal{P} - \mathcal{X} B_p B_p^T \mathcal{P} - \mathcal{P} B_p B_p^T \mathcal{X} + \mathcal{X} B B^T \mathcal{X} \\ + (B_p^T \mathcal{P} + \mathcal{F} C_p)^T (B_p^T \mathcal{P} + \mathcal{F} C_p) < 0 \end{aligned} \tag{13}$$

Based on this equation, the eigenvalue of Σ_c shifted to the left half-plane depend on the reduction of α. A step-wise procedure to choose the optimal value of α is proposed in [23] and is given below.

A. Iterative linear matrix inequality

For a given plant Σ_p (with realization A_p, B_p, C_p) is stabilizable via output feedback controller given in (3). The following step-wise procedure is adopted to obtain optimal α

- Step 1: Select $\mathcal{Q} > 0$ and resolve $\mathcal{P}$ from $A_p^T \mathcal{P} + \mathcal{P} A_p - \mathcal{P} B_p B_p^T + \mathcal{Q} = 0$ Set $i = 1$ and $\mathcal{X}_i = \mathcal{P}$
- Step 2: For $\mathcal{P}_i$, $\mathcal{F}$ and α_i. Lessen α_i focus on succeeding LMI constraints

$$\begin{bmatrix} \begin{matrix} (A_p^T \mathcal{P} + \mathcal{P} A_p - \mathcal{X} B_p B_p^T \mathcal{P} - \\ \mathcal{P} B_p B_p^T \mathcal{X} + \mathcal{X} B B^T \mathcal{X} - \alpha_i \mathcal{P}) \end{matrix} & (B_p^T \mathcal{P} + \mathcal{F} C_p)^T \\ (B_p^T \mathcal{P} + \mathcal{F} C_p) & -I \end{bmatrix} < 0 \tag{14}$$

$$\mathcal{P}_i = \mathcal{P}_i^T > 0 \tag{15}$$

Denote α_i^* as the minimum value of α_i

- Step 3: If $\alpha_i^* \geq 0$, then $\mathcal{F}$ is a stabilizable output feedback gain.
  Stop
- Step 4: Minimize trace $(\mathcal{P}_i)$ subject to LMI constraints (14)-(15) with $\alpha_i = \alpha_i^*$. Denote $\mathcal{P}_i^*$ as the $\mathcal{P}_i$, that minimize trace $\mathcal{P}_i$
- Step 5: If $||\mathcal{X}_i - \mathcal{P}_i^*|| < \delta$, go to step 6 else set $i = i + 1$ and $\mathcal{X}_i = \mathcal{P}_{i-1}^*$, then go to step 2
- Step 6: The system may not be stabilizable via output feedback gain.
  Stop

V. VIRTUAL ACTUATOR DESIGN

The design procedure to select the value of the virtual actuator gains $\mathcal{M}$ and $\mathbb{N}$ and to have little change among the behaviour of nominal and reconfigured systems is discussed in this section. It can be seen from generalizing virtual actuator equations (5)-(6) and reconfigured close-loop system given in (7)-(9), the stability depends on the proper choice of $\mathcal{M}$ such the σ_{min} $(A_p - B_f \mathcal{M})$ lies in the left half-plane. In this paper, the LQR approach is adopted to compute the value of $\mathcal{M}$ that satisfies the Algebraic Riccati Equation (ARE) given as

$$A_p^{'} \mathbb{P} - \mathbb{P} A + \mathbb{Q} + \mathbb{P} B_f \mathbb{R}^{-1} B_f^{'} \mathbb{P} = 0 \tag{16}$$

The optimal controller gain is evaluated using

$$\mathcal{M} = \mathbb{R}^{-1} B_f^{'} \mathbb{P} \tag{17}$$

where $\mathbb{P}$ is the symmetric positive definite matrix obtained using ARE.

To match the behaviour of the reconfigured system with that of the nominal system under the commanded input ω, $\mathbb{N}$ is selected so as

$$B_\Delta = B_p - B_f \mathbb{N} = 0 \tag{18}$$

holds.

In this paper, the matrix $\mathbb{N}$ is chosen through a zero-pole placement technique [24]. The goal is to have y_f become equal to y_p that is y_Δ converges to zero. According to (5) the equilibrium of output deviation is

$$y_\Delta = C_p (sI - A_p + B_f \mathcal{M})^{-1} (B_p - B_f) \mathbb{N} u_p \tag{19}$$

The goal is to vanish steady-state output deviation y_Δ for all constant input; it requires

$$0 = C_p (sI - A_p + B_f \mathcal{M})^{-1} (B_p - B_f) \tag{20}$$

Since (20) is the same as the transfer function at $s = 0$, it corresponds to the placement of input-decoupling zeros at zero. Eq. (20) has a solution in $\mathbb{N}$ if

$$\begin{aligned} rank(C_p (A_p + B_f \mathcal{M})^{-1} B_p) = \\ rank(C_p (A_p + B_f \mathcal{M})^{-1} (B_p, B_f)) \end{aligned} \tag{21}$$

If the solvability is confirmed, then the matrix $\mathbb{N}$ can be found by inverting (17) such that

$$\mathbb{N} = \mathcal{H}^{+} C_P (A_p - B_f \mathcal{M})^{-1} B_p \tag{22}$$

where $\mathcal{H} = C_P(A_p - B_f\mathcal{M})^{-1}B_p$ and right pseudo inverse $\mathcal{H}^{+}$ is constructed such that $\mathcal{H}\mathcal{H}^{+}\mathcal{H} = \mathcal{H}$.

VI. Simulation Results

This section presents the simulation results to control β and ϕ of aircraft that suffer from actuator failure due to damage that occurs in its vertical-tail. The information of the failed actuator is expected to be known. The expression of state coefficients of the plant is obtained from [13] and given as

$$A_p = \begin{bmatrix} -0.11 & 0 & -673 & 32.20 \\ -3.53 & -0.84 & 0.31 & 0 \\ 3.65 & -0.04 & -0.25 & 0 \\ 0 & 1 & 0.03 & 0 \end{bmatrix}$$

$$B_p = \begin{bmatrix} 0 & 0 & 0.0041 & 0.0032 \\ 0.22 & 0.35 & 0.126 & 0.027 \\ 0.02 & 0.02 & -0.275 & -0.2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C_p = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Where $u_p = [\delta_{a_o}, \delta_{a_{in}}, \delta_{r_u}, \delta_{r_l}]'$ is the control input vector that is considered in the simulation.

The nominal controller is designed based on ILMI using the CVX toolbox in MATLAB. To achieve output tracking, $\mathbb{N}$ is calculated to guarantee zero steady-state error at step input that is

$$\mathbb{N} = -\frac{1}{C_p(A_p - B_p\mathcal{F}C_p)B_p}$$

The nominal controller gain is computed as

$$\mathcal{F} = \begin{bmatrix} 9.31 & 1.04 \\ 1.04 & 1.57 \\ 2.15 & 0.20 \\ 4.22 & 0.94 \end{bmatrix}$$

A. Tracking performance under the nominal condition

Testing controller enactment under normal condition (without actuator failure), a 12^o and 0^o amplitude signal is smeared to the roll slip and side slip, respectively. The simulation result in Fig. 4 shows the defined tracking of the roll slip angle. The roll angle trajectory commanded input thru satisfactory delay while variation for side-slip angle is within the limit of ± 0.1 which is observed as zero steady-state error.

B. Simulation results under actuator failure

In this simulation, the vertical-tail damage occurs, and there is no upper rudder control surface, and all the hydraulic lines pull apart. Therefore, the control surfaces associated with HA fails to activate. Since the inboard aileron and lower rudders remain fortified by DRAS. So, the switching device takes EHA into the loop, and the two control surfaces equipped with EHA can still be operational. Therefore, the input distribution matrix B_p changes to B_f as

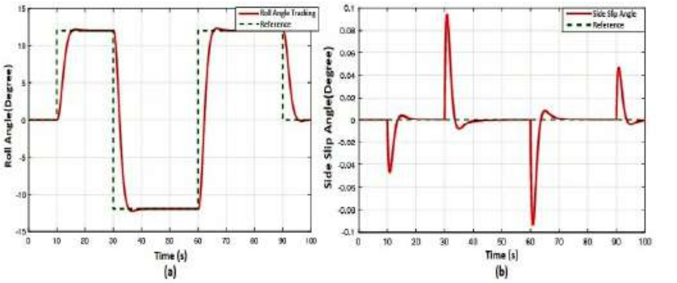


Fig. 4. System dynamical performance under the normal condition at square input

$$B_f = \begin{bmatrix} 0 & 0 & 0 & 0.0032 \\ 0 & 0.35 & 0 & 0.027 \\ 0 & 0.02 & 0 & -0.2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Due to a faulty actuator, Fig. 5 shows the degradation in system performance. As grasped by Fig. 5(a), the roll slip angle cannot trail input and leads to sizable steady-state error. Also, in Fig. 5(b), the fluctuation in the side-slip angle's output reaches over the prescribed level of tolerance.

C. Simulation results of control reconfiguration after actuator failure

To reconfigure the controller structure after the occurrence of the fault to the system, only EHA provides the required control input and brings the system to the nominal condition. To overcome the slower dynamics of EHA, a first-order delay system $u_{EHA} = e^{-0.1t} u_{HA}$ is announced, whereas, u_{EHA} is the input to EHA and u_{HA} is the current input to HA. To provide tolerance against this severe failure condition, a virtual actuator is brought into operation. Choosing the parameter of the virtual actuator as

$$\mathcal{M} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1.27 & -1.31 & -13.61 & -0.95 \\ 0 & 0 & 0 & 0 \\ 5.2 & 0 & -0.24 & 0.30 \end{bmatrix}$$

Such that $\sigma_{min}(A_p - B_f\mathcal{M})$ is stable whereas $\mathbb{N}$ can be chosen using (22). The roll angle and side-slip angle are recovered from the simulation plots to their nominal position (see Fig. 6) with a virtual actuator inserted in the middle of the faulty plant and the nominal controller.

VII. Conclusion

The paper presents the fault-tolerant control strategy subject caused by vertical-tail damaged aircraft driven by dissimilar redundant actuation system. First, designing the nominal output feedback controller using iterative linear matrix inequality. A virtual actuator circuit is induced between the faulty plant and the nominal controller to facilitate fault-tolerant control

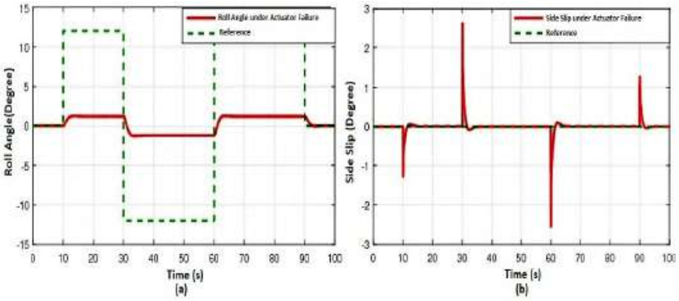


Fig. 5. Degradation of dynamical performance under actuator failure

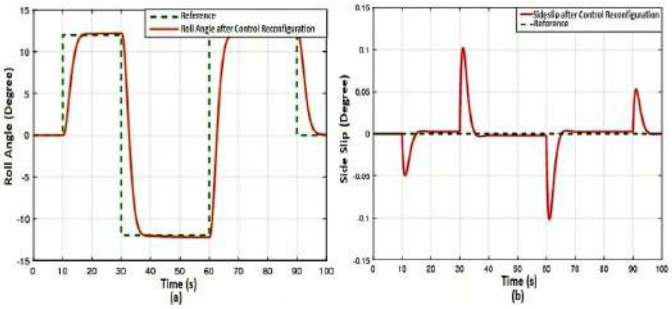


Fig. 6. Dynamical system performance after control reconfiguration

capability. The simulations are performed subject to HA failure due to severe damage. Due to dissimilar redundancy, the control surfaces associated with EHA provides the desired flight path in severe failure condition.

References

[1] L. Crider, “Control of commercial aircraft with vertical tail loss”, Proceedings of AIAA 4th aviation technology, integration and operation (ATIO) forum, Chicago, lllinois, September, 2004.

[2] J.D. Boskovic, R. Prasanth, and R. Mehra, “Retrofit fault-tolerant flight control design under control effect or damage”, Journal of Guidance, Control and Dynamics, vol. 30(3), pp. 703–12, 2007.

[3] J. Wang, Z. Li, and Z. Peng, “Modeling and analysis of the dissimilar redundant actuator system”, Machine Tool and Hydraulics, vol. 36(6), pp. 79–81, 2008.

[4] Y. Fu and Y. Pang “Design and working mode analysis of dissimilar redundant actuator system”, Journal of Beijing University of Aeronautics and Astronautics, vol. 38(4), pp. 432–437, 2012.

[5] W. Karam and J. Mare, “Force control of a roller screw electormechanical actuator for dynamic loading of aerospace actuators”, International conference on fluid power and motion control, Bellingham, pp. 515–28, 2008.

[6] Y. Hitachi “Damage-tolerant flight control system design for propulsion controlled aircraft”, Dissertation, University of Toronto, 2009.

[7] J. Zhao, B. Jiang, P. Shi, and Z. He, “Fault tolerant control for damaged aircraft based on sliding mode control scheme”, International journal of innovative computing, information and control, vol. 10(1), pp. 293–302, 2014.

[8] X. Li and H. Liu, ”A passive fault tolerant flight control for maximum allowable vertical tail damaged aircraft”, The Journal of Dynamic Systems, Measurement and Control, vol. 34(3), pp. 1625–32, 2012.

[9] J. Cieslak, D. Henry, A. Zolghadri, and P. Goupil, “Development of an active fault-tolerant flight control strategy”, Journal of Guidance, Control and Dynamics, vol. 31(1), pp. 135–47, 2008.

[10] M. A. Ashraf, S. Ijaz, Y. Zou, and M. T. Hamayun, “An integral sliding mode fault tolerant control for a class of non-linear Lipschitz systems,” IET Control Theory Appl., vol. 15, no. 3, pp. 390–403, 2021.

[11] D. Ye and G.H. Yang, “Adaptive fault-tolerant tracking control against actuator faults with application to flight control”, IEEE Transactions on Control Systems Technology, vol. 14(6) pp. 1088–96, 2006.

[12] R. E. Bavili, M. J. Khosrowjerdi, and R. Vatankhah, ”Active Fault-Tolerant Controller Design using Model Predictive Control”, Control Engineering and Applied Informatics, vol. 17, pp. 68 -76, 2015.

[13] J. Wang, S. Wang, X. Wang, C. Shi, and M. M. Tomovic, ”Active fault-tolerant control for vertical tail damaged aircraft with dissimilar redundant actuation system”, Chinese Journal of Aeronautics, vol. 29, pp. 1313-1325, 2016.

[14] W. Xingjian, W. Shaoping, Y. Zhongwei, and Z. Chao, ”Active fault-tolerant control strategy of large civil aircraft under elevator failures”, Chinese Journal of Aeronautics, vol. 28(6), pp. 1658–1666, 2015.

[15] X. Sun, X. Wang, Z. Zhou, and Z. Zhou, “Active Fault-Tolerant Control Strategy for More Electric Aircraft under Actuation System Failure,” Actuators, vol. 9, no. 4, pp. 122, Nov. 2020.

[16] S. Ijaz, L. Yan, M.T. Hamayun, and C. Shi, “Active fault tolerant control scheme for aircraft with dissimilar redundant actuation system subject to hydraulic failure”, Journal of the Franklin Institute, vol. 356(3), 2019.

[17] Z. Huixuan, S. Qinglin, and C. Zengqiang, ”Equivalent sliding mode fault tolerant control based on hyperbolic tangent function for vertical tail damage”, Journal of Southeast University, vol. 36(2), pp. 152-162, June 1, 2020.

[18] K. Ahmadi, D. Asadi, F. Pazooki, ”Nonlinear L1 adaptive control of an airplane with structural damage,’ Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, vol. 233(1), pp. 341-353, 2019.

[19] D. Sun, R. Choe, E. Xargay, and N. Hovakimyan, ”An L1 Adaptive Backup Flight Control Law for Transport Aircraft with Vertical-Tail Damage”, AIAA Guidance, Navigation and Control Conference, California, USA, 2015.

[20] Z. Zhou, S. Wang, and X. Wang, ”Fault Tolerant Control Strategy Based on Actuation Switch Mechanism for More-electric Aircraft with Vertical Tail Damage”, IEEE International Conference on Mechanics and Automation, Takamatsu, Japan, 2017.

[21] W. Jun, W. Shaoping, W. Xingjian, S. Cun, and M. M. Tomovic, ”Active fault tolerant control for vertical tail damaged aircraft with dissimilar redundant actuation system”, Chinese Journal of Aeronautics, vol. 29(5), pp. 1313–1325, 2016.

[22] J. Lunze and T. Steffen, ”Control reconfiguration after actuator failures using disturbance decoupling methods”, IEEE Transactions on Automatic Control, vol. 51(10), 2006.

[23] Y. Cao, J. Lam, and Y. Sun. ”Static output feedback stabilization: an ILMI approach”, Automatica, vol. 34(12), pp. 1641 -1645, 1998.

[24] T. Steffen, ”Control Reconfiguration of Dynamical System”, Linear Approaches and Structural Tests, Springer-Verlag, Heidelberg, pp. 110-120, 2005.

Development of Churn Prediction Model using XGBoost – Telecommunication Industry in Sri Lanka

Prasanth Senthan
Department of Physical Sciences and Technology, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
sprasanth@appsc.sab.ac.lk

RMKT Rathnayaka
Department of Physical Sciences and Technology, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kapilar@appsc.sab.ac.lk

Banujan Kuhaneswaran
Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
bhakuha@appsc.sab.ac.lk

BTGS Kumara
Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kumara@appsc.sab.ac.lk

***Abstract*— Maintaining a customer base at a feasible rate is considered important in most business organizations since customers are the precious asset of the business sector. It is a vital task to retain the customers at a steady level in any business enterprise for the overall stability of its business activities. Sufficient pieces of evidence in this connection have been gathered to prove that the telecommunication industry is the most affected field of business by the tendency of the customers to shift towards alternative service providers. Therefore, a distinctive effort has been made to design this specific forecasting method is carried out utilizing a combination of a properly approachable method aiming at clarifying the probability of the above-mentioned tendency of the clients seeking an alternative service provider in the industry. In this attempt, a data set that included 10, 000 postpaid consumer particulars including 20 attributes were taken for this research for a thorough analysis of this aggravating issue in the telecommunication industry. In the end, a satisfying outcome was witnessed and certain clarification was made out of the 10,000 subscribers 4888 showed positive attitudes and 5112 indicated negative to the churning behavior. Besides, this specific data set was subjected to complete verification in comparison to certain supervised machine learning algorithms such as Decision tree, Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Along with this, ensemble techniques such as Random Forest, Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost) also have been considered. Subsequently, an assurance was made that XGBoost possessed the ability to bring out the maximum and précised accuracy of 82.90%. Eventually, a hyperparameter tuning had been performed with XGBoost. As a result, an assurance was acquired that XGBoost showed an upsurge in the previously obtained accuracy from 82.90% to 83.13%.**

***Keywords*— XGBoost, AdaBoost, Churn Prediction**

I. INTRODUCTION

The customers of any business enterprise are the major component of its progress and success. They are the ones who contribute immense boost for the continuous development and reputation of these companies. So, it is obvious that these enterprises largely rely on their customers. Therefore, it is essential for them to preserve their customer base and to uphold their reputation in the competitive business background. Otherwise, there are chances for the regular clients to churn from the existing company or their service providers. As a consequence of this, the productivity and the valuable status of the company may severely be damaged. Here, the word "Customer Churn" means the termination of the relationship of a particular client or clients with the organization. For example, the individual who is a usual customer of a business establishment "A" could perhaps discontinue having the services and select another establishment named "B" [1] [2]. Depending on the contemporary condition of the organization, numerous reasons may be cited for the above feature. Customer churn often occurs in most organizations anywhere in the world. They may be industries like banking, finance companies, insurance companies, or telecommunication services. It is risky and unreliable to identify the reasons for such churning behaviors of the customers and to avoid these situations because the consistency and the progress of the company are straightly influenced by this attempt. So, it is inevitable for the company to have a practical "Churn Prediction" method for safeguarding the company from being suffered by the behavior of customer churning. A properly planned churn prediction does have an important part and makes a significant impact on handling customer churn.

To provide satisfactory communication, the telecommunication industry has various service providers in its possession. Almost all the people in the world have subscriptions to telecommunication and at the same time, they opt to use multiple systems. As the subscribers tend to switch on to an alternative service provider due to dissatisfaction, customer churn is very common to see in this industry. The reasons for this tendency may be poor customer service, high usage of service, and very low switching cost [3]. What has been mentioned above are the traditional causes. However, this particular attrition behavior is also caused by some personal intentions that have been recognized recently in this field. If an attempt has to be made to find out the reasons for the above, the following features stand as proof. They are namely "Disappointment based on facilities provided by the service provider at the primary on entering the firm for their needs" and "The failure of the enterprise in supplying the expected service to their clients", though they might be somewhat impractical on a particular situation. Besides, another noticeable destructive feature in the telecommunication industry is that it faces an inevitable customer churn of 40% every year [4]. In consequence of these harmful circumstances, this industry employs various unique methods to survive in the competitive market. These attempts may aim to sustain the continuous presence of the customers at a steady level. It is usually presumed that retaining the subscribers who are already present in the industry is more advisable than seeking new ones because the latter burden the company with unbearable spending. The

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

expenditure to attract new subscribers forces the company to bear nearly 5 times higher expenses [5][6].

Since it is very costly to attract fresh customers into the company, it is truly advisable to keep the ones who are already there in the business. Further, it is said that the existing long-term customers are the cash – cow of the company, the entrepreneurs ought to make sure about their continuous existence there. It should not be forgotten that to retain the existing customers most companies use a huge amount of their annual budget for “Customer Churn Prediction”. A properly planned research in this connection could help the organization detect the most apparent churners so that the organization concerned can resist the issue of churning and would be able to retain the existing clients. If this happens, the precious manpower would not be wasted.

Besides, a few more models of this nature came into being previously that were very similar to the churn prediction within the companies of telecommunication services. Here too, the following models of algorithms were taken into consideration. They are namely – ANN, Decision tree, Logistic regression, SVM, and Bayesian models (Naïve Bayes). In most of the previous efforts of developing “Churn Prediction”, a contrasting and comparing approach was carried out making use of the basic algorithms that are accustomed in the field. In this effort, the ensemble algorithms were not considered (XGBoost, AdaBoost). Besides, in the greater number of such cases, it has been noted that only a last and ultimate suggestion was shown in creating prediction models using techniques of greater accuracy devoid of the implementation of the graphical user interface.

The sole purpose of this prediction effort was to foresee the number of clients who tend to shift to another service provider with the use of some unverified machine learning algorithms. This effort also aimed at creating an effective churn prediction model to manage the deliberate churning behavior of the customers. To proceed with this research, exactly 10,000 post-paid customer records were collected from a telecommunication company in Sri Lanka. Identifying such churners in a telecommunication company is indeed a difficult challenge using the data of prepaid customers who have many options for their services. This is because there is no reliable source or agreement available for such identification. In the process of this study, 20 attributes were made available. Some of them are “Statistics of Billing”, “Complaint Details”, “Call Duration Details and so on. Further, a user-easing way was made use of in the effort of identifying churners and consequently, a Graphical User Interface (GUI) was established for this purpose. Accordingly, all the above-mentioned “Supervised Algorithms” such as Decision tree, Logistic Regression, Random Forest, SVM, XGBoost, AdaBoost, ANN were properly subjected to proper contrasting and comparison. Consequently, a highly accurate technique was used to make an effective churners’ prediction model. For this endeavor, the cross-validation was purposefully and thoroughly done with all the aforementioned techniques.

II. Literature Review and Related Works

Keramati et al. have accomplished research using the binomial logistic regression algorithm to pinpoint the features that mostly influence customer churn. These features are namely number of service failures, number of complaints, amount of service usage switching cost, and some other customer demographic factors. Further, the results showed that client/customer status (active or non-active) facilitates the connection between the attrition behavior of customers and the reason for attrition. The non-appearance of greatly influencing factors and pressurization between various operators are the major limitations found in this study. For example, the service cost is a powerful factor of customer churn because of the limitation of pressure. Also, the service cost lies between multiple operators [1].

Sharma et al. have carried out research using the Neural Network method to predict the churning customers in prepaid telecommunication services. This research has identified the factors that are necessary to be considered to improve an accurate and consistent churn prediction model. There are 20 factors taken into consideration and the model was developed by examining 2,427 customers. Some of these features include state, account length, area code, and overall day minutes, etc. To evade the overtraining of the model, a randomly chosen ratio of the data set was trained. The model has produced results with 92.35% accuracy in the prediction of churn behavior. But, there was a restriction in poor performance and it took some extra time to give out the expected results [2].

Keramati et al. have made a hybrid model consisting of Decision Tree, SVM, ANN, and K-Nearest Neighbor (KNN). 95% accuracy in churn prediction was obtained from this recommended model. The accuracy value is efficiently achievable here. As an added feature of this method, a novel approach has been suggested to pinpoint the main factors that need to be adjusted in the churn prediction model. The outcomes of the research have revealed that the hybrid methodology which is comprised of four methodologies depicted above produced higher accuracy when in comparison with isolated algorithms performed in this process. Finally, a completely new approach has been suggested for dimensionality reduction [5].

A hybrid churn prediction model using bagging and boosting methods has been prepared by Fathian et al. During this research, a technique incorporated with specific predictive modeling was taken into consideration. To name them would be ANN, Self-Organizing Maps (for clustering), SVM, the algorithm named KNN, and Decision Tree. In this process, the PCA (Principal Component Analysis]) was utilized, aiming at the total elimination of the unnecessary features. The comparison between the existing models was performed, based on the following 5 factors namely specification, accuracy, sensitivity, F-measure, and AUC. The results of the research have shown that merging the clustering techniques depicted above with bagging and boosting has created better results concerning accuracy. This attempt had some difficult experiences while trying to obtain a better classification [8].

During yet another research, Olle et al. have made a hybrid churn prediction model with the amalgamation of Logistic regression, Voted Perceptron, and clustering techniques. The model had achieved a little more leap forward to produce very high accuracy results compared to the results obtained by the previously mentioned isolated algorithms. The accuracy of this model is approximately 99%. 23 variables had been taken to analyze and 2,000 subscriber details were assembled from Asian telecommunication. The model had undergone training and testing against a relatively smaller number of subscriber data. Therefore, it is suggested that this particular model has feasible to be enhanced in the future in dealing with larger data sets [9].

Preeti et al. researched a customer churn prediction with the adaption of Logistic regression and Decision tree techniques. This study evidenced that choosing the relevant and correct set of attributes would produce high accuracy results. Hence, they have ascertained that how Logistic regression has identified the factors that affect the defection rate of customers. A model with higher précised accuracy is not needed if it is incapable of describing the causes for the attrition behavior. Besides, the study provides some strategic pathways (comparative analysis) to select the finest algorithm to recommend the prediction model [10].

Umayaparvathi et al. have developed a churn prediction for the telecommunication industry. During this study, an easy interpretation of churn prediction methodology was proposed. Further, the efficiency of some predictive modeling methods corresponding to the subject was subjected to the analysis here. They are Gradient Boosting, KNN, Decision Tree, Ridge Regression Classifier, SVM, Random Forest, and Logistic Regression. In order to increase the accuracy of the prediction model, each attribute in the data set was assessed to find its significance in churn behavior. Openly adaptable churn data set from two telecommunication companies were taken to validate the performance of this model. The outcome of the research indicated that Gradient Boosting is the best classifier in the churn prediction [11].

A novel churn prediction was modeled by Gregory using a prevailing algorithm named as XGBoost. This model was prepared to take part in the “WSDM Cup Churn Challenge competition” in 2018. Music streaming data given by KKBOX (One of the foremost music streaming services in ASIA) have been used in this study. These data comprise user log details, transactions, and statistics about members. Log loss is the utility used in this research to assess the performance of the prediction model. The model developed with XGBoost outperformed all the models exhibited during this competition. In order to increase the accuracy of the model, the Light GBM (Light Gradient Boosting Machine) library and XGBoost primary based model were integrated. The final results of the research have ascertained that feature engineering is a significant part of boosting the accuracy of the model [12].

Umayaparvathi et al. have completed yet another survey concerning churn prediction for the telecommunication industry. This study was mostly focused on the analysis of freely and commonly available telecommunication data sets, methods used in emerging the prediction models, and the metrics used to assess the performance of the model. In addition to the previous researchers, this study conveyed the features of the considered data sets and classified the most noticeable features. Further, the performance metrics analyzed in this attempted study were Confusion matrix, Accuracy, Precision, Recall, F1-score, AUC curve, and Lift curve [13].

J. Vijaya et al. also designed a model of hybrid churn prediction by mingling both supervised and unsupervised learning techniques. K-Means clustering, Decision tree, SVM, Linear discriminant analysis, Naive Bayes, and KNN have been adapted to carry out this study. 50,000 subscriber details with 230 features were subjected to this particular analysis. This analysis contains 190 numerical features and 40 nominal features. The results of this study depicted that the model developed with the mixture of SVM and K-means algorithm has brought out a better outcome in accuracy, sensitivity, and consistency [14].

Sabbeh attempted to predict the churn behavior of customers by associating ten varieties of predictive modeling methods. They are Discriminant Analysis, SVM, Decision Trees (CART), Ada boost, Random Forest, Stochastic Gradient Boosting, KNN, Multi-Layer perceptron, Naïve Bayesian, and Logistic regression. 3,333 subscriber details were taken for analysis throughout this study. Finally, the following outcomes were noticeably acquired. They are as follows. Both the AdaBoost and Random Forest (Ensemble techniques) nearly produced similar outcomes with 96% of accuracy, Multi-Layer perceptron and Support Vector Machine produced results with 94% accuracy in churn behavior, Linear Discriminant Analysis, and Logistic Regression revealed the accuracy of 86.7% and at last 90%, 88% accuracy level was obtained by the Decision Tree and Naïve Bayes respectively. Only the comparison between single predictive modeling techniques was taken into consideration. This is the main shortcoming of this study [15].

Özer Çelik et al. have researched “Comparing Techniques Used in Customer Churn Analysis” by associating XGBoost, Random Forest, SVM, Logistic Regression, ANN, Naive Bayes, KNN, and Weibull Time To Event Recurrent Neural Network (WTTE-RNN). During this research, the following observations were made. Firstly, the deep learning technique can be utilized for a large amount of data, and secondly, the ensemble machine learning techniques can be used for a small portion of data. In order to produce prominent results in prediction. Besides, with the help of the Cox Regression technique, highly distributed independent features towards the dataset can be easily identified. The main shortcoming found in this study was the way of adapting the dataset. For instance, the separate dataset is used for machine learning techniques and deep learning techniques. In other words, a unique dataset could have been used for the machine learning technique and another dataset for the deep learning technique separately [16].

On the basis of related works mentioned above, several techniques have been employed in the development of the churn prediction model so far. Rather than using basic techniques in this regard, the use of ensemble techniques gives better accuracy in most instances. Besides, it can be a strong foundation to develop a churn prediction model using the “Ensemble Techniques” favoring the researchers in the future on the Sri Lankan context as well as international level.

III. Methodology

Fig. 1 shows the overall methodological framework used during this research to develop the final prediction model. Each step has been explained in detail below.

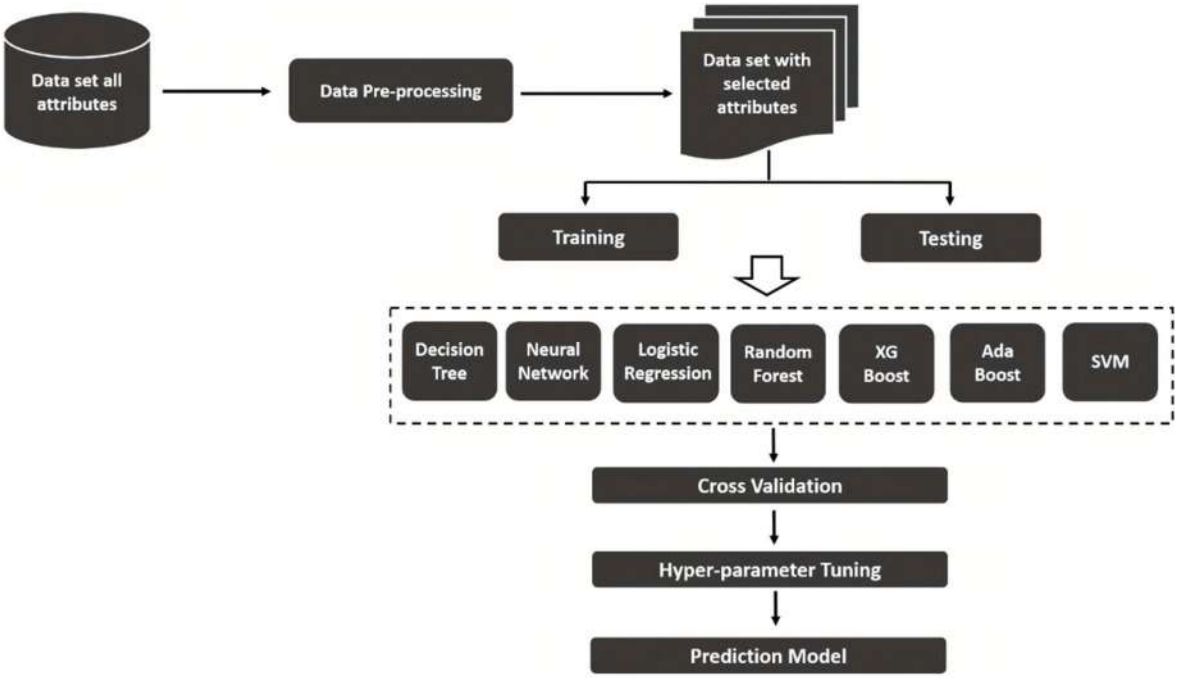


Fig. 1. Overall methodological framework

A. Data set and Requirement analysis

Requirement analysis or the prerequisite analysis would be certainly a leap ahead for examining anticipations of the consumers for a new or modified creation. Pinpointing exactly the major necessities of the particular analysis very correctly is indispensable before the making of any software project or model. Being inclined in this endeavor, many approaches can be made to find more evidence in this regard. To name a few are 'Face to Face Meetings', 'Brainstorming, 'Surveys', 'Web Scraping', and 'Data Collection' from a Web API.

In the course of the above-mentioned effort, a due consultation was made with the reputed telecommunication industry. During this survey, subscriber particulars from exactly 10,000 post-paid subscribers for the year 2019 were collected and systematically studied. The content of 20 attributes along with the target attribute of the dataset and their descriptions are represented by the Table I have given here.

B. Data Pre-processing

This is a task that is aimed at carrying out "Data Pre-Processing" and removing the unfinished, noise-producing, or unreliable data from the system. This process is considered to be the most essential phase in originating such a forecasting structure to find out the churn behavior among the customers. In the course of this strategy, the "Python Pandas Library" was made use of, as an aid to achieve the above-stipulated target. The following tasks were executed under the above function.

a) Storing the raw data into a Comma Separated Values format (CSV) file.

Initially, the required dataset has been obtained in an excel sheet format. To Perform the rest of the processes conveniently the dataset has been converted into a CSV file format.

b) Loading the CSV with pandas.

c) Checking for number of columns and rows

d) Removing unnecessary columns

The data set contained a column named "National Identity Card Number" (NIC) which helps to identify each customer uniquely. It has been removed from the data set because it is not useful for the prediction purpose.

e) Checking for null and duplicate values

The null and duplicate values make the final prediction model to be overfitted. To overcome this issue the null and duplicate values have been replaced with the mean value of the corresponding column.

f) Converting all variables to a common type

For instance, the total number of complaints and the number of negative feedbacks sent by the customers have been presented in the integer data type. Meanwhile, download bandwidth, monthly bill, and most of the other features have been presented in the double format with decimal points. To conclude, the integer datatype features have been converted into the double data type.

g) Outlier detection and replacement

i. Initially, outlier values for each column have been identified. (First Quartile (Q1), Third Quartile (Q3), and Inter Quartile range (IQR) values of all columns have been used in this regard.)

ii. Replacement of random values within quartile range for outlier detected cells.

iii. Validation of each column value is within the quartile range limit.

iv. Creation of new CSV without the outlier values.

h) Normalizing the data set between the fixed ranges.

i. Min Max Scaling technique has been adopted in this regard.

ii. Normalized Value = (Actual Value – Minimum Value) / (Maximum Value – Minimum Value)

b) Names of the Attributes & Their Descriptions

No	Attribute	Description	Repre-sentation
1	Monthly Bill	The average amount of payment done by the customers	Sri Lankan Rupees
2	Billing complaint resolve time	The time duration used to resolve billing complaint	Minutes
3	Billing complaint count	Number of complaints concerning bills	Numeric
4	Promotions	The Short Message Services provided by the company to the customers about their products, services, and so on	Numeric
5	Hotline call time	Total number of minutes taken for hotline call duration in a specified period	Minutes
6	Hotline call count	Total number of hotline calls in a specified period	Numeric
7	Arcade visit time	Total number of minutes visit taken to visit the arcade	Minutes
8	Arcade visit time waiting time	Total number of minutes taken to wait at the arcade	Minutes
9	Arcade visit count	Total number of visits made to the arcade	Numeric
10	Negative feedbacks customer sent	The count of negative feedbacks given by a customer	Numeric
11	Positive feedbacks customer sent	The count of positive feedbacks given by a customer	Numeric
12	Complaint resolve duration	Amount of time used to resolve complaints	Minutes
13	Total complaints	Number of complaints made by a customer	Numeric
14	Complaint breach count	Count of complaints not resolved within a given time	Numeric
15	Coverage related complaint duration	Amount of time taken for complaints regarding coverage made by a customer	Minutes
16	Coverage related complaint count	Complaints made regarding coverage	Numeric
17	Adjustment charges	Alternative charges for additional packages	Sri Lankan Rupees
18	Download bandwidth	Average Bandwidth	Megabytes per second
19	Data used for download	The total amount of data used	Megabytes per second
20	Data Charges	Total payment for data usage	Sri Lankan Rupees
21	Churn [Target]	Churn represents by – 0, Non-churn represent by - 1	Boolean

C. Feature Engineering and Model Generation

After completing the data abstraction and enhancement segment, the dataset with all the attributes has been fed into the selected supervised machine learning algorithms and encountered a problem of model overfitting. After a certain amount of analysis, the subsequent phase would be pinpointing the finest set of attributes that could offer a significant contribution to the data set. The Pearson correlation analysis technique was followed for the above purpose. By the utilization of the aforementioned technique, highly correlated attributes towards the target variable were removed. This was done because the highly correlated variables would not help improve the model score further. Then the dataset has been reshaped according to the significant attributes and fed into the selected supervised algorithms.

As the next step of identifying the most suitable algorithm consisting of better accuracy, this procedure comes to an end with the preparation of the final or last forecasting model. Besides, to identity the churners and non-churners a user-friendly GUI has been developed by adapting the algorithm which provides the highest accuracy.

IV. Results and Discussion

A. Results of detecting the extremely co-related features

During this process, a thorough analysis was carried out to find out the relationship between the prevailing attributes and the target attribute. Pearson correlation analysis was used for this purpose.Fig.2 shows the numerals and the complete score obtained by comparing and contrasting the relationship between each variable with the target variable [‘churn’]. As Fig.2 shows, the highly correlated features that contain an absolute score of 0.5 and more were identified.

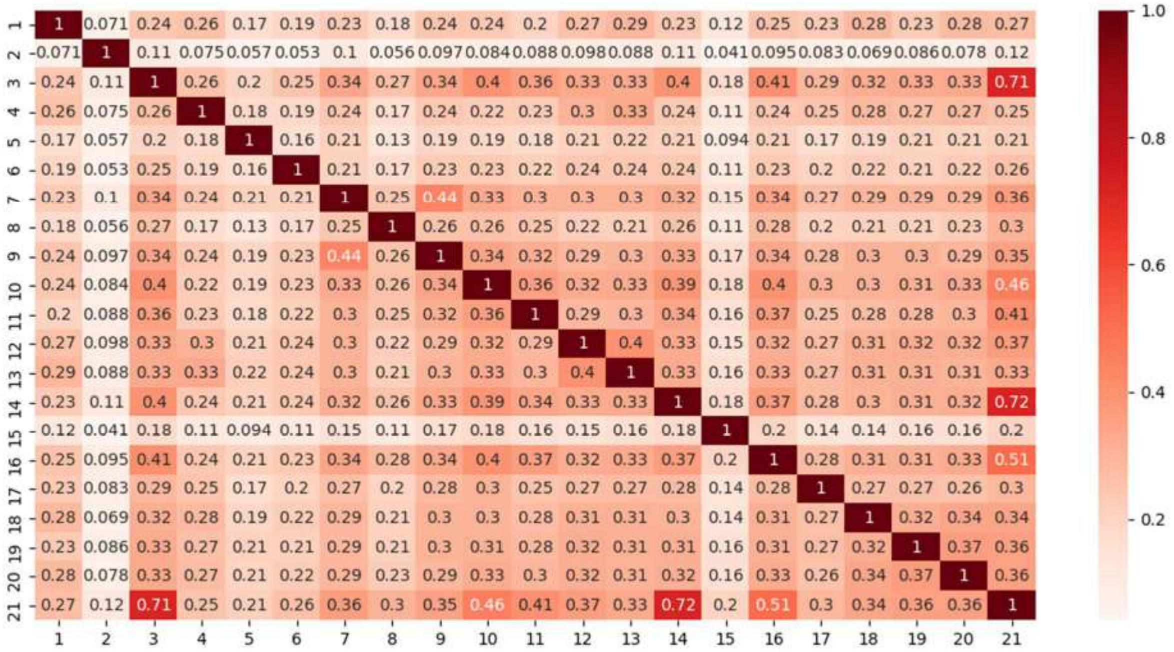
With the results obtained from above, further analysis also was done to recognize 17 out of 20 attributes provide a certain amount of correlation with the target variable. Consequently, the removal of the rest of the highly correlated attributes from the data set was successfully carried out.

After finding out only the important attributes, the whole dataset was classified into two non-equivalent sections. These two sections are namely training and testing. Then, it is fed into the particular supervised machine learning techniques (Random Forest, XGBoost, AdaBoost, Logistic Regression, SVM, ANN, and Decision Tree) to identify the best one to originate the final forecasting model.

The possible results of the performances during the forecasting task were duly evaluated. For this purpose, several methods such as Accuracy, Mean Squared Error (MSE), Confusion Matrix, Mean Absolute Error (MAE), Precision, Recall, and F1-Score were used. The main procedure taken into consideration here was the K-fold cross-validation. [1][2].

B. Comparison of performance

The discussion below is based on the results got during the comparison of performances concerning several algorithms in forecasting the churners and non-churners in the industry of telecommunication. This type of discussion has paved a dependable way to follow a feasible approach towards a “Supervised Learning Technique” in machine learning because the target variable and the highly connected variables were tested in this lengthy investigation. Numerous algorithms were adapted and compared against certain evaluation metrics in the course of this prediction process. Shown below is the additional investigation of this task.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1	0.071	0.24	0.26	0.17	0.19	0.23	0.18	0.24	0.24	0.2	0.27	0.29	0.23	0.12	0.25	0.23	0.28	0.23	0.28	0.27
2	-0.071	1	0.11	0.075	0.057	0.053	0.1	0.056	0.097	0.084	0.088	0.098	0.088	0.11	0.041	0.095	0.083	0.069	0.086	0.078	0.12
3	0.24	0.11	1	0.26	0.2	0.25	0.34	0.27	0.34	0.4	0.36	0.33	0.33	0.4	0.18	0.41	0.29	0.32	0.33	0.33	0.71
4	0.26	0.075	0.26	1	0.18	0.19	0.24	0.17	0.24	0.22	0.23	0.3	0.33	0.24	0.11	0.24	0.25	0.28	0.27	0.27	0.25
5	0.17	0.057	0.2	0.18	1	0.16	0.21	0.13	0.19	0.19	0.18	0.21	0.22	0.21	0.094	0.21	0.17	0.19	0.21	0.21	0.21
6	0.19	0.053	0.25	0.19	0.16	1	0.21	0.17	0.23	0.23	0.22	0.24	0.24	0.24	0.11	0.23	0.2	0.22	0.21	0.22	0.26
7	0.23	0.1	0.34	0.24	0.21	0.21	1	0.25	0.44	0.33	0.3	0.3	0.3	0.32	0.15	0.34	0.27	0.29	0.29	0.29	0.36
8	0.18	0.056	0.27	0.17	0.13	0.17	0.25	1	0.26	0.26	0.25	0.22	0.21	0.26	0.11	0.28	0.2	0.21	0.21	0.23	0.3
9	0.24	0.097	0.34	0.24	0.19	0.23	0.44	0.26	1	0.34	0.32	0.29	0.3	0.33	0.17	0.34	0.28	0.3	0.3	0.29	0.35
10	0.24	0.084	0.4	0.22	0.19	0.23	0.33	0.26	0.34	1	0.36	0.32	0.33	0.39	0.18	0.4	0.3	0.3	0.31	0.33	0.46
11	0.2	0.088	0.36	0.23	0.18	0.22	0.3	0.25	0.32	0.36	1	0.29	0.3	0.34	0.16	0.37	0.25	0.28	0.28	0.3	0.41
12	0.27	0.098	0.33	0.3	0.21	0.24	0.3	0.22	0.29	0.32	0.29	1	0.4	0.33	0.15	0.32	0.27	0.31	0.32	0.32	0.37
13	0.29	0.088	0.33	0.33	0.22	0.24	0.3	0.21	0.3	0.33	0.3	0.4	1	0.33	0.16	0.33	0.27	0.31	0.31	0.31	0.33
14	0.23	0.11	0.4	0.24	0.21	0.24	0.32	0.26	0.33	0.39	0.34	0.33	0.33	1	0.18	0.37	0.28	0.3	0.31	0.32	0.72
15	0.12	0.041	0.18	0.11	0.094	0.11	0.15	0.11	0.17	0.18	0.16	0.15	0.16	0.18	1	0.2	0.14	0.14	0.16	0.16	0.2
16	0.25	0.095	0.41	0.24	0.21	0.23	0.34	0.28	0.34	0.4	0.37	0.32	0.33	0.37	0.2	1	0.28	0.31	0.31	0.33	0.51
17	0.23	0.083	0.29	0.25	0.17	0.2	0.27	0.2	0.28	0.3	0.25	0.27	0.27	0.28	0.14	0.28	1	0.27	0.27	0.26	0.3
18	0.28	0.069	0.32	0.28	0.19	0.22	0.29	0.21	0.3	0.3	0.28	0.31	0.31	0.3	0.14	0.31	0.27	1	0.32	0.34	0.34
19	0.23	0.086	0.33	0.27	0.21	0.21	0.29	0.21	0.3	0.31	0.28	0.32	0.31	0.31	0.16	0.31	0.27	0.32	1	0.37	0.36
20	0.28	0.078	0.33	0.27	0.21	0.22	0.29	0.23	0.29	0.33	0.3	0.32	0.31	0.32	0.16	0.33	0.26	0.34	0.37	1	0.36
21	0.27	0.12	0.71	0.25	0.21	0.26	0.36	0.3	0.35	0.46	0.41	0.37	0.33	0.72	0.2	0.51	0.3	0.34	0.36	0.36	1

Fig. 2. Correlation between variables of the data set

As depicted before in this research, 10,000 subscriber details with 17 attributes and one target variable were incorporated for deep analysis. With the randomly selected datasets, 70% was included for training and 30% was taken for testing. In Table I and Table II "Evaluation Metric" results and "Confusion Metric" results of various models are presented.

TABLE I. PERFORMANCE OF CLASSIFIERS AGAINST CERTAIN EVALUATION METRICS

Models	Accuracy (%)	MSE (%)	ASE (%)	Precision (%)	Recall (%)	F-1 (%)
Decision Tree	75.00	25.00	25.00	74.87	73.75	74.31
Neural Network	82.40	17.60	17.60	75.45	96.14	84.55
Logistic Regression	77.36	22.63	22.63	73.23	84.84	78.61
Random Forest	78.50	21.50	21.50	74.49	85.38	79.56
XGBoost	82.90	17.10	17.10	74.48	99.04	85.03
AdaBoost	82.03	17.96	17.96	74.34	96.73	84.07
SVM	79.76	20.23	20.23	72.61	94.28	82.04

TABLE II. CONFUSION METRICS OF THE CLASSIFIERS

Models	Confusion Matrix			
	TN	FP	FN	TP
Decision Tree	1165	364	386	1085
Neural Network	1027	470	58	1445
Logistic Regression	1073	456	223	1248
Random Forest	1099	430	215	1256
XGBoost	**1030**	**499**	**14**	**1457**
AdaBoost	1038	491	48	1423
SVM	1006	523	84	1387

In addition to the results obtained from the aforementioned tables, the K-fold cross-validation technique was performed with 5 folds to find the best technique to deal with any proportion of testing and training data. (Fig 3)

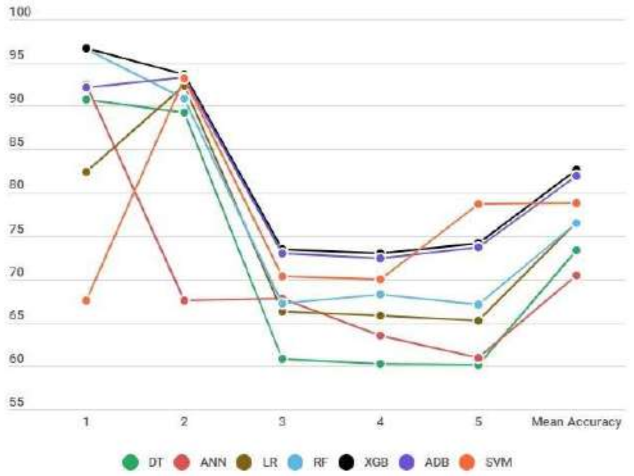


Fig. 3. Five-Fold cross-validation accuracies of selected algorithms

The aforementioned graph shows the accuracies obtained by the algorithms for each fold along with mean accuracy. All in all, the accuracy declines after the 2nd fold over the period up to the 3rd fold, whereas the accuracy shows some optimum level with ensemble techniques. Even though there is a slight increase in accuracy shown by SVM at the 5th fold, the highest mean accuracy is obtained only by XGBoost.

As a final attempt, it was noted and ensured that with all sorts of analyses completed so far, XGBoost has surpassed all other techniques with higher and better accuracy which is 82.90 % showing a low error rate of 17.1 %. Therefore, a more confirmed assumption can be made that the XGBoost is the most suitable algorithm in the course of attaining the final and ultimate model of prediction. Furthermore, to obtain an even better accuracy out of what already received, Grid Search CV hyper-parameter tuning was done with XGBoost and this effort has brought about 83.13% of accuracy in the forecasting of churning behavior. It is also well noted that during the process of tuning, “Learning Rate”, “Max Depth”, “Min_Child_Weight”, “Gamma” and “Colsample By Tree” were taken into attention because hyper-parameters and tuning had been carried out with optimum values.

It can be concluded that the ultimate prediction model was designed with the use of XGBoost along with a GUI implementation and it is sufficiently efficient to forecast the churners and non-churners relatively earlier.

V. Conclusion and Future Works

It can be thus concluded that forecasting the attrition behavior on the part of the clients is compulsory for almost all the manufacturing or service sectors. Start-ups too can be included in this importance. As an experimental procedure, the management of the churning attitude of the clients in the telecommunication industry has been effectively assessed in the process of this research. As it has become severely competitive, there is a tendency on the part of the users of the communication facilities to leave certain service providers and opt for another one. Therefore, as mentioned above, this model will enable an organization to identify the possibility of the churning and the reasons for such behavior on time and in advance. Further in this investigation, an attempt had been made to employ several supervised machine learning techniques. As a result of this effort, evaluation has been successfully carried out in connection with various performance metrics.

Besides, this study has made an effort to adapt several supervised machine learning techniques here and so, the evaluation was easily done concerning different performance metrics. Besides, a thorough comparison incorporating the algorithms such as Decision tree, Neutral Networks, Logistic Regression, Random Forest, SVM, XGBoost, and AdaBoost, has duly been done. Subsequently, correct cross-validation on these algorithms was also performed by utilizing a very suitable method to develop the most expected final prediction model in this field. As a reward, a relatively increased accuracy of 82.90% was produced by XGBoost, and at the same time, the least error rate of 17.10 % was gained by the XGBoost. As the last effort, a most practical prediction model was successfully created. The above effort has brought about the advantage of having the capability to predict the probability of the churners and non-churners very correctly. A further expansion of this model can be made available through the development of a combined model of churn prediction in the telecommunication industry. So, this particular model can be considered as the most appropriate prediction model which could be used in several other companies free of charge. Further, it was felt that when the hyper-parameter tuning was used before executing cross-validation, the accuracy shown by the prediction done in the supervised learning techniques would be greater. So, it is certain that anyone willing to carry out similar research in the future can depend on this effective churn prediction model when they attempt to combine or associate the supervised machine learning method and the unsupervised machine learning technique.

VI. Acknowledgement

Sincere gratitude goes to Mr. Siva Palaninathan, Senior Research Engineer from Canada for the valuable guidance and timely support in the completion of this research.

VII. REFERENCES

[1] Keramati and Ardabili. “Churn analysis for an Iranian mobile operator,” in Telecommunications Policy, pp.344-356, May 2011.

[2] Sharma, A., and Kumar Panigrahi, P. (2011). A Neural Network-based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications, 27(11), pp.26-31, August 2011.

[3] Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, and Aatif Kamal. “ Telecommunication Subscribers' Churn Prediction Model Using Machine Learning,” in IEEE International Conference on Digital Information Management (ICDIM), September 2013.

[4] Essam Shaaban, Yehia Helmy, and Ayman Khedr. “A Proposed Churn Prediction Model,” in International Journal of Engineering. Research and Applications (IJERA), Vol. 2, pp.693-697, 2012.

[5] A. Keramati, R. Jafari-Marandia, M. Aliannejadib, I. Ahmadianc, M. Mozaffaria and U. Abbasia,d. “Improved churn prediction in telecommunication industry using data mining techniques,” in Applied Soft Computing, pp. 994–1012, August 2014.

[6] Ying Huang and Tahar Kechadi. “An effective hybrid learning system for telecommunication churn prediction,” in Expert Systems with Applications, pp.5635–5647, 2013.

[7] Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., and Faris, H. (2015). Hybrid Data Mining Models for Predicting Customer Churn. International Journal of Communications, Network and System Sciences, 08(05), pp.91-96, April 9, 2015.

[8] Fathian, M., Hoseinpoor, Y. and Minaei-Bidgoli, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. Kybernetes, 45(5), pp.732-743, 2016.

[9] Georges D. Olle Olle and Shuqin Cai, " A Hybrid Churn Prediction Model in Mobile Telecommunication Industry," International Journal of e-Education, e-Business, e-Management and e-Learning vol. 4, no. 1, pp. 55-62, 2014.

[10] Preeti K. Dalvi, Siddhi K. Khandge and Ashish Deomore, Aditya Bankar. “Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression,” in Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[11] V. Umayaparvathi and K. Iyakutti, "Attribute selection and Customer Churn Prediction in the telecom industry," in International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, 2016, pp. 84-90.

[12] Gregory, Bryan. (2018).” Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data. “[Accessed: 20-Mar-2019].

[13] V. Umayaparvathi and K. Iyakutti, " A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics,” in International Research Journal of Engineering and Technology (IRJET), Vol. 3, Issue 04, April 2016

[14] J. Vijaya and E. Sivasankar, “Improved Churn Prediction Based on Supervised and Unsupervised Hybrid Data Mining System,” in Information and Communication Technology for Sustainable Development, 2017.

[15] Sahar F. Sabbeh, “Machine-Learning Techniques for Customer Retention: A Comparative Study,” in International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, 2018.

[16] O. Celik and U. O. Osmanoglu, “Comparing to Techniques Used in Customer Churn Analysis,” J. Multidiscip. Dev., vol. 4, no. 1, pp. 30–38, 2019.

Torque control using metaheuristic optimization for optimal energy consumption of a BLDCM

Rania Majdoubi
LCS Laboratory, faculty of sciences,
Mohammed V University in Rabat
rania_majdoubi@um5.ac.ma

Lhousasaine Masmoudi
LCS Laboratory, faculty of sciences,
Mohammed V University in Rabat
lhmasmoudi@gmail.com

Abderrahmane Elharif
Mechanical laboratory, faculty of
sciences, Mohammed V University in
Rabat
elharifa@gmail.com

***Abstract*— The emergence of electrical energy in the field of vehicle traction remains a major preoccupation of electro technicians today after a long absence, owing to the collective commitments and responsibilities of the states and automobile manufacturers as well as the advances made in various technological and scientific fields and more precisely the optimization of the electrical energy consumed by the motor. In this paper, we present a method of control of a Brushless Direct Current Motor (BLDCM) during the traction of an autonomous robot. This is done by optimizing the power consumed by the motor to find the optimal current and then find the optimal torque to function the machine. Then we optimize the energy of the traction to find the optimal slip ratio and the optimal linear velocity for an optimal operation that allows us to reuse the lost energy.**

***Keywords*— autonomous robot, BLDCM, optimization, optimal current, optimal slip ratio, power consumed, traction.**

I. Introduction

With a strong increase in demand for the use of electrical vehicles, the consumption of electrical energy has remained higher in recent years. Energy savings not only make it possible to quickly recover the additional cost of investing in electrical equipment (e.g. reactive energy compensation, electricity consumption monitoring devices or the use of a high-efficiency motor), but also to ensure the competitiveness of a company [1].

In this context, the problem of optimization comes to solve this challenge [2, 3]. Optimization problems in electrical engineering present several difficulties related to the user's needs [4, 5] (search for a global solution, reliability and precision of the solution, diversity of the problems treated, available calculation time, ...), as well as to the characteristics of the optimization problem [6, 7] (non-linear, derivatives not easily accessible, ...), and to the important calculation time. The resolution of such difficulties has been subject of numerous works using several optimization methods [8, 9]. In order to determine the most appropriate way to solve the problem of optimization of the energy consumed by the engine, it is necessary to inspect all the optimization methods to determine which ones are the most appropriate for the problems that the authors consider. the choice of the optimization method is made according to several criteria: the type of variables related to the problem, the knowledge of the mathematical model that describe the problem, the knowledge of the evolution of the objective function on the solution spaces, or the possibility of using a population of solutions to perform the optmal solution. Several optimization methods and algorithms are descussed in literature [10, 11]. Some authors use combinatorial optimization to solve the desired objective using either the exact method [12, 13] or the approximate method [14]. This last, uses either the heuristic method [15] or the metaheuristic method [16]. Other authors use continuous optimization [17] to solve the problem either through the linear method [18] or the nonlinear method which is based on the local gradient method [19] or the local gradient free method [20] or also the classical global gradient method [21] which brings us to the neighborhood or distributed methaheuristic method. Others use the evolutionary algorithm to solve the optimization problem using the genetic algorithm [22], also others use the swarm intelligence algorithm [23], and finally some authors use the hybrid optimization methods to solve a given problem [24], either using serial hybrid method, pararallel hybrid method or insertion hybrid method.

The objective of this paper is to optimize the energetic consumption of Brushless Direct Current Motor (BLDCM) using a cascading optimization. The first is relative to the optimization of the power used to generate the mechanical energy and the other is the optimization of the traction power used to navigate the agriculture robot « Agri Eco robot » [25, 26] in a rough environment, in order to gain the energy, and reuse it to aliment other equipment. In this perspective, we will present in the second paragraph the motor modeling, the power efficiency of the motor, and the traction efficiency of the wheel. The obtained relations in terms of power efficiency and traction efficiency are optimized in order to find the optimal current and optimal slip ratio that held to the optimal linear velocity for an optimal function of the robot, this will be discussed in the third part. The last paragraphe is dedicated to the result of this approach that will be implimented and validated using Matlab Simulink Software. At the end of this paper, we will summarize the results that we have reached, to be then reused in the next papers.

II. Description of the model

A. The motor modeling

The main objective of this paragraph is to develop a continuous model of the BLDCM [27]. The motor connection devices is given in the schematic as follows:

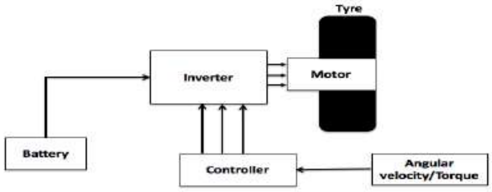


Fig. 1 Motor connection devices

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

It is assumed that the motor is governed by the following conditions:

- The three-phases stator are symmetric, this means that their resistances R and inductances L, are equal in the three phases;
- The induction field distribution created by the magnet is purely trapezoidal;
- The Electromotive Forces (EMFs) in the gap are trapezoidal distributed;
- The Magnetic Circuit is assumed to be unsaturated, which allows flux to be expressed as a linear current function;
- The Magnetic circuit is assumed to be perfecty laminated; this means that the hysteresis effect and the Foucault currents are neglected;
- Skin effect and temperature effect are all neglected.

The Electrical Equations that govern the operation of the BLDCM [28]. If the motor is without salience, the voltages at the terminals of the three stator phases are written according to Ohm's law:

$$\begin{bmatrix} V_a = Ri_a + L\frac{di_a}{dt} + e_a \\ V_b = Ri_b + L\frac{di_b}{dt} + e_b \\ V_c = Ri_c + L\frac{di_c}{dt} + e_c \end{bmatrix} \quad (1)$$

In which,

$$\begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} = p\Phi_m\omega \begin{bmatrix} tra(\Theta_e) \\ tra(\Theta_e - \frac{2\pi}{3}) \\ tra(\Theta_e + \frac{2\pi}{3}) \end{bmatrix}$$

$$tra(\Theta_e) = \begin{bmatrix} 1 & si\ 0 < \Theta_e < \frac{2\pi}{3} \\ 1 - \frac{6}{\pi}(\Theta_e - \frac{2\pi}{3})\ si & \frac{2\pi}{3} < \Theta_e < \pi \\ -1 & si\ \pi < \Theta_e < \frac{5\pi}{3} \\ -1 + \frac{6}{\pi}(\Theta_e - \frac{5\pi}{3})\ \ si & \frac{5\pi}{3} < \Theta_e < 2\pi \end{bmatrix}$$

where,

V_a, V_b, V_c : Voltages between the end of the phase winding and the middle of the Direct Current voltage source.

i_a, i_b, i_c: The stator phase currents.

e_a, e_b, e_c: The back-electromotive forces (back-EMF) in the stator phases.

Φ_m: The maximum flux produced at the stator.

p: The number of pole pairs.

Θ_e : The electrical position of the rotor that relate the mechanical position by the equation.

$$\Theta_e = p\Theta \quad (2)$$

The conversion of electrical energy into mechanical energy in the brushless direct current motor (BLDCM) is governed by the following relationship derived from the Theorem of Momentum:

$$C_{em} - C_r = f_v\omega + J\frac{d\omega}{dt} \quad (3)$$

where,

$$C_{em} = \frac{1}{\omega}(e_a i_a + e_b e_b + e_c e_c) \quad (4)$$

where,

J is the inertia of the rotor, the mechanical speed of rotation of the rotor,

C_{em} the motor torque provided by the stator,

C_r the load resisting torque,

f_v the viscous friction coefficient.

The projection of this equation into the two-phase frame is obtained using the projection into the Park frame whose equations is given by:

$$P(\Theta) = \frac{2}{3}\begin{pmatrix} \cos(p\Theta) & \cos(p\Theta - \frac{2\pi}{3}) & \cos(p\Theta + \frac{2\pi}{3}) \\ -\sin(p\Theta) & -\sin(p\Theta - \frac{2\pi}{3}) & -\sin(p\Theta + \frac{2\pi}{3}) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad (5)$$

The magnetic equation is defined as follows:

$$\Phi_d = L_d i_d + \Phi_m \quad (6)$$

$$\Phi_q = L_q i_q \quad (8)$$

Thus, the electrical equation is defined as follows:

$$V_d = Ri_d + L_d\frac{di_d}{dt} - p\ \omega L_q i_q \quad (9)$$

$$V_q = Ri_q + L_q\frac{di_q}{dt} + p\ \omega L_d i_d + p\omega\Phi_m \quad (10)$$

The mechanical equation is defined as follows:

$$C_{em} - C_r = f_v\omega + J\frac{d\omega}{dt} \quad (11)$$

In which,

$$C_{em} = \frac{3}{2}p\ i_q\Phi_m \quad (12)$$

where,

V_d,V_q: Voltages projected at Park's reference frame.

(d, q): Symbols refers to the direct and quadratic axis.

L_d , L_q: Direct inductance and quadratic inductance.

For the control of this type of motor, we apply the principle of vector control, which is identical to the control of a Direct Current motor with separate excitation. However, the Park reference frame mark must be used.

The d-axis component of the stator current acts as the excitation and allows the flux value in the machine to be adjusted. The q-axis component acts as the armature current and controls the torque.

Thus, the strategy consists in imposing the current i_q at a value corresponding to the desired torque C_{em} while maintaining zero current i_d, in order to work at maximum torque. The speed regulation is done in cascade by imposing the desired value of the speed on channel q.

In the following we will consider that L_d= L_q= L

B. The energitic efficiency of the motor

Let the energy distribution chain of the motor be defined as follows:

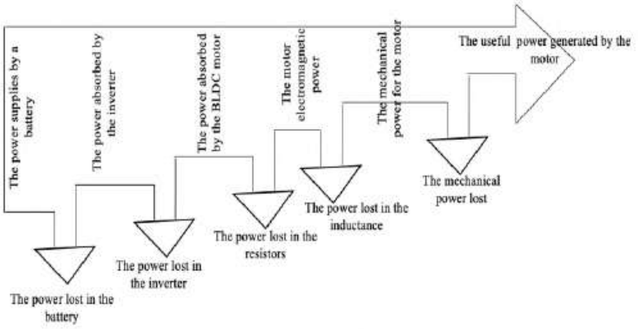

The power supplies by a battery
The power absorbed by the inverter
The power absorbed by the BLDC motor
The motor electromagnetic power
The mechanical power for the motor
The useful power generated by the motor
The power lost in the battery
The power lost in the inverter
The power lost in the resistors
The power lost in the inductance
The mechanical power lost

Fig. 2 The energy distribution chain obtained from the battery to run the engine

By neglecting the ohmic and inductive voltage across the stator windings, the following expressions are obtained:

$$P_c = i_s V_s \qquad (13)$$

Where,

$$i_s = \sqrt{i_d^2 + i_q^2} \leq i_n$$

$$V_s = \sqrt{(p\omega L i_q)^2 + (p\omega)^2(\Phi_m + L i_d)^2} \leq V_n$$

i_n: Maximum value of the current that can be absorbed by the BLDCM.

V_n: Maximum output voltage of the inverter depends on the DC bus voltage at its input.

Hence, the power consumed by the motor is defined as follows.

$$P_c = \sqrt{i_d^2 + i_q^2}\sqrt{(p\omega L i_q)^2 + (p\omega)^2(\Phi_m + L i_d)^2} \leq i_n V_n (14)$$

While regulating direct and quadratic current to work at maximum torque, we obtain i_d=0.

$$P_c = p\omega i_q \sqrt{(L i_q)^2 + (\Phi_m)^2} \leq i_n V_n = P_n \qquad (15)$$

In which, $\boldsymbol{P_n}$ is the nominal power of the motor.

The effective power required to function the motor is defined as follows:

$$P_u = C_{em}\omega \qquad (16)$$

Hence,

$$P_u = \frac{3}{4} p\, i_q \Phi_m \omega \qquad (17)$$

The energetic efficiency of the motor is defined in equation as follows.

$$\eta = \frac{P_u}{P_c} = \frac{1}{2} * \frac{\frac{3}{2}\Phi_m}{\sqrt{(L i_q)^2 + (\Phi_m)^2}} \qquad (18)$$

C. The traction efficiency of the wheel

Let the energy distribution chain of the wheel of the robot be defined as follows:

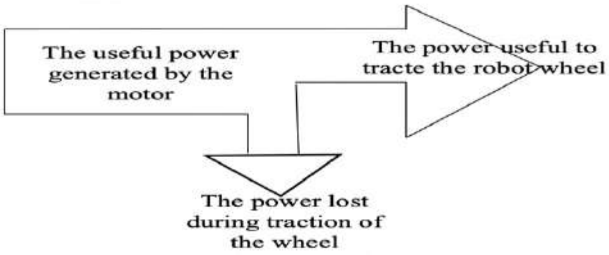

The useful power generated by the motor
The power useful to tracte the robot wheel
The power lost during traction of the wheel

Fig. 3 The energy distribution chain obtained from the rotation of the motor to move the wheel

We assume that the robot is towed by a single wheel defined as follows

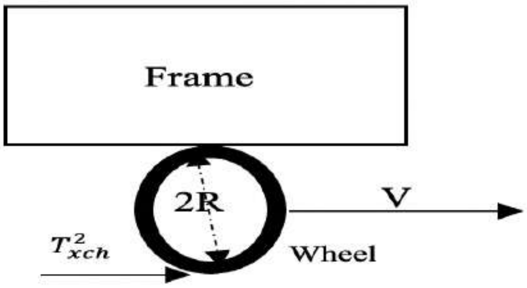


Fig. 4 The frame of the robot under one wheel

The power provided by the motor is defined as follows:

$$P_u = \frac{3}{4} p\, i_q \Phi_m \omega \qquad (19)$$

The power required for the traction of the vehicle:

$$P_t = T_{xch}^2 \times V \qquad (20)$$

With T_{xchi}^2 is the longitudinal tractive force [29], [30], [31]:

$$T_{xch}^2 \approx R w \tau_{m,t} \theta_f \qquad (21)$$

Where:

$$\tau_{m,t} = (C + (\left(\frac{k_c}{w} + k_\emptyset\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q \left(R\omega\theta_f\right)^q) \tan(\emptyset))(1 - \exp\left(\frac{R\,\theta_f s)}{K}\right) \qquad (22)$$

And:

$$\theta_f = \left(\frac{2^n m_t g}{R^n w\left(\frac{k_c}{w} + k_\emptyset\right)}\right)^{1/(2n+1)} \qquad (23)$$

V is the linear velocity of the wheel:

$$V = R\omega(1 - s) \qquad (24)$$

Where:

w is the width of the wheel

k_c is the cohesion module;

$k_\emptyset$ is the friction module;

n is the sinkage exponent;

K_q is the sinkage speed constant;

q is the soil propriety constant;

$\tau_{m,t}$ is the maximum of tangential shear stress.

C is cohesion;

ϕ is the angle of internal friction;

K is shear deformation modulus;

s is the slip ratio of the wheel.

Hence

$$\eta_t = \frac{P_t}{P_u} = \frac{4}{3} * \frac{R^2 w \tau_{m,t} \theta_f (1-s)}{p\, i_q \Phi_m} \qquad (25)$$

III. Optimization using metaheuristic

A. The optimal current and torque to optimize the energy consumed by the machine

Mathematically, the optimal problem is to maximize the machine's output for better machine efficiency

The objective function is mention as follows:

Maximize ($\eta = \frac{1}{2} * \frac{\frac{3}{2}\Phi_m}{\sqrt{(L i_q)^2 + (\Phi_m)^2}}$) involves minimizing $(f(i_q) = \sqrt{\left(L i_q\right)^2 + (\Phi_m)^2}\ )$

Hence, the heuristic function under study is defined as follows:

minimize $(f(i_q) = \sqrt{\left(L i_q\right)^2 + (\Phi_m)^2}\ )$

Subject to:

$$0 < i_q < i_n$$

$$V_s = p\omega\sqrt{\left(L i_q\right)^2 + (\Phi_m)^2} \leq V_n$$

Given variables:

$$\omega, i_q, V_n, i_n, L, \Phi_m$$

Once set all variables except i_q

There we find:

$$i_q < i_{qopt}$$

where,

$$i_{qopt} = \frac{1}{\sqrt{2}}\frac{\Phi_m}{L} \qquad (26)$$

B. The optimal slip ratio for an optimal control of the energy consumed by the motor

Mathematically, the optimal problem is to maximize the traction efficiency of the Brushless Direct Current motor inserted in the wheel for a better energetic efficiency of the wheel consumption.

Hence,

Maximize $\left(\eta_t = \frac{4}{3} * \frac{R^2 w \tau_{m,t} \theta_f (1-s)}{p\, i_q \Phi_m} \right)$ involves maximizing $(g(\mathrm{s}) = \tau_{m,t}(1-s))$

Hence, the heuristic function under study is defined as follows:

Maximize $(g(\mathrm{s}) = \tau_{m,t}(1-s))$

Subject to:

0< s < 1 While acceleration

Given variables:

$$C, k_c, k_\emptyset, \theta_f, \emptyset, K, \omega, K_q, \mathrm{R}, \mathrm{s}$$

Once set all variables except s

Hence, we find:

$$s_{opt} = {(K + (LambertW(0, \exp\left(\frac{R\,\theta_f}{K}\right)) + \frac{R\,\theta_f}{K} - 1))} \Big/ {R\,\theta_f} \quad (27)$$

From the point of view, the slip ratio of the wheel is defined as follows:

$$s_{opt} = \frac{R\omega - V_{opt}}{R\omega} \quad (28)$$

Hence, the linear velocity of the wheel is:

$$V_{opt} = R\omega(1 - s_{opt}) \quad (29)$$

IV. Result of the optimization design

A. Overal optimization design

The algorithm elaborated to optimize the energy consumption of the motor is illustrated in the flowchart as follows:

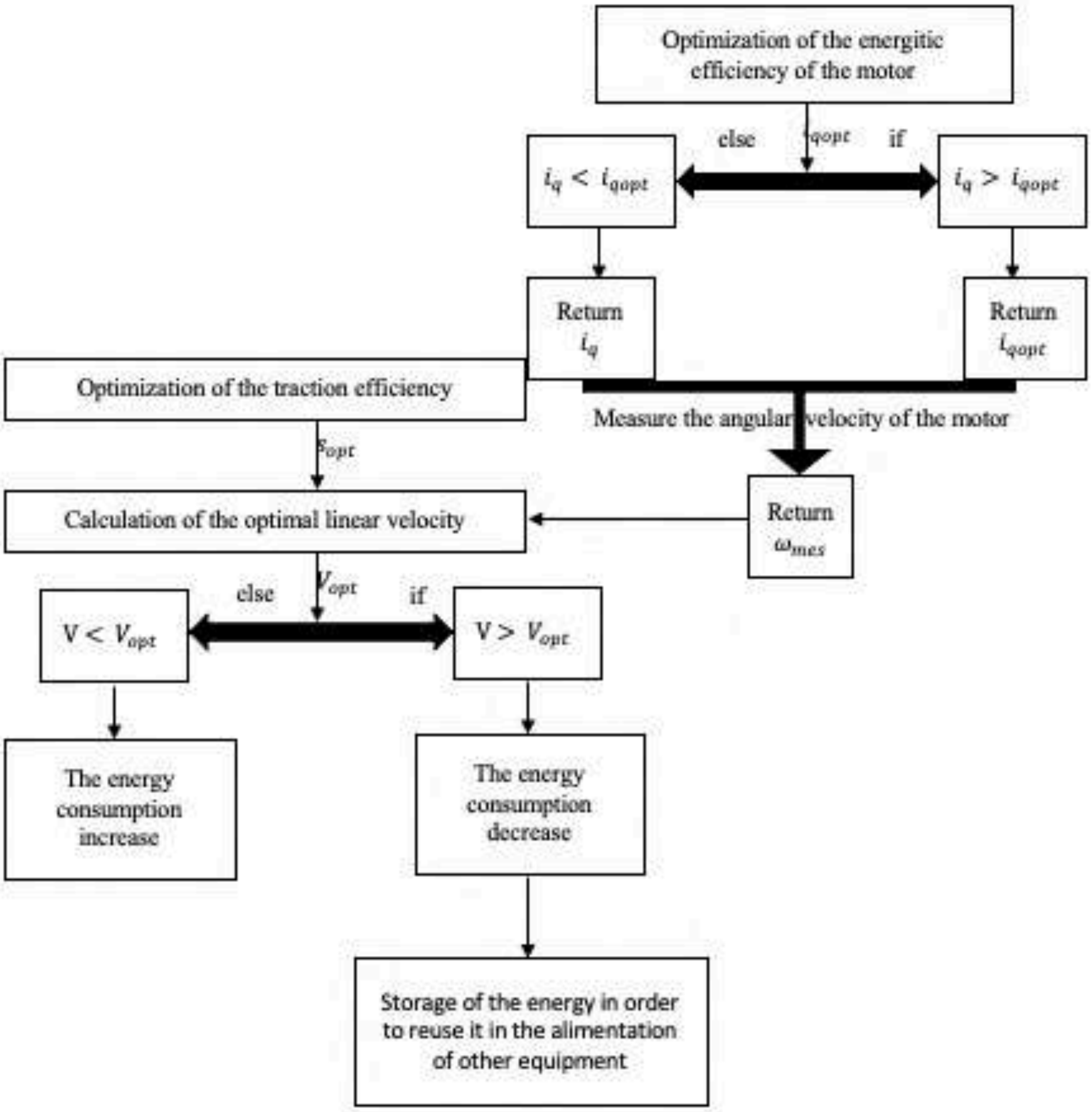

Optimization of the energitic efficiency of the motor
i_{qopt}
else
if
$i_q < i_{qopt}$
$i_q > i_{qopt}$
Return i_q
Return i_{qopt}
Optimization of the traction efficiency
Measure the angular velocity of the motor
s_{opt}
Calculation of the optimal linear velocity
Return ω_{mes}
V_{opt}
else
if
V < V_{opt}
V > V_{opt}
The energy consumption increase
The energy consumption decrease
Storage of the energy in order to reuse it in the alimentation of other equipment

Fig. 5 Flowchart of the optimization algorithm

B. *Result of the modeling system*

The mathematical modelling of the motor is validated under Matlab Simulink Software as shown as follows:

In which,

$$i_{qref} = \begin{Bmatrix} i_q & if\ i_{qref} < i_{qopt} \\ i_{qopt} & if\ \ i_{qref} > i_{qopt} \end{Bmatrix} \quad \text{and} \quad i_{dref} = 0 \quad /$$

To maximize the torque generated by the motor.

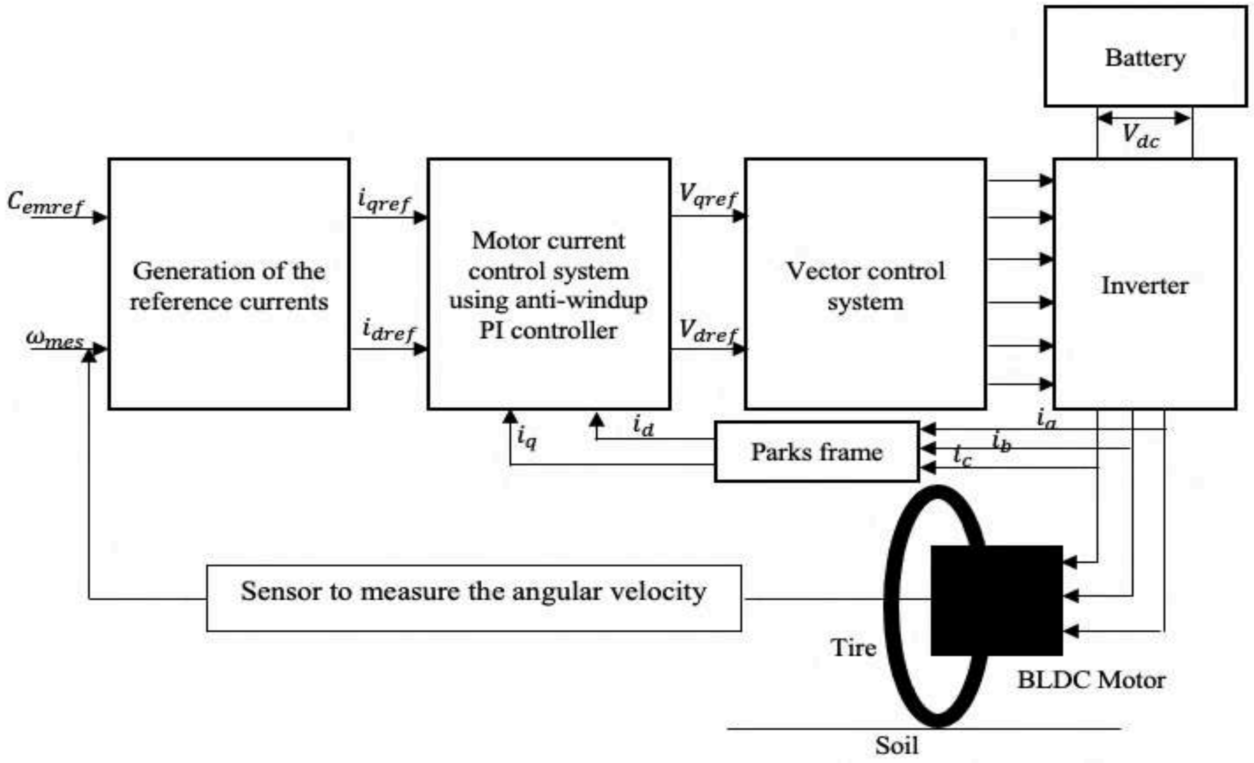


Fig. 6 The closed loop torque regulation of the BLDCM

The characteristics of the motor is defined in the table as follows:

TABLE I. BLDC MOTOR PARAMETERS

Nominal power (W)	250
Dc link voltage (V)	36
Maximum rotor-flux (Wb)	1.255 e-2
Viscous friction coefficient (N$mrad^{-1}s^{-1}$)	1.6e -3
Phase resistance (mΩ)	500
Phase inductance (mH)	0.68
Maximum speed (rpm)	3000
Pairs of poles	4
Moment of inertia (Kgm^2)	0.06

The soil parameters are defined as follows:

TABLE II. SOIL PARAMETERS

The cohesion module (KN/m^{n+1})	14.5
The friction module (KN/m^{n+2})	705.22
The sinkage exponent	0.36
The viscous friction coefficient (KPa)	4.76
The sinkage speed constant	0
The soil propriety constant	0.1
The angle of internal friction (Degree)	31.5
The shear deformation modulus (Cm)	1,2

The geometrical parameters of the robot wheel are defined in the table as follows:

TABLE III. WHEEL AND VEHICLE PARAMETERS

The mass of the vehicle (Kg)	84
The gravitational acceleration (m/s^2)	9.8
The wheel radius (m)	0.20
The wheel width (m)	0.05

Hence, the figures as follows gives us the result of the approach elaborated previously:

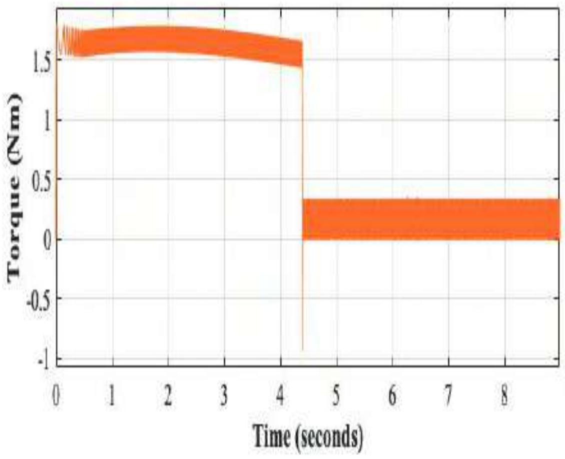


Fig. 7 *The torque imposed* during regulation

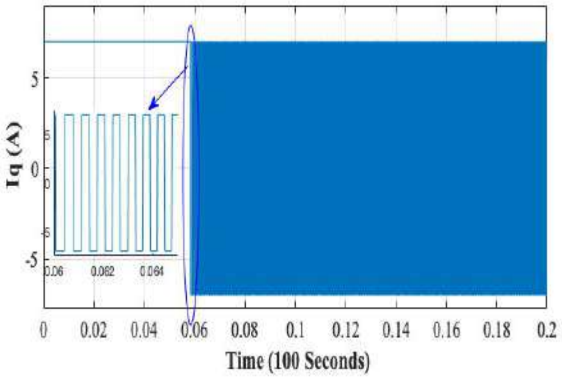


Fig. 8 The quadratic current response as a function of time

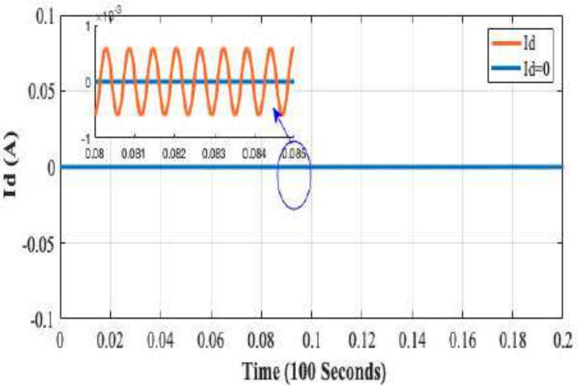


Fig. 9 The direct current during regulation to have maximum torque

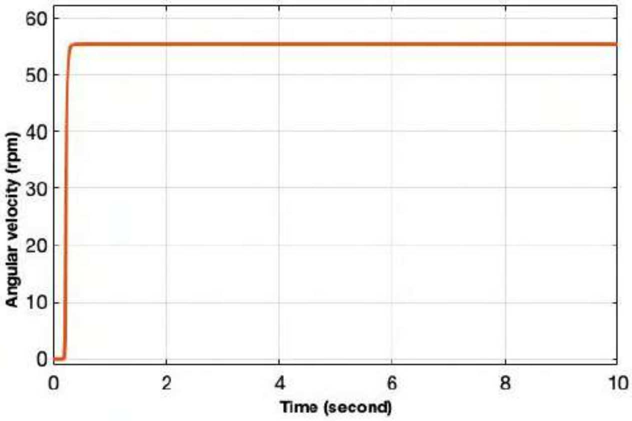


Fig. 10 The measured angular velocity of the wheel while controlling the torque

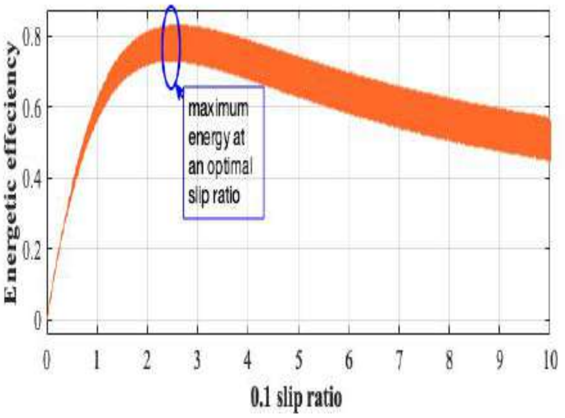


Fig. 11 The energetic efficiency of the motor as a function of slip ratio

From the figure 11, the optimal slip ratio is about 0.27, and the measured angular velocity in steady state is about 55 rpm, hence we obtain a linear velocity equal to 0.84 m/s. We can conclude that if we exceed a velocity equal to 0.84 m/s, the energy consumption decrease. Hence, we can storage the energy and reuse it to aliment other robot equipment.

To have a maximum torque we have regulate the direct current to be zero and the result of the regulation is appropriate.

V. CONCLUSION

In this paper, we propose an algorithm to optimize the energy consumption of the vehicle using a cascading optimization, once use the optimization of the current to control the torque generated by the motor, and the other use the optimization of the slip ratio to obtain an optimal velocity achieved with minimal energy consumed under different driving profiles. In this perspective, we have modelized the Brushless Direct Current Motor (BLDCM) with Trapezoidal Back-Electromotive Force (back-EMF) and its mathematical model, as well as the control of this kind of machine, using the park transformation to apply the algorithm proposed for the optimization. This assumption is validated using Matlab Simulink Software.

Optimizing the energy consumed while controlling the torque generated by the driven wheel of the robot navigating a deformable soil (while sliding) in real time, will be the subject of further works.

ACKNOWLEDGMENT

The authors of this paper are thankful to the Ministry of Higher Education and Scientific Research of Morocco (MESRSFC), and the National Center for Scientific and Technical Research of Morocco (CNRST) for financing this project.

REFERENCES

[1] F. willemien. visser@TELECO.-P. f. Willemien VISSER LTCI (Laboratoire commun en Traitement et Communication de l'Information), (2017, October), UMR 5141 CNRS-TELECOM ParisTech INRIA (Institut National de Recherche en Informatique et Automatique, *Willemien VISSER LTCI (Laboratoire commun en Traitement et Communication de l'Information)*. 2017.

[2] W. Tarnowski, "Present-day problems and methods of optimization in mechatronics," *Acta Mech. Autom.*, vol. 11, no. 2, pp. 154–165, 2017, doi: 10.1515/ama-2017-0024.

[3] P. Duysinx and O. Br, "Optimization of mechatronic systems : application to a modern car equipped with a semi-active suspension," *6th World Congr. Struct. Multidiscip. Optim.*, no. June, pp. 1–10, 2005.

[4] J. Jiang, G. Ding, J. Zhang, Y. Zou, and S. Qin, "A Systematic Optimization Design Method for Complex Mechatronic Products Design and Development," *Math. Probl. Eng.*, vol. 2018, pp. 1–14, 2018, doi: 10.1155/2018/3159637.

[5] M. Yoshimura and A. Takeuchi, "Concurrent Optimization of Product Design and Manufacturing Based on Information of Users' Needs," *Concurr. Eng.*, vol. 2, no. 1, pp. 33–44, 1994, doi: 10.1177/1063293X9400200104.

[6] D. N. Kumar, "Introduction and Basic Concepts : Optimization Problem and Model Formulation," *Optim. Methods*, no. 2, 2016.

[7] H. Mosallaei and Y. Rahmat-Samii, “Nonuniform Luneburg and two-shell lens antennas: Radiation characteristics and design optimization,” *IEEE Trans. Antennas Propag.*, vol. 49, no. 1, pp. 60–69, 2001, doi: 10.1109/8.910531.

[8] C. Pil and H. Haruhiko, “a a a,” vol. 1, no. 3, pp. 191–203, 1996.

[9] V. Zhmud, L. Dimitrov, and O. Yadrishnikov, “Calculation of regulators for the problems of mechatronics by means of the numerical optimization method,” *2014 12th Int. Conf. Actual Probl. Electron. Instrum. Eng. APEIE 2014 - Proc.*, pp. 739–744, 2015, doi: 10.1109/APEIE.2014.7040784.

[10] Y. He and J. McPhee, “Multidisciplinary design optimization of mechatronic vehicles with active suspensions,” *J. Sound Vib.*, vol. 283, no. 1–2, pp. 217–241, 2005, doi: 10.1016/j.jsv.2004.04.027.

[11] F. Roos, H. Johansson, and J. Wikander, “Optimal selection of motor and gearhead in mechatronic applications,” *Mechatronics*, vol. 16, no. 1, pp. 63–72, 2006, doi: 10.1016/j.mechatronics.2005.08.001.

[12] J. Puchinger and G. R. Raidl, “Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification,” *Lect. Notes Comput. Sci.*, vol. 3562, no. PART II, pp. 41–53, 2005, doi: 10.1007/11499305_5.

[13] M. Gasse, D. Chételat, N. Ferroni, L. Charlin, and A. Lodi, “Exact combinatorial optimization with graph convolutional neural networks,” *arXiv*, no. NeurIPS, 2019.

[14] C. Stummer, “Comment on ‘Approximative solution methods for multiobjective combinatorial optimization’ (Ehrgott/Gandibleux),” *Top*, vol. 12, no. 1, 2004.

[15] A. H. Halim and I. Ismail, “Combinatorial Optimization: Comparison of Heuristic Algorithms in Travelling Salesman Problem,” *Arch. Comput. Methods Eng.*, vol. 26, no. 2, pp. 367–380, 2019, doi: 10.1007/s11831-017-9247-y.

[16] W. B. Qiao and J. C. Créput, “Massive 2-opt and 3-opt moves with high performance gpu local search to large-scale traveling salesman problem,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11353 LNCS, pp. 82–97, 2019, doi: 10.1007/978-3-030-05348-2_8.

[17] X. S. Yang, “Multiobjective firefly algorithm for continuous optimization,” *Eng. Comput.*, vol. 29, no. 2, pp. 175–184, 2013, doi: 10.1007/s00366-012-0254-1.

[18] Q. Xiong and A. Jutan, “Continuous optimization using a dynamic simplex method,” *Chem. Eng. Sci.*, vol. 58, no. 16, pp. 3817–3828, 2003, doi: 10.1016/S0009-2509(03)00236-7.

[19] M. Arakawa and I. Hagiwara, “Nonlinear Integer, Discrete and Continuous Optimization Using Adaptive Range Genetic Algorithms,” no. November, 2014.

[20] J. Lampinen and I. Zelinka, “Mixed integer-discrete-continuous optimization by differential evolution,” *Proc. 5th Int. Conf. Soft Comput.*, pp. 71–76, 1997.

[21] G. CORRIEU, M. LALANDE, and C. ROUSSEL, “Méthode simplifiée pour calculer la récupération optimum d’énergie sur un pasteurisateur de lait à plaques,” *Lait*, vol. 61, no. 605–606, pp. 233–249, 1981, doi: 10.1051/lait:1981605-60614.

[22] S. Jin, M. Zhou, and A. S. Wu, “Sensor network optimization using a genetic algorithm,” *7th World Multiconference Syst. Cybern. Informatics*, pp. 1–6, 2003, [Online]. Available: http://www.cs.ucf.edu/~ecl/papers/0307.sci.sjin.pdf.

[23] C. Blum and X. Li, “Swarm Intelligence in Optimization.”

[24] B. Yassin, A. Lahcen, and E. S. M. Zeriab, “Hybrid optimization procedure applied to optimal location finding for piezoelectric actuators and sensors for active vibration control,” *Appl. Math. Model.*, vol. 62, pp. 701–716, 2018, doi: 10.1016/j.apm.2018.06.017.

[25] R. Majdoubi, L. Masmoudi, and A. Elharif, “Mechanical Study and Design of An Ecological Mobile Agriculture Robot,” presented at The International Conference on Micro and NanoSatellites., Rabat, Morocco, 2018.

[26] R. Majdoubi, L. Masmoudi, A. Elharif, and M.K. Ettouhami ,“Etude et conception d’un robot mobile écologique dédié à la pulvérisation des fraises sous serre,” presented at the Journées d’Etudes *Tech. 2018 l’Association Fr. Mécanique des Matériaux*, 2018.

[27] Majdoubi, R., Masmoudi, L., Bakhti, M., Elharif, A., Jabri, B. (2020). Parameters estimation of bldc motor based on physical approach and weighted recursive least square algorithm. Int. J. Electr. Comput. Eng., 11(1): 133-145. https://doi.org/10.11591/ijece.v11i1.pp133-145.

[28] Majdoubi, R., Masmoudi, L., Bakhti, M., Jabri, B. (2021). Torque control oriented modeling of a brushless direct current motor (BLDCM) based on the extended park's transformation. Journal Européen des Systèmes Automatisés, Vol. 54, No. 1, pp. 165-174. https://doi.org/10.18280/jesa.540119

[29] K. Yoshida and G. Ishigami, “Steering characteristics of a rigid wheel for exploration on loose soil,” *2004 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, vol. 4, no. August, pp. 3995–4000, 2004, doi: 10.1109/iros.2004.1390039.

[30] L. Ding, H. Gao, Z. Deng, K. Yoshida, and K. Nagatani, “Slip ratio for lugged wheel of planetary rover in deformable soil: Definition and estimation,” in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, Dec. 2009, pp. 3343–3348, doi: 10.1109/IROS.2009.5354565.

[31] R. Majdoubi and L. Masmoudi, "Eco-design of a mobile agriculture robot based on classical approach and FEM creteria," *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, 2021, pp. 978-982, doi: 10.1109/ICCCIS51004.2021.9397234.

Mechatronics Application for a Smart Inhaler

Patricia Enrique
University of Waterloo
Waterloo, Canada
psenriqu@uwaterloo.ca

Dorothy Yang
University of Waterloo
Waterloo, Canada
dcyang@uwaterloo.ca

Matthew Rose
University of Waterloo
Waterloo, Canada
mtrose@uwaterloo.ca

Steven Ding
University of Waterloo
Waterloo, Canada
steven.ding@uwaterloo.ca

James Tung
University of Waterloo
Waterloo, Canada
james.tung@uwaterloo.ca

Andrew Kennings
University of Waterloo
Waterloo, Canada
akennings@uwaterloo.ca

***Abstract*—Metered dose inhalers (MDI) are used to manage or to provide quick relief to asthmatics. The improper use of an inhaler can result in the administration of an incorrect medication dosage, reducing the effectiveness of the device and resulting in higher treatment costs for the patient. The design process of an inhaler sleeve is outlined, intended to assist in the use of a standard MDI.**

The purpose of the sleeve is to monitor the user's technique during the four most critical steps of their inhaler use. These steps include shaking the device before use, coordinating the canister actuation with their inhalation, inhaling at the correct rate, and holding their breathing for the correct amount of time.

The final design consists of a 3-D printed sleeve and a mobile application that guide the patient through the process. Lights are integrated into the sleeve indicating to the user which step they are on and for how long. Sensors are able to monitor the inhaler throughout its use and provide feedback by sharing the data collected to a mobile app. From there, users are shown what steps they are preforming correctly, and which ones need further improvement.

A prototype printed in polylactic acid (PLA) acts as a reference for the accuracy of the 3-D model created. A finite element analysis conducted on the model indicates that the PLA material has sufficient strength for the product; however, opting to print in polyethylene terephthalate glycol (PETG) will allow the inhaler to fit into the sleeve more easily and will reduce the concentrated stresses and deformations seen on the sleeve.

Keywords*— *3-D printing, asthma, inhalation, inhaler, mechatronics, medicine, monitor, puffer, respiration.

I. INTRODUCTION

Asthma is a chronic disease that impacts the airways in the lungs, causing inflammation and constricting the bronchial tubes which restricts the amount of air that can reach the lungs. It affects 1 in 13 people worldwide and is the leading chronic disease in children [1]. Asthma can have several triggers such as exercise, strong scents, or allergic triggers including pollen or pet dander [2]. An asthmatic encountering these triggers can experience asthma symptoms such as wheezing, shortness of breath and chest tightness [3].

The symptoms can be managed using long-term control and quick-relief asthma medications. The medications are dispensed using one of two types of inhalers: the metered dose inhaler (MDI), or the dry powder inhaler (DPI). The most commonly used inhaler is the MDI which is illustrated in Fig. 1 along with its corresponding components [4]. When the user presses on the canister, the medicine is dispensed out of the MDI as a measured puff of fine mist that is then inhaled [5].

Research reveals that up to 94% of asthmatics demonstrate incorrect techniques when using their MDI [6]. Improper inhaler techniques can lead to the administration of an incorrect medication dosage, reducing the benefits of the inhaler. Further, this creates an economic loss due to wasted medication and increased emergency hospital visits for higher cost treatments. The estimated annual losses in Ontario, Canada alone are approximately $141 million [7].

The purpose of this paper is to present a device capable of monitoring the four most common mistakes practiced by asthmatics when using and MDI. These mistakes are as follows [8]:

- Inadequate shaking and mixing of inhaler before use.
- Failure to coordinate actuation of the MDI or canister press on inhalation.
- Incorrect inspiratory flow rates.
- Too short a period of breath holing after inhalation.

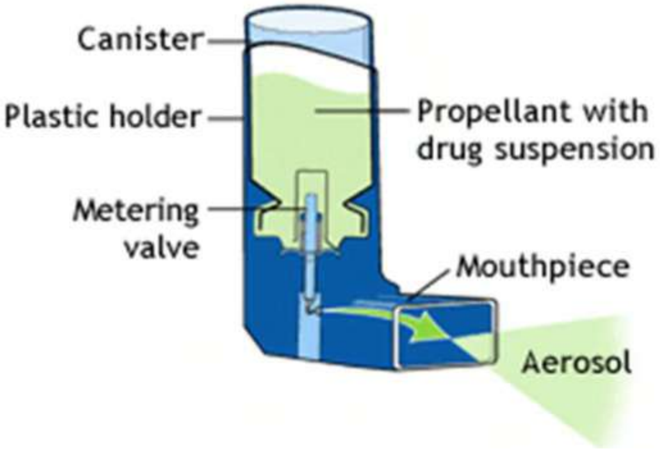


Figure 1. Labelled components of a metered dose inhaler (MDI) [4]

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

The main cause for improper inhaler use technique is the ineffective education practices currently in place. Patients learn how to use an MDI through diagrams, instructions and demonstrations by their healthcare provider. After receiving the demonstration, it is up to the patient to continue to practice proper inhaler techniques.

Consequently, a need exists to provide a solution that helps asthma patients practice consistently proper inhaler techniques on their own to increase the effectiveness of asthma medications.

The developed design is able to monitor the four mistakes mentioned previously and further provide feedback to the user. The feedback on the inhaler use technique is provided to the user both through real-time feedback and after their MDI use. The main goal of the sleeve is to provide a more satisfactory user experience and to reduce the wasted medication.

II. Proposed Solution

A. Constraints and Criteria

The main constraints for the designed device are outlined in Table I and are related to the ability to monitor the inhaler use mistakes and the ergonomics of the design. In addition to the constraints, several criteria are also considered. The product usability determines the chances that a patient will adopt the device. Creating a design that is simple and straightforward to use motivates an asthmatic to frequently use the product and develop the proper inhaler techniques.

The discreetness of the device further encourages its use by the user. The patient should feel comfortable using the product when needed despite their environment. This consideration lends to avoiding designs that are loud and obtrusive which may draw the attention of people nearby.

Implementation and the environmental impact of the proposed design will determine the success of the final product. A device that is easily adapted to current MDIs is more accepted by patients and will result in less environmental waste than developing a unique inhaler. Developing a reusable design will further minimize the environmental impact of the product.

To determine the optimal design for the problem presented, various solutions are compared using decision matrices, and one is selected to meet the constraints and criteria outlined. Each matrix lists the options in the left-hand column and the criteria in the top row. Each criterion is associated with a weight depending on its importance. The options are scored in terms of the criteria with a high score indicating a good option. The option with the highest total after scoring all the criterion is determined to the optimal choice.

TABLE I. Constraints in consideration for the device

Characteristic	*Value*	*Comments*
Product cost	≤ $150 CAD	Based on similar products currently on the market
Device mass with inhaler	≤ 250 gr	Inhaler accounts for 30 gr
Device Dimensions	≤ 10x10x10 cm	Maximum device volume is 1000 cm^3
Battery life	≥24 hrs	
Shaking acceleration detection	≤ 80 m/s^2	Required values to provide the user with valuable feedback
Inhalation speed detection	≤ 100 L/min	
Actuation force detection	≤ 50 N	

B. Design Options

For the overall design, three possible options are considered. The first option is a full inhaler redesign, replacing the current MDI, where the medication canister alone is inserted into. The second is a sleeve attachment that is designed around the standard MDI. Finally, a cap attachment is considered that connects to the top of the standard MDI. Comparing the three designs using the decision matrix shown in Table II, results in the sleeve attachment as the optimal option. This design is compatible with existing MDIs and more compact than the cap, providing additional space for sensors.

C. User Feedback

Four possible feedback methods are considered to provide the patient with real-time feedback. Visual feedback is achievable through the use of lights or a display screen, auditory feedback is limited to a speaker, and haptic feedback is provided through a vibration motor. The optimal user feedback method is determined from the decision matrix in Table III and results in the visual feedback method. This method is more compact and provides minimal public disturbances.

D. Shake Detection

Proper inhaler shaking technique verification is determined through the use of an accelerometer and gyroscope to measure the acceleration of the device. The benefits of this detection method is its accuracy, compact size, and its availability in premade development boards.

E. Inhalation Detection

A method to detect inhalation is necessary to ensure that the user is inhaling at the appropriate rate and for the correct duration. Two options are compared to detect inhalation: an air flow sensor, and a differential pressure sensor. Both options make use of the hollow MDI body by attaching to the top of the sleeve and allowing air to flow through the inhaler body to the sensor. The differential pressure sensor is selected by using the decision matrix in Table IV due to the size and expected cost.

F. Canister Actuation

Three options are compared to ensure proper canister actuation and coordination, each are further categorized as automatic or manual actuation. The automatic options actuate the canister for the patient when inhalation is detected to ensure correct actuation timing. This method is completed by using a motor or a spring to press down the canister. The manual option indicates to the user when to actuate using the selected feedback method. Actuation is detected through the use of a thin-film pressure sensor located in the bottom of the device which reads the actuation force. The optimal actuation method is determined by using the decision matrix seen in Table V. The manual option, using an indicator and feedback, resulted in the optimal choice due to the size, cost, and ease of implementation.

III. Detailed Design

The final design integrates the selected sensors into a sleeve to monitor the user's inhaler technique and to provide feedback during and after its use.

A. *Detection of Common User Mistakes*

Before using the MDI, the patient must properly shake the device to mix the medication in the canister. Failure to properly complete this step can result in the inaccurate administration of the medication and reduce the effectiveness of the MDI. To validate proper inhaler shaking technique, an accelerometer and gyroscope are integrated into the design. This sensor combination is used to recognize that the inhaler is being shaken with an appropriate amount of force and with the correct orientation. Experimental tests consisting of shaking an MDI are completed to determine the threshold of force that indicates that the inhaler is properly being shaken. An accelerometer and gyroscope are attached to an MDI before being shaken by hand to collect the results summarized in Fig. 2. The test is repeated multiple times with little deviation from the data presented. It can be seen that a reading of ± 60 m/s^2 indicates that the inhaler is shaken correctly with peaks of ± 8 m/s^2. The sensors must be calibrated to indicate that the user is shaking the inhaler at the correct acceleration by utilizing an accelerometer that can measure an acceleration of at least 8 m/s^2.

Following the shaking of the MDI, the user is required to simultaneously actuate the inhaler and breathe in the medication. Two common mistakes occur at this step: failure to coordinate the actuation of the canister and inhalation, and failure to inhale at the correct speed. These errors can result in the incorrect dose administration and in the incorrect speed at which the medication reaches a user's lungs. To prevent these common mistakes, two additional sensors are implemented. The first is a thin film pressure sensor which recognizes when and for how long the canister is actuated. Reasearch has identified that the force required to actuate the canister to be around 40N [9]. The thin film pressure sensor is integrated into the sleeve design to measure a force between 30N to 40N indicating that the patient is administering the medication.

During the actuation of the canister, the user is required to breath in the medication slowly for 5 seconds. The optimal rate of inhalation is between 20-30L/min [10] and needs to be monitored by the differential pressure sensor. Due to the small pressure difference, an orifice plate is designed to generate a more dramatic pressure difference that is then detectable by the differential pressure sensor. The sensor should be able to detect a flow rate of 0-100L/min. Bernoulli's equation combined with information provided by [11] are referenced to determine the dimensions of the orifice plate required for the corresponding pressure difference. Once integrated into the sleeve, the orifice plate combined with the differential pressure sensor are able to detect the user's inhalation rate.

Following the inhalation of the medication, the patient is required to hold their breath for 10 seconds to ensure the medication reaches their lungs and is given time to work. A simple timer integrated into the design is able to indicate to the user when the time has been reached.

TABLE II. Design Options Decision Matrix

	Integration with Existing Market	***Size***	***Ease of Implementation***	***Reusability***	***Total***
Weight	0.4	0.3	0.15	0.15	1
Housing	1	3	1	3	1.95
Sleeve	2	2	3	2	2.15
Cap	3	1	2	1	1.95

TABLE III. User Feedback Method Decision Matrix

	Public Disturbance	***Size***	***Ease of Implementation***	***Ease of Use***	***Total***
Weight	0.2	0.5	0.1	0.2	1
Lights	2	4	4	2	3.2
Screen	3	1	1	4	2
Speaker	1	3	3	3	2.6
Vibration	4	2	2	1	2.2

TABLE IV. Inhalation Detection Decision Matrix

	Cost	***Size***	***Ease of Implementation***	***Total***
Weight	0.2	0.5	0.3	1
Air flow sensor	1	1	2	1.3
Differential pressure sensor	2	2	1	1.7

TABLE V. Canister Actuation Decision Matrix

	Accuracy	***Size***	***Ease of Implementation***	***Cost***	***Total***
Weight	0.35	0.35	0.15	0.15	1
Motor	3	2	2	1	2.2
Indicator and Feedback	1	3	3	3	2.3
Spring	2	1	1	2	1.5

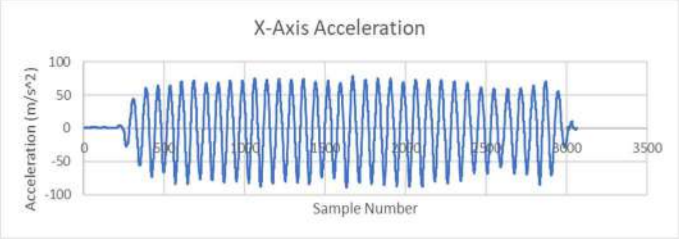


Figure 2. Summary of results gathered from threshold force experiments. The x-axis represents time over which samples are collected

B. *Power*

To power the electronic circuit, a 3.7V, 30mAh rechargeable battery is kept in low power mode until the user starts the device by pressing a button. Setting the standby mode to low power will allow the device to be ready for more than 70 days. Assuming that the product is used for 30 seconds every day, the battery will last for approximately 20 days before requiring to be recharged.

C. Providing Real Time User Feedback

Reinforcing proper inhaler techniques through real-time feedback is achieved by the use of 4 LED lights configured similar to that shown in Fig. 3. The first LED, shown in blue, indicates the step that the patient is on. Each step: shaking the inhaler, actuating the canister, breathing in, and holding the breath are represented by a different colour. This LED indicates to the user the amount of time that they should be spending on each step.

The remaining three LEDs indicate to the patient how well they are preforming the task. If the inhaler is being shaken too gently or the inhalation rate is too slow, the far left LED will turn on. Conversely, if the inhaler is being shaken too aggressively or the inhalation rate is too rapid, the far right LED will turn on. Ideally, the middle green LED will be turned on throughout the use of the device to indicate that the user is following proper MDI techniques.

D. Providing Post Inhaler Usage Feedback

Further improvement of inhaler use techniques is provided to the patient after each use of the product. This is achieved through the connected mobile application which has an appearance and user flow similar to that displayed in Fig. 4. Patients are able to connect their sleeve to the app on their phone which then provides an overview of how to correctly apply inhaler techniques. Also, an explanation of the sleeve is provided so that patients can confidently use the device. The data collected by the sensors on the sleeve design is shared to the app via a Bluetooth module. Patients can access the data of their recent uses to see how well they are applying the inhaler techniques. Additionally, when initially connecting the sleeve to the app, patients will be asked if the canister is new or not. If it is not new, the approximate dosage remaining can be entered so that the app can monitor when the canister is close to being empty and will need to be replaced.

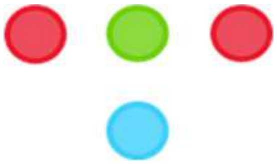
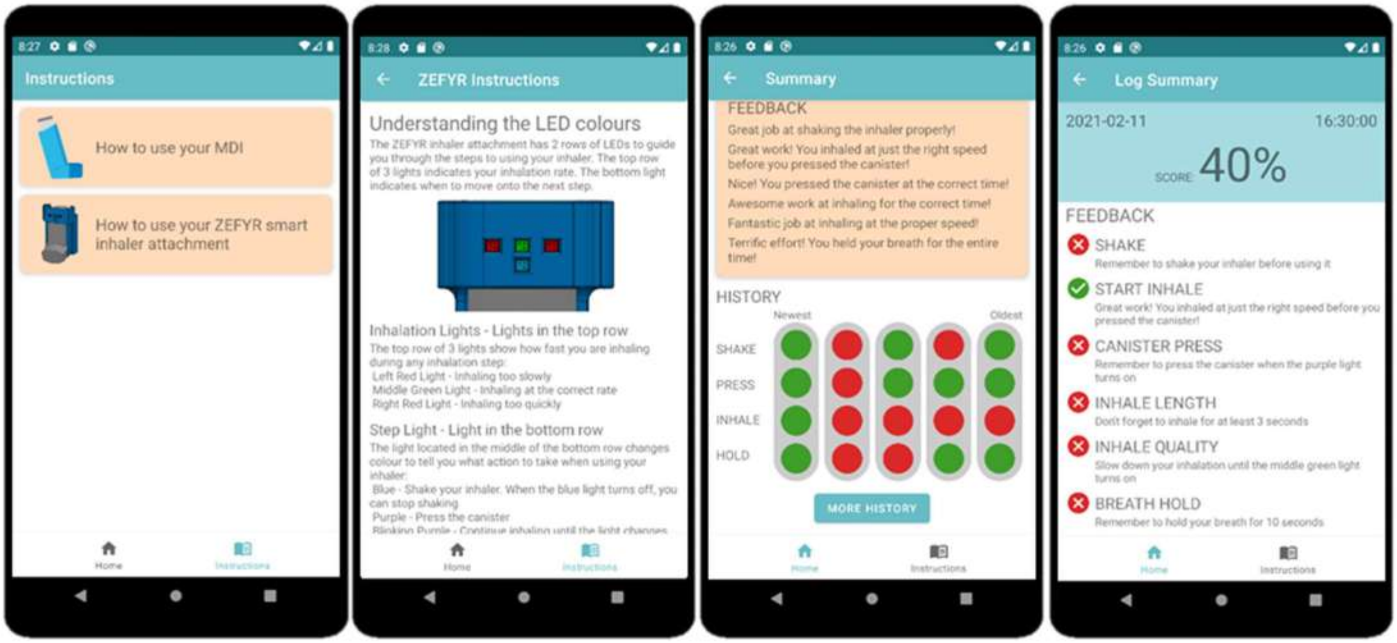


Figure 3. LED light configuration for real-time user feedback

E. Prototype

The final design that is able to integrate all the components necessary to meet the objectives and constraints outlined, is shown in Fig. 5 (a). The sleeve, shown in teal, is designed to fit onto the back, sides, and bottom of a standard MDI. The dimensions accommodate a fixed transition fit, allowing the patient to press the MDI into the sleeve without the possibility of the inhaler falling out. The shape is optimized for a comfortable use and to prevent the inhaler from being excessively bulky. The back of the sleeve is dropped below the top of the canister to allow the user to easily access the canister for actuation.

The approximate location of each sensor and component are identified in Fig. 5 (b) and (c). The development board with integrated Bluetooth module, accelerometer, and gyroscope, and the battery are placed on the back of the sleeve to prevent interference with the use of the inhaler. The touch pressure sensor is located between the sleeve and the MDI body under the inhaler. This is the optimal location as the force applied to the canister by the user is fully transferred onto the sensor. The differential pressure sensor and corresponding orifice plate are placed close to the top of the MDI body to detect the flow rate of the air flowing from the mouthpiece through the hollow MDI body and to the sensor. This location capitalizes on the hollow MDI body and allows the design to be more compact and not as bulky. The indication LEDs are implemented into the front of the sleeve to be visible for the patient during use.



Understanding the LED colours

The ZEFYR inhaler attachment has 2 rows of LEDs to guide you through the steps to using your inhaler. The top row of 3 lights indicates your inhalation rate. The bottom light indicates when to move onto the next step.

Inhalation Lights - Lights in the top row

The top row of 3 lights show how fast you are inhaling during any inhalation step:
Left Red Light - Inhaling too slowly
Middle Green Light - Inhaling at the correct rate
Right Red Light - Inhaling too quickly

Step Light - Light in the bottom row

The light located in the middle of the bottom row changes colour to tell you what action to take when using your inhaler:
Blue - Shake your inhaler. When the blue light turns off, you can stop shaking
Purple - Press the canister
Blinking Purple - Continue inhaling until the light changes

FEEDBACK

Great job at shaking the inhaler properly!
Great work! You inhaled at just the right speed before you pressed the canister!
Nice! You pressed the canister at the correct time!
Awesome work at inhaling for the correct time!
Fantastic job at inhaling at the proper speed!
Terrific effort! You held your breath for the entire time!

HISTORY

	Newest				Oldest
SHAKE	🟢	🔴	🟢	🔴	🟢
PRESS	🟢	🔴	🟢	🟢	🟢
INHALE	🟢	🔴	🔴	🔴	🔴
HOLD	🟢	🔴	🔴	🟢	🟢

2021-02-11 16:30:00

SCORE 40%

FEEDBACK

☒ SHAKE
Remember to shake your inhaler before using it

✓ START INHALE
Great work! You inhaled at just the right speed before you pressed the canister!

☒ CANISTER PRESS
Remember to press the canister when the purple light turns on

☒ INHALE LENGTH
Don't forget to inhale for at least 3 seconds

☒ INHALE QUALITY
Slow down your inhalation until the middle green light turns on

☒ BREATH HOLD
Remember to hold your breath for 10 seconds

Figure 4. Mobile application user flow available to patients after use

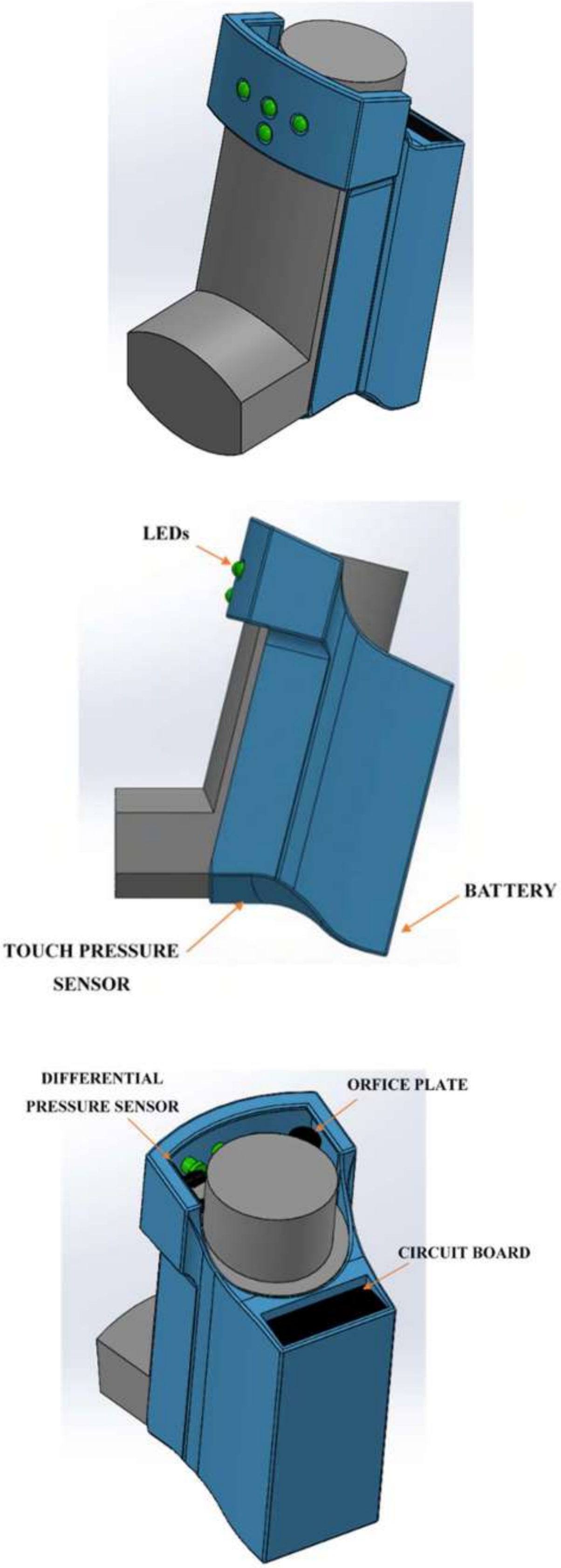


Figure 5. The final sleeve design; (a) final sleeve, (b) and (c) sensor and circuit board locations

A prototype of the sleeve is 3-D printed using PLA and can be seen in Fig. 6. The weight of the prototype with the MDI meets the constraint by falling below 250g. The sleeve fits the MDI properly and indicates appropriate sizing of the model. For future iterations, modifications will be made to increase the ergonomics of the design while maintaining the aesthetic and functionality. Further adjustments to the sensor locations will be made to optimize the sleeve space and to keep the design as compact as possible. Ideally, the final product will be 3-D printed using PETG to allow for more flexibility in the device.

IV. Experimental Results

To perform an analysis on the sleeve design, the model is imported into Abaqus CAE 2020 and the material properties of the product are applied. The original design is 3-D printed in standard PLA filament. This material has a Young's modulus of 4.0 GPA, a Poisson's ratio of 0.3, and a density of 1.3g/cm^3 [12].

The analysis is performed in two steps: the initial step and the loading step. The initial step applies the boundary conditions of the sleeve. The product is designed for the user to place their thumb in the curved profile at the bottom. An average thumb is able to cover the entire curvature allowing that surface to act as a fixed boundary condition. The rest of the sleeve is able to deform freely when the force is applied.

The second step entails applying the loads and forces to the model that will cause the deformations and stress concentrations. The canister that is actuated by the user contacts the MDI through a narrow nozzle. This allows the actuation force to be modelled by a concentrated force vector on the sleeve. To properly orient the force vector parallel to the sleeve walls, the base is partitioned into sections and a new datum is created. A visual representation of the applied force vector and boundary conditions can be seen in Fig. 7.

After defining the boundary conditions and the concentrated force, a mesh is able to be defined for the design. Due to the irregular shape of the sleeve, a tetrahedral element shape is used for the mesh. The fine meshed design is shown in Fig. 8 with a total of 2686 elements.

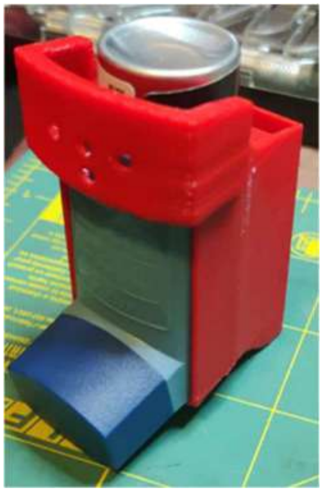


Figure 6. 3-D printed prototype of the final sleeve design model

To determine the effect of the material type on the deformation and stress concentration of the sleeve, a second model is generated. The set-up is identical except that the material properties applied correspond to PETG which has a Young's modulus of 2.0 GPA, a Poisson's ratio of 0.4, and a density of 1.27g/cm^3 [13].

The analysis conducted on the models made from PLA and PETG demonstrates a large deformation and stress concentration at the location where the concentrated force is applied. The values for the maximum stress and deformation in the sleeve depend on the type of mesh used for the analysis. In both cases, a coarse mesh and a fine mesh were used for comparison. The results of the analysis are summarized in Table VI. Since the design that the mesh is applied to remains the same for both models, the elements created for the fine and coarse meshes are the same. The elements in the coarse mesh analysis is 1075 and the elements in the fine mesh analysis is 2686.

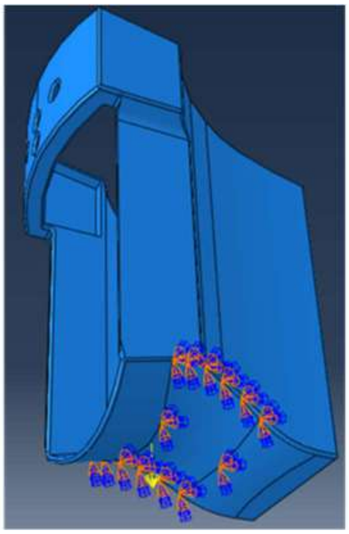


Figure 7. Force vector and boundary conditions applied to sleeve design model

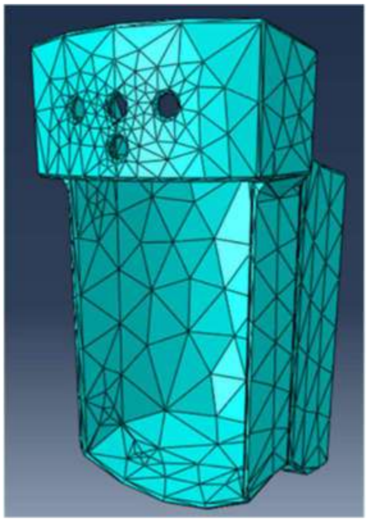


Figure 8. Fine mesh applied to the sleeve design model

TABLE VI. SUMMARY OF FINITE ELEMENT ANALYSIS RESULTS

Model Material: PLA			
Mesh	***Maximum Stress (MPa)***	***Maximum Displacement (mm)***	***Runtime (s)***
Fine	533	0.195	11
Coarse	136	0.0795	9
Model Material: PETG			
Mesh	***Maximum Stress (MPa)***	***Maximum Displacement (mm)***	***Runtime (s)***
Fine	0.375	0.000316	10
Coarse	0.0975	0.000129	9

The deformed model appears the same in both materials with varying magnitudes of stress and deformation. The contour plot for the stress concentration, percentage, and deformation, in millimeters, of the PLA design with a fine mesh is shown in Fig. 9. A deformation scale factor of 47 is applied to visualize the results.

The study conducted demonstrates that the sleeve design made from PETG is best suited to reduce the deformation and stress concentration in the product. To prevent failure from the concentrated force of actuating the MDI canister, the final sleeve should be produced using PETG. This will assist in developing a solution to the improper inhaler use that is durable and effective.

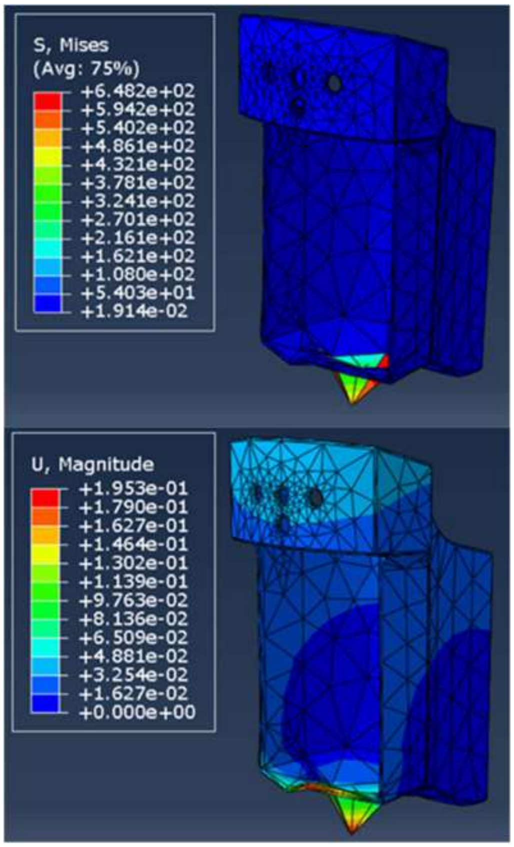


Figure 9. Stress concentration and deformation results (respectively) from the finite element analysis conducted on the final sleeve design model made from PLA

V. Conclusions

A sleeve design described aims to reduce the errors associated with the use of an MDI. Various sensors are implemented into the sleeve to monitor the four most common mistakes of the inhaler use. The patient is able to receive real-time feedback to aid in administering the proper medication dosage and to enforce proper inhaler use techniques. Additionally, a mobile app is developed to provide the user with further information on their technique and the areas in which they require improvement.

A 3-D printed prototype of the sleeve is created and is able to house a development board with an integrated accelerometer and gyroscope to measure the shaking force and orientation of the inhaler. The coordination between the canister actuation and the patient's inhalation is observed with a thin film pressure sensor located at the bottom of the sleeve. A differential pressure sensor is included in the design to monitor the inhalation rate of the patient when administering the medication. LED lights are implemented onto the front of the sleeve to provide the user with real-time feedback.

Conducting a finite element analysis on the sleeve model illustrates that the prototype can benefit from a change in material type from PLA to PETG. This modification allows for a reduction in the maximum deformation and the maximum stress concentration seen by the model. Reducing these values results in a more durable product which requires to be replaced less frequently, reducing maintenance cost and the environmental impact of the device.

In the current stage of development, the smart inhaler has a cost of $135 CAD to account for the sensors, LEDs, and the printing material. The prototype has a weight is 64.2 gr. and a volume of approximately 171 cm^3.

As the sleeve continues to be developed, alternate manufacturing options are being researched to determine an optimal material for the product. Additionally, the opportunities presented by analyzing the data collected by the sleeve is under development. Further advances in the mobile app may result in the ability to predict the likelihood of a patient experiencing the need to use the device given a certain environment.

Acknowledgment

The team would like to thank the Department of Mechanical and Mechatronics Engineering at the University of Waterloo for providing financial support and the resources necessary to develop the smart inhaler prototype.

References

[1] Asthma and Allergy Foundation of America, "Asthma Facts and Figures," Asthma and Allergy Foundation of America, Jun 2019. [Online]. Available: https://www.aafa.org/asthma-facts. [Accessed 5 December 2020].

[2] Asthma Canada, "What are asthma triggers?," Asthma Canada, [Online]. Available: https://asthma.ca/get-help/asthma-triggers/. [Accessed 5 December 2020].

[3] Asthma Canada, "What is asthma?," Asthma Canada, [Online]. Available: https://asthma.ca/get-help/understanding-asthma/. [Accessed 5 December 2020].

[4] Infinity Pediatrics, "Metered Dose Inhaler (MDI or pMDI)," [Online]. Available:http://infinitypediatrics.ca/wp-content/uploads/2015/01/Infinity-Pediatrics-MDI-no-spacer-Device-Info.pdf. [Accessed 5 December 2020].

[5] B. Montgomery, "Your asthma puffer is probably contributing to climate change, but there's a better alternative," The Conversation, 25 March 2018. [Online]. Available: https://theconversation.com/your-asthma-puffer-is-probably-contributing-to-climate-change-but-theres-a-better-alternative-92874. [Accessed 5 December 2020].

[6] L. Jahedi, S. R. Downie, B. Saini, H.-K. Chan and S. Bosnic-Anticevich, "Inhaler Technique in Asthma: How Does It Relate to Patients' Preferences and Attitudes Toward Their Inhalers?," Journal of Aerosol Medicin and Pulmonary Drug Delivery, vol. 3, no. 1, pp. 42-52, 2017.

[7] A. S. Ismaila, A. P. Sayani, M. Marin and Z. Su, "Clinical, economic, and humanistic burden of asthma in Canada: a systematic review," BMC Pulmonary Medicine, vol. 13, no. 1, p. 70, 2013.

[8] E. R. McFadden, "Improper patient techniques with metered dose inhalers: Clinical consequences and solutions to misuse," *The Journal of Allergy and Clinical Immunology,* vol. 96, no. 2, pp. 278-283, 1995.

[9] A.-M. Ciciliani, "Handling forces for the use of different inhaler devices," *International Journal of Pharmaceutics,* 5 April 2019.

[10] P. Haidl, S. Heindl, K. Siemon, M. Bernacka and R. M. Cloes, "Inhalation device requirements for patients' inhalation maneuvers," Respiratory Medicine, vol. 118, no. 1, pp. 65-75, 2016.

[11] Menon, E. Shashi. (2015). Transmission Pipeline Calculations and Simulations Manual - 12.23 Venturi Meter. (pp. 456-467). Elsevier. Retrieved from

https://app.knovel.com/hotlink/pdf/id:kt00U8PSR6/transmission-pipeline/meters-and-venturi-meter

[12] Giang, Ken, "PLA vs. ABS: What's the difference?," 3D HUBS, [Online]. Available: https://www.3dhubs.com/knowledge-base/pla-vs-abs-whats-difference/ [Accessed 6 December 2020].

[13] Giang, Ken, "CNC machining in PET," 3D HUBS, [Online]. Available: https://www.3dhubs.com/cnc-machining/plastic/pet/ [Accessed 6 December 2020].

Design of Piezoelectric Energy Harvesting Shoes for Charging Phones

Shahriar Khan
Dept of EEE
Indepencent University, Bangladesh
Dhaka, Bangladesh
skhan@iub.edu.bd

Muhit Kabir Sarneabat
Chemical Engg Division
Institute of Engineers, Bangladesh
Dhaka, Bangladesh
smkabir87@gmail.com

***Abstract*—In spite of great advances in cell phone technology, the rapid discharge of phone batteries remains a widespread problem. One solution may be recharging phones with movements of the body, such as with energy harvesting shoes. Soft-soled impact-absorbing shoes are best for the feet and for general health. Energy harvesting from shoes have been of interest for decades, but the technology is still in the research phase. Piezoelectric voltage generation at shoes is made more feasible with recent advances in flexible piezoelectric materials. The characteristics of energy conversion at shoes have been studied. Shoes connected through wires on the legs to power electronics can charge a phone at the waist. An extra battery source at the power electronics would be an inconvenience, and a simple full wave rectifier with filter is proposed for initial testing. The resulting large time constant implies the need for a large capacitor. The design considerations in this paper may be the basis of further research and experimentation, and possible commercial implementation.**

Keywords—energy, phone, phone, harvesting, shoe, power electronics, piezoelectric, piezoelectricity, walking, rectifier, filter, capacitor, battery.

I. Introduction

In spite of recent rapid advances in smart phones, there has been relatively slow progress in battery technology, and rapid depletion of phone batteries remains a major problem. Over the last two decades, microprocessors for smart phones have shown exponential growth keeping up with Moore's Law, but Lithium ion batteries have shown slower development.

Electric sockets for recharging may rarely be available in public areas, away from the home. The electric grid may be less available in remote locations and in less-developed countries. Stranded hikers may be calling for help with the last bit of charge on their cell phones. Foot soldiers in remote locations may need additional sources for recharging their portable electronics.

Power banks for charging phones have limited backup charge and require prolonged connection to a power socket. Solar powered phone chargers are likely to be bulky and not portable. These point to the possibility of recharging phones with body movements, such as with energy harvesting shoes.

A. Well-Cushioned Shoes Good for the Feet

From thousands (millions) of years of selective evolution, the feet of Man have mostly evolved for walking on the soft grass of grasslands, or the soft decaying plant matter on the forest floor. The human foot is not well adapted to the hard, smooth and uniform pavements of today.

Soft-cushioned shoes are comfortable, and good for the feet and for general health [1, 2]. Much or most of the energy of walking or running is dissipated in these shoe soles. The absorption of impact means that the mechanical energy of running is being largely converted to heat energy at the soles. At least some of this mechanical energy may instead be converted to electrical energy for charging electronic devices such as the phone.

B. Harvesting Energy at Shoes

Harvesting the energy of body movement for charging mobile and wireless electronics has been proposed for long [3]. The build-up of static electricity at devices at shoe soles has been suggested for harvesting energy [4]. Another possibility is electromagnetic voltage generation at the shoe sole, such as with an armature and field [5, 6]. But the electromagnetics may be cumbersome and not durable enough to be placed at a shoe sole.

Shoe outputs of many volts have been reported in the literature. A relatively high one watt of power has been reported from electromagnetic energy harvesters.

II. Piezoelectricity in Shoes

Of all the proposed methods for energy harvesting in shoes, piezoelectricity is the most common [7, 8, 9, 10]. One option is to use piezoelectric bending beams as shoe inserts [11].

Piezoelectric ceramics for a 90 kg person showed 0.4% or 1.43 mW of walking energy can be harvested [12].

Experimental results from 2009, report a "6-layer heel footwear harvesters have an average power output of 9 mW/shoe at walking speed of 4.8 km/h." Another study reports 55.6 µJ of energy, and peak power of 1.6 mW generated with each footstep.

One study reported walking or running generating 10-20 µJoules per step [13].

Piezoelectricity at the shoe is further supported by recent developments in flexible piezoelectric materials [14]. Piezoelectric powered shoes have been proposed to power a GPS device [15].

A. Power Electronics Interface

With a piezoelectric shoe energy harvester, what type of power electronics is to be used? The piezoelectricity has special voltage and current outputs requiring specific design to interface with the phone [16, 17].

It has been proposed the power electronics and a battery be kept inside the shoe sole [5]. But this would be problematic as it would increase complexity of the shoe and

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

would not be rugged and durable. Also, different phones may require different power electronics.

A piezoelectric cantilever device was implemented and tested together with its electronics [18]. This required a microprocessor, which would be difficult without an extra battery. A Buck converter has been proposed [19], but this too has the requirement of having an independent, separate, battery supply.

This paper proposes keeping the power electronics away from the shoe at the waist (or pocket) level, and to not require any separate intermediate battery.

A separate independent battery may be an added inconvenience and has been avoided in the design. The simple full-wave diode rectifier with filter configuration is proposed here for initial testing and for proof-of-principle.

III. Charging and Battery specifications

The charging requirements and the battery specifications of phones will determine the design of power electronics interface with the piezoelectric shoe.

Lithium Ion batteries for phones are mostly rated at 3.8 V, and may start working at 3.6 - 3.7 V. Other charger voltages are in the region of 5 V,

Two safety precautions include that a battery with less than 1.8 V should never be recharged. Secondly, the charger voltage should not exceed 4.25 V. A battery below 3.0 V needs a trickle current, before its voltage rises to 3.4 V.

A phone with a 4 inch screen may consume 0.75 w power from a battery of 5-6 watt hours capacity. Under practical usage, a battery may last for no more than 5-6 hours of usage.

Mobile phone batteries range from about 700 mA-hrs to 1200 mA-hrs, whereas smart phone batteries are in the range of 1000 - 1300 mA-hrs. The internal resistance of the battery may be about 0.15 ohms. After 300-400 charging-discharging cycles, the internal resistance may double to 0.3 ohms.

IV. Energy Absorbing Characteristics

The energy absorbed by the soles will be the sum of the energy absorbed by the rubber, and the energy converted by the piezoelectric material. The energy absorbed by the rubber will depend on the material and its energy absorbing characteristics.

If the shoe sole acts like a perfect spring, and no energy is converted to piezoelectricity, the energy absorbed during the down step would be completely released during the upstep (figure below).

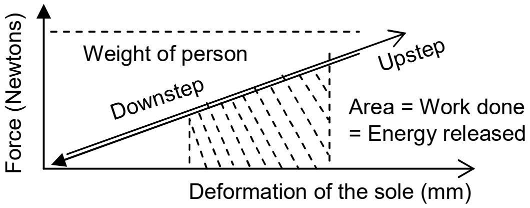


Fig 1. If the shoe acts like a perfect spring, the energy absorbed during down step would be completely released during the up step. Harvested energy will be zero.

The mechanical energy during downstep and during upstep will be about half the force multiplied by distance: We consider a person of 70 kg, deforming a sole by 1 cm at every step.

$$\textit{Work during Downstep} = \textit{Work during Upstep}$$

$$= \frac{1}{2}\ \textit{Weight}\ \text{x}\ \textit{Deformation of sole}$$

$$\textit{Work done} = 0.5 \text{ x } 70 \text{ kg x } 9.8 \text{ x } 0.1 = 3.43 \text{ Joules}$$

In comparison to the spring above, a rubber band, shows some mechanical hysteresis (figure below). The energy lost to hysteresis, or converted to heat, is the difference in the energy used for stretching and the energy released from contraction.

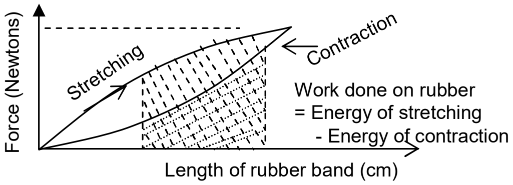


Fig 2. The stretching and contraction of a rubber band shows energy lost to hysteresis (converted to heat) as the work done during stretching minus the work released during contraction.

The energy absorbed or released equals the area under the curve above.

$$\textit{Energy absorbed or released} = \int F(x) \text{ x } dx$$

If the material of the sole has a hysteresis effect such as above, the work done during downstep will be greater than the energy during upstep. The shoe sole may recover slowly from deformation, meaning the force during downstep would be greater than during upstep.

The energy of hysteresis would be the area within the curve (figure below).

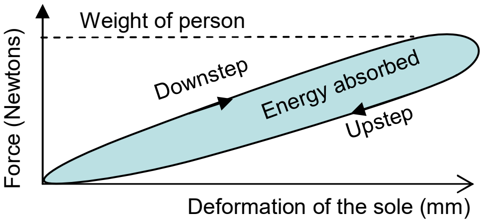


Fig 3. The area in the hysteresis curve will be the work done in the shoe sole.

If very little energy from the sole is returned during upstep, the total energy converted in the shoe is almost half of the energy in the rectangle (figure below).

$$\textit{Energy available} = \frac{1}{2} \text{ x } \textit{Weight} \text{ x } \textit{Deformation of sole}$$

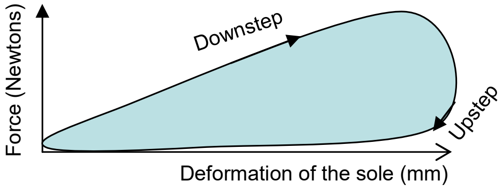


Fig. 4 . If due the hysteresis, the sole exerts little force during upstep, the energy converted would be 0.5 x weight x deformation of sole.

In case the hysteresis effect is very strong during downstep and upstep (from shoe material and energy harvesting), the energy conversion per cycle will be the close to the full area of the rectangle below

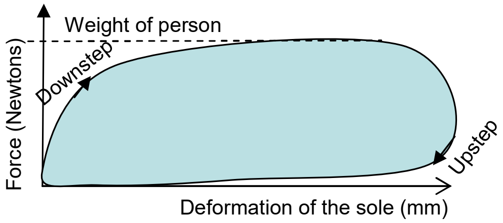


Fig 5 . In the extreme case, the force-deformation curve (hysteresis effect) will take up almost the full rectangular area (maximum energy harvesting).

Assuming maximum conversion of the energy, the energy available with each stepping cycle would be the area of the rectangle above:

$$W = MgH = 70 \text{ kg x } 9.8 \text{ m/s}^2 \text{ x } 0.01 \text{ m}$$

$$W = 6.86 \simeq 7 \text{ Joules}$$

Assuming each step taking 0.5 seconds, the power available during walking would be:

$$P = \frac{W}{T} = \frac{7\text{J}}{0.5\,\text{sec}} = 14 \text{ watt}$$

Higher energy and power may be obtained from greater depth of the shoe sole and with faster walking or running. The energy from jogging or running may be of the order of five times (estimated) greater than for walking.

There are few practical methods of conversion of this mechanical energy to electrical energy. Piezoelectric materials can hardly be made to deform by the above one cm from the movement of stepping.

A low efficiency of 10 % would give 1.4 watt in conversion of walking to electrical energy.

V. Force and Voltage from Piezoelectricity

The option of piezoelectricity for shoes is supported by recent developments in flexible piezoelectric materials.

Layers of piezoelectric material may be placed in the heels and the rest of the soles, lowering the chance of breakage (figure below). The work done will depend largely on the speed of walking or running.

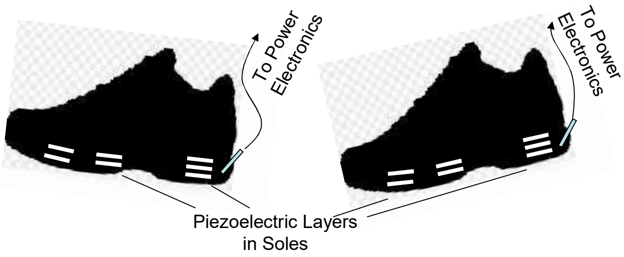


Fig 6. Similar piezoelectric layers may be distributed in both soles of shoes, minimizing chance of breakage. Connections may be in series, so as to maximize voltage.

Identical piezoelectric devices may be placed in both shoes for maximizing energy harvesting.

Series or parallel connection of the above piezoelectric layers in the soles will be largely decided by the needs of the power electronics. However, a series connection appears best, as the voltage would be maximized.

A. Pressure and Generated Voltage

Considering the stepping patterns and speeds of walking, the forces on the soles (heels) and the resulting deformation may be as follows.

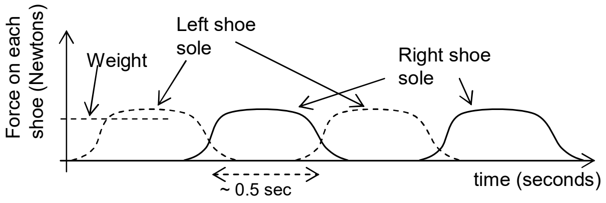


Fig 7. Force (in Newtons) upon each shoe during a simple walking motion

The generated piezoelectric voltages may be proportional to the force, at the right and left shoe (with no output current).

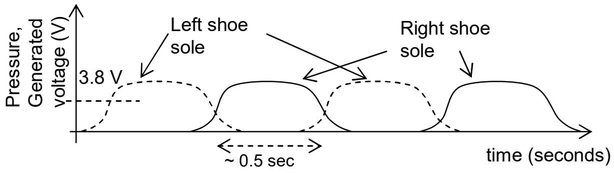


Fig 8. Pressure and voltage generated in the left and right shoes (without flow of current).

When the walking changes to running, the foot kicks up the body. The body will be in the air, and there will be a momentary increase in force at the sole (figure below). The body may stay suspended for about less than a second, during which the force will be zero. As the other foot lands on the ground, there will be an sharp force of impact.

All forces may generate proportional voltage at the piezoelectric material.

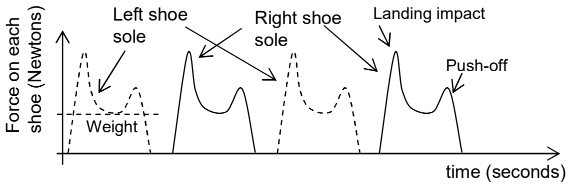


Fig 9. Force upon each shoe for slow running or jogging. Spikes are seen during the landing impact and during the push-off.

VI. Overall Structure

In deciding the overall structure, the power electronics may be placed at the waist or at the shoe. Placing at the waist close to the phone appears best at this time, as the shoe may be too compact, and present a source of additional wear and tear. Also, different phones may require different power electronic circuits.

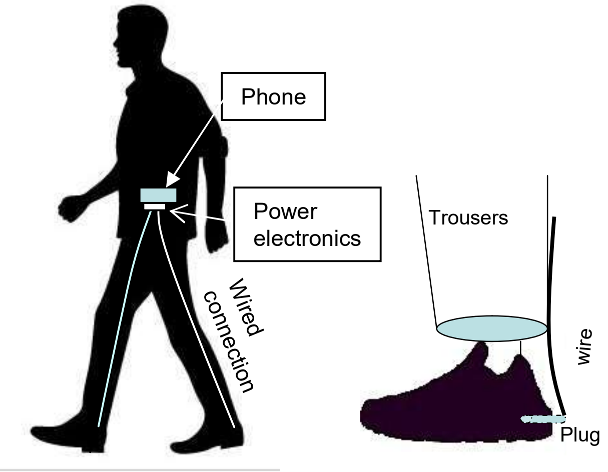


Fig. 10. Wires connect from the piezoelectric soles to the power electronics and then to the phone. A rugged wired connections avoid breakage from running.

Behind the heal of the shoe, there may be a rugged plug, connecting the wire to the power electronics. The wires and connections attached to the legs must also be rugged, in order to withstand walking or running over long distances.

VII. Power Electronics

A switched mode converter in the power electronics will require an additional power source or battery, which would be inconvenient.

For harvesting energy from the generated voltage, a current flow is required. The discharging current from the piezoelectric material may show the pulsed waveform below.

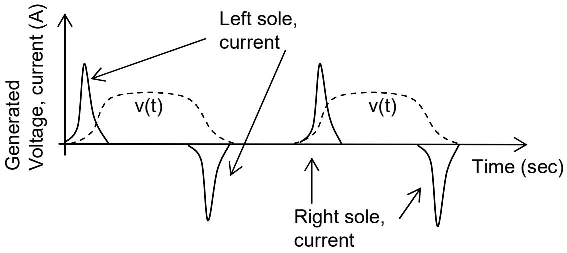


Fig. 11. For piezoelectric shoes, the current may be positive during the downstep, and negative during the upstep.

The alternating current from the piezoelectric material can be passed through a full-wave rectifier and then filtered to give a mostly DC waveform.

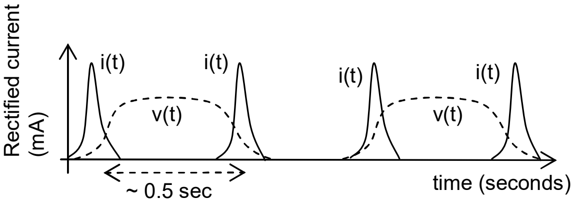


Fig. 13. After passing through a full-wave rectifier, the pulsed current will be in the same direction at intervals of about 0.5 second.

Passing through a full wave rectifier and a filter will give a DC voltage and current with a small ripple, suitable for charging a phone. The two shoes may be placed in series, effectively adding the voltages (figure below).

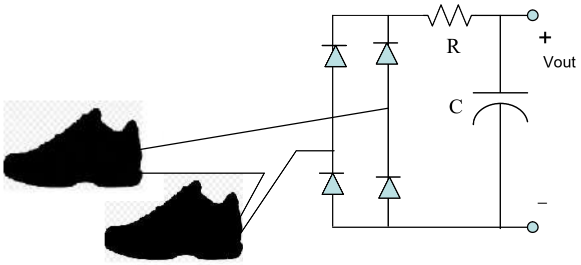


Fig. 12. The piezoelectric materials of the shoe can be connected in series with a full-wave rectifier and a filter.

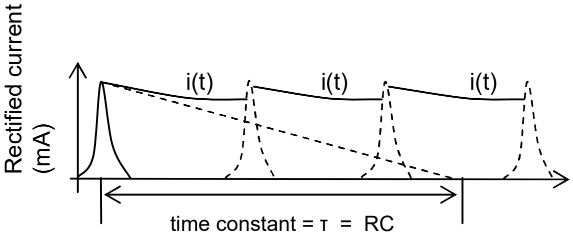


Fig. 14. DC output with ripple after passing through a filter. The time constant may be too large (compared to conventional electronics), meaning the capacitor may need to be large.

The diode rectifier/s may be compact enough to be placed at the shoe or at the plug. However, the filter may require a large capacitance, which may be too bulky to place in or near the shoes.

A. Filter Design

The time constant of the filter is the length of intersection on the time axis of a tangent from the curve (figure above), and will equal the product of the resistance *R* and the capacitance *C*.

$$\tau = RC \text{ seconds}$$

In order to keep the ripple value low, the time constant τ may have to be several times larger than the stepping time. With stepping time of the order of 0.5 second, the time constant may be

$$\tau = RC = 2 - 4 \text{ seconds.}$$

This time constant of 2 - 4 sec is much larger than the order of milliseconds encountered in conventional power electronics. This implies that either *R* or *C* or both must be unusually large. Having a large *R* may be undesirable as it would dissipate too much energy. A large capacitor *C* may be too large and bulky. The electrolytic capacitor may be the best option, as it may have a large value (at the cost of low voltage withstanding capability).

A capacitor of the order of 1 μfarad, would imply a resistance of the order of 2 - 4 Mohms, which would still be too large, consuming the energy generated.

New capacitor technologies may allow capacitors of 1 millifarad, requiring a resistor of the order of 2 - 4 Kilo-ohm.

The voltage and current characteristics of the phone during charging will affect the specifications of *R* and *C* and the rest of the power electronics. The design of *R* and *C* and the filter are clearly design issues that must be overcome.

B. Other Considerations

The voltage generated by the shoe will be proportional to the force during the step, or the weight of the person. This implies that the given circuits will require a minimum threshold of a person's weight, below which the battery charging voltage cannot be reached.

In reality, the power electronics might have added complexity, such as for not charging when the phone battery voltage is less than 1.8 V.

There is the possibility of having intermittent charging of the phone, with pulses every 0.5 seconds, lasting less than 0,1 seconds. However, such pulsed charging may not be accepted by modern phones and their electronic circuitry.

VIII. Conclusion

In spite of great advances in microprocessors and communication technology, the rapid discharge of batteries remains a major problem for phones. Backup charging devices such as power banks have limitations such as bulkiness, and limited ability to hold charge.

The well-cushioned thick-soled shoe provides the human foot an appropriate interface to absorb the impact of walking or running on hard pavement. This raises the possibility of introducing new technologies for recharging phones by harvesting the energy of shoes. Energy harvesting shoes have been of interest for decades, but have not come close to commercial feasibility. Generators with armatures and magnets have been proposed for shoes, but the associated mechanics may be too cumbersome and non-durable. Piezoelectricity at the soles will be simpler in construction. A major challenge is finding a flexible and durable piezoelectric material for the shoe, capable of months of rugged use.

The power electronics for converting the piezoelectric voltage and current at the shoe to the DC for the phone is an additional challenge. The power electronics may be placed in the pocket or the waist region, in close proximity to the phone. A full-wave rectifier and a filter appear the best for initial testing. A large capacitor may be required for the filter prior to the phone.

As different phones have different charging voltage and current requirements, different power electronic circuits may be required for different phones.

In spite of the clear challenges, the design considerations in this paper may be a basis for further research, development and experimentation.

IX. References

[1] H. Chiu, T. Shiang, D. Lin, "Cushioning properties of Shoe-surface Interfaces in different impact energies." International Symposium on Biomechanics in Sports, 2001

[2] Shahriar Khan, Health and Disease According to Darwinian Evolution, ISBN:978-984-33-6163-9, S. Khan, Dhaka, Bangladesh, Feb 2015.

[3] J. A. Paradiso and T. Starner, "Energy scavenging for mobile and wireless electronics," in IEEE Pervasive Computing, vol. 4, no. 1, pp. 18-27, Jan.-March 2005,

[4] T. C. Hou, Y. Yang, H. Zhang, J. Chen, L. J. Chen, Z. L. Wang. "Triboelectric nanogenerator built inside shoe insole for harvesting walking energy." Nano Energy. 2013, Sep 1;2(5):pp. 856-62.

[5] Xu Borui, Li Yang, "Force Analysis and Energy Harvesting for Innovative Multi-functional Shoes" Frontiers in Materials , Vol. 6, 2019, pages 221

[6] Shahriar Khan, Electrical Energy Systems, Third Edition, ISBN: 978-984-33-7638-1, S. Khan, Dhaka, Bangladesh, Dec. 2019.

[7] H. Kalantarian, M. Sarrafzadeh, "Pedometers Without Batteries: An Energy Harvesting Shoe," IEEE Sensors Journal, vol. 16, no. 23, Dec.1, 2016, pp. 8314-8321,

[8] J. Kymissis, C. Kendall, J. Paradiso and N. Gershenfeld, "Parasitic power harvesting in shoes," Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215), Pittsburgh, PA, USA, 1998, pp. 132-139.

[9] P. Chaudhary, P. Azad, "Energy Harvesting Using Shoe Embedded with Piezoelectric Material." Journal of Elec Materi 49, November 2020, pp. 6455–6464.

[10] A. Gupta, A. Sharma, Piezoelectric Energy Harvesting via Shoe Sole, International Journal of New Technology and Research (IJNTR), ISSN:2454-4116, Volume-1, Issue-6, October 2015 Pages 10-13.

[11] Xin Y, Li X, Tian H, Guo C, Qian C, Wang S, Wang C. Shoes-equipped piezoelectric transducer for energy harvesting: A brief review. Ferroelectrics. 2016 Mar 15;493(1), pp.12-24.

[12] Turkmen AC, Celik C. Energy harvesting with the piezoelectric material integrated shoe. Energy. 2018 May 1;150, pp. 556-64.

[13] R. Meier, N. Kelly, O. Almog and P. Chiang, "A piezoelectric energy-harvesting shoe system for podiatric sensing," 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 2014, pp. 622-625.

[14] C. Dagdeviren, P. Joe, O. L. Tuzman, K. Park, K. J. Lee, Y. Shi, et al "Recent progress in flexible and stretchable piezoelectric devices for mechanical energy harvesting, sensing and actuation," Extreme Mechanics Letters, Vol. 9, Part 1, 2016, pp 269-281.

[15] A. Gatto and E. Frontoni, "Energy Harvesting system for smart shoes," 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA), Senigallia, Italy, 2014, pp. 1-6,, 28 May 2013

[16] Shahriar Khan, Semiconductor Devices and Technology, Third Edition, ISBN: 978-094-33-5983-4, Dhaka, Bangladesh, June 2018,.

[17] Shahriar Khan, DC Circuits and Transients; 4th Edition, ISBN 978-984-33-3560-9, S. Khan, Dhaka, Bangladesh, 2012,.

[18] E. Camilloni, M. Carloni, M. Giammarini, M. Conti, "Energy harvesting with piezoelectric applied on shoes," Proceedings Volume 8764, VLSI Circuits and Systems VI; 2013.

[19] N. Ahmad, M. T. Rafique, R. Jamshaid , "Design of Piezoelectricity Harvester using Footwear," 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Kuala Lumpur, Malaysia, December 2019.

High Gain DC-DC Converter for Three-Wheeler Electric Vehicles

Abdullah Al Mamun
Dept. of Electrical & Electronic Eng.
Independent University of Bangladesh
Dhaka, Bangladesh
abdullahalmamun@iut-dhaka.edu

Sultanul Arfin
Dept. of Electrical & Electronic Eng.
Islamic University of Technology
Dhaka, Bangladesh
sultanularfin@iut-dhaka.edu

Shahriar Khan
Dept. of Electrical & Electronic Eng.
Independent University of Bangladesh
Dhaka, Bangladesh
skhan@iub.edu.bd

***Abstract*—There is an ongoing increase in development of lightweight electric vehicles. In developing countries, electric three-wheelers in use are mostly limited to low voltage drive systems. They are powered directly from the limited battery voltage without any step-up conversion before connection to the motor controller. Due to low voltage and low power DC motor drives, speed and torque characteristics of the vehicle suffers when compared to other EVs with similar battery storage capacity. This paper presents a novel power electronic drive system of step-up DC-DC converter composed of cascaded boost and SC-Zeta topology, capable of voltage gain as high as 40 times at a duty cycle of 0.8. The boosted output voltage is given to a three phase inverter for control of traction motors up to 4 KW of power which will enable existing three-wheelers to use high performance drive system. The proposed converter is simulated & compared with other existing high gain converters in literature, and the output results are presented graphically.**

Keywords*— electric vehicle, EV, rickshaw, high gain, step-up, hybrid, DC-DC, converter, boost, switched capacitor, SC, zeta, DC-AC, three phase, inverter, motor, control

I. Introduction

With the rise of electric vehicles in the US, there is an ongoing increase in development of lightweight and mini electric vehicles and three wheelers. In developing countries electric three-wheelers, also known as auto-rickshaws, in use are mostly limited to low voltage drive systems. They are powered directly from the limited battery voltage without any step-up conversion before connecting to the motor controller. Owing to low voltage and low power DC motors, the speed and torque characteristics of the vehicle suffers greatly when compared to other EVs with similar battery storage capacity. Adding high powered motors into a low voltage system adds cost and complexity in the vehicle power electronics. A 3 KW DC motor will require currents in ranges of hundreds of amperes to be driven from a 12V – 48V DC supply. Not only are such high current controllers prohibitively expensive for three-wheelers, high currents will cause significant I^2R power losses in the motor as well as all the circuit components. Due to the complexity associated with high current drive systems, most conventional electric three-wheeler and auto-rickshaws tend to be driven by low voltage and low power traction motors, and the maximum torque and speed obtained from such system is severely restricted [1-3].

Most modern EVs in developed regions use three-phase AC motors operating at voltages above 250V. This choice of voltage allows the motor to be powerful enough without drawing excessive currents. To address the poor speed and power performance of existing three-wheelers in developing regions, it is desirable to use motors with voltage ratings in the range of 300V – 600V AC [4]. In this paper, we propose a novel power electronic system for conventional electric three-wheelers that will enable them to use high voltage and high performance three-phase AC motors powered from low voltage but high capacity battery packs. The designed system is divided into two stages, a hybrid DC-DC converter and a three-phase DC-AC inverter. The proposed hybrid converter is composed of a boost converter cascaded with a modified switched capacitor (SC) Zeta converter. This novel combination is responsible for providing necessary high step-up gain and continuous input and output current. Theoretical analysis is performed to derive expressions for voltage gain, average currents, voltage and current stress of components, power loss and efficiency of the converter circuit. Simulation results are presented to show the performance of the designed converter in comparison to existing non-isolated high gain step-up converters in literature.

II. Background Theory

Operating three-phase traction motors of 300V and above, only from 12V – 48V battery input, will require extreme duty cycles for conventional converters, making them impractical. There has been a number of literatures dedicated to high gain step-up voltage conversion techniques [5-8]. Most solutions use transformers or coupled inductors to get the required voltage boost. Use of transformers or coupled inductors where isolation is not required only adds to weight, cost and unnecessary power loss [5]. Switched-capacitor (SC) and switched-inductor (SL) structures proposed in [6] allows various conventional converters to generate additional voltage gain. The capacitors in a SC structure are arranged in such a way that when the converter switches between on and off states, the capacitors are connected in parallel during the charging cycle, and discharged in series during the discharge cycle to provide high voltage gain ratio [9-10].

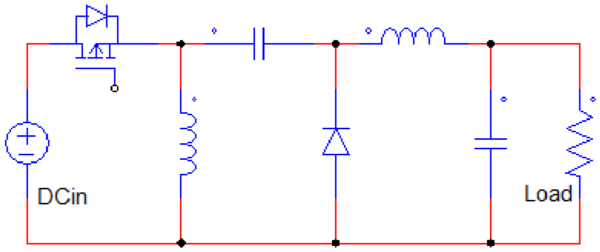


Fig. 1. A conventional step-up Zeta converter

The topology of a conventional Zeta converter, as shown in Fig. 1, allows us to implement a SC structure to obtain hybrid SC-Zeta as in [9]. Although the SC-Zeta has higher gain than the conventional converters, it is still unable to provide the required voltage gain within a reasonable duty cycle. So an improved step-up DC-DC converter capable of the required step-up gain must be designed. To keep enough margin for adjustment of duty cycle and regulation of output voltage to control output power, it is necessary to design a novel converter with the required voltage gain capable of operating within a reasonable duty cycle.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

III. Proposed System

In this paper we propose a cascaded boost-SC-zeta hybrid DC-DC converter, shown in Fig. 2, working in tandem with a three-phase DC-AC inverter, shown in Fig. 3. The proposed hybrid DC-DC converter can boost the voltage sufficiently to be fed into the inverter to generate three-phase AC as required by the high performance motor of the electric three-wheelers and other equivalent EVs. The output of this inverter can be regulated to control power and speed of the motor for optimum performance. The proposed DC-DC converter can step-up the input voltage as high as 180 times at duty cycle of 0.9 while keeping both input and output current continuous.

A. Topology of our proposed hybrid DC-DC converter

The proposed DC-DC converter circuit is shown in Fig. 2, which consists of a boost converter cascaded with a modified switched capacitor (SC) zeta. The SC structure consists of two capacitors *C2* and *C3* and two diodes *D2* and *D3*. Similar to existing zeta converter, our designed circuit contains an inductor L_0 and a capacitor C_0 at the output. The presence of an input inductor *L1* and output inductor L_0 provides the additional advantage of maintaining continuous input and output currents.

B. Principles of operation of the designed circuit

The operating states of our designed DC-DC converter can be divided into two states: on-state and off-state. The designed converter circuit is considered to be operating under continuous conduction mode (CCM) in both of the states.

1) On-state ($0 \leqslant t \leqslant DTs$)

In this state, switch *S1* and switch *S2* are simultaneously turned on, while *D1, D2* and *D3* diodes are turned off. The current flow during on-state is shown in Fig. 4. Inductor *L1* is parallel to input V_{in} and stores energy directly from the source. The current I_{L1} increases linearly with the ratio V_{in}/L_1. The inductor L_2 is being energized by the capacitor *C1* as it discharges. The two capacitors *C2* and *C3* that are connected in series, are now being discharged while output inductor L_o is being energized and output capacitor C_o is charged.

2) Off-state ($DTs \leqslant t \leqslant Ts$)

The switch *S1* and switch *S2* are simultaneously turned off in this time interval, while the diodes *D1*, *D2* and *D3* are in forward bias, turning them on. Fig. 5 shows the current flow during this state. In this time period, capacitor *C1* is being charged by inductor *L1*, and inductor *L2* releases its stored energy to capacitors *C2* and *C3*. While *C1*, *C2* and *C3* capacitor charges, C_o is being discharged. Inductor L_o helps to supply continuous current to the load and capacitor C_o helps to maintain constant voltage at the output.

IV. Theoretical Analysis

Some assumptions are made to simplify the theoretical analysis of this circuit. All the components in this converter circuit are considered to be ideal and lossless. To ensure operation in CCM mode, *L1*, *L2 and* L_o inductor values are assumed to be large enough. *C2* and *C3* capacitor values are equal. *C1* and output capacitor C_0 values are large enough to ignore the output voltage ripple. Expressions for voltage gain, average currents, voltage and currents stress of components of this converter is derived in this analysis. Losses in the components are only considered during power loss and efficiency analysis.

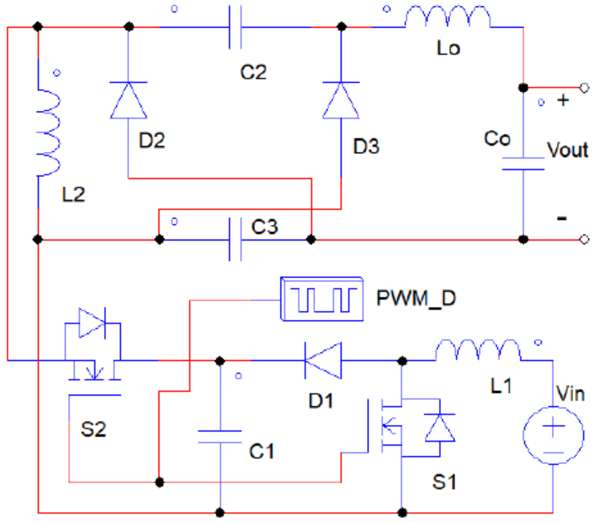


Fig. 2. Proposed Boost-SC-Zeta Hybrid DC-DC Converter Circuit

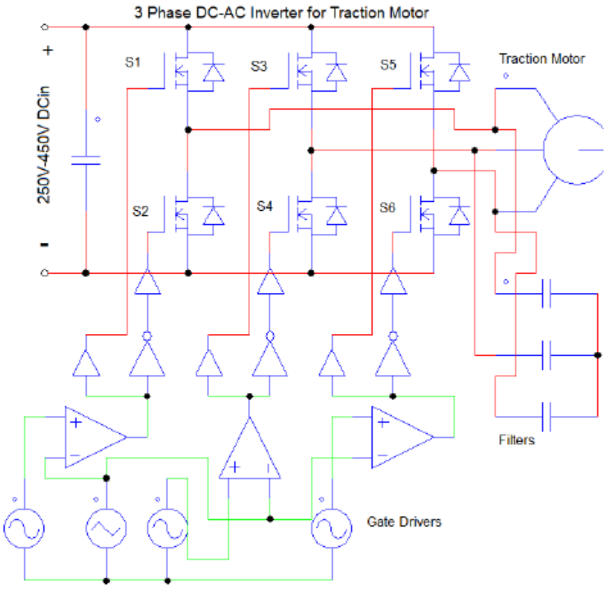


Fig. 3. 3-Phase DC-AC Inverter for Traction Motor Drive System

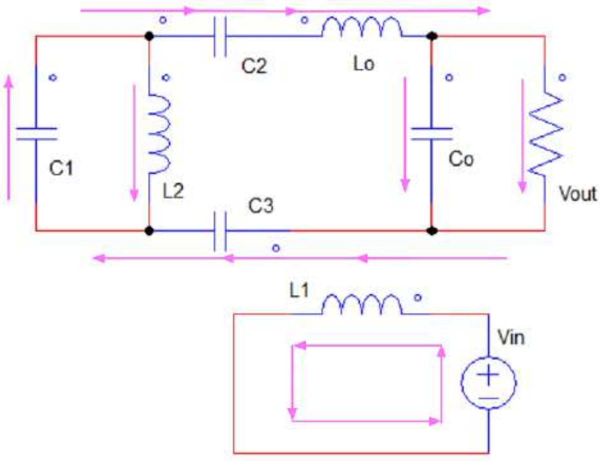


Fig. 4. Current Flow in the proposed hybrid DC-DC converter during on-state of switches *S1* and *S2* in time period $0 \leqslant t \leqslant DT_s$

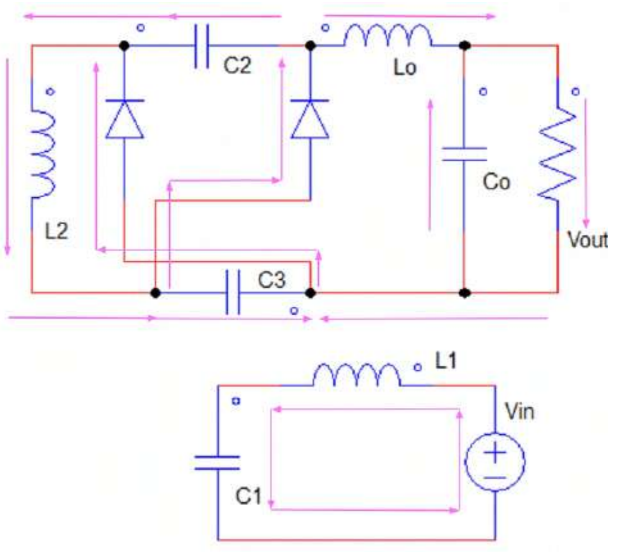


Fig. 5. Current flow of the proposed converter circuit at off-state in time period $DT_s \leqslant t \leqslant T_s$

A. Voltage gain analysis

For operation during the on-state for time $0 \leq t \leq DT_s$, the following equations for the inductors *L1, L2 and L_0* can be obtained:

$$V_{L1} = V_{in}$$

$$V_{L2} = V_{C1}$$

$$V_{L0} = 2V_{C2,3} + V_{C1} - V_{out}$$

For operation during off-state of the switch during time period $DT_s \leq t \leq T_s$, the following equations could be obtained:

$$V_{L1} = V_{C1} - V_{in}$$

$$V_{L2} = V_{C2,3}$$

$$V_{L0} = V_{out} - V_{C2,3}$$

By using the above equations and applying voltage-second balance principles on the inductor *L1*, we get:

$$\frac{1}{T_s}\int_0^{DT_s} V_{L1_{[On]}}\,dt + \int_{DT_s}^{T_s} V_{L1_{[Off]}}\,dt$$

$$V_{in}D - (V_{C1} - V_{in})(1-D) = 0$$

$$V_{C1} = V_{in}\frac{1}{1-D} \qquad (1)$$

Applying voltage-second balance principles on inductor *L2*, we get:

$$\frac{1}{T_s}\int_0^{DT_s} V_{L2_{[On]}}\,dt + \int_{DT_s}^{T_s} V_{L2_{[Off]}}\,dt$$

$$V_{C1}D - V_{C2,3}(1-D) = 0$$

$$V_{C2,3} = V_{C1}\frac{D}{1-D}$$

$$V_{C2,3} = V_{in}\frac{D}{(1-D)^2}$$

By Applying voltage-second principles on inductor $L_{0,}$ we get:

$$\frac{1}{T_s}\int_0^{DT_s} V_{L0_{[On]}}\,dt + \int_{DT_s}^{T_s} V_{L0_{[Off]}}\,dt$$

$$(2V_{C2,3} + V_{C1} - V_{out})D \\ - (V_{out} - V_{C2,3})(1-D) = 0$$

$$V_{out} = V_{C1}\frac{2D}{1-D}$$

Substituting V_{C1} in the above equation, the final output voltage could be written as:

$$V_{out} = V_{in}\frac{1}{1-D} \times \frac{2D}{1-D}$$

$$V_{out} = V_{in}\frac{2D}{(1-D)^2} \qquad (4)$$

The gain of the designed converter is derived to be:

$$G = \frac{V_{out}}{V_{in}} = \frac{2D}{(1-D)^2}$$

B. Analysis of Input Current

$$I_{in} = I_{out}\frac{V_{out}}{V_{in}}$$

$$I_{in} = I_{out}\frac{2D}{(1-D)^2} \qquad (5)$$

$$I_{in} = \frac{V_{out}}{R}\frac{2D}{(1-D)^2}$$

$$I_{in} = \frac{V_{in}}{R}\left[\frac{2D}{(1-D)^2}\right]^2 \qquad (6)$$

C. Analysis of average current

Considering the current flow in the capacitor C_o to be negligible, we get:

$$I_{L0} \cong I_{OUT} = \frac{V_{OUT}}{R} = \frac{V_{IN}}{R}\frac{2D}{(1-D)^2} \qquad (7)$$

$$I_{C2,3_{ON}} = I_{L0} = \frac{V_{IN}}{R}\frac{2D}{(1-D)^2} \qquad (8)$$

By applying the principles of charge-second balance on capacitors *C2* and *C3* to yield:

$$\frac{1}{T_S}\left[\int_0^{DT_S} I_{C2,3_{ON}}\,dt + \int_{DT_S}^{T_S} I_{C2,3_{OFF}}dt\right] = 0$$

$$I_{C2,3_{ON}}D - I_{C2,3_{OFF}}(1-D) = 0$$

$$I_{C2,3_{OFF}} = I_{C2,3_{ON}}\frac{D}{1-D}$$

$$I_{C2,3_{OFF}} = \frac{V_{IN}}{R}\frac{2D^2}{(1-D)^3} \tag{9}$$

Average currents through the inductor *L1* is:

$$I_{L1} = I_{IN} = \frac{V_{IN}}{R}\left[\frac{2D}{(1-D)^2}\right]^2 \tag{10}$$

The average current when capacitor *C1* is off:

$$I_{C1_{OFF}} = I_{IN} = \frac{V_{IN}}{R}\left[\frac{2D}{(1-D)^2}\right]^2 \tag{11}$$

By applying charge-second balance principles on capacitor *C1*, we get:

$$\frac{1}{T_S}\left[\int_0^{DT_S} I_{C1_{ON}}\,dt + \int_{DT_S}^{T_S} I_{C1_{OFF}}dt\right] = 0$$

$$I_{C1_{ON}}D - I_{C1_{OFF}}(1-D) = 0$$

$$I_{C1_{ON}} = I_{C1_{OFF}}\frac{1-D}{D}$$

$$I_{C1_{ON}} = \frac{V_{IN}}{R}\frac{4D}{(1-D)^3} \tag{12}$$

Average currents through the inductor *L2* can be derived as follows:

$$I_{L2_{ON}} = I_{C1_{ON}} - I_{C2_{ON}}$$

$$I_{L2_{ON}} = \frac{V_{IN}}{R}\frac{4D}{(1-D)^3} - \frac{V_{IN}}{R}\frac{2D}{(1-D)^2}$$

$$I_{L2_{ON}} = \frac{V_{IN}}{R}\frac{2D(1+D)}{(1-D)^3} \tag{13}$$

$$I_{L2_{OFF}} = I_D + I_{C2_{OFF}}$$

$$I_{D2} = I_{C2_{OFF}} + I_{L0} = \frac{V_{IN}}{R}\frac{2D}{(1-D)^3} \tag{14}$$

$$I_{L2_{OFF}} = \frac{V_{IN}}{R}\frac{2D(1+D)}{(1-D)^3} \tag{15}$$

$$I_{L2_{OFF}} = I_0\left(\frac{1+D}{1-D}\right) \tag{16}$$

D. Analysis of current stress of switches

Current stress of the switch *S1 (I_{S1})* and switch *S2 (I_{S2})* can be expressed as:

$$I_{S1} = I_{in} = \frac{V_{in}}{R}\left[\frac{2D}{(1-D)^2}\right]^2 \tag{17}$$

$$I_{S1} = I_0\frac{2D}{(1-D)^2} \tag{18}$$

$$I_{S2} = I_{C1_{[On]}} = \frac{V_{in}}{R}\frac{4D}{(1-D)^3} \tag{19}$$

$$I_{S2} = I_0\frac{2D}{1-D} \tag{20}$$

E. Analysis of voltage stress on switches

Voltage stresses on switch *S1* (V_{S1}) and switch *S2* (V_{S2}) can be obtained as follows

$$V_{S1} = V_{in} + V_{L1}$$

$$V_{L1} = V_{C1} - V_{in} = \frac{V_{in}}{1-D} - V_{in} = V_{in}\frac{D}{1-D} \tag{21}$$

$$V_{S1} = V_{in} + V_{in}\frac{D}{1-D}$$

$$V_{S1} = V_{in}\frac{1}{1-D} \tag{22}$$

$$V_{S2} = V_{C1} + V_{L2}$$

$$V_{S2} = V_{in}\frac{1}{1-D} + V_{in}\frac{D}{(1-D)^2}$$

$$V_{S2} = V_{in}\frac{1}{(1-D)^2} \tag{23}$$

F. Power loss and efficiency analysis

For efficiency estimation conduction and parasitic resistances can be defined as:

- Conduction resistance of switches: R_{DS}
- Conduction resistance of diodes: R_{Fx}
- Diode threshold voltage: V_{Fx}
- Inductor ESR: R_{Lx}
- Capacitor ESR: R_{Cx}

1) For Switch

Power loss in switch *S1* during conduction is expressed as:

$$P_{R_{DS_1}} = r_{DS1}I_{S1}{}_{RMS}^{2} = r_{DS1}I_0^2\frac{4D^2}{(1-D)^2} \tag{24}$$

$$P_{SW1} = f_sC_SV_{S1}^2 = f_sC_SV_{in}^2\frac{1}{(1-D)^2} \tag{25}$$

$$P_{S1} = P_{r_{DS1}} + \frac{1}{2}P_{SW1} \tag{26}$$

Power loss in switch *S2* during conduction is expressed as:

$$P_{R_{DS_2}} = r_{DS2}I_{S2}{}_{RMS}^{2} = r_{DS2}I_0^2\frac{4}{(1-D)^2} \tag{27}$$

$$P_{SW2} = f_sC_SV_{S2}^2 = f_sC_SV_{in}^2\frac{1}{(1-D)^4} \tag{28}$$

$$P_{S2} = P_{r_{DS2}} + \frac{1}{2} P_{SW2} \qquad (29)$$

2) For inductors

The power loss of inductors *L1*, *L2* and *Lo* is derived to be:

$$P_{L1} = R_{L1} I_{L1}^2 = R_{L1} I_0^2 \left[\frac{2D}{(1-D)^2}\right]^2 \qquad (30)$$

$$P_{L2} = R_{L2} I_{L2}^2 = R_{L2} I_0^2 \left[\frac{1+D}{1-D}\right]^2 \qquad (31)$$

$$P_{L0} = R_{L0} I_{L0}^2 = R_{L0} I_0 2 \qquad (32)$$

3) For Capacitors

Power loss of capacitors *C1*, *C2* and *C3* is expressed as:

$$P_{C1} = r_{c1} I_{C1_{ON}}^2 + r_{c1} I_{C1_{OFF}}^2$$

$$P_{C1} = r_{C1} \left[\frac{V_{in}}{R} \frac{4D}{(1-D)^3}\right]^2 + r_{C1} \left[\frac{V_{in}}{R} \frac{4D^2}{(1-D)^4}\right]^2 \qquad (33)$$

$$P_{C1} = r_{C1} I_0^2 \left[\frac{2}{1-D}\right]^2 + r_{C1} I_0^2 \left[\frac{2D}{(1-D)^2}\right]^2 \qquad (34)$$

$$P_{C2,3} = r_{c2,3} I_{C2,3_{ON}}^2 + r_{c1} I_{C2,3_{OFF}}^2$$

$$P_{C2,3} = r_{C2,3} \left[\frac{V_{in}}{R} \frac{2D}{(1-D)^2}\right]^2 + r_{C2,3} \left[\frac{V_{in}}{R} \frac{2D}{(1-D)^3}\right]^2 \qquad (35)$$

$$P_{C2,3} = r_{C2,3} I_0^2 + r_{C2,3} I_0^2 \left[\frac{D}{1-D}\right]^2 \qquad (36)$$

4) For Diodes

Power loss in diodes *D1*, *D2* and *D3* is derived to be:

$$P_{D1} = R_{F1} I_0^2 \frac{4D^2}{(1-D)^4} + V_{F1} I_0 \qquad (37)$$

$$P_{D2,3} = R_{F2,3} I_0^2 \frac{1}{(1-D)^2} + V_{F2,3} I_0 \qquad (38)$$

5) Total Loss

Total loss of power and efficiency of the designed converter can be written as follows:

$$P_{LOSS} = \sum (P_{S_{1,2}} + P_{L_{0,1,2}} + P_{C_{1,2,3}} + P_{D_{1,2,3}}) \qquad (39)$$

$$\eta = \frac{P_{OUT}}{P_{OUT} + P_{LOSS}} \qquad (40)$$

V. Simulation Results

For simulation of the designed converter, all the active and passive components in our converter circuit were chosen to be ideal and lossless. Input voltage V_{in} is fixed at 12V DC and the switching frequency is set to be 100 kHz. To ensure operation in CCM mode, values for the inductors *L1*, *L2* and *L0* are chosen to be 0.1 mH, 0.2 mH and 0.5 mH respectively. The values for capacitor *C1* is 220 µF, *C2*, *C3* are equal and 1 µF each, and *C0* is 2.2 µF. Capacitor values are chosen to minimize output ripple voltage.

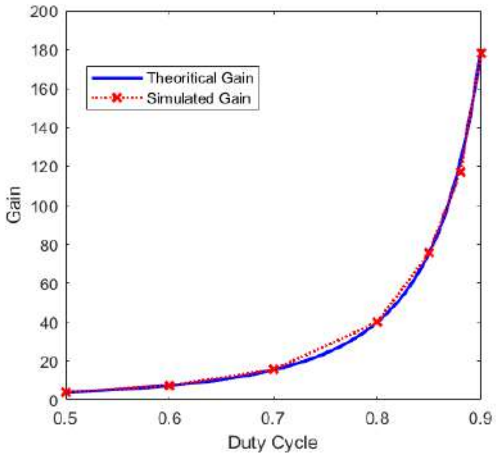


Fig. 6. Graph of theoretically derived gain and software simulated gain against duty cycle

A. Theroretical Gain versus Simulated Gain

The theoretical gain of the circuit *2D/(1-D)*^{2} as derived in (4), is compared with simulation results in Fig. 6. The graph clearly shows that the theoretically derived gain and simulated gain are in agreement with each other which proves the validity and correctness of operation of the proposed converter circuit.

B. Comparison between existing high gain hybrid DC-DC converters

The designed converter is compared to existing non-isolated high gain hybrid DC-DC converters in the literature. From Fig. 7, it is clear that the designed circuit is capable of achieving significantly higher gain compared to existing non-isolated step-up hybrid topologies.

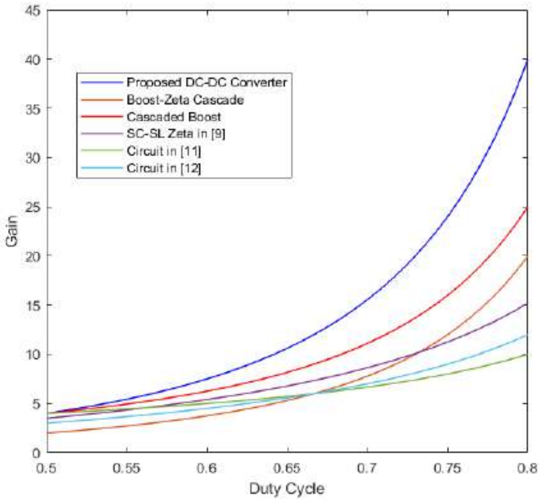


Fig. 7. Comparison of our designed hybrid converter with other existing non-isolated high gain DC-DC converters in the literature

C. Simulation of DC-AC inverter output

The boosted output voltage from the designed DC-DC converter in is fed into a 3-phase DC-AC inverter. Line to line voltages of the inverter output is plotted in Fig. 8. The frequency and magnitude of output voltage from the inverter can be changed for optimal control of the vehicle motor.

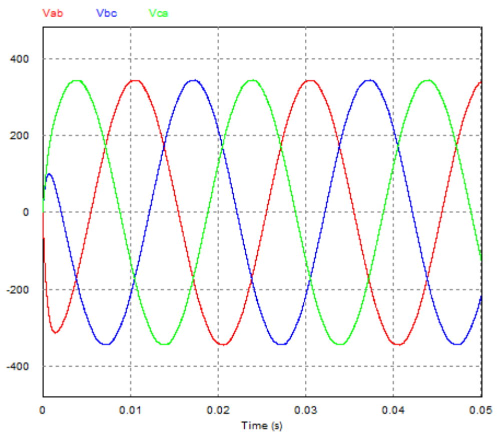


Fig. 8. Line to line ouput voltage of the 3-phase inverter

D. Simulation of Motor Output Performance

A three-phase AC induction motor was connected to the designed system with the converter at 80% duty cycle and inverter at 50 Hz to observe the performance of the motor. A graph of motor speed in RPM is plotted in Fig. 9. Speed up to 3000 RPM can be obtained by operating inverter at 150 Hz.

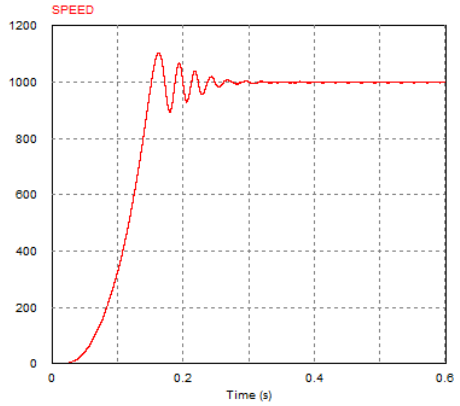


Fig. 9. Simulation of motor speed in RPM against time

TABLE I. COMPARISON BETWEEN PROPOSED COVERTER CIRCUIT AND EXISTING NON-ISOLATED HYBRID DC-DC CONVERTERS IN LITERATURE

	Proposed DC-DC Converter	**Circuit in [9]**	**Circuit in [12]**	**Circuit in [7]**
Switches	2	1	1	1
Diodes	3	5	3	2
Capacitor	4	3	5	4
Inductor	3	3	3	3
Total Component	12	12	12	10
Voltage Stress	$\frac{V_{in}}{(1-D)^2}$	$\frac{V_{in}(1+D)}{1-D}$	$\frac{V_0+3}{3}$	$\frac{V_0+2V_{in}}{2V_{in}}$
Gain	$\frac{2D}{(1-D)^2}$	$\frac{(3+D)D}{1-D}$	$\frac{3D}{1-D}$	$\frac{2D}{1-D}$
Gain, D = 0.8	40	15.2	12	8

VI. CONCLUSION

The increased interest in electric vehicles and their motor drives have inspired this study. In developing countries such as Bangladesh, electric three-wheelers are the most common form of EVs currently in use. There is the backdrop of home-grown solutions to local problems and needs for better electric three-wheelers and equivalent EVs. To address the limited performance of conventional three-wheelers, a high gain step-up hybrid DC-DC converter is designed and analyzed in this paper. From the theoretical and simulation results, it is concluded that the designed converter is capable of achieving voltage gain as high as 40 times at a duty cycle of 0.8 and 180 times at a duty cycle of 0.9, significantly higher than existing solutions. The ability to generate such high step-up voltage gain by the proposed converter will allow existing electric three-wheelers and auto-rickshaws to implement traction motors with voltage ratings and performance level comparable to modern EVs, which will greatly improve the vehicle experience. In addition, the ability to deliver up to 4 KW of power with continuous input and output current makes this converter well-suited for applications that requires non-isolated high step-up DC-DC power conversion.

REFERENCES

[1] Shahriar Khan, Electrical Energy System, Fourth Edition, S. Khan, Dhaka, Bangladesh, February 2021.

[2] C. C. Chan and K. T. Chau, "An overview of power electronics in electric vehicles," in IEEE Transactions on Industrial Electronics, vol. 44, no. 1, Feb. 1997, pp. 3-13.

[3] Y. Jiang and M. Krishnamurthy, "Performance evaluation of AC machines for propulsion in a range extended electric auto rickshaw," 2012 IEEE Transportation Electrification Conference and Expo (ITEC), Dearborn, MI, 2012, pp. 1-6.

[4] Z. Amjadi and S. S. Williamson, "Power-Electronics-Based Solutions for Plug-in Hybrid Electric Vehicle Energy Storage and Management Systems," in IEEE Transactions on Industrial Electronics, vol. 57, no. 2, Feb. 2010, pp. 608-616.

[5] R. D. Middlebrook, "Transformerless DC-to-DC converters with large conversion ratios." IEEE transactions on Power Electronics 3, no. 4, 1988: pp. 484-488.

[6] B. Axelrod, Y. Berkovich, and A. Ioinovici, "Switched capacitor/switched-inductor structures for getting transformerless hybrid DC–DC PWM converters." IEEE Transactions on Circuits and Systems I: Regular Papers 55, no. 2 , 2008, pp. 687-696.

[7] M. R. Banaei, and H. A. F. Bonab, "A High Efficiency Non-Isolated Buck-Boost Converter Based on ZETA Converter." IEEE Transactions on Industrial Electronics, 2019.

[8] Shahriar Khan, Semiconductor Devices and Technology, Third Edition, ISBN: 978-094-33-5983-4, S. Khan, Dhaka, Bangladesh, June 3, 2018.

[9] S. Arfin, A. Al Mamun, T. Chowdhury and G. Sarowar, "Zeta based Hybrid DC-DC Converter using Switched Inductor and Switched Capacitor Combined Structure for High Gain Applications," 2019 IEEE International Conference on Power, Electrical, and Electronics and Industrial Applications (PEEIACON), Dhaka, Bangladesh, 2019, pp. 1-4.

[10] S. Arfin, A. Al Mamun, S. N. Mahmood and G. Sarowar, "Ćuk Derived SC-SL Based High Step Down Hybrid DC-DC Converter," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2019, pp. 1-5.

[11] L. Yang, T. Liang and J. Chen, "Transformerless DC–DC Converters With High Step-Up Voltage Gain," IEEE Transactions on Industrial Electronics, vol. 56, no. 8, Aug. 2009, pp. 3144-3152.

[12] M. R. Banaei and H. A. F. Bonab, "A Novel Structure for Single-Switch Nonisolated Transformerless Buck–Boost DC–DC Converter," IEEE Transactions on Industrial Electronics, vol. 64, no. 1, pp. 198-205, Jan. 2017.

Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection

Fahima Hossain [1, *]
Department of Computer Science
& Engineering
Jagannath University
Dhaka-1100, Bangladesh
minda.fahima25@gmail.com

Mohammed Nasir Uddin [2]
Department of Computer Science
& Engineering
Jagannath University
Dhaka-1100, Bangladesh
nasir.jnu.cse@gmail.com

Rajib Kumar Halder [3, *]
Department of Computer Science
& Engineering
Jagannath University
Dhaka-1100, Bangladesh
rajib.cse1346@gmail.com

***Abstract*— Spam and non-spam email identification are one of the most challenging tasks for both email service providers and consumers. The spammers try to spread misleading facts through irritating messages by attracting user's attention. Several spam identification-models have previously been proposed and tested but the recorded accuracy has shown that further work in this direction is needed to achieve improved accuracy, low training time, and less error rate. In this research work, we have proposed a model that classifies the e-mail into spam and ham. DBSCAN and Isolation Forest are used to identify the extreme values outside of the specific range. Heatmap, Recursive Feature Elimination, and Chi-Square feature selection techniques are used to select the effective features. The proposed model is implemented in both machine learning and deep learning to establish a comparative analysis. Multinomial Naïve Bayes (MNB), Random Forest (RF), K-Nearest Neighbor (KNN), Gradient Boosting (GB) are used to introduce ensemble method in machine learning implementation. Recurrent Neural Network (RNN), Gradient Descent (GD), Artificial Neural Network (ANN) for deep learning implementation. An ensemble method is constructed to combine multiple classifiers' output. The ensemble methods allow producing better prediction accuracy compared to a single classifier. Our proposed model obtained an accuracy of 100%, AUC=100, MSE error = 0 and RMSE error = 0 for machine learning implementation and accuracy of 99%, loss value= 0.0165 for deep learning implementation based on an email spam base dataset collected from the UCI machine learning repository.**

Keywords— DBSCAN, Isolation Forest, Feature Selection, Classification, Machine Learning, Deep Learning

I. Introduction

Electronic mail is a system used by electronic devices to exchange messages between sender and receiver. Email operates via computer networks, mainly the Internet. Email services are operated by a client server architecture where web-based email is the client and POP3 (Post Office Protocol 3), IMAP (Internet Message Access Protocol), MAPI (Messaging Application Programming Interface) are the email servers. In the developing world, email is commonly used by companies, governments, and non-governmental organizations. The number of global e-mail users reached approximately 3.823 billion, 3.930 billion, and 4.037 billion in 2018, 2019 and 2020 respectively [1]. But the overwhelming nature of unsolicited mail has complicated its use. Unsolicited email spam, also known as junk email, is an email that is sent in bulk. Spam email has an attractive look and most of them contains tempting pictures or text to draw the attention of users. Some spam emails are often used as an advertising tool to deliver advertisements to a large number of target users via the internet. While spam emails of this kind are often not harmful, just waste users time. In addition, spam emails play a vital role in phishing where a fake links are embedded in emails by spammers and directs the users to fake web site [7]. So, it is necessary to identify which emails are spam and which are ham to protect the email users from spammers. In case of spam identification, both machine learning and deep learning model must be able to decide if the sequence of words contained in an email is more similar to spam emails or not.

The main objective of this research work is to design a two-dimensional identification system to predict spam based on email dataset and appropriate preprocessing mechanism to make the data suitable for better classification.

The rest of the paper is organized as follows: section II gives details of the existing works; section III illustrates details about data and methodology. In section IV, we describe the evaluation, validation, experiments. And the last two sections are about conclusion and future work.

II. Related Works

Shreyasi Sinha, Isha Ghosh, and Suresh Chandra Satapathy et. al [2], proposed a model based on backpropagation and backpropagation with momentum to perform spam detection. The authors optimized the model using SGO (Social Group Optimization) to improve the classification performance. Neural network is used to work with any types of data (text, audio, image etc.) for both classification and clustering purpose. The use of backpropagation has one disadvantage that is, it needs more iterations and thus increases computation time. The authors didn't introduce any idea about removing outliers as outliers have a great impact on textual dataset and normalization technique often increases variation among the data points. They have added only 2 hidden layers. Less number of hidden layers do not always produce better performance.

Vashu Gupta, Aman Mehta, Akshay Goel, Utkarsh Dixit et. al [3], presented a spam detection model based on the combinations of classifiers such as Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, and Decision Tree and produced an ensemble method using voting

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

classifier. They used voting classifier to offer more accurate prediction accuracy than the individual classifiers. This system didn't handle the noisy data point.

Deepika Mallampati, K. Chandra Shekar et. al [4], built an efficient technique to filter spam email using SVM classifier to work with non-linear data. The authors introduced the use of kernel function in SVM. The kernel function is introduced to separate non-linear data by converting data into a higher dimensional feature space. But the use of kernel function often brings a number of difficulties such as it elevates the training time and memory size requirement by creating excessive support vectors. Furthermore, using only a single classifier do not always gives better predictive accuracy.

Haoyu Wang, Bingze Dai et. al [5], developed a predictive scheme using Bayesian linear regression and random forest regression for numerical prediction on spam dataset. The authors achieved better accuracy and lower MSE for random forest regression algorithm than the Bayesian linear regression model. In this work, there is still a need to extract more relevant attributes for developing a spam detection model.

Sunday Olusanya Olatunji [6], proposed support vector machines-based model to differentiate spam and non-spam emails. The author attempted to find optimal parameter set to improve mode's performance. As the performance of SVM depends on its parameter, so a parameter search algorithm has been implemented to apply to SVM. SVM has been deployed in this work as it can work with minimum number of samples to give better predictive accuracy. They have put all the efforts to improve only the detection accuracy. This system has a time complexity to in training phase. A detection system should not only be concerned about predictive accuracy but it should also perform timely.

Rozita Talaei Pashiri, Yaser Rostami et al. [7], emphasized on building a spam detection model based on feature selection method. They selected the relevant features using the Sin-Cosine Algorithm (SCA). The authors had been able to reduce the error in feature selection compared to MLP (Multi-Layer Perceptron). The features were updated using metaheuristic algorithm before applying to MLP. In this work, a single neural network is used that is a drawback.

III. Methodology

This section illustrates a detailed description about the methodology of this research work and the model consists of five steps.

1. **Data Collection:** In this work, a spam-base dataset has been collected from UCI (University of California, Irvine) machine learning repository that contains total 58 attributes where 57 are independent features and one is dependent feature [8].

TABLE I. Information on spam-base dataset.

Attribute Types	Fifty five real and continuous attributes; two integer and continuous attributes; one nominal attribute
Number of Instances	4601
Number of spam instances	1813
Number of non-spam instances	2788
Used for	Classification
Characteristics of Dataset	Multivariate
Number of class in target feature	2 (0 denotes the email as non-spam and 1 represents the email as spam)

2. **Data Cleaning:** Data cleaning means making the data processable for a machine to build an efficient machine learning model.

   I. **Converting to integer:** The data points that are continuous and floating point are converted to integer to increase processing flexibility. The floating-point data values create additional variation (range) among data points. Thus, this variation increases computational complexity. Computational complexity decreases machine's predictive performance.

   II. **Removing Outliers:** Outliers are removed from the dataset after converting the data from continuous and floating point into integer. Outliers are the data points whose distribution are different from the normal data points. These outliers affect the mean, standard deviation, skewness and distribution of the dataset. It causes biasness in machine learning model. So, these outliers should be eradicated from the dataset to stop being affected by these reasons. Outliers can be removed from the dataset using various methods. In this research work, two methods DBSCAN and Isolation Forest are applied to remove outliers.

3. **Feature Engineering:** Feature engineering is a process which is executed to construct a subset of features from the original feature set. It reduces training time, overfitting and improves applicability and accuracy of machine learning model. In this research work, three well-known feature selection techniques are implemented to extract the effective features.

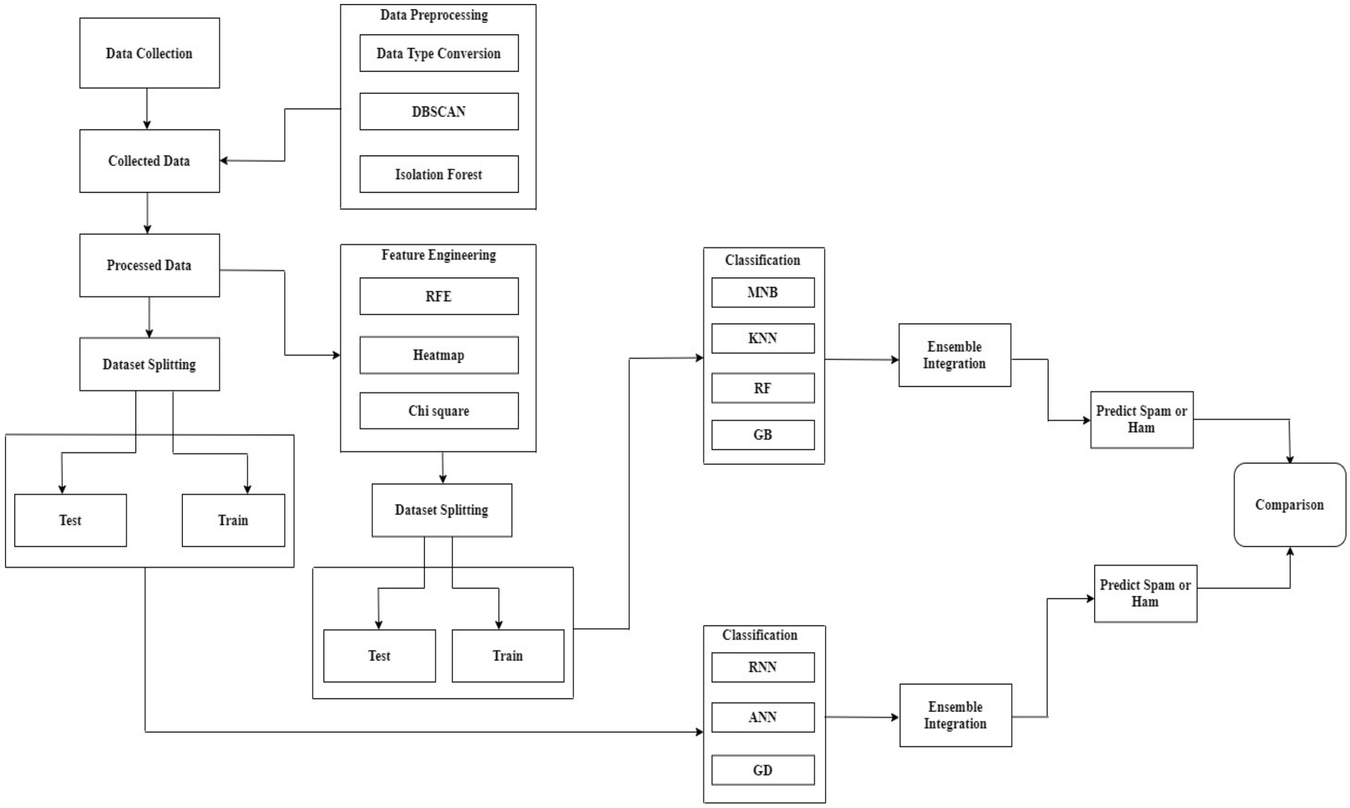


Fig.2. Proposed architecture for spam detection.

4. **Data Splitting:** In this step, the dataset is divided into 80% training set and 20% testing set. Training set will be used to train the model and testing set will be used to validate the performance of the model.

5. **Classification & Prediction:** Both machine learning and deep learning techniques are implemented for classifying the email. The training and testing set are provided to the classification algorithms to classify the email as spam or ham. Four classification algorithms: Multinomial Naïve Bayes, K-Nearest Neighbor, Random Forest and Gradient Boosting are implemented to integrate the output from these base classifiers that is integrated using an ensemble integration method named Stacking. Artificial Neural Network, Recurrent Neural network and Gradient Descent are implemented to get a combined outcome.

   a. **Multinomial Naïve Bayes:** Multinomial Naïve Bayes is an extension of Naïve Bayes where each feature in a feature vector is assigned a weight. Feature vector stores the number of occurrences of a feature. It works comparatively faster than the Naïve Bayes algorithm. It is mainly useful for textual documents to perform word count. To deal with new words, it uses a smoothing technique called Laplacian smoothing (determines how many times the new word appeared). This algorithm works in two steps: training and testing.

   b. **K-Nearest Neighbor:** K-Nearest Neighbor is a type of supervised learning algorithm which can be used for classification and regression purposes. It classifies the new samples depending on the similarity with the training samples.

   c. **Random Forest:** Random forest is a boosting algorithm which is an ensemble learning method. Random forest creates many classification trees. It uses no pruning strategy.

   d. **Gradient Boosting:** Gradient Boosting is also an ensemble method that combines the output from weak learner and produce a model with better accuracy. It builds regression trees starting from a single leaf. Instead of creating trees for each of the attribute it creates trees for all the observations. Then split into two groups the predictor. The predictor is selected in a way that reduces residual error [11]. In the next step, this algorithm tries to minimize the error and iterates continuous to make decision trees.

   e. **Recurrent Neural network (RNN):** RNN is a type of ANN which finds the characteristics of data and from the characteristics extract the pattern. Using these patterns, RNN predict on given data. RNN has one input layer, one output layer and a number of hidden layers. The number of hidden layers vary as per the requirements of a model. Each layer of RNN stores information about the previous layer. It has a memory in the hidden layer to store this information. Same biases and weights are given to each of the layer to convert output from input.

f. **Gradient Descent:** Gradient Descent is an optimization algorithm which tries to find an optimal solution by going through iteratively to reduce the loss function. It is mostly used when we want to reduce the cost function. It tries to fit a line like the one in linear regression. In gradient descent intercept is provided for a random guess. With the value of the intercept, it calculates the predicted value.

g. **Artificial Neural Network (ANN):** ANN, also known as Feed Forward Neural Network is combined with a group of perceptrons in its each layer. Activation function in ANN provides the ability to process non-linear function. This function maps the inputs to outputs. The inputs are converted to output using feed forward method and the inputs are updated using backpropagation method.

IV. Performance Analysis

We got 2504 and 4147 instances from original dataset by applying DBSCAN and Isolation Forest respectively. Using DBSCAN algorithm for data preprocessing, Heatmap selects 29 attributes, Chi-Square selects 44 attributes, Recursive Feature Elimination selects 44 attributes from 58 attributes. The dataset which is processed using Isolation Forest, Heatmap selects 27 attributes, Chi-Square selects 44 attributes and Recursive Feature Elimination selects 22 attributes from 58 attributes. Various measures are used to evaluate the performance of the research work. Confusion matrix, Accuracy, precision, recall, sensitivity, specificity and F1-score are calculated for machine learning techniques. Loss function and accuracy is calculated to measure the performance of the deep learning techniques. Confusion Matrix is a table which separates classes in binary classification problem. Elements of a confusion matrix are False Negative (FN), False Positive (FP), True Negative (TN) and True Positive (TP).

Confusion Matrix after applying DBSCAN,

$$\begin{bmatrix} 630 & 0 \\ 0 & 247 \end{bmatrix}$$

Confusion Matrix after applying Isolation Forest,

$$\begin{bmatrix} 499 & 0 \\ 0 & 331 \end{bmatrix}$$

AUC-ROC is a curve that is used to validate a classification model. In this visualization TPR (True Positive Rate) and FPR (False Positive Rate) is represented along two axes in a graph where X-axis denotes FPR values and Y-axis denotes TPR values. A model with higher AUC means the model has a better chance to classify 0’s as 0’s and 1’s as 1’s.

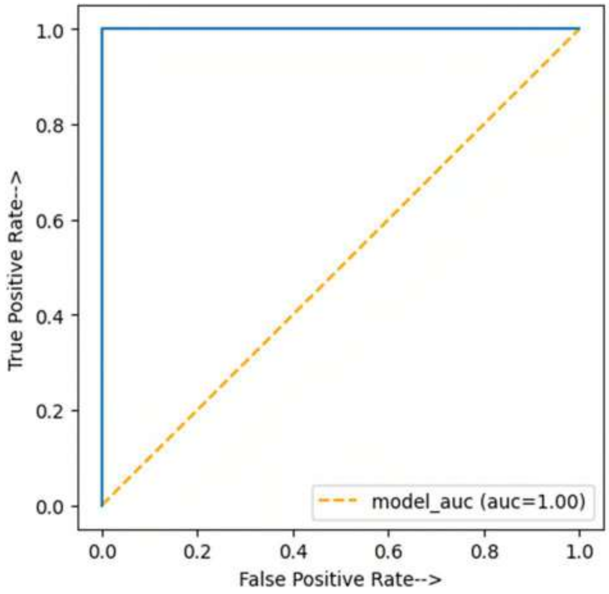


Fig. 3. AURUC Curve for proposed model with ML (Machine Learning).

The proposed machine learning model obtained AUC =1.00. That means, this model has 100% possibility to classify emails as spam or ham.

Result analysis based on Machine Learning algorithms:

TABLE I. Result based on DBSCAN.

Parameter Names	Values
Training Time	1.7619 secs
Testing Time	0.0718 secs
Accuracy	100.0 %
Precision	1.0
Recall	1.0
F1-score	1.0
Execution Time	1.4535 secs

TABLE II. Result based on Isolation Forest.

Parameter Names	Values
Training Time	2.6927 secs
Testing Time	0.0901 secs
Accuracy	100.0 %
Precision	1.0
Recall	1.0
F1-score	1.0
Execution Time	2.2244 secs

TABLE III. Result based on DBSCAN preprocessing and single classifier.

Feature Selection Method	Classifier	Accuracy
Heatmap	MNB	73.0539 %
Heatmap	KNN	82.6347 %
Heatmap	RF	88.8224 %
Heatmap	GB	89.6208 %
Chi-Square	MNB	83.2335 %
Chi-Square	KNN	84.6307 %
Chi-Square	RF	92.4152 %
Chi-Square	GB	93.2136 %
RFE	MNB	77.6447 %
RFE	KNN	85.2295 %
RFE	RF	91.4172 %
RFE	GB	93.0139 %

TABLE IV. Result based on Isolation Forest preprocessing and single classifier.

Feature Selection Method	Classifier	Accuracy
Heatmap	MNB	64.8193 %
Heatmap	KNN	77.1084 %
Heatmap	RF	89.5181 %
Heatmap	GB	85.5422 %
Chi-Square	MNB	76.6265 %
Chi-Square	KNN	77.1084 %
Chi-Square	RF	90.8434 %
Chi-Square	GB	88.9157 %
RFE	MNB	72.7711 %
RFE	KNN	77.9518 %
RFE	RF	88.9157 %
RFE	GB	88.1928 %

TABLE V. Result based on DBSCAN and heatmap feature selection.

Parameter Names	Values
Training Time	1.8365 secs
Testing Time	0.0625 secs
Accuracy	90.8245 %
Precision	0.9099
Recall	0.7934
F1-score	0.8475

TABLE VI. Result based on Isolation Forest and heatmap feature selection.

Parameter Names	Values
Training Time	2.7809 secs
Testing Time	0.1327 secs
Accuracy	86.9879 %
Precision	0.8847
Recall	0.7689
F1-score	0.8228

TABLE VII. Result based on DBSCAN and Chi-Square feature selection.

Parameter Names	Values
Training Time	2.0886 secs
Testing Time	0.0658 secs
Accuracy	91.4893 %
Precision	0.8959
Recall	0.8285
F1-score	0.8609

TABLE VIII. Result based on Isolation Forest and Chi-Square feature selection.

Parameter Names	Values
Training Time	3.1713 secs
Testing Time	0.1267 secs
Accuracy	90.7631 %
Precision	0.8775
Recall	0.8900
F1-score	0.8837

TABLE X. Result based on DBSCAN and RFE feature selection.

Parameter Names	Values
Training Time	2.0736 secs
Testing Time	0.0638 secs
Accuracy	92.2872 %
Precision	0.9215
Recall	0.8037
F1-score	0.8585

TABLE XI. Result based on Isolation Forest and RFE feature selection.

Parameter Names	Values
Training Time	2.7022 secs
Testing Time	0.1287 secs
Accuracy	89.6386 %
Precision	0.8659
Recall	0.8712
F1-score	0.8685

Result analysis based on Deep Learning algorithms:

TABLE XII. Result based on DBSCAN and Isolation Forest.

Outlier Detection Technique	Deep Learning Algorithms	Loss Value	Accuracy
DBSCAN	Recurrent Neural Network (RNN)	0.6803	92.42 %
Isolation Forest	Recurrent Neural Network (RNN)	0.0450	99.28 %
DBSCAN	Gradient Descent	0.2568	89.42 %
Isolation Forest	Gradient Descent	0.0165	99.28 %
DBSCAN	Artificial Neural Network (ANN)	0.2469	91.42 %
Isolation Forest	Artificial Neural Network (ANN)	0.1333	97.95 %

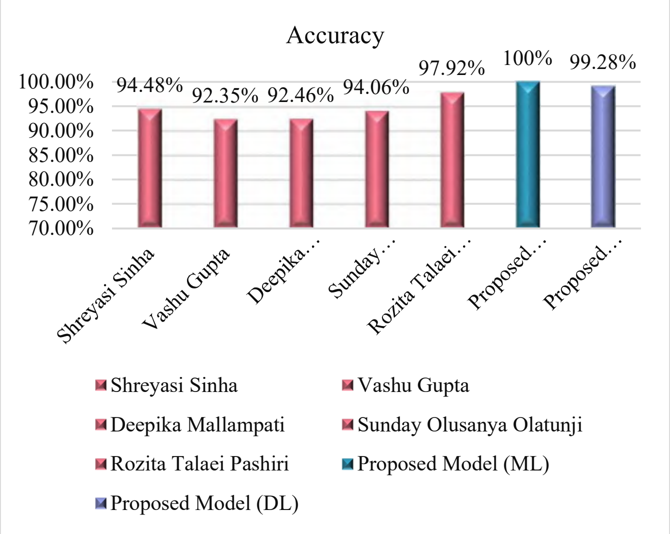


Fig. 4. Accuracy comparison with existing systems.

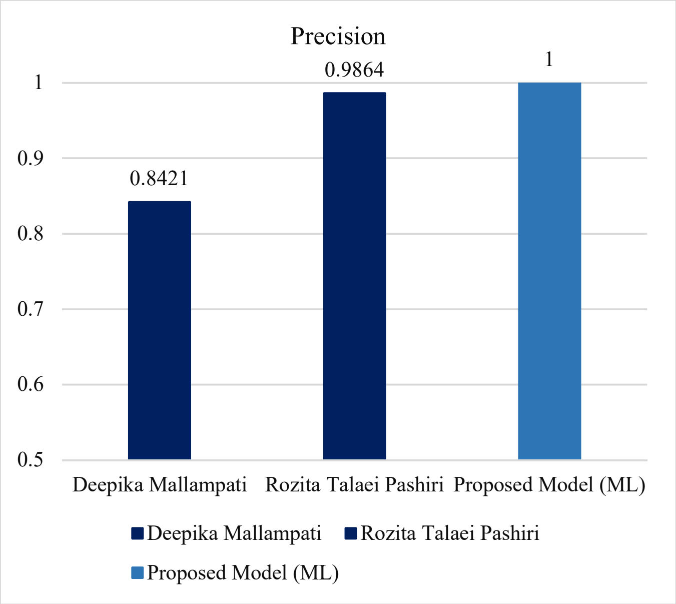


Fig. 5. Precision comparison with existing systems.

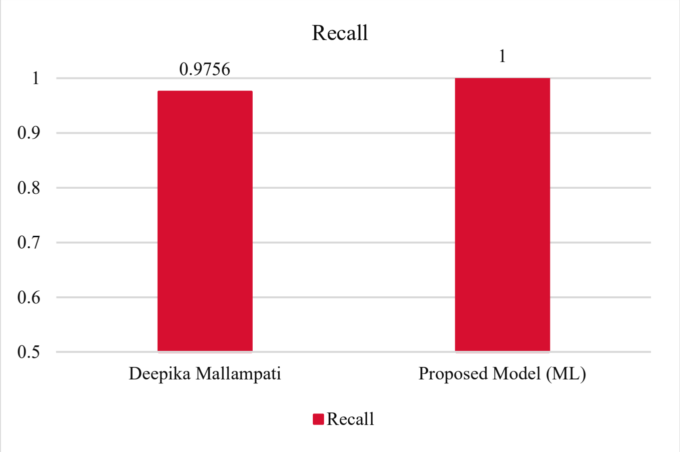


Fig. 6. Recall comparison with existing systems.

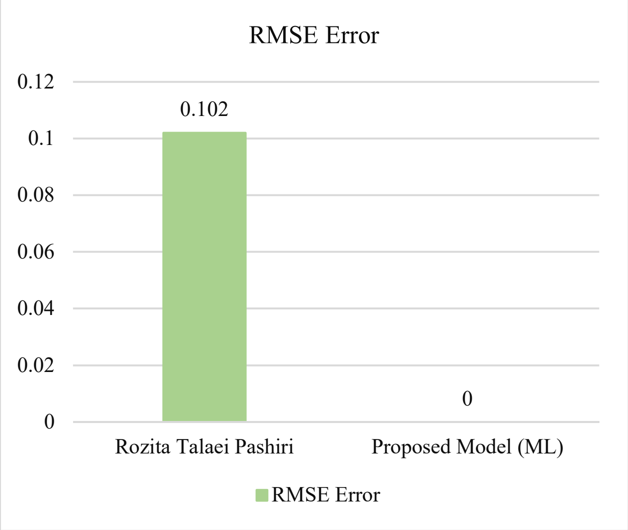


Fig. 7. RMSE error comparison with existing systems.

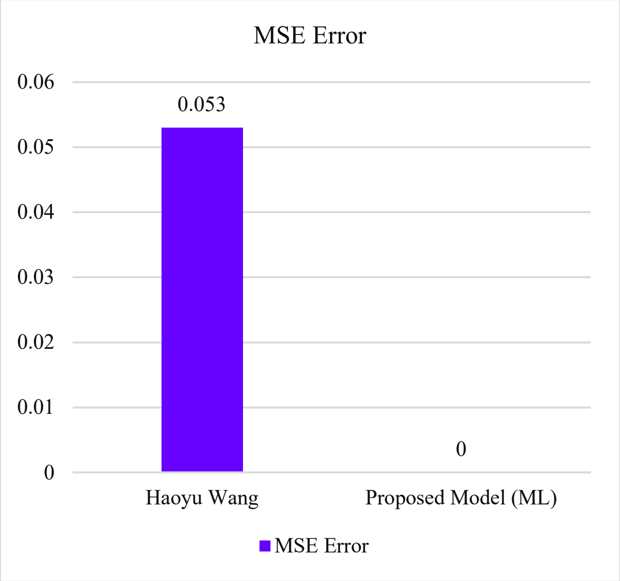


Fig. 8. MSE error comparison with existing systems.

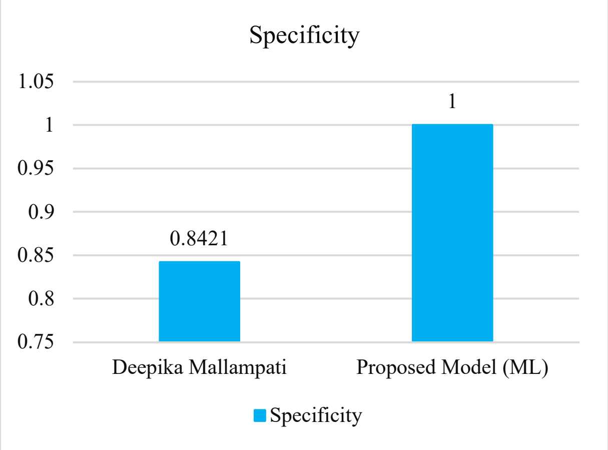


Fig. 9. Specificity comparison with existing systems.

Conclusion and Future Work

Spam filtering systems secures privacy for the email users. It analyzes the email content to take proper measures on blocking and deleting the identified spam email. In this research work, we have observed the classification performance with and without outliers and irrelevant features.

In this comparative study, we observed that Machine Learning algorithms perform better than deep learning algorithms for tabular dataset to identify spam e-mails. This model helps effectively to detect spam. In future, we will work for image spam, blank spam, backscatter spam detection and we will work with text data direct from text content instead of tabular data using natural language processing.

References

[1] "Number of e-mail users worldwide 2024 | Statista," *Statista*, 2021. [Online]. Available: https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/. [Accessed: 28- Feb- 2021].

[2] S. Sinha, I. Ghosh and S. Satapathy, "A Study for ANN Model for Spam Classification," *Advances in Intelligent Systems and Computing*, pp. 331-343, 2020. Available: 10.1007/978-981-15-5679-1_31 [Accessed 27 February 2021].

[3] V. Gupta, A. Mehta, A. Goel, U. Dixit and A. Pandey, "Spam Detection Using Ensemble Learning," *Harmony Search and Nature Inspired Optimization Algorithms*, pp. 661-668, 2018. Available: 10.1007/978-981-13-0761-4_63 [Accessed 27 February 2021].

[4] D. Mallampati, K. Chandra Shekar and K. Ravikanth, "Supervised Machine Learning Classifier for Email Spam Filtering," *Innovations in Computer Science and Engineering*, pp. 357-363, 2019. Available: 10.1007/978-981-13-7082-3_41 [Accessed 27 February 2021].

[5] H. Wang, B. Dai and D. Yang, "A Comparative Study of Two Different Spam Detection Methods," *Communications in Computer and Information Science*, pp. 95-105, 2019. Available: 10.1007/978-981-15-1304-6_8 [Accessed 27 February 2021].

[6] S. Olatunji, "Improved email spam detection model based on support vector machines," *Neural Computing and Applications*, vol. 31, no. 3, pp. 691-699, 2017. Available: 10.1007/s00521-017-3100-y [Accessed 27 February 2021].

[7] R. Talaei Pashiri, Y. Rostami and M. Mahrami, "Spam detection through feature selection using artificial neural network and sine–cosine algorithm," *Mathematical Sciences*, vol. 14, no. 3, pp. 193-199, 2020. Available: 10.1007/s40096-020-00327-8 [Accessed 27 February 2021].

[8] "UCI Machine Learning Repository: Spambase Data Set," *Archive.ics.uci.edu*, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Spambase. [Accessed: 25- Feb- 2021].

[9] H. Zhang, B. Liu, P. Cui, Y. Sun, Y. Yang and S. Guo, "An Outlier Detection Algorithm for Electric Power Data Based on DBSCAN and LOF," *Proceedings of the 9th International Conference on Computer Engineering and Networks*, pp. 1097-1106, 2020. Available: 10.1007/978-981-15-3753-0_110 [Accessed 25 February 2021].

[10] S. Tripathy and L. Sahoo, "Improved Method for Noise Detection by DBSCAN and Angle Based Outlier Factor in High Dimensional Datasets," *Lecture Notes in Electrical Engineering*, pp. 213-221, 2019. Available: 10.1007/978-981-13-8715-9_27 [Accessed 25 February 2021].

[11] G. Madhukar Rao and D. Ramesh, "A Hybrid and Improved Isolation Forest Algorithm for Anomaly Detection," *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pp. 589-598, 2020. Available: 10.1007/978-981-15-7234-0_55 [Accessed 25 February 2021].

[12] M. Togbe et al., "Anomaly Detection for Data Streams Based on Isolation Forest Using Scikit-Multiflow," *Computational Science and Its Applications – ICCSA 2020*, pp. 15-30, 2020. Available: 10.1007/978-3-030-58811-3_2 [Accessed 25 February 2021].

[13] Y. Sun and J. Platoš, "Text Classification Based on Topic Modeling and Chi-square," *Advances in Intelligent Systems and Computing*, pp. 513-520, 2020. Available: 10.1007/978-981-15-3308-2_56 [Accessed 25 February 2021].

[14] K. Rao and G. Rao, "Ensemble learning with recursive feature elimination integrated software effort estimation: a novel approach," *Evolutionary Intelligence*, 2020. Available: 10.1007/s12065-020-00360-5 [Accessed 25 February 2021].

[15] B. Gaye and A. Wulamu, "Sentiment Analysis of Text Classification Algorithms Using Confusion Matrix," *Communications in Computer and Information Science*, pp. 231-241, 2019. Available: 10.1007/978-981-15-1922-2_16 [Accessed 26 February 2021].

[16] M. Asghar, A. Ullah, S. Ahmad and A. Khan, "Opinion spam detection framework using hybrid classification scheme," *Soft Computing*, vol. 24, no. 5, pp. 3475-3498, 2019. Available: 10.1007/s00500-019-04107-y [Accessed 26 February 2021].

[17] J. Yoon, "Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach," *Computational Economics*, vol. 57, no. 1, pp. 247-265, 2020. Available: 10.1007/s10614-020-10054-w [Accessed 28 February 2021].

A Hypothesis Testing tool for the comparison of different Cyber-Security Mitigation Strategies in IoT

Asterios Mpatziakas, Stavros Papadopoulos, Anastasios Drosou, Dimitrios Tzovaras
Information Technologies Institute
Centre for Research and Technology (CERTH)
Thessaloniki, Greece
{ampatziakas, spap, drosou, Dimitrios.Tzovaras@iti.gr} @iti.gr

***Abstract*—Internet of Things (IoT) is a field with tremendous growth that already shows great impact in numerous domains. Simultaneous with this development is the need for better Cyber-security: IoT systems are attacked by various adversaries targeting IoT services, platforms and networks, which can have disruptive consequences. These attacks can be countered using multiple strategies with different effects to the system. The following paper, proposes a novel approach based on Machine Learning and Statistical Hypothesis Testing, which allows the security operator to investigate how using different strategies affects various KPI related to the security of the IoT network and if the KPI resulting from modifications to a mitigation strategy are statistically different when compared to those occurring from a starting mitigation action set.**

***Index Terms*—Internet of Things, Machine Learning, Hypothesis Testing, Cyber-Security, Attack Mitigation**

I. Introduction

From Industry 4.0 to smart cities, the growth of the Internet of Things (IoT) ecosystem has been empowering tremendous advancements and extensive impact in multiple domains while more industrial and commercial opportunities can be foreseen in the future. In tandem with this expansion, has been the rise of attacks against IoT systems, attacks which can potentially become as pervasive as the spread of IoT applications. Successful attacks against the IoT can cause disruption of services provided, breach of sensitive or private data, huge monetary losses, damage to property and even physical harm [1].

Designing secure IoT systems is the first step towards guarding them from malicious acts. However, new exploits and vulnerabilities are constantly discovered, so a robust attack detection system combined with methods and tools to mitigate new threats is an extremely important part to preserve the soundness of any IoT system. This is a challenging task since there are multiple attacks targeting different layers of the IoT network, and each of these attacks can be counter-measured by multiple mitigation actions [2].

The plethora of possible attack responses, creates the need of tools that can help the network operator investigate how different mitigation actions can affect the network so that she can make proper decisions for its safeguarding. The outcomes of different mitigation strategies can be quantified by using carefully selected Key Performance Indicators (KPI). Such a tool is described and evaluated in the remainder of this paper.

A. Key Contributions

This paper presents a novel approach to distinguish differences between sets of mitigation actions used to counter an attack or threat against an IoT network. It allows the Security Operator of the IoT network to modify a set of existing mitigation actions and see the impact of the changes. More specifically, the KPI values resulting from different mitigation sets are used to ascertain clusters of such sets using a machine learning algorithm. Then, the difference between these clusters, is evaluated by means of a p-value provided by a method based on Statistical Hypothesis Testing. To our knowledge, this is a novel approach not suggested elsewhere in available the literature.

The rest of the paper is structured as follows: Section II briefly presents relevant literature and Section III presents the methodology used to formulate the proposed approach while section IV contains the results of experiments that were performed to validate the proposed method. Finally, section V contains the conclusions of this paper along with future research directions and aims.

II. Related Work

This section briefly presents the use of Hypothesis Testing based algorithms in the context of IoT networks through examples found in recently published literature, along with details concerning four KPIs used in this work to quantify the effects of one or more mitigation actions in the system.

A. Statistical Hypothesis Testing in the context of IoT networks

Hypothesis testing in the context of statistics, is the use of a sample of data to evaluate the plausibility of a hypothesis concerning the distributions of the sample data. Numerous applications of algorithms based on Hypothesis testing have been successfully applied to IoT domain specific problems.

This work is supported by the European Unions Horizon 2020 Research and Innovation Program through the SerIoT project under Grant Agreement No. 780139.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Such methods have been extensively employed for attack detection, for example: In [3] such an algorithm is used to detect a Link flooding attack on an IoT Network, in [4] Hypothesis Testing is used against spectrum sensing data falsification in cognitive IoT Networks while in [5] Li et. al use it to empower an distributed attack detection System.

Examples of other uses are numerous: Hypothesis Testing based algorithms were used to perform polling for the values of multiple KPIs in an fog based IoT sensor network [6] , to facilitate protocols for the authentication of IoT devices [7] or manage the privacy of Smart Energy Meters [8]. Moreover, decentralized methods have been proposed, to preserve IoT sensor energy consumption [9], to safeguard privacy [10] or even preserve robustness in the case of existing noise and occurence small data sets [11]. To our knowledge, no other hypothesis-based application for evaluating different mitigation strategies in IoT has been presented in the available literature.

B. Security Related Key Performance Indicators for the selection of Mitigation Actions

The following section presents four Key Performance Indicators (KPIs), selected to be used in this report as the metrics to be describe the mitigation actions that has been deployed to secure an IoT network. These were selected, based on the results of a extensive literature review, which is available in [12].

The first KPI is Common Vulnerability Scoring System (CVSS), which is an open Industry standard for assessing the severity of cyber-security vulnerabilities. This is accomplished by the use of a score between [0,10] with 10 representing a vulnerability with the highest severity. The score is calculated using predefined values and equations, available in [13]. Using CVSS, security experts can easily share discovered vulnerabilities via public databases such as the National Vulnerability Database [14]. Moreover, any attack vector can be easily translated to a vulnerability in the device it affects. In section IV the following formula is used for the calculation of the CVSS score for a number of mitigation actions to be applied to the system: $CVSS = 10 - mean(CVSS_{all\ detected\ vulnerabilities})$.

The second KPI is named Return on Response Investment (RORI), which is used to calculate an index associated to a set of the mitigation actions . This KPI can be used to evaluate optimal plans by ranking them based on their efficiency in stopping potential attacks, and simultaneously preserve the best possible service for users. The authors of [15] provide the formula used in this paper to calculate RORI:

- The financial cost expected to occur annualy, in the absence of applying a mitigation strategy is refered as Annual Loss Expectancy (ALE).
- Risk Mitigation (RM) estimates the coverage of one more actions in the mitigation of an attack.
- The cost expected to occur due to the application of mitigations is called Annual Response Cost (ARC) while
- The fixed cost associated to the system infrastructure, regardless of the application of a mitigation strategy is called Annual infrastructure value (AIC).

Then,

$$RORI\ Score = \frac{ALE * RM - ARC}{ARC + AIC}.$$

The third KPI is named Vulnerability Coverage (VC). VC of a mitigation action cm_i , is defined as the number of vulnerabilities it covers when applied divided by the number of total vulnerabilities so VC $\in [0,1]$ [16]. Disjoint VC includes the vulnerabilities covered by a single countermeasure, whereas joint VSC refers to the vulnerabilities covered by multiple countermeasures.

Finally , the Deployment Cost KPI evaluates the deployment costs of the mitigation actions by considering deployment time, consumed resources and the importance of the device that is affected by the countermeasure as assessed by the network security operator [17]. To calculate it, three quantities are required. First, Deployment Time (DT) which is measured in milliseconds. This is the time required for a mitigation action to be deployed. It can be assessed using historical data and be dynamically updated. Then, the Device Importance (DI) is needed, which is arbitrarily assessed by the network security operator considering the specifics of each use case. A value is assigned to each device, where Device Importance $\in [0,1]$. The last quantity needed is Resource Consumption (RC), can be imputed either by measurements or an arbitrarily chosen ranking scheme can be used based on the network operators' expertise i.e $RC = \{$Very low: 1, Low:2, Medium : 3 , High :4 , Very High : 5$\}$. Finally the KPI value is calculated by the following formula: $Deployment\ Cost = DT * DI * RC$.

III. Proposed Model

In the following section, a method to distinguish between different sets of mitigation actions is presented: More specifically, the KPI values resulting from different mitigation sets are used to cluster the them. Then, the difference between these clusters is evaluated by means of a $p-$value provided by a Monte Carlo (MC) based method named Statistical Significance of Clustering (SigClust) using Soft Thresholds. Figure 1 presents a high level overview of the proposed tool.

A brief presentation of the SigClust method with Soft Thresholds, as shown in [18] follows: Let $X = [x_1, x_2, \cdots, x_n]$, x $\in \mathbb{R}^d$, be a data-set of n observations each containing the values of d different KPIs.The method starts from the null hypothesis that the data of X come form a single multivariate Gaussian distribution $N(\mu, \Sigma)$. A test level α, e.g. $\alpha = 0.95$ is pre-specified to finally test the Hypothesis.

Let C_1 and C_2 be two disjoint sets resulting from the application of a clustering of the data points contained in X i.e. $C_1 \cup C_2 = \{1, 2, \cdots, n\}$. Then, an indicator of the strength of the clusters can be attained by the Cluster Index (CI), that is used as the test statistic of the method:

$$CI = \frac{\sum_{k=1}^{2} \sum_{i \in C_k} ||x_i - \bar{x}^{(k)}||^2}{\sum_{i=1}^{n} ||x_i - \bar{x}^{(k)}||^2}, \qquad (1)$$

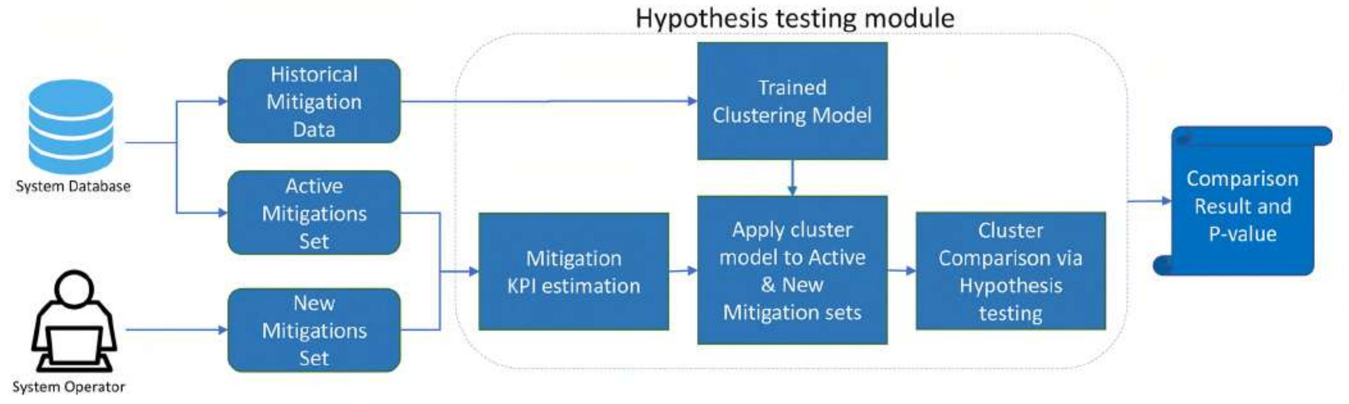

Hypothesis testing module
System Database
System Operator
Historical Mitigation Data
Active Mitigations Set
New Mitigations Set
Trained Clustering Model
Mitigation KPI estimation
Apply cluster model to Active & New Mitigation sets
Cluster Comparison via Hypothesis testing
Comparison Result and P-value

Fig. 1. High level overview of the Hypothesis Testing Tool.

where $\bar{x}^k$ is the mean of cluster $k \in [1,2]$ while $\bar{x}$ is the overall mean. Estimate values $(\hat{\lambda}_1, \cdots, \hat{\lambda}_d)$ for the eigenvalues of Σ must be computed. Let $(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_d)$ be the eigenvalues of the sample covariance matrix.The covariance matrix Σ can be written as

$$\Sigma = \Sigma_0 + \sigma_N^2 I \tag{2}$$

for a low rank positive semi-definite matrix Σ_0. Let W_0 be a positive semi-definite matrix W_0 with rank(Σ_0)=rank(W_0). Then the precision matrix C can be defined, where

$$C \equiv \Sigma^{-1} \equiv (\Sigma_0 + \sigma_N^2 I)^{-1} = \frac{1}{\sigma_N^2} I - W_0 \tag{3}$$

To estimate Σ, the negative log-likelihood is minimized to using C and the sample covariance

$$\tilde{\Sigma} = \frac{(X - \bar{X})(X - \bar{X})^T}{n}, \tag{4}$$

to yield,

$$argmin_c[-log|C| + trace(C\tilde{\Sigma})], \tag{5}$$

subject to 3 and $C, W_0 \succeq 0$. Let $M \geq 0$, be a tuning parameter. An additional constrained is set to control the signal versus the noise of the data:

$$trace(W_0) \leq M \tag{6}$$

Finally,

$$\hat{\lambda}_j = \begin{cases} \tilde{\lambda}_k - \tau, \text{if } \tilde{\lambda}_k > \tau + \sigma_N^2 \\ \sigma_N^2, \text{if } \tilde{\lambda}_k \leq \tau + \sigma_N^2 \end{cases} \tag{7}$$

where τ is obtained by solving

$$\sum_{k=1}^{d} \left(\frac{1}{\sigma_N^2} - \frac{1}{(\tilde{\lambda}_k - \tau)_+} \right)_+ \tag{8}$$

Using the computed eigenvalues, the theoretic optimal CI value is obtained by

$$TCI = 1 - \frac{2}{\pi} \frac{max((\hat{\lambda}_1, \cdots, \hat{\lambda}_d))}{\sum_{i=1}^{d} \lambda_i} \tag{9}$$

The rest of the process is comprised by the following four steps:

1) Initially, data from the null distribution is simulated: $(x_1, ..., x_d)$ are independent with $x_j \sim N(0, max(\hat{\lambda}_j, \hat{\sigma}_N^2))$.
2) Then, this data is clustered using the k-means algorithm with k = 2; the corresponding CI value are calculated.
3) By repeating this process a large number of times, an empirical distribution of the values of CI is obtained. Using the CI values obtained by the simulation, calculate a $p-$value for the CI value of X.
4) Finally, a conclusion can be derived based on test level α.

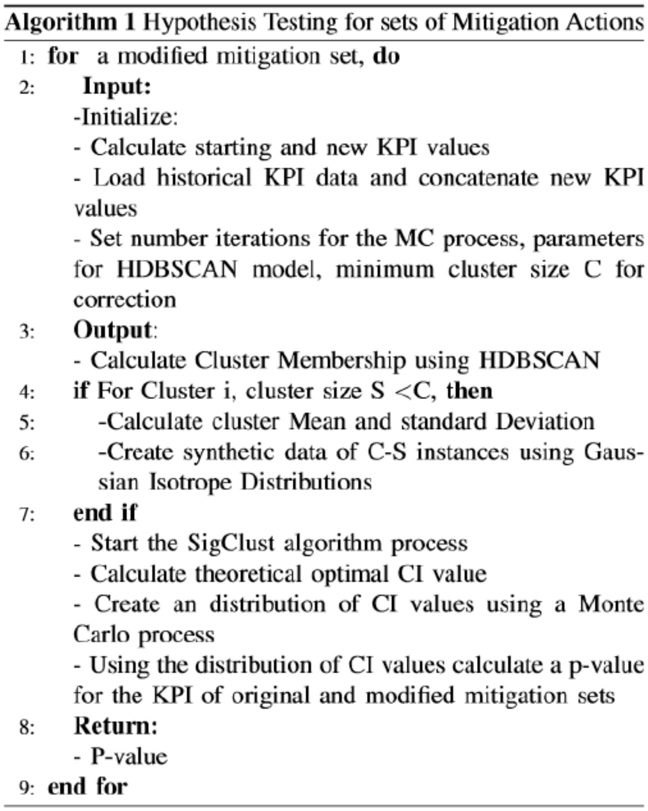
By using the p-value from the final step the following hypothesis is answered:

- H_o The clusters come from the same distribution ($p \leq 0.05$) or
- H_1 The clusters come from a different distribution ($p \geq 0.05$).

To apply this method, to distinguish the difference between different mitigation action sets the following procedure followed: Let $X = [x_1, x_2, \cdots, x_n]$, $x \in \mathbb{R}^d$, be a data set of n historical observations each containing the values of d different KPIs, with x_n being the latest observation. Let x_{n+1} be the KPI values occurring from modifying the mitigation actions that produce x_n. Some clustering algorithm is applied to all data, resulting to a partition of clusters $C = \{C_1, C_2, \cdots C_j\}, j \leq n$ for all data points.

Let C_A, C_B be the clusters, in which the data-points x_n and x_{n+1} belong to. If the size of cluster size S is smaller than a predefined minimum size C, then the following correction is applied: The Mean and Standard Deviation of each KPI is calculated. These are used as input to create and $C - S$ synthetic data points for the cluster using a Gaussian Isotrope Distribution i.e. a Gaussian Distribution where the covariance matrix is represented by the simplified matrix $\Sigma = \sigma^2 I$.

Proceeding the process, in the case when $C_A = C_B$ the case is trivial and no testing is required since the points belong to the same cluster. Else, let $x_{cluster} \subset X$ be the subset of all



Algorithm 1 Hypothesis Testing for sets of Mitigation Actions

```
for  a modified mitigation set, do
    Input:
    -Initialize:
    - Calculate starting and new KPI values
    - Load historical KPI data and concatenate new KPI values
    - Set number iterations for the MC process, parameters for HDBSCAN model, minimum cluster size C for correction
    Output:
    - Calculate Cluster Membership using HDBSCAN
    if For Cluster i, cluster size S <C, then
        -Calculate cluster Mean and standard Deviation
        -Create synthetic data of C-S instances using Gaussian Isotrope Distributions
    end if
    - Start the SigClust algorithm process
    - Calculate theoretical optimal CI value
    - Create an distribution of CI values using a Monte Carlo process
    - Using the distribution of CI values calculate a p-value for the KPI of original and modified mitigation sets
    Return:
    - P-value
end for
```

points that belong either in C_A or C_B. Then the Statistical Significance of Clustering methodology described above is applied as is, answering the question: "Are C_A or C_B different in terms of their underlying distribution and is this difference statistically significant?".

IV. Performance Evaluation

The following section contains the evaluation of the performance of the Hypothesis Testing Tool under two different scenarios. The first scenario is used to showcase the approach in a data-set with clear cluster memberships. Additionally, it is used to investigate the sensitivity of the approach to the change of different parameters of the clustering algorithm and the underlying data set size.

The second scenario, investigates the operations and the effectiveness of the proposed method to a data set with noise and no clear cluster membership. Table III shows the steps of the algorithm used for Hypothesis Testing.

The Hypothesis Testing tool was written in Python. The experiments using the Jackstraw method where performed with the implementation described in [19] using the R Statistical Language.

A. Experimental setup

For the clustering of the KPI values, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [20] algorithm was used. This machine learning algorithm has been shown to operate well in the presence of data with noise and outliers, successfully recognizing clusters with arbitrary shapes, different sizes and dissimilar densities.

Two sets of four KPI values are examined: One is completely synthetic with lack of noise while the second is obtained by simulating a scenario where 150 devices are under attack, i.e. 30 of each type of the devices presented in table I.

The HDBSCAN algorithm is heavily infected by choice of a parameter that indicates the minimum number of members a cluster must have to be considered valid. To determine this number an arbitrarily chosen integer m which is the max number of minimum cluster membership was chosen. Then for cluster minimum membership number $M = \{2, 3, \cdots, m\}$, HDBSCAN was used to cluster the data and calculate the value of the Variance Ratio Criterion score V [21]: a higher value indicates clusters are dense and well separated.

The proposed method, is compared to a similar SoA method originating from the field of Bio-Informatics, termed the Jackstraw method [19] as implemented in [22]. This method calculates the cluster membership significance instead of difference between clusters. It produces a p-value to answer the following Hypothesis:

- H_o The point examined belongs to the same distribution with the other members of thesame cluster ($p \leq 0.05$) or
- H_1 The point comes from a cluster with a different distribution ($p >$0.05).

To compare two clusters using the Jackstraw method, for the experimental results presented, the following assumption was used: two clusters have statistically significant difference only if the Jackstraw algorithm correctly ascertains cluster membership for the cluster members of both clusters.

B. Experimental Results

In all experiments, it is assumed that a data point i.e. the current KPI values of the system belong to a cluster and then a data point representing the modified version of the original is tested for differences, occurring by a set of modified mitigation actions.

C. Experiment 1: Synthetic data-set from three isotropic distributions

In this experiment, the proposed Hypothesis Testing tool is applied to synthetic data with clear cluster membership to investigate the sensitivity of the SigClust and HDBScan algorithms in changes to their underlying parameters: Three clusters are used as the most trivial case. Additionally, experiments with different numbers of iterations for the Monte Carlo process were carried out and the run-time for each is reported, to ascertain an appropriate number for the second experiment.

The synthetic data set was created by sampling 3 discrete isotropic Gaussian distributions with pre-specified centres and deviations, shown in Table II. Data-sets with different sizes were created, to determine the data set size needed for the method to produce correct results. Experiments with the following data set size were performed N =[200, 500, 1000, 2500, 3000, 4000, 5000, 10000] while in each case the clusters were equally sized. For each data-set, the Monte Carlo process for

TABLE I
MITIGATION ACTIONS AND RELEVANT INFORMATION REQUIRED TO CALCULATE THE KPIS SHOWN IN THE EXPERIMENTAL RESULTS

Device Name	Device Import.	Loss	CVSS ID	CVSS	Mitigation Action	Mitigation Resources	Mitigation Time	Deployment Cost
Open Vswitch	50	40	CVE 2017 9265	7.5	Honeypot	2	0.25	25
					Block	1	0.2	10
					Blacklist	1.5	0.2	15
					Block Port	1.8	0.2	18
			CVE 2018 17205	5	Honeypot	2	0.25	25
					Block	1	0.2	10
					Blacklist	1.5	0.2	15
					Block Port	1.8	0.2	18
IP camera	25	20	CVE 2018 19080	4.3	Honeypot	2	0.25	12.5
					Block	1	0.2	5
					Blacklist	1.5	0.2	7.5
					Block Port	1.8	0.2	9
			CVE 2018 19081	10	Honeypot	2	0.25	12.5
					Block	1	0.2	5
					Blacklist	1.5	0.2	7.5
					Block Port	1.8	0.2	9
			CVE 2018 19082	7.5	Honeypot	2	0.25	12.5
					Block	1	0.2	5
					Blacklist	1.5	0.2	7.5
					Block Port	1.8	0.2	9
ONOS SDN	75	40	CVE 2018 1000615	5	Honeypot	2	0.25	37.5
					Block	1	0.2	15
					Blacklist	1.5	0.2	22.5
					Block Port	1.8	0.2	27
			CVE 2018 12691	4.3	Honeypot	2	0.25	37.5
					Block	1	0.2	15
					Blacklist	1.5	0.2	22.5
					Block Port	1.8	0.2	27
Windows PC	100	60	CVE 2019 1368	2.1	Honeypot	2	0.25	50
					Block	1	0.2	20
					Blacklist	1.5	0.2	30
					Block Port	1.8	0.2	36
			CVE 2019 1359	9.3	Honeypot	2	0.25	50
					Block	1	0.2	20
					Blacklist	1.5	0.2	30
					Block Port	1.8	0.2	36
Thermostat	25	60	CVE 2017 14020	9.3	Honeypot	2	0.25	12.5
					Block	1	0.2	5
					Blacklist	1.5	0.2	7.5
					Block Port	1.8	0.2	9

TABLE II
MEAN AND STANDARD DEVIATION FOR EACH KPI PER CLUSTER FOR SCENARIO 1 AND DATA POINTS USED FOR COMPARISON

Cluster	Deployment Cost	Vulnerability Coverage	CVSS Score	ROSI Score
A	12±1	70±1	3±1	2600±1
B	15±1	52±1	5±1	2800±1
C	17±1	40±1	2±1	3000±1
Data Points	point A1: (DC: 15.5, VC: 60, CVSS: 5, ROSI: 2750), point A2: (DC: 15, VC: 52, CVSS: 5, ROSI: 2800), point B1: (DC: 12, VC: 70, CVSS: 3, ROSI: 2600), point C1: (DC: 17, VC: 40, CVSS: 2, ROSI: 3000)			

clustering Significance was run for i iterations where i = [5, 25, 100, 300, 500, 1000].

For all cases, the expected result i.e. statistical significance when the clusters came from the same clusters, was found when the data-set size was larger than N = 1000, for any and all number of iterations. However, for N $\leq$ 1000, the algorithm cannot correctly distinguish between members of the same cluster.

For the experiments three different cases are examined: In the first case, both the original and the modified data point belong to the same cluster (Cluster A). In the second and third cases, the two data points belong to different clusters: A and B, A and C respectively. In all cases, the clusters have an equal point of members.

Table III contains the results for N = 500. In the case where we examine two points from the same cluster, all results are False i.e. the test fails to distinguish that the points belong to the same cluster. However, the opposite holds when comparing points from different clusters: for all different numbers of the iterations, the algorithm correctly indicates that statistical significance is found i.e. the points come from different clusters.

D. Experiment 2: Simulated data-set

In this experiment, the tool is applied to to a data set with noise and no clear cluster membership. This data set was produced by applying the mitigation selection algorithm presented in [12], on the mitigation rules found in table I. The data

TABLE III
EXPERIMENT 1, CLUSTERING SIGNIFICANCE RESULTS FOR N = 500.

Comparison	Iterations	p-value	Run Time (s)
Original Point A1 - Modified Point A2	5	0	0.24693
	25	0	0.84057
	50	0	1.58204
	100	0	7.21314
	300	0	9.7017
	500	0.008	25.0076
	1000	0.004	53.44258
Original Point A1 - Modified Point B1	5	0	0.239022
	25	0	0.239022
	50	0	2.066215
	100	0	3.73554
	300	0	12.16492
	500	0	19.0975
	1000	0	41.70072
Original Point A1 - Modified Point C1	5	0	0.24693
	25	0	0.84057
	50	0	1.58204
	100	0	7.21314
	300	0	9.7017
	500	0	18.29001
	1000	0	36.39336

TABLE IV
CLUSTER COMPARISON RESULTS FOR EXPERIMENT 2, WITH THE PROPOSED CORRECTION, THE ORIGINAL DATA AND A SOA METHOD

Results of proposed method with Correction			
Ground Truth \ Assessment	Same Cluster	Different Cluster	Accuracy
Same Cluster	8	1	**95.74%**
Different Cluster	1	36	

Results of proposed method with original data only			
Ground Truth \ Assessment	Same Cluster	Different Cluster	Accuracy
Same Cluster	2	8	80.43%
Different Cluster	1	35	

Results of Jackstraw method			
Ground Truth \ Assessment	Same Cluster	Different Cluster	Accuracy
Same Cluster	9	0	89.13%
Different Cluster	5	32	

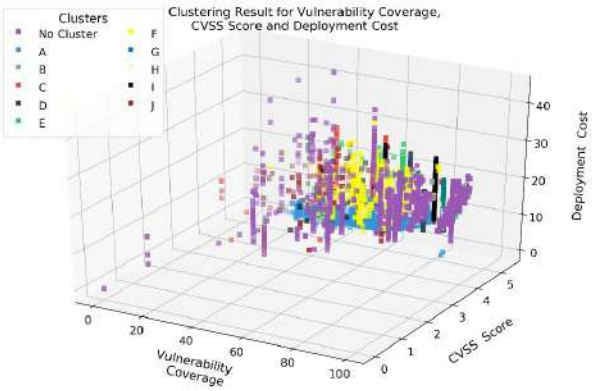


Fig. 2. Figure shows the distribution of the ten clusters found in experiment 2 in terms of the Vulnerability Coverage, CVSS score and Deployment Cost KPIs.

TABLE V
NUMBER OF POINTS PER CLUSTER FOR EXPERIMENT 2

Cluster	Number of points
Not Assigned	1016
A	3006
B	165
C	171
D	304
E	328
F	464
G	226
H	257
I	173
J	179

examined in this experiment, was found to optimally contain ten clusters, with a minimum of 150 members (variance V = 9240.615), while a lot of the points were determined to be noise i.e. not assigned to any of the clusters. Figure 2 shows the distribution of the data and the clusters found, while Table V contains number of the points belonging to each cluster. The data-set has 6300 unique data points, in some cases not exceeding the threshold of N >1000, determined to be needed in experiment 1.

For this experiment, all clusters were cross-examined for significance (47 unique pairs), using both the Sigclust and the Jackstraw methods. The algorithm was initially tested without the cluster size correction: It correctly assessed only 80.43 %: The majority of the assessments is correct if different clusters are compared but erroneous assessments occur when members of the same cluster were compared.

However, based on the fact there are clusters with membership count lower than 1000, the cluster size correction was applied: For each such cluster, enough synthetic data using Gaussian isotropic distributions are created to reach membership threshold $C = 1000$. This data was only used for the case of a comparison of a cluster with itself and not for clustering or for inter-cluster comparison . Using the correction, the number of correct assessments rose to 95.65 %. The proposed method with the correction outperforms both the method without the correction and the SoA method which achieves 89.13 % accuracy. Finally, the F_1 scores for each method are: 0.8889 for the corrected method, 0.3077 for the method without a correction and 0.7619 for the Jackstraw methods.

V. CONCLUSIONS

This paper presented a novel mechanism to ascertain the difference between a starting and a modified set of mitigation actions was developed. This mechanism allows the system operator to explore the effects of different mitigation plans in the system in terms of four cyber-security KPIs using statistics and machine learning tools.

To reiterate, experiments show that the proposed tool achieves a score of 95.78% accuracy, in discerning whether a statistically significant difference exists between different mitigation plans using a data set with numerous outliers. The algorithm outperforms another similar SoA method by 6.61% in terms of Accuracy.

The Hypothesis testing tool is currently being integrated with the system developed in the SerIoT project [1] and more specifically with an approach using Software Defined Networking (SDN) controllers, an automated Mitigation Engine and a Visual Analytics Dashboard that allows a user-friendly overview and operation in the case where manual intervention to the system is required.

Finally, we plan to further augment the proposed method in order to improve its performance and usability by introducing two enhancements: The first enhancement, will be a mechanism that will automatically fine-tune the model used to cluster the mitigation action set. The second enhancement will involve experimentation using ensemble methods to combine the results of the Sigclust and the Jackstraw models to obtain a method with improved performance.

Acknowledgment

This work is supported by the European Unions Horizon 2020 Research and Innovation Program through the SerIoT project under Grant Agreement No. 780139. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

References

[1] G. Baldini, P. Fröhlich, E. Gelenbe, J. Hernández-Ramos, M. Nowak, S. Nowak, S. Papadopoulos, A. Drosou, and D. Tzovaras, "Iot network risk assessment and mitigation: the seriot approach," in *Security Risk Management for the Internet of Things: Technologies and Techniques for IoT Security, Privacy and Data Protection*, J. Soldatos, Ed., 2020, ch. 5, pp. 88–104.

[2] Y. Lu and L. D. Xu, "Internet of things (iot) cybersecurity research: A review of current research topics," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2103–2115, 2019.

[3] A. Allakany, G. Yadav, K. Paul, and K. Okamura, "Detection and mitigation of lfa attack in sdn-iot network," in *Web, Artificial Intelligence and Network Applications*, L. Barolli, F. Amato, F. Moscato, T. Enokido, and M. Takizawa, Eds. Cham: Springer International Publishing, 2020, pp. 1087–1096.

[4] J. Wu, C. Wang, Y. Yu, T. Song, and J. Hu, "Sequential fusion to defend against sensing data falsification attack for cognitive internet of things," *ETRI Journal*, vol. 42, no. 6, pp. 976–986, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.2019-0388

[5] F. Li, R. Xie, Z. Wang, L. Guo, J. Ye, P. Ma, and W. Song, "Online distributed iot security monitoring with multidimensional streaming big data," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4387–4394, 2020.

[6] R. Kassab, O. Simeone, and P. Popovski, "Fog-based detection for random-access iot networks with per-measurement preambles," 2020.

[7] M. Walshe, G. Epiphaniou, H. Al-Khateeb, M. Hammoudeh, V. Katos, and A. Dehghantanha, "Non-interactive zero knowledge proofs for the authentication of iot devices in reduced connectivity environments," *Ad Hoc Networks*, vol. 95, p. 101988, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870519304895

[8] A. Ukil, S. Bandyopadhyay, and A. Pal, "Iot-privacy: To be private or not to be private," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2014, pp. 123–124.

[9] A. Tarighati, J. Gross, and J. Jaldén, "Decentralized hypothesis testing in energy harvesting wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4862–4873, 2017.

[10] M. Sun and W. P. Tay, "On the relationship between inference and data privacy in decentralized iot networks," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 852–866, 2020.

[11] M. R. Leonard, M. Stiefel, M. Fauß, and A. M. Zoubir, "Robust sequential testing of multiple hypotheses in distributed sensor networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4394–4398.

[12] H. S. Project, "Deliverable d4.5: Unsupervised iot-ready engine for threat mitigation," SerIoT: Secure and Safe Internet of Things, Tech. Rep., 2020.

[13] F. C. S. I. Group, "Common Vulnerability Scoring System version 3.1 Specification Document," FIRST - Forum of Incident Response and Security Teams, Cary, USA, Standard, Jun. 2019.

[14] H. Booth, D. Rike, and G. Witte, "The national vulnerability database (nvd): Overview," National Institute of Standards and Technology, Tech. Rep., 2013.

[15] G. G. Granadillo, H. Débar, G. Jacob, C. Gaber, and M. Achemlal, "Individual countermeasure selection based on the return on response investment index," in *International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security*. Springer, 2012, pp. 156–170.

[16] A. Shameli-Sendi, H. Louafi, W. He, and M. Cheriet, "Dynamic optimal countermeasure selection for intrusion response system," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 5, pp. 755–770, 2018.

[17] F. Li, Y. Li, Z. Yang, Y. Guo, L. Yin, and Z. Wang, "Selecting combined countermeasures for multi-attack paths in intrusion response system," in *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, 2018, pp. 1–9.

[18] H. Huang, Y. Liu, M. Yuan, and J. S. Marron, "Statistical significance of clustering using soft thresholding," *Journal of Computational and Graphical Statistics*, vol. 24, pp. 975 – 993, 2015.

[19] N. C. Chung, "Statistical significance of cluster membership for determination of cell identities in single cell genomics," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/08/07/248633

[20] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[21] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[22] N. C. Chung, "Statistical significance of cluster membership for unsupervised evaluation of cell identities," *Bioinformatics*, vol. 36, no. 10, pp. 3107–3114, 03 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa087

.

Roadmap of Security threats between IPv4/IPv6.

Fadi Abusafat
Information system Department-Algorithm centre.
University of Minho.
Guimaraes, Portugal.
[0000-0001-8821-9549]

Tiago Pereira
Information system Department-Algorithm centre.
University of Minho.
Guimaraes, Portugal.
[0000-0001-5075-6189]

Henrique Santos
Information system Department-Algorithm centre.
University of Minho
Guimaraes, Portugal.
[0000-0001-5389-3285]

***Abstract*— The idea of the Internet of Things is to connect every physical device with internet. Each device should be presented by a unique address of Internet Protocol. There are two versions of IP known by IPv4 and IPv6. IPv4 assumed to cover whole network interfaces. Since the appearance of IoT, number of connected devices increased sharply and IPv4 could not afford this enormous numbers. Therefore, the solution came by introduced IPv6 to afford this massive number in IoT environment. Despite IPv4 provided interoperability with several types of protocols, robust and easy implementation but it's vulnerable for several kinds of attacks. Therefore, IPv6 introduced while having security protocols such as IPsec, neighbour discovery protocol and Secure Neighbour Discovery Protocol. However, due to IPv4 still working as well as updated with security components, there is need to interact between both versions of protocol. Therefore, there is need for interacting mechanism between both them. These mechanisms made implantation of IPv6 complicated and consumed more resources. Besides, there are threats for attacks. Therefore, security features in IPv6 are not enough and there is need for a new defense line that can secure network services and IP from attacks. Intrusion Detection mechanism considered good mechanism to provide protection due to it works based on two levels Host and Network. Besides, it uses several security approaches such as signature-based and network-based. However, considering this tool as it's in IoT environment will not bring the light of security in IoT. Therefore, it should be developed while considering features of IoT devices to secure IoT environment. To achieve this, we have to investigate every feature, component and operation in network. This review paper aims to analyse difference between IPv4/IPv6 and point out security threats related to IP in both versions. Also, identifying main security threats to IPv6.**

Keywords*— *IPv4, IPv6, IP Security threats, ICMPv6, Network Layer attacks.

I. INTRODUCTION.

The 21st century is considered the time of development of technology due to the appearance of Internet-Of-Things (IoT). The main core of IoT is to associate physical devices in human life through connections over internet. This association provided several facilities for human activities. IoT has several application such as Smart City, Smart Home, Smart Grids and Smart Buildings [1]. Ideally, it should apply with every industrial fields. This association leads to massive financial gains. Several studies estimate the financial impact of IoT from 2018 to 2023 to have increased from 249.4 to 2,0030.1 Million Dollar respectively [2]. This massive impact came from massive numbers of newly connected devices that interact and share data over internet. Studies suggest the estimated number of connected devices over internet increased from 7 to 21.5 billion device from 2018 to 2025 [3]. However, devices need Internet Protocol/Transmission Control Protocol IP/TCP to exchange data. The main advantage of IP/TCP is to allocate unique addresses to recognise of each device. Since the introduction of IoT, IP has been changed sharply from version 4 (IPv4) to version 6 (IPv6). This amendment introduced several security challenges for security mechanism as well as security attacks due to IPv4 still working and there is need to interacts between both versions while these mechanisms have interoperability challenges and vulnerable for network layer attacks [4]. In this review paper, we are going to draw a roadmap of potential security threats based on IPv4 and IPv6 in IoT/Smart City environment. To achieve this purpose, I am going to use my knowledge in Pentesting as well as most updated works toward it. Updated work will be collected from strong conference and journals. Ideally, there are three main questions to be highlighted through this study:

A. What are differences between IPv4 and IPv6?

B. What are security threats for IPv4 and IPv6?

C. What are specific threats for IPv6?

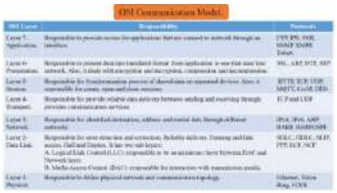
II. IPv4 vs IPv6

A. *Open Source Interconnection model*

The Open Source Interconnection (OSI) model is considered to be the updated model in communication and interacting data between devices. It consists from seven layers known by Application, Presentation, Session, Transport, Network, Data Link and physical layer. Application layer is responsible for providing access to applications that are in connection with an internet. Presentation layer is responsible to present data into translated formats. Session layer is responsible for creating, opening and closing sessions in order to share data between

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

devices. Transport layer is responsible for processing message delivery between senders and receivers. Network layer is responsible for identifying addresses, destinations, and routes of data through different networks. Data link layer is responsible for error detection, corrections, link access, framing and reliable delivery. While the physical layer is responsible to define physical network. Since IP is located in network layer, we can assume rules for IP. Fig1, shows several protocols in each layer [5], [6]



OSI Communication Model

OSI Layer	Responsibility	Protocols
Layer 7- Application	Responsible to provide access for applications that are connect to network through an interface.	FTP, IPS, SSH, SNMP, XMPP, Telnet
Layer 6- Presentation	Responsible to present data into standard format from application to one that used into network. Also, it deals with encryption and decryption, compression and decompression.	SSL, [illegible]
Layer 5- Session	Responsible for Synchronization process of shared data on separated devices. Also, it responsible for create, open and close sessions.	HTTP, TCP, UDP, MQTT, CoAP, DDS
Layer 4- Transport	Responsible for provide reliable data delivery between sending and receiving through provides communication services	TCP and UDP
Layer 3- Network	Responsible for identified destinations, address and routed data through different networks	IPv4, IPv6, ARP, RARP, [illegible]
Layer 2- Data Link	Responsible for error detection and correction, Reliable delivery, Framing and link access, Half and Duplex. It has two sub layers. A. Logical Link Control (LLC): responsible to be an intermediate layer between MAC and Network layer. B. Media Access Control (MAC): responsible for interaction with transmission media.	SDLC, HDLC, SLIP, PPP, LCP, NCP
Layer 1- Physical	Responsible to define physical network and communication topology.	Ethernet, Token Ring, FDDI

Fig 1, OSI communication model.

B. *Length of Address.*

The length/size of IPv4 is 32 bit while it is 128 bit in IPv6. Therefore, the number of available address in IPv4 and IPv6 are 2^{32} and 2^{128} respectively. The format of address used in IPv6 is alphanumeric hexadecimal notation while it is numeric dot decimal notation in IPv4. Prefix notation is 24 in IPv4 while it's 48 in IPv6. Also, IPv6 is represented by four hexadecimal digits in eight groups while in IPv4 is presented by three numeric dot decimal in four groups. Beside IPv6 supports auto configuration while in IPv4 support Dynamic Host Configuration Protocol (DHCP) or manual configuration. Table 1, shows comparison between IPv4 and IPv6 based on address features [7].

Table 1 address features between IPv4 and IPv6.

IPv4	IPv6
Length: 32 bit	Length : 128 bit
Available address: 2^{32}	Available address : 2^{128}
Format: numeric dot decimal notation	Format: Alphanumeric hexadecimal notation
Prefix notation : 24	Prefix notation : 48
Represented by: Four hexadecimal digits with eight groups	Represented by: Three numeric dot decimal in four groups.
Supports: DHCP and Manual configuration.	Supports: Auto configuration.

C. *Header.*

Header in IPv4 consists of 14 fields. Firstly, the version value is 4 bits. Internet Header Length (IHL) have varying sizes between 20 to 60 bytes and is used to avoid errors. The type of services (ToS) used to provide quality of services such as Voice over IP (VoIP). Explicit Congestion notification (ECN) is optional, and it used to notify senders or receivers network updates. Total length size is 16 bits and is used to point out the size of total datagram. The size can be ranged between 20 to 65535 bytes. Since this, fragmentation process has been introduced to deal with packet that have size bigger than this range. Identification (ID) is a feature used to identify fragment of IP uniquely. Flags are used to control and identify fragment. There are three flags known by Bit 0, Bit 1 and Bit 2 which are used for reserved, do not fragment and more fragments respectively. Fragment Offset has a length of 13 bit and is used to specify the offset of a fragment relative to the beginning of IP datagram. Time to live (TTL) has length 8 bit and used to present maximum time of datagram will be live on the internet. TTL is measured in seconds and it range between 0-255. In the case TTL value is zero then datagram will be removed. Protocol is used to present used Protocol in portion of datagram such as 6 present TCP and 17 present UDP. Checksum of header has 16 bit and is used to check errors in header. It is used to compare values of header checksum at each hop and discards packets in case of mismatch. Source Address has 32 bit and it present sender of data. Destination Address has 32 bit and present destination of sent packet Options it used for settings related for security, route, time-stamp and usually used when value of IHL is set to more than 5 [7]–[9].

It is unlikely for IPv4, IPv6 has less than 8 fields. Starting by version and it represent the version of IP and has 4 bits. Traffic Class has 8 bits and is divided into two main parts. First one consists of first six bits and is used to make router familiar with kind of services that should be provided. Secondly, last two bits and used for ECN. Flow label has 20 bits and designed for real-time media and streaming. Also, it used to maintain sequential flow of packets. This help router to identify particular packet belonging to specific flow of information. Therefore, it helps to avoid reordering of packets. Payload length has 16 bits and used to inform router with size of information that packet have. The size can be up to 65535 bytes and it will set for zero in case the maximum size is exceeded. Next header has 8 bits and used to indicate the type of extension header or Upper layer in case header is not present. Hop limit has 8 bits and used to stop looping packet in network. It is the equivalent to TTL in IPv4. Source address and destination address have 128 bits [7]–[9].

D. *Quality of Service.*

Quality of Service (QoS) is a set of requirements that are used to ensure proper delivery for packets. Ideally several parameters are used to construct metrics of QoS such as bandwidth, transmitted data, delay, lost data, received data and other parameters. Therefore, to evaluate the QoS between IPv4 and IPv6, will be based onto fields that indicate proper delivery. In IPv4, the flow of packets will be based on source, destination, ports and type of protocol in transport layer. However, these parameters could be affected due to fragmentation and encryption process. While in IPv6, the flow of packets based on previous fields plus flow labelled which is pre-defined in header. Flow labelled consists of 20 bits. The field of 8 bits for traffic class and used to distinguish between classes or priorities of IPv6 packets. This distinction made by source node and router. Flow labelled provided several advantages such as reduce average time for processing in router in network, reduce delays of packet, reduce the use of resources that caused by frequent change route [10].

E. *Auto Configuration.*

The main aim of this feature is to connect devices such as PC to internet automatically without need for manual configuration or software. Also, it provides a unique IP address to overcome scalability issues. This feature is an improvement of Link Layer Discovery Protocol (LLDP) which uses a set of attributes to discover neighbour devices. The set of attributes known by Type Length Value (TLV) which consists of type, length, and descriptions of value [11]–[14].

Dynamic Host Configuration Protocol (DHCP) is used for automatic configuration of devices in network. It is used elements in configurations such as IP, subnet mask, gateway, and other information. Generally, this process consists of discover, offer, request and Knowledge. There are some similarities on functionality of DHCP in IPv4 and IPV6. Firstly, the components of DHCP are DHCP Client, DHCP Server and DHCP Relay. These components are not changed in both IPv4 and IPv6. DHCP client is a device on a network that utilise a DHCP protocol to get network configuration. DHCP server is a component that provides a network configuration to DHCP client. This server is configured by a network administrator with network parameters to meet client needs. DHCP relay also known by DHCP relay agent and it used to pass messages in DHCP client and DHCP server are in different network. Secondly, scopes and leases. Scopes is a group if information that are used to configure device on network while lease is determine how long device on network can use that configuration. Finally, both use four messages to provide basic configuration of device on network. These configurations are discover message/solicit message, offer message/advertise message, request message/request message and acknowledgment messages/ replay message [11]–[14].

There are differences between DHCP for IPv4 and IPv6. Firstly called a reservation. In IPv4, MAC address been used to obtain IP address while in IPv6 used DHCP unique identifier (DUID) to allocated IP address. This mechanism is more sophisticated. However, reservation always require updating, as IPv4 is based on MAC while in IPv6 is based on DUIDs. The second difference is stateful and stateless. IPv6 have these two methods to configure devices on the network. Stateful stores device configuration while stateless not. In stateful DHCP sever knows IP address for all devices on network while in stateless it does not record any IP address and devices use router advertisement message to configure itself with an IP. A part of this IP is configured by device itself. In IPv4, it's not possible for a device to configure a part of IP for itself due to limitation on number of usable addresses. Second difference is Broadcast and Multicast. IPv6 uses multicast rather than broadcast. Broadcast packets goes to all devices on the network once the device is loaded on a network. This consumes more resources if the network has many devices. IPv6 uses multicast which it sent packets for selected devices on network rather than all devices. This reduce traffic on network [11]–[14].

F. *Mobility.*

The main idea behind Mobile IP protocol (MIP) is to keep devices connected to the internet while device in continuous mobility. In IPv4, mobility mechanism consists from three functional units known by home agent (HA), forgiven agent (FA) and mobile node (MN). Every MN has permanent home address allocated from home network but when it moves out, it gets a temporary address (CoA) which is used to identify MN in visited network. However, routing issue cause delays in mobility in IPv4. While IPv6 provide a great support for mobility due to its uses two IP address known by home address and CoA. However, routing has been improved in IPv6 [15].

G. *Security.*

IPv6 loaded with IP security series (IPsec) for security, authentication and data integrity provides authentication on header and encapsulating security payload extension (ESP). Also, it is designed based on end-to-end encryption and it supports more-secure name resolution. Besides, the secure neighbour discover protocol (SEND) added extra security features to neighbour discovery protocol (NDP) which is responsible for discovering other node on local link. However, NDP is not secure but SEND secure it with cryptographic method. Updated IPv4 include IPsec features. Therefore, security is different but not that big [4] [7].

III. SECURITY THREATS TOWARDS IPV4 AND IPV6.

In this section, we are going to investigate Cyber security attacks based on bothIP addresses. We would like to mention new sophisticated attacks are multi-use which means it will be based on IP and other services. However, we are going to classify attacks based onto previous features and operations in IPv4 and IPv6..

A. *Fragmentation attacks*

Fragmentation process is used when sent packet in more than maximum size therefore, attacks related for this process are:

1) *Ping of death.*

The main aim of this attack is to destroy services on destination machine. Idyllically, this attack used ping feature to create a small fragment and when these fragments assemble at destination, they exceed the max size of IP packet of 65535 bytes. This attack belong for Denial of Service (DoS) but it utilised connection features to be conducted [16].

2) *Drop attack.*

This attack based on reassemble rules of fragmentation. One of these rules is to indicate location of fragment to reassemble successfully at destination. Hence, hackers utilise this rule through sending fragment with overlap in order to make destination node unable to reassemble sent packet [16].

3) Overlapping.

The main aim of this attack is to gain access based on TCP flags. It sends first fragment with TCP flag in order to reach the destination. The second fragment is sent with different value of TCP flag. This fragment is not blocked due to verification conducted only at the first fragment and when both fragments reach the destination, flag of first fragment will be over written with value of second fragment [16].

4) UDP and ICMP attack.

The main aim of this attack is to consume resources through sending UDP or ICMP packets bigger than network MTU [16].

B. Routing attacks.

Routing is a process to identify path of traffic inside or outside the network. It works in both versions of IP. Therefore, there is a threat for several attacks such as:

1) Flood attack.

This attack is based on sending large amount of traffic that make the destination unable to process sent packets. This attack belong for DoS. This attack utilises several types of protocols such as TCP, UDP and ICMP [17]–[19].

2) Sniffing attack.

This attack aims to capture traffic while being sent through a network. It has many aims such as stealing confidential data or dropping some packet. The best example of this attack is Man in the Middle (MiTM) which is based on fool router and source to make traffic passed through it [17]–[19].

3) Fake attack.

This attack is based on introducing a fake device or access point which is not authorised to be inside a network. Once this fake device is installed inside a network, it can pass traffic through it in order to steal data [17]–[19].

4) ARP spoofing.

ARP protocol is used to enable network communication between devices. It is used to map MAC and IP address and this mapping information stored in ARP table. This attack is based on sending wrong ARP messages over local networks to connect hacker MAC addresses with legitimate devices. Hence, hacker devices will obtain IP address and start spoofing, modifying and blocking communication. This attack belongs for MITM [17]–[19].

IV. Security Threats towards IPv6.

There are some threats toward IPv6 such as:

1) ICMP Threat.

IPv6 networks use ICMP message to conduct some important mechanisms such as router discovery when router respond for end node with router solicitation message (RS) with router advertisement (RA). This information saved for a time in routing tables. Therefore a threat here hackers could fool victims with RA messages to present itself as a router. Therefore, hackers can steal and see traffic. This attack conducted by MiTM tools. Unlikely in IPv4, blocking messages of ICMP is common development of secure features in IPv4 network [7].

2) Fragmentation process.

In IPv6, fragmentation processes are denied by an intermediate node and it conducts only by source node. The minimum recommendation size for MTU is 1280 bytes. Some security features recommend discarding all fragment with less than that value except if it’s in the last round of flow. Using fragments, an attacker establishes that port numbers are not in the first fragment. This helps in overcoming security mechanisms and in order to send massive numbers of small fragments which cause system crashes. Therefore, it is recommended to limit number of fragments [20].

3) Transition mechanism.

IPv4 and IPv6 protocols still coexist. Therefore, there is a need for compatible transmission in order to avoid risk of failure in internet connection. Despite the core of IPv6 is to provide improvements of IPv4 but different between both protocols resulting in two completely different protocols. This case compatibility problem means IPv4 hosts and routers will not be in a position to directly manage IPv6 neither IPv6 will directly manage Ipv4. Therefore, there is a need for transition mechanisms such as tunnelling, and dual-stack configuration. Tunnelling is capable of dealing with address selection and DNS resolution but it increases routing process and consumes more memory and CPU. Once it increased routing process, it will be vulnerable for routing attacks. Translation is easy to implement, works with private address and configured NAT node but it shows administration challenges due to its complexity and requires extra configuration which causes slow packet flow. Finally, it poses security threats with NAT due to it keeping sessions to apply address and port transition for inbound and outbound traffic. So, in case an injection of unknown packet came from inside network, a new session will be created. Tunnelling mechanism allows IPv6 packets to transport over IPv4 but it causes problems such as delay loaded CPU to perform encapsulation [21]. To clarify these threats, we use NAT64 which used to translate between IPv6 and IPv4. It consists from three main parts known by NAT64 prefix, DNS64 server and NAT64 router. Let’s presume, two networks A and B needs to communicate. Network A is a network IPv6-based and Network B is a network IPv4-based. In network A, there is a device need to communicate with a Website in network B. The first step is a device in network A communicate with DNS64 server asking about IPv6 for website in network B. Suppose, DNS64 server does not have record about this website. So, it will communicate IPv6 DNS server asking about it. IPv6 DNS server communicate with IPv4 DNS server about address of website. IPv4 DNS server replies with address of website in Network B due to it is located at the same network. Then IPv6 DNS server forward it to DNS64 which will do prefixing for it in hexadecimal. After that, it will forward it for a device which it used to communicate with NAT64 router which be the main components between both networks. NAT64 router made translation between IPv6 and IPv4 header. Finally, translated IPv4 packet will be forward it for website to conduct communication. This whole process is very complicated and need high resources such as CPU and routing process. Hence, it will be vulnerable for several attacks such as sniffing, DoS and routing.

4) *Secure Neighbour Discovery Protocol.*

ICMPv6 include Neighbor discovery Protocol (NDP). It is designed for several services in IP such as multicast, NDP, and neighbor discover (ND). It uses several messages such as router solicitation, router advertisement, neighbor solicitation, and neighbor advertisement. However, security in NDP is based on its scope and without securing NDP, IPv6 is still vulnerable for several attack such as MiTM, rouge and replay.

Secure Neighbour Discovery protocol (SEND) introduced to protect NDP and make IPv6 safe protocol. However, deployment of SEND is not easy and it computes intensively besides massive bandwidth consumption. So, IPv6 is still vulnerable for these attacks [22].

5) *IPsec.*

Internet Protocol Security (IPsec) is used to secure network packets at IP through enables cryptographic. It used widely in build Virtual Private Network (VPNs) by establishing Internet Key Exchange Protocol (IKE). IKE consists from two versions, each one in different mode and phases. Also it uses several authentication methods and configuration options. In case pair keys refused at different versions and modes in IKE, can introduce bypass authentication and made network vulnerable for authentication attack [23].

V. FUTURE WORK.

The main aim of this review paper is to point out security threats for IP in both versions. I plan to use this knowledge in my PhD research which is titled "identify security architecture compromised Intrusion Detection Mechanisms (IDS) to detect major attack in IoT and Smart City context". IDS is considered a promising security mechanism but not suitable to apply as it currently stands and should be improved, therefore we need to recognise threats related to IPv4 and IPv6 [24].

VI. CONCLUSION.

IPv6 have been introduced in order to overcome of IPv4 challenges in IoT context which it's limited to number of address while scalability issue is very clear in IoT. Besides, it introduced to overcome security issues in IPv4 through adapting several security protocols such as IPsec and SEND. However, existing of IPv4 introduced requirement of interoperability issue with IPv6. Several mechanism been introduced to server this purpose such as tunneling. These mechanism consumed many resources and this made network vulnerable for several types of network attacks. Besides, security features in IPv6 such as IPsec is also vulnerable for authentication and MiTM attacks.

There are several feature of IoT nodes, one of them is lightweight due to only capability to send small size of packet. Therefore, adapting IPv6 while holding translation mechanism that consumed plenty of resource introduced challenges for interoperability in IoT devices. Therefore, proposed solution for adapting IPv6 in IoT network, should be lightweight.

REFERENCES

[1] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Bus. Horiz.*, vol. 58, no. 4, pp. 431–440, 2015, doi: 10.1016/j.bushor.2015.03.008.

[2] F. S. Market, "IoT in Banking and Financial Services Market," 2020. https://www.marketsandmarkets.com/Market-Reports/iot-banking-financial-services-market-172304505.html (accessed Feb. 20, 2021).

[3] K. L. Lueth, "State of the IoT 2018: Number of IoT devices now at 7B-Market accelerating Market Update Global number of Connected Devices: 17B," 2018. https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/ (accessed Aug. 02, 2020).

[4] S. Praptodiyono, R. K. Murugesan, I. H. Hasbullah, C. Y. Wey, M. M. Kadhum, and A. Osman, "Security mechanism for IPv6 stateless address autoconfiguration," *Proc. 2015 Int. Conf. Autom. Cogn. Sci. Opt. Micro Electro-Mechanical Syst. Inf. Technol. ICACOMIT 2015*, pp. 31–36, 2016, doi: 10.1109/ICACOMIT.2015.7440150.

[5] A. H. Alhamedi, V. Snasel, H. M. Aldosari, and A. Abraham, "Internet of things communication reference model," *2014 6th Int. Conf. Comput. Asp. Soc. Networks, CASoN 2014*, pp. 61–66, 2014, doi: 10.1109/CASoN.2014.6920423.

[6] M. Bagga, P. Thakral, and T. Bagga, "A study on IoT: Model, communication protocols, security hazards countermeasures," *PDGC 2018 - 2018 5th Int. Conf. Parallel, Distrib. Grid Comput.*, pp. 591–598, 2018, doi: 10.1109/PDGC.2018.8745984.

[7] M. Shrivastava, "Threats and Security Aspects of IPv6," *Glob. J. Comput. Technol. Vol*, vol. 1, no. 2, pp. 51–55, 2015, [Online]. Available: https://www.researchgate.net/profile/Manish_Shrivastava8/publication/280568665_Threats_and_Security_Aspects_of_IPv6/links/55ba56b608aed621de0ace20.pdf .

[8] E. Durdaği and A. Buldu, "IPV4/IPV6 security and threat comparisons," *Procedia - Soc. Behav. Sci.*, vol. 2, no. 2, pp. 5285–5291, 2010, doi: 10.1016/j.sbspro.2010.03.862.

[9] D. G. Chandra, M. Kathing, and D. P. Kumar, "A comparative study on IPv4 and IPv6," *Proc. - 2013 Int. Conf. Commun. Syst. Netw. Technol. CSNT 2013*, no. June, pp. 286–289, 2013, doi: 10.1109/CSNT.2013.67.

[10] O. J. S. Parra, A. P. Rios, and G. L. Rubio, "Quality of service over IPV6 and IPV4," *7th Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2011*, pp. 4–7, 2011, doi: 10.1109/MACE.2011.6040165.

[11] Y. Cui, Q. Sun, K. Xu, W. Wang, and T. Lemon,

"Configuring IPv4 over IPv6 Networks: Transitioning with DHCP," *IEEE Internet Comput.*, vol. 18, no. 3, pp. 84–88, 2014, doi: 10.1109/MIC.2014.49.

[12] J. Montavont, C. Cobarzan, and T. Noel, "Theoretical analysis of IPv6 stateless address autoconfiguration in low-power and lossy wireless networks," *Proc. - 2015 IEEE RIVF Int. Conf. Comput. Commun. Technol. Res. Innov. Vis. Futur. IEEE RIVF 2015*, pp. 198–203, 2015, doi: 10.1109/RIVF.2015.7049899.

[13] H. Rafiee and C. Meinel, "A secure, flexible framework for DNS authentication in IPv6 autoconfiguration," *Proc. - IEEE 12th Int. Symp. Netw. Comput. Appl. NCA 2013*, pp. 165–172, 2013, doi: 10.1109/NCA.2013.37.

[14] ITfreetaining, "Key Concepts Both protocols use DHCP Client / Relay / Server," 2019. https://www.youtube.com/watch?v=YDqUZJnB14g (accessed Feb. 25, 2021).

[15] D. Le, Y. Yao, Y. Jin, and M. Zhu, "Modelling and performace analysis of mobility on CNGI," *Proc. - 2011 IEEE Int. Conf. Comput. Sci. Autom. Eng. CSAE 2011*, vol. 4, pp. 733–737, 2011, doi: 10.1109/CSAE.2011.5952949.

[16] M. Bouabdellah, N. Kaabouch, F. El Bouanani, and H. Ben-Azza, "Network layer attacks and countermeasures in cognitive radio networks: A survey," *J. Inf. Secur. Appl.*, vol. 38, pp. 40–49, 2018, doi: 10.1016/j.jisa.2017.11.010.

[17] A. K. Abdelaziz, M. Nafaa, and G. Salim, "Survey of routing attacks and countermeasures in mobile ad hoc networks," *Proc. - UKSim 15th Int. Conf. Comput. Model. Simulation, UKSim 2013*, pp. 693–698, 2013, doi: 10.1109/UKSim.2013.48.

[18] R. K. Kapur and S. K. Khatri, "Analysis of attacks on routing protocols in MANETs," *Conf. Proceeding - 2015 Int. Conf. Adv. Comput. Eng. Appl. ICACEA 2015*, pp. 791–798, 2015, doi: 10.1109/ICACEA.2015.7164811.

[19] M. Karthigha, L. Latha, and K. Sripriyan, "A Comprehensive Survey of Routing Attacks in Wireless Mobile Ad hoc Networks," *Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020*, pp. 396–402, 2020, doi: 10.1109/ICICT48043.2020.9112588.

[20] M. Mavani and L. Ragha, "Security Implication and Detection of Threats due to manipulatingIPv6 Extension Headers," *Annu. IEEE India Conf.*, 2013.

[21] A. S. Ahmed, R. Hassan, and N. E. Othman, "Security threats for IPv6 transition strategies: A review," *2014 4th Int. Conf. Eng. Technol. Technopreneuship, ICE2T 2014*, vol. 2014-Augus, no. July 2020, pp. 83–88, 2015, doi: 10.1109/ICE2T.2014.7006224.

[22] A. S. Ahmed, R. Hassan, and N. E. Othman, "Secure neighbor discovery (SeND): Attacks and challenges," *Proc. 2017 6th Int. Conf. Electr. Eng. Informatics Sustain. Soc. Through Digit. Innov. ICEEI 2017*, vol. 2017-Novem, pp. 1–6, 2018, doi: 10.1109/ICEEI.2017.8312422.

[23] D. Felsch, M. Grothe, J. Schwenk, A. Czubak, and M. Szymanek, "the Dangers of Key Reuse: Practical Attacks on Ipsec Ike 27 Th Usenix Security Symposium," *Proc. 27th USENIX Secur. Symp.*, pp. 1–25, 2018.

[24] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in Internet of Things," *J. Netw. Comput. Appl.*, vol. 84, no. January, pp. 25–37, 2017, doi: 10.1016/j.jnca.2017.02.009.

IoT-Based Synergistic Approach for Poultry Management System

Moses Oluwafemi Onibonoje, *Member, IEEE*
Department of Electrical/Electronics and Computer Engineering
Afe Babalola University
Ado-Ekiti, Nigeria
onibonojemo@abuad.edu.ng

Abstract—**Poultry farming has contributed immensely to global food security and the economy. Its produces are favourites and hugely subscribed, due to the uniqueness of their nutrients to all categories of people and the alternatives they provide to other high-cholesterol proteins. The increase in the world's population will continuously stretch for an increase in demands for poultry products. A smart way to ensure continuous production and increased yields in various farms is to adopt automated and remote management of poultries. This paper modelled and developed a collaborative system using the synergistic wireless sensor network technology and the internet of things. The system integrated resourcefully selected wireless sensors, mobile phone, other autonomous devices and the internet to remotely monitor and control environmental parameters and activities within the farm. Parameters such as temperature, humidity, water level, food valve level, ammonia gas, illumination are sensed, benchmarked against selected thresholds, and communicated wirelessly to the sink node and the internet cloud. The required control actions can also be initiated remotely by the administrator through messages or command signal. Also, the various parameters and actions can be read or documented in real-time over the web. The system was tested and evaluated to give an average of about 93.7% accuracy in parameters detection and 2s delay in real-time response. Therefore, a modelled system has been developed to provide robust and more intuitive solutions in poultry farming.**

Keywords—internet of things, early chicken, monitoring system, poultry, wireless sensors.

I. Introduction

Food security and food quality enhancement is an important global issue being addressed by the sustainable development goals (SDG). The increasing world population has also pressurized the food production capacity to be drastically increased and improved upon, to cater for the survival of people. The past few years have experienced increased awareness around the globe about the need for increased agricultural yields and food safety. Poultry farming has contributed immensely to the economies of many nations and global food security. Poultry has been described by [1], [2] to include a range of domestic bird species. Poultry farming identifies specifically with the agro-food sector for raising chicken for meat and eggs. However, food security is directly affected by poultry farming due to consumer issues linked to availability for meeting demands, affordability, accessibility, and consumption safety [3]–[5]. Good productivity and profitability in the sector have a crucial impact on national revenues and incomes to the farmers [6]. The profitability of a poultry farm is a measure of the revenue fewer costs, and dependent on the efficient economics of productivity [7]. Human civilization and industrial revolutions have introduced synergistic technologies to help the agri-food sector to grow and thrive [8]. Nations are adopting new measures and protocols to digitalize food production and automate agricultural practices. Beyond the impending possibilities from traditional ways of managing poultries and others, food production continues to present the need to incorporate innovative disruptions for automated systems [9], [10]. Poultry management systems characteristically contain three basic separate functions which include sensing or monitoring, decoding and decision-making, and necessary intervention [11].

Internet-connected objects can collaborate with other autonomous devices, and with themselves to execute tasks in poultry management and food production. Internet of things (IoT) has been described as a revolutionary technology associated with devices such as sensors and personal computers being connected to communication infrastructure and the internet via software solutions for self-sufficient actions devoid of human interference. In recent times, the IoT capacity being utilized by machinery, modern tools, information technologies, and emerging communication, is a major consideration for the agri-food sector. The aim is to bolster the food regime with automation, connectivity, digitalization and efficient use of resources [12], [13]. IoT is poised for a far-reaching inference on the agri-food sector and becomes an important part of a smart poultry farm. Implementing IoT technology in smart poultry production will incorporate internet-based devices to autonomously act, while humans are only to intervene at a point requiring a higher intelligent level [14]–[19]. The communication capability between sensors and various equipment used on the farm remains the main advantage of IoT in the process of process and tasks automation, to enhance management efficiency. The feasible communication types within the IoT network include device-device, human-device, and device-human communications. A simple communication signal can initiate the automation of different procedures within the system. Presently, most nodes of sensors within the IoT network are homogenous, thereby making the required communication architecture very similar to that of a wireless sensor network.

The new complex IoT-based multi-device communications are meant to face greater challenges associated with heterogenous nodes such as limitation in power supply, storage, processing units, and the inconsistency of the network access. The multi-device challenges can be addressed through instrumented algorithms, thereby allowing a higher capability of farm automation and processes such as precision feeding instrumentation in which feeds are administered based on prior weight determination of the broilers [15]. The application of radiofrequency tags is also very useful in poultry systems to monitor the feeding frequency, the frequency of egg-laying, and health status detailing. The data on the birds can then be used to decide and initiate control of the previously measured parameters. Therefore, an IoT infrastructure can be implemented in the poultry farms by ensuring the processes and methods for data protection, data overseeing and governance to ensure continuous data integrity and quality [20], [21]. Existing poultry management systems are attributed to automated condition monitoring, but manual intervention methods. The

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

manual feeding process has its unique limitations which include injurious over-feeding, and delayed supply of feeds. This paper aims at combining a synergistic approach in combining autonomously designed devices with the internet to solve ecological monitoring, decision-making, and interventionist action automatedly.

II. Materials and Methods

The poultry management system is modelled and developed using resourcefully selected devices and validated algorithms.

A. System Model

The multiple function decision-making model [22] is adopted in designing the poultry management system. The intervention provided by the system at any particular time is a subject of the decision-making pitch arising from multiple monitored input parameters.

The system is considered as a large entity with m subsystems and n major actions being taken. Assuming b_{ij}^k is the value of the input environmental factor i being monitored in initiating a unit action j, and r_j^k is the total output of action j in the subsystem k. Consider p_i as the final relevance of factor i in the entire system. Therefore, the representative equation of the entire system is as stated in equation (1).

$$\sum_{k=1}^{m} \sum_{j=1}^{n} b_{ij}^k r_j^k + p_i = \sum_{k=1}^{m} r_i^k \qquad i = 1,2, \dots, n \qquad (1)$$

Assuming d_j^k represents the total output of action j in subsystem k as determined by the decision-making of the subsystem. Meanwhile, if all $d_j^k (k = 1, 2, , m, j = 1, 2, \dots . n)$ is the solution to equation (1), the decision-making of the entire system should agree with the collective decision made by the subsystems which set the decision-making plan of the entire system as $r_j^k = d_j^k (k = 1, 2, , m, j = 1, 2, \dots . n)$. However, this is observed not to always be true as d_j^k does not often satisfy equation (1), thereby making the requirement for a further procedure necessary.

Meanwhile, if any of the solutions of $d_j^k (k = 1, 2, , m, j = 1, 2, \dots . n)$ minimizes all the differences amongst r_j^k and d_j^k, a satisfactory solution of the entire system can be established. Weighted measure multi-criteria decision-making technique is hereby proposed as a mathematical model solution as highlighted in equations (2) and (3).

$$\min \left[\sum_{k=1}^{m} \sum_{j=1}^{n} \frac{1}{2} c_j^k (r_j^k - d_j^k)^2 \right] \qquad (2)$$

$$\sum_{k=1}^{m} \sum_{j=1}^{n} b_{ij}^k r_j^k + p_i = \sum_{k=1}^{m} r_i^k \qquad (3)$$

where $c_j^k (k = 1, 2, , m, j = 1, 2, \dots . n)$ are the weighting coefficients, while $c_j^k \geq 0$ and $\sum_{k=1}^{m} \sum_{j=1}^{n} c_j^k = 1$.

B. Hardware Design

The system has the main objective of monitoring some ecological parameters and initiating interventionist corrective measures, while the system status is related through the internet. For instance, a low-level feed could be an indication for an automatic supply of feed from the food shelf and a message to place an order to the feed suppliers. Also, a high-level of sensed ammonia gas and humidity is a signal for automated increased ventilation rate and the need to evacuate chicken droppings/wastes. The block diagram of the system hardware is as shown in Fig 1.

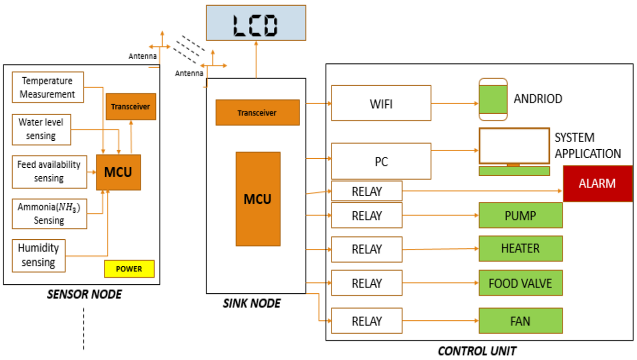


Fig. 1. The hardware block diagram of the system

The hardware unit of the system was designed with wireless and autonomous sensor nodes to monitor temperature, humidity, air quality, water level and feed availability in interested poultry space, and communicate the processed data via their transceiver to the access node. The access node communicates the data to the internet where it can be accessed in real-time. The interventionist actions being initiated by the access node and the cloud, based on the data from the sensor nodes are also being relayed in real-time to the necessary stakeholders and administrator of the poultry system.

For accurate calibration of the temperature and humidity requirement in the system, the thermal neutral zone of the chicks/birds is required. This is often required weekly. However, this changes on an approximately daily basis due to poultry growth from 'placing' to 'lifting', hence the need for the system to cover the daily weight average. For instance, the bodyweight difference of chicks from day 7 to day 13 is often significance, just preceding the weekly change occurrence. A 7-day old chick can weigh 190 – 200 grams while weighing 490 – 500 grams at 14 days.

The system consists of two heterogeneous sensor nodes. The first node contains the temperature, humidity and air quality sensors, while the second node had the water level and feed availability sensors which measured its required parameters with a 1kg load cell. The hardware devices are selected from the comparative analysis of some alternatives and decided upon based on resourceful considerations. DHT21 AM2301 was selected for temperature and humidity sensing having analyzed its features against LM series, TMP series, SHT series, and DHT variants for temperature, and HR202 series, SENS-HYD2, SHT series and other DHT variants for humidity.

This study has selected MQ137 as its gas sensor. It took into account the possibility of some gases in the air such as CO_2, H_2S, CH_4, NH_3, but it focused on the presence of the ammonia gas (NH_3) due to the poultry droppings and waste. Ammonia gas affects the growth of chickens and can cause several diseases like hand foot disease, mouth disease, bird flu etc. The threshold value of ammonia gas in the air is 40%. The capacitive 2-probe sensor was selected for water detection, while a weight sensor was chosen for the feed sensing. Arduino Nano was selected for the microcontroller unit due to size, compatibility in voltage and current requirement, and the grove pins. The microcontroller board is based on the ATmega328, having 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analogue inputs, a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button. XBee RF module was selected for the transceiver unit after comparing the ZigBee, Bluetooth, and WIFI communication network standards. The modules are designed to meet IEEE 802.15.4/Zigbee standards. This standard defines a dual physical (PHY) layer and media access (MAC) layer, which is 2.4Ghz and 868/915Mhz. The Xbee has the capability of communicating these modules with the microcontroller (MCU) through a universal asynchronous receiver/transmitter. The General Packet Radio Service (GPRS) module was selected for mobile connectivity, having compared it with the Enhanced Data rate for GSM Evolution (EDGE) and High-Speed Packet Access (HSPA) technologies.

C. *Software Design*

The algorithm of the system was designed and developed in two sections, to report the changes in parameters (temperature, humidity, water level, feed availability, ammonia gas level) in the poultry farm; and initiate the appropriate control measure to normalize the changes as designed with the overall system. The algorithm flowchart of the system working model is as shown in Fig 2. The flowchart mainly highlights the specific operational details of the system, while the configuration details of the transceivers and microcontroller modules are accordingly generic. The system was developed to cover the three main units: the monitoring sensor node, the network coordinating node, the internet and mobile unit, and the control unit.

III. Results and Discussion

The implemented IoT-based poultry management system incorporated the proposed model, and the resulted hardware and software units. The resulted system nodes are as shown in Fig 3. The sensor node monitoring the temperature, humidity and ammonia gas is shown in Fig 3(a), while the sensor node measuring the water level and food availability is shown in Fig 3(b). The critical temperature for layers is 20°C. For every 1°C lower than 20°C, the birds require an extra 1.5 g of feed per day. The most efficient temperatures for layers are between 20 – 24°C. When temperatures rise above 24°C, shell quality and egg weight will reduce. The critical temperature for broilers and rearing birds is highly dependent on age. The load cell measures the quantity of feed in grams and was calibrated to measure the food weight in the percentage of the standard value. As the feed is being gradually consumed, the digital scale measures the reduction in the weight and once the quantity of feed reduced beneath 25% of the standard which is the maximum quantity of feed the load cell can accommodate, it will turn the servo motor which in-turn opens the nozzle for grains to pour back into the plate. Also, as water is being consumed, the digital scale measures the reduction in the weight and once the quantity of water reduced beneath 25% of the maximum water capacity, the solenoid valve opens the tap for the water chamber to be refilled.

The ventilation unit and the coordinating node with the internet gateway and the display LCD are as shown in Figs 3(c) and 3(d) respectively, while the prototype system unit during testing is shown in Fig 3(e). The average parameters reading and the corresponding intervention action taken by the system is as highlighted in Table 1, while the user application interface result connected to the internet cloud and the corresponding indicated result on the LCD at around 16:20 hours on Wednesday 6th of January, 2021 is as shown in Fig 4. It can be deduced that the Temperature was higher during the day and lower at night and the relative humidity was higher in the night time and lower in the day. The bulb is controlled based on the temperature, if the temperature dropped below 24°C the bulb automatically came ON and it remained ON until the temperature rose above 35°C then it went OFF after which the fan was turned ON.

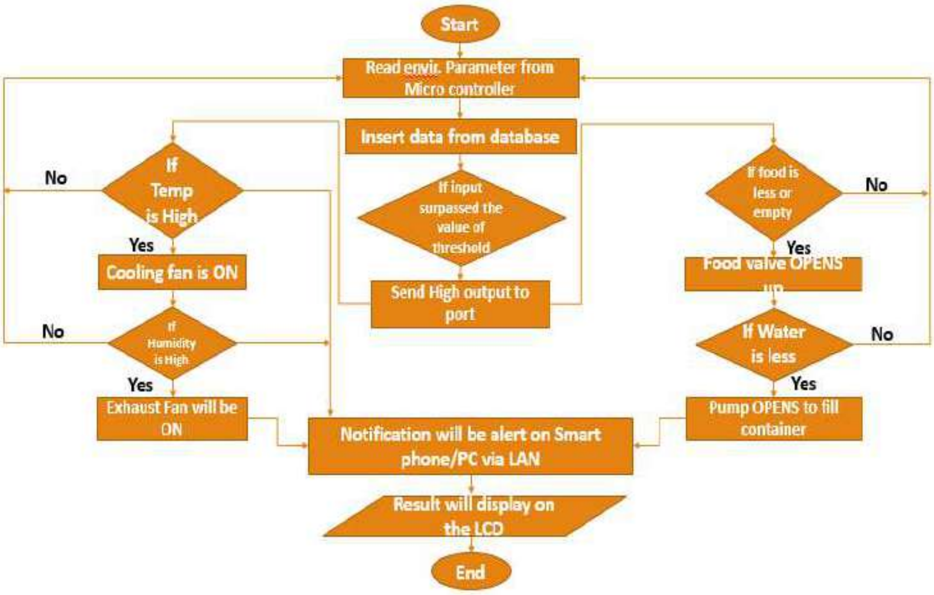

Start
Read envir. Parameter from Micro controller
Insert data from database
If Temp is High
No
Yes
Cooling fan is ON
If Humidity is High
No
Yes
Exhaust Fan will be ON
If input surpassed the value of threshold
Send High output to port
If food is less or empty
No
Yes
Food valve OPENS up
If Water is less
No
Yes
Pump OPENS to fill container
Notification will be alert on Smart phone/PC via LAN
Result will display on the LCD
End

Fig. 2. The algorithm flowchart of the system

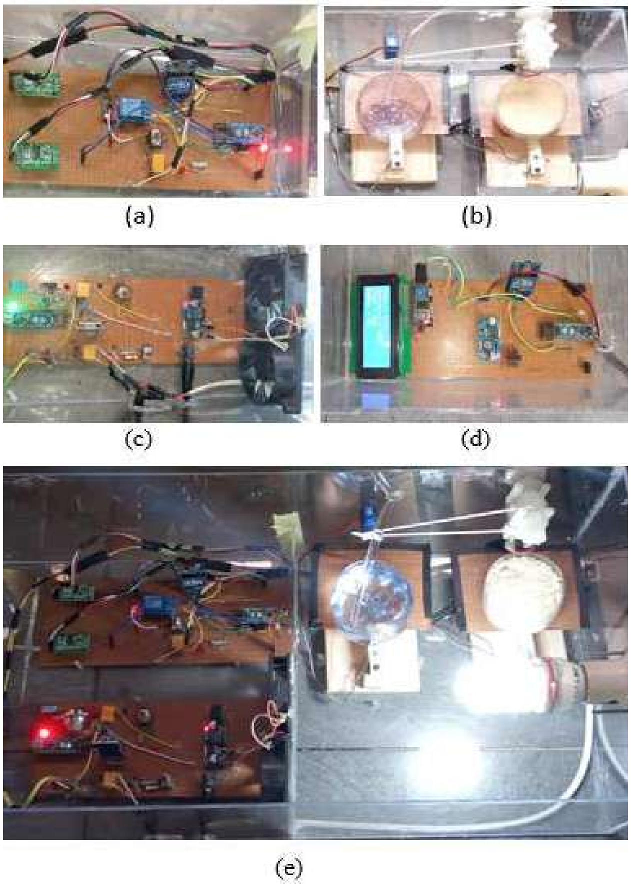


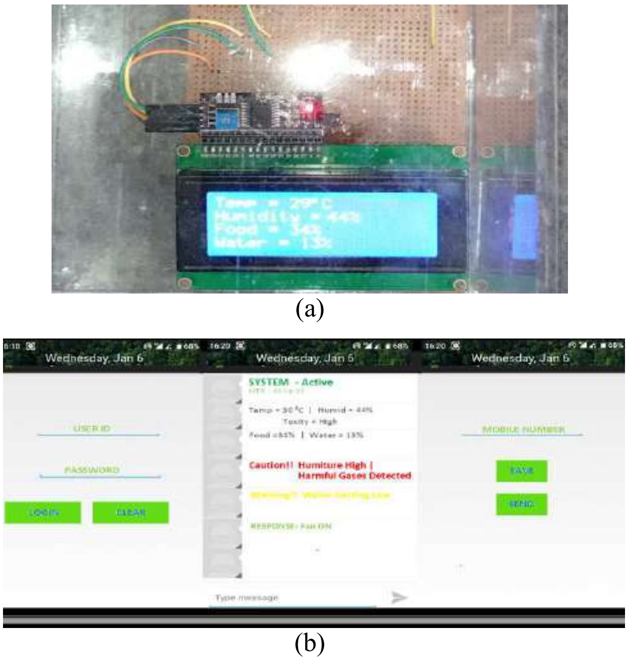
Fig. 3. The system hardware units (a) sensor node 1, (b) sensor node 2, (c) the fan unit, (d) coordinating node, and (e) the prototype system under testing.

TABLE I. AVERAGE TEMPERATURE AND HUMIDITY READING WITH THE CORRESPONDING SYSTEM RESPONSE

Period (GMT)	Measurement		System Response
	Average temperature (^0C)	*Average relative humidity (%)*	
00:00 - 02:00	25	63	No response
02:00 - 04:00	24.	67	Bulb turns ON
04:00 - 06:00	23.7	70	Bulb stays ON
06:00 - 08:00	23.5	68	Bulb stays ON
08:00 – 10:00	25	58	Bulb turns OFF
10:00 – 12:00	27.3	50	No response
12:00 – 14:00	29.7	46	No response
14:00 – 16:00	31	43	Fan turns ON
16:00 – 18:00	30	44	Fan stays ON
18:00 – 20:00	28.8	47	Fan turns OFF
20:00 – 22:00	26.9	53	No response
22:00 – 24:00	26	58	No response

The smart poultry management system was integrated with an MQ2 gas sensor senses toxic smoke, dust and gases. However, ammonia was the most important gas to measure. If the concentration of the gas was below recommended levels whereby the system was designed to have an upper critical level for ammonia as 20 ppm. The Alarm comes ON whenever the threshold value is triggered, this prompts the farmer to re-evaluate the nutrient of the feed given to the chickens. The Fan was also controlled by the MQ2 gas sensor to achieve optimization of the system ventilation whenever the system faces certain irregularities of been smoky, dusty or increased levels of ammonia.

The indicated results on the system local LCD and that of the user application interface are almost the same, with about 2s delay response time.


Temp = 29° C
Humidity = 44%
Food = 34%
Water = 13%

Wednesday, Jan 6

SYSTEM - Active

Temp = 30 °C | Humid = 44%
Toxity = High
Food =84% | Water = 13%

Caution!! Humiture High | Harmful Gases Detected

Warning!! Water Getting Low

RESPONSE: Fan ON

Fig. 4. The interface results (a) LCD, (b) Application interface

IV. CONCLUSION

A precision poultry management system based on IoT synergistic approach has been developed. The system monitors and regulates the condition of the poultry chickens through the internet in real-time. This study synergized among sensor technology, wireless communication, internet-of-things and automation in the implementation of the system. The achieved aim was to improve the health and growth of the birds by preventing infectious diseases among the birds, ensure increased egg production, reduce the gaseous pollutants, reduce human intervention, and maximize the overall profit. The system consists of resourcefully selected node components and internet-enabled autonomous devices and networks. This study has modelled and developed the management system to enable the farmers or administrators of the poultry farms to relate with or control the conditioning parameters of the birds based on the temperature, humidity, water level values, food dispensing, and ammonia gas coefficients in the poultry farm. In further works, the system would be more robust by incorporating energy rechargeability and portability to the nodes. Also, the proposed model would be validated and evaluated to depict the real-timeliness of the developed system. Furthermore, the system portends a possibility for integrating more than one poultry farm in the same neighbourhood and monitored over the same internet space and infrastructure.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Victor Ijeh during the experimental set-up and data collection.

REFERENCES

[1] N. S. N. Abd Aziz, S. M. Daud, R. A. Dziyauddin, M. Z. Adam, and A. Azizan, "A Review on Computer Vision Technology for Monitoring Poultry Farm--Application, Hardware, and Software," *IEEE Access*, 2020.

[2] R. Ribeiro, M. Teixeira, A. L. Wirth, A. P. Borges, and F. Enembreck, "A learning model for intelligent agents applied to poultry farming," in *International Conference on Enterprise Information Systems*, 2015, vol. 2, pp. 495–503.

[3] J. Bongaarts, "Food and Agriculture Organization of the United Nations: the state of food and agriculture: agricultural trade and poverty: can trade work for the poor?," *Popul. Dev. Rev.*, vol. 33, no. 1, pp. 197–198, 2007.

[4] E. L. N. Kaswati, A. H. Saputro, and C. Imawan, "Examination system of chicken meat quality based on hyperspectral imaging," in *Journal of Physics: Conference Series*, 2020, vol. 1528, no. 1, p. 12045.

[5] C. Okinda *et al.*, "Egg volume estimation based on image processing and computer vision," *J. Food Eng.*, vol. 283, p. 110041, 2020.

[6] M. O. Onibonoje, N. Nwulu, and P. N. Bokoro, "Food 4.0: An Introduction," in *Artificial Intelligence and IoT-Based Technologies for Sustainable Farming and Smart Agriculture*, IGI Global, 2021, pp. 83–100.

[7] H. El Bilali, G. O. Palmisano, F. Bottalico, G. Cardone, and R. Capone, "Food security and nutrition in agro-food sustainability transitions," *Food Secur. Nutr.*, pp. 57–86, 2021.

[8] M. O. Onibonoje and N. Nwulu, "Synergistic Technologies for Precision Agriculture," in *Artificial Intelligence and IoT-Based Technologies for Sustainable Farming and Smart Agriculture*, IGI Global, 2021, pp. 123–139.

[9] M. Costantini, V. Ferrante, M. Guarino, and J. Bacenetti, "Environmental sustainability assessment of poultry productions through life cycle approaches: A critical review," *Trends Food Sci. Technol.*, 2021.

[10] M. O. Onibonoje, N. I. Nwulu, and P. N. Bokoro, "A wireless sensor network system for monitoring environmental factors affecting bulk grains storability," *J. Food Process Eng.*, vol. 42, no. 7, 2019, doi: 10.1111/jfpe.13256.

[11] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming--a review," *Agric. Syst.*, vol. 153, pp. 69–80, 2017.

[12] J. Miranda, P. Ponce, A. Molina, and P. Wright, "Sensing, smart and sustainable technologies for Agri-Food 4.0," *Comput. Ind.*, vol. 108, pp. 21–36, 2019.

[13] M. Aleksandrova, "IoT in Agriculture: Five Technology Uses for Smart Farming and Challenges to Consider," 2018.

[14] J. Astill, R. A. Dara, E. D. G. Fraser, B. Roberts, and S. Sharif, "Smart poultry management: Smart sensors, big data, and the internet of things," *Comput. Electron. Agric.*, vol. 170, p. 105291, 2020.

[15] W. Sarachai, P. Ratnapinda, and P. Khumwichai, "Smart notification system for detecting fan failure in evaporative cooling system of a poultry farm," in *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, pp. 296–299.

[16] A. Nasiri, M. Omid, and A. Taheri-Garavand, "An automatic sorting system for unwashed eggs using deep learning," *J. Food Eng.*, vol. 283, p. 110036, 2020.

[17] M. O. Onibonoje, N. I. Nwulu, P. N. Bokoro, and S. L. Gbadamosi, "An IoT-Based Approach to Real-Time Conditioning and Control in a Server Room," in *IEEE IDAP*, 2019, pp. 1–6.

[18] M. O. Onibonoje, N. I. Nwulu, and P. N. Bokoro, "An Internet-of-Things Design Approach to Real-Time Monitoring and Protection of a Residential Power System," 2019, doi: 10.1109/SEGE.2019.8859879.

[19] M. O. Onibonoje and T. O. Olowu, "Real-time remote monitoring and automated control of granary environmental factors using wireless sensor network," 2018, doi: 10.1109/ICPCSI.2017.8391925.

[20] X. Zhuang and T. Zhang, "Detection of sick broilers by digital image processing and deep learning," *Biosyst. Eng.*, vol. 179, pp. 106–116, 2019.

[21] J. O. Bandele, M. O. Onibonoje, and A. O. Aladeloba, "Non-adaptive decision thresholds for gain saturated FSO links limited by weak to strong turbulence and pointing errors," *Int. J. Eng. Res. Africa*, vol. 46, 2020, doi: 10.4028/www.scientific.net/JERA.46.63.

[22] W. Shu-ning and C. Ting, "The Multiple Criteria Decision-Making Concept Applied to Input-Output Analysis," *IFAC Proc. Vol.*, vol. 20, no. 9, pp. 607–609, 1987.

A Multi-Objective Model Approach to an IoT-Based Granary Monitoring System

Moses Oluwafemi Onibonoje, *Member, IEEE*
Department of Electrical/Electronics and Computer Engineering
Afe Babalola University
Ado-Ekiti, Nigeria
onibonojemo@abuad.edu.ng

Adebayo Tunbosun Ogundipe
Directorate of Information Communication Technology
Afe Babalola University
Ado Ekiti, Nigeria
bosundipe@abuad.edu.ng

Adedayo Olukayode Ojo
Department of Elect/Elect. and Computer Engineering
Afe Babalola University
Ado-Ekiti, Nigeria
ojoao@abuad.edu.ng

Kehinde Adeniji
Department of Elect/Elect. and Computer Engineering
Afe Babalola University
Ado-Ekiti, Nigeria
adenijika@abuad.edu.ng

***Abstract*—A granary is designed to manage and control different environmental factors to ensure the desirable quality of the stored bulked grains, hence an effective monitoring system. This paper aims at establishing a model for an effective granary system to achieve multi-objective balance on real-time monitoring, minimum deployable nodes, coverage efficiency, longer node life, reduced cost and reduced power consumption. Multi-variable linear regression analysis was used to develop the mathematical model which formed the basis for formulating the objective function of the generic algorithm requisite. The simulated result of the model was validated by using a weighted sum method. The optimized model by using the available constraints is a template for a developed network of wireless nodes approach to establish a promising and resourceful representation of an operable granary monitoring system design.**

Keywords*— *granary, model, monitoring system, multi-objective design, optimization

I. Introduction

In recent times, researchers have established the possible and tremendous gains of applying wireless sensor networks in the condition monitoring of stored agricultural produces. Considerably, grains contribute well to the global economy and food security. Adequate monitoring is a necessity to ensure the desired quality of grains across the entire value chain, especially storage [1]. A granary is effectively designed to store grains, and also manage and control the different environmental factors that can affect the desirable quality of the stored bulked grains. Losses during storage can have relative economic impact on small-scale farmers and the nation, causes mainly due to poor storage infrastructure. The important factors to be monitored and controlled during storage include relative humidity, temperature, moisture, carbon gases and others [2]. Storage of agricultural produces is a very crucial component of ensuring food security. A poor and unscientific design of storage bins cannot guarantee the desired storage objective of the produces. Cereals and grains are the major agri-food products and income source for the dwellers of most developing countries. Globally, the annual losses during the storage of the products are being estimated at 30% [3], [4]. Beyond the losses due to other factors and effects, the quality of the post storage grains is an important area of consideration. The diffusion of moisture contents from warmer region of a storage bin to the colder end as a result of the temperature gradients is the main cause of grain deterioration. Therefore, effective monitoring of grain quality requires a distributed detection of the temperature gradients and other parameters at the different locations within the bin [5], [6]. In effect, an efficient storage setup demands to be designed to cater for both the multiple factors and the constant varying temperature gradients within it.

Various designs and measures are being developed to provide the needed schemes to solving the problems associated with granaries as being illustrated in Figure 1 [1]. A pilot test is good a measure to provide the experimental details of the desired designs due to the large variability of the involved parameters. However, on-site natural experiments present a time-consuming, irreproducible and expensive feasibility. Also, most devices for grain monitoring have been discovered to generally under-predict parameters for freshly dried grains. This is a basis established from the assumption that most devices are relatively sensitive to the outside layer of the grains. Consequently, a veritably developed mathematical model is the out-to-out scheme to be simulated, validated and implemented as the solution [7].

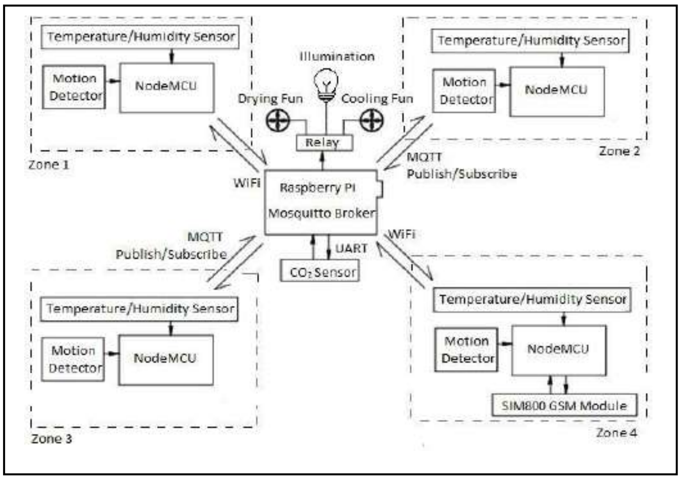


Fig. 1. A typical distributed scheme of granary monitoring [1]

Designing granaries and the attendant monitoring systems to minimize grain losses requires the requisite prediction of the temperature and moisture distribution within the bins. The prediction will also assist in developing and evaluating the strategies necessary for ventilation or aeration [8], [9]. Predictive models are particularly based of simplified observations and assumptions, and embedded in mechanistic theories when acceptably substantiated. This is recognized in [10] by developing a probabilistic model to generate a strain in predicting the other behaviour of the same species using maize grain extracts. The model is useful for introducing

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

strategies to improve grain storage conditions and preventing economic losses. Meanwhile, to establish a practical predictive model to cater for the generation of mycotoxin in the post-harvest period of storage is difficult, because of the constant changes in stored grains temperature and humidity, and the complexity of the granary environment. Also, multi-fungal grain loading which enables interaction among different fungi can lead to nonconformity between the modelled and real-life acquired results [11]. Other literature have also indicated that most methods and models employed in diagnosing the condition of grains during storage could not provide sufficient information that adequately determine the required parameters. Mathematical models for convective wheat drying, and the drying characteristics prediction is reported by [12]. In many of the cases, the analysis of the models are implemented by numerical methods, which foists some restrictions on their real-time application for detecting defects in grain mass. The discrete-element approach has been discussed to be suitable in modelling particulate structures like silos and bins. Calibration and validations have been introduced alongside the simplifications approach of numerical models, in reference to obtaining consistent results for comparison with the experimental outputs [13]–[15]. The implementation of derivable expert systems from the models of technological ecosystems yields the new approach to solving the dire problems [16], [17].

A system to adequately predict, monitor and control the various parameters to ensure a safe storage of grains should be designed using a multi-objective scheme [2]. The existing systems in the monitoring of the grains conditions do not solve issues related to inefficient coverage, high power consumptions, high rate of monitoring node failure, and data acquisition delays. An efficient solution will be required to overcome these issues and ensure optimal trade-off among them. Wireless sensor networks (WSNs) have been designed into a promising field as granary and greenhouse monitoring systems. Mathematical models involve the representation of the real-world problems with mathematical equations, and the solutions predicted from the simulated and validated results of such equations. The issues and the representative parameters to be considered in designing a monitored granary are many, therefore, the solutions to be proposed should be multi-objective in nature [18]. Different models related to specific features of WSN had been discussed [19], [20]. The various existing WSN models concentrate on addressing at most two or three of the several issues at stake. Hence, this study presents a multi-objective approach to designing an environmentally monitored granary system using the genetic algorithm. Genetic algorithm effectively adopts the useful info within a set of solutions to produce new solutions with enhanced performance. The remaining parts of the work is structured under the sections that include: methodology, results and discussion, conclusion and the references.

II. Methodology

The design optimization adopted in this study is the artificial genetic approach as earlier being proposed by Goldberg [21], [22]. Design optimization involves the process of determining the minimum/maximum of the objective function parameters, and must satisfy certain set of specific constraints. In this multi-objective optimization problem, all the vectors components that store the different objectives are being optimized simultaneously. The pilot bin was constructed with a capacity of one tonne of maize grains. The internal configuration was designed for monitoring nodes localization which was determined using Zigzag pattern deployment algorithm by [23]

A. *Mathematical Model*

The model equations of the multivariable linear regression (MVLR) analysis used here were incorporated in the adopted generic algorithm. Assuming there are '*r*' independent variables, having a functional relationship with the dependent variable *Y*. Therefore, the dependent variable can be approximated by using [24].

$$Y_i = \propto_0 + \propto_{1i}\, x_{1i} + \propto_{2i}\, x_{2i} + \cdots + \propto_{ri}\, x_{ri} + \in, \quad i = 1, 2, 3 \ldots . n \qquad (1)$$

In the determination of the threshold interests, the grain quality *Q* is the dependent variable with the main independent variables taken basically as moisture content *m*, temperature *t* and other factors represented as *g*. Therefore, the variables are related in equation (2).

$$Q = \propto_0 + \propto_1\, m + \propto_2\, t + \propto_2\, g \qquad (2)$$

For control measures, the two major means in controlling the uneven airflow, temperature and the humidity concentration are the operable heater, *H* and fan, *F* within the designed storage bin. The operation of the two control components is dependent upon all the variables in equation (2) as expressed in the form shown in equation (3) and (4)

$$H(x) = \propto'_0 + \propto'_1\, m + \propto'_2\, t + \propto'_2\, Q \qquad (3)$$

$$F(x) = \propto''_0 + \propto''_1\, m + \propto''_2\, t + \propto''_2\, Q \qquad (4)$$

By substituting the regression coefficients in Table I, the final model equations to determine the ultimate conditions for the control measures can be determined as,

$$H(x) = 237.5692 + 0.3451m - 0.0453t - 0.2101Q \qquad (5)$$

$$F(x) = 86.7592 + 0.1512m - 0.0118t - 0.0817Q \qquad (6)$$

$$Q(x) = w_1 \times H(x) + w_2 \times F(x) \qquad (7)$$

TABLE I. Experimentalvalues in chromosomes reproduction and objectives functions of the optimized process

Initial Population of the chromosomes	**Objective function *H(x)***	**Objective function *F(x)***	**Initial channel number**	**Quality function value *Q(x)***	**Probability in selection *P(x)***	**Random number set**	**Final channel number**	**Cumulative probability in selection *P'(x)***
231456545612987	235.34	85.441	1	163.54	0.1541	0.453	3	0.1541
632657451123854	238.68	88.431	2	167.77	0.1553	0.675	2	0.3094
843222345624427	245.56	89.218	3	169.56	0.1632	0.143	2	0.4726
951239874352356	242.56	86.754	4	165.11	0.2210	0.743	5	0.6936
754219765992316	248.21	87.381	5	170.21	0.2304	0.563	4	0.9240

Using the weighted function method to obtain the objective function, this can be expressed as shown in equation (7), where c and w_2 are the weight functions. The conformity of minimizing the effect of the uneven air-flow through the constant use of the fan is much higher than the necessity for intermittent drying process through the circulation of drying air. Hence, w_1 and w_2 can take values of 0.1 and 0.9 respectively.

B. Genetic Algorithm Supporting Approach

The genetic algorithm (GA) approach resolves optimization problems by replicating the nature of evolution of life. GA randomize the survival of the fittest within a set of string structures and exchange of information used to form searched algorithms. Repeatedly across generations, different new sets of artificial strings are created from the evolution of the old sets of strings with the fittest features. GA works on the contents of reproduction, mutation and crossover operators. The step-by-step GA approach used in this work follows a pattern as expressed, using an *‘n’* number of populations with 1-bit chromosomes randomly generated.

- The fitness *F(x)* was calculated for each chromosome *‘x’*
- Using three operators: selection, mutation and crossover; *‘n’* offspring was created from the current population
- A different new population is adapted to replace the current population
- The iteration “i” – “iii” continues repeatedly to achieve the termination criterion.

In validating the model for both the correlation coefficient and the MVLR, the analysis of variance (ANOVA) approach is the best choice [25]. The square of the correlation coefficient (R^2) relates to the significance (or otherwise) of the link between the objective function and the design input variables. The aim of the design in using the MVLR is to ensure a large value of the R^2 in between 0 and 1. The coefficients and constants of the accepted model were expressed using the multiple regression analysis in terms of air temperature (drying) by adapting the Page Model.

- $R^2 = 0.713$ for the modelling equation of heating temperature
- $R^2 = 0.727$ for the modelling equation for drying air flow

The closeness of the values of R^2 to unity in both cases validates the model approach.

III. Results and Discussion

The different genetic algorithm input parameters were varied during test-runs being carried out on the GA for both the sensing unit and the control units. Better results were achieved with more tests being run. The GA parameters are as shown in Table II. The grains temperature within varying bin column height were recorded across 4 *hours* for air velocities of 0.25 *m/s* and 0.1 *m/s* respectively. The simulated results for the control heating temperature and fan airflow for the composite nodes within the monitoring system are as shown in Figs 2 – 3.

The final solutions of the optimization front by Pareto method is as shown in Figure 4. It is observed that the optimization model can find different solutions based on the constituent variable parameters of the quality of grains in focus.

TABLE II. THE RESULTING GA PARAMETERS

Details	Values
Population	45
Iteration number	100
Probability (Cross-over)	0.870
Probability (Mutation)	0.032

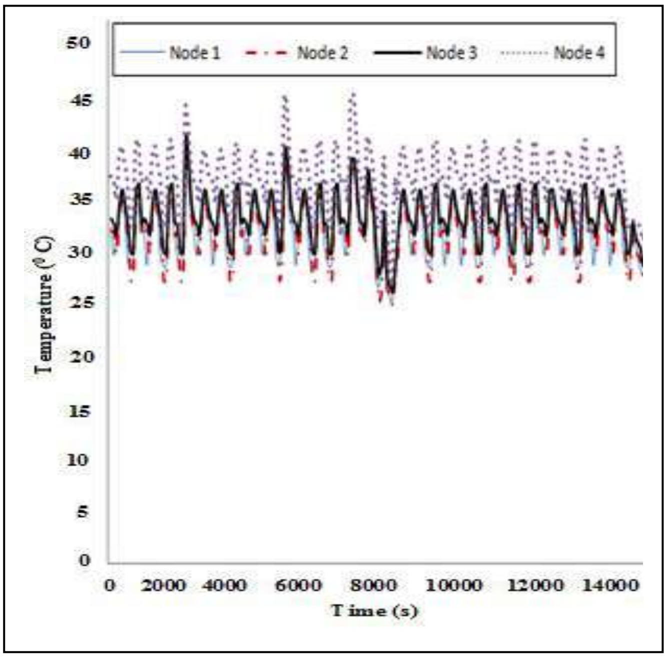


Fig. 2. The simulated heating temperature variation

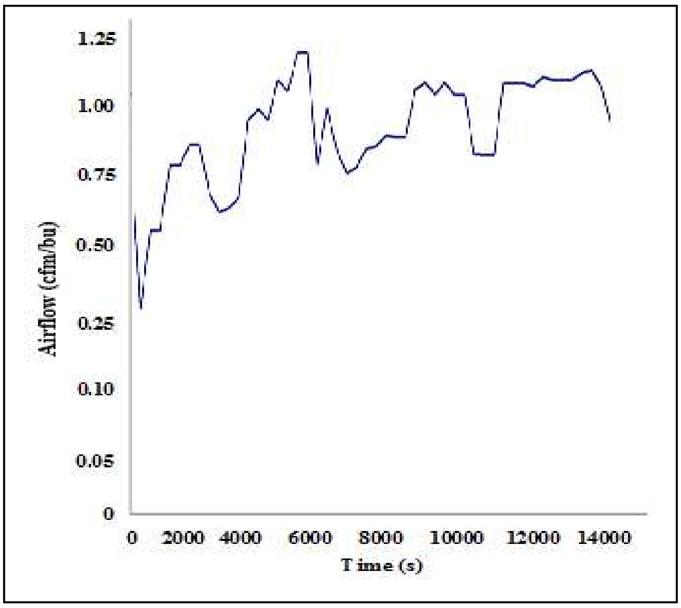


Fig. 3. The simulated air-flow variation

IV. Conclusion

In this paper, a multi-objective optimization technique has been proposed for granary systems. The approach focusses on the possibilities of monitoring and controlling the parameters affecting the quality of stored grains. The airflow constraint from the controlling fan and the mitigating heating temperature are the parameters of interest in the optimization process of the model. The genetic approach combines the multiple objectives into a singular scalar function and randomly specify for each selection. The model has a good application for storage in modular forms and suitable for middle-level farmers and marketers of grains.

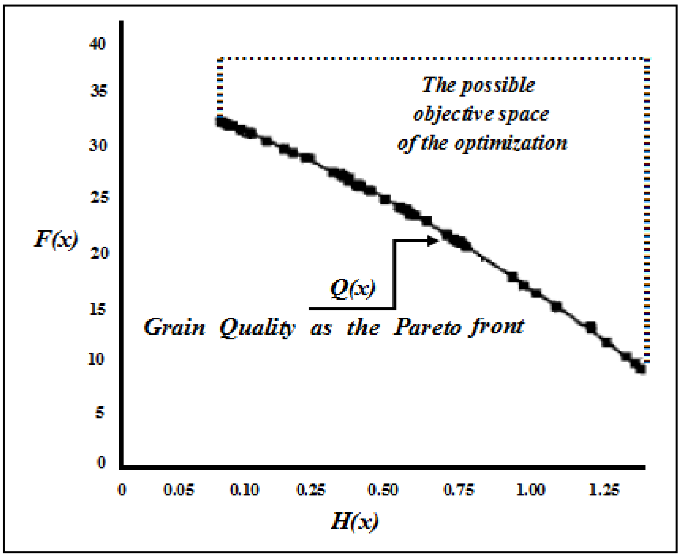


Fig. 4. The simulated result of the optimization front of the variables

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Victor Ijeh during the experimental set-up and data collection.

REFERENCES

[1] T. B. Tedla, J. J. L. Bovas, Y. Berhane, M. N. Davydkin, and P. S. James, "Automated Granary Monitoring And Controlling System Suitable For The Sub-Saharan Region," *Int. J. Sci. Technol. Res.*, vol. 8, no. 12, pp. 1943–1951, 2019.

[2] M. O. Onibonoje, N. I. Nwulu, and P. N. Bokoro, "A wireless sensor network system for monitoring environmental factors affecting bulk grains storability," *J. Food Process Eng.*, vol. 42, no. 7, 2019, doi: 10.1111/jfpe.13256.

[3] K. M. S. Banga, S. Kumar, N. Kotwaliwale, and D. Mohapatra, "Major insects of stored food grains," *IJCS*, vol. 8, no. 1, pp. 2380–2384, 2020.

[4] A. B. Abass, G. Ndunguru, P. Mamiro, B. Alenkhe, N. Mlingi, and M. Bekunda, "Post-harvest food losses in a maize-based farming system of semi-arid savannah area of Tanzania," *J. Stored Prod. Res.*, vol. 57, pp. 49–57, 2014.

[5] L. I. Quemada-Villagómez *et al.*, "Numerical Study to Predict Temperature and Moisture Profiles in Unventilated Grain Silos at Prolonged Time Periods," *Int. J. Thermophys.*, vol. 41, no. 5, pp. 1–28, 2020.

[6] N. Bluvshtein, E. Villacorta, C. Li, B. C. Hagen, V. Frette, and Y. Rudich, "Early detection of smoldering in silos: Organic material emissions as precursors," *Fire Saf. J.*, p. 103009, 2020.

[7] M. O. Onibonoje, J. O. Bandele, and T. E. Fabunmi, "Modelling and Implementation for Airflow and Temperature Distribution in a Small-Scale Granary," *Int. J. Eng. Res. Africa*, 2020.

[8] B. Daviron, C. Perrin, and C.-T. Soulard, "History of urban food policy in Europe, from the ancient city to the industrial city," in *Designing Urban Food Policies*, Springer, Cham, 2019, pp. 27–51.

[9] M. O. Onibonoje, K. S. Alli, T. O. Olowu, M. A. Ogunlade, and A. O. Akinwumi, "Resourceful selection-based design of wireless units for granary monitoring systems.," *ARPN J. Eng. Appl. Sci.*, vol. 11, no. 23, pp. 13754–13759, 2016.

[10] M. P. Jiang *et al.*, "Predictive model of aflatoxin contamination risk associated with granary-stored corn with versicolorin A monitoring and logistic regression," *Food Addit. Contam. Part A*, vol. 36, no. 2, pp. 308–319, 2019.

[11] D. Garcia, A. J. Ramos, V. Sanchis, and S. Mar\'\in, "Predicting mycotoxins in foods: a review," *Food Microbiol.*, vol. 26, no. 8, pp. 757–769, 2009.

[12] F. Hammami, S. Ben Mabrouk, and A. Mami, "Modelling and simulation of heat exchange and moisture content in a cereal storage silo," *Math. Comput. Model. Dyn. Syst.*, vol. 22, no. 3, pp. 207–220, 2016.

[13] Á. Ramirez-Gómez, "The discrete element method in silo/bin research. Recent advances and future trends," *Part. Sci. Technol.*, vol. 38, no. 2, pp. 210–227, 2020.

[14] Y. Menni, A. Azzi, C. Zidani, and B. Benyoucef, "Numerical analysis of turbulent forced-convection flow in a channel with staggered l-shaped baffles," *J. New Technol. Mater.*, vol. 277, no. 5615, pp. 1–12, 2016.

[15] Y. Menni, A. Azzi, and A. J. Chamkha, "Turbulent heat transfer and fluid flow over complex geometry fins," in *Defect and Diffusion Forum*, 2018, vol. 388, pp. 378–393.

[16] I. Lutsyk, Y. Franko, V. Rak, I. Lutsyk, R. Leshchii, and O. Potapchuk, "Mathematical modeling of energy-efficient active ventilation modes of granary," in *2019 9th International Conference on Advanced Computer Information Technologies (ACIT)*, 2019, pp. 105–108.

[17] R. Kalinichenko and V. Voytyuk, "Mathematical Modeling of Teplomass-exchange Processes of High-Temperature Thermo-Processing of Grain Materials," *Sci. Her. NULES Ukr. Ser. Tech. Energy*, vol. 275, pp. 59–67, 2017.

[18] M. O. Onibonoje, A. O. Ojo, and T. O. Ejidokun, "A mathematical modelling approach for optimal trade-offs in a wireless sensor network for a granary monitoring system.," *Int. J. Technol.*, vol. 10, no. 2, pp. 212–218, 2019.

[19] O. Tymchenko, K. Zelyanovsky Michałand Szturo, and O. O. Tymchenko, "Mathematical models for specialized and sensory networks of wireless access," *Tech. Sci. Warm. Maz. Olsztyn*, 2016.

[20] M. O. Onibonoje, "Modeling and analysis of modulation schemes on antenna space blocks time code," *Int. J. Sci. Technol. Res.*, vol. 8, no. 9, pp. 2246–2251, 2019.

[21] P. Vijian and V. P. Arunachalam, "Modelling and multi objective optimization of LM24 aluminium alloy squeeze cast process parameters using genetic algorithm," *J. Mater. Process. Technol.*, vol. 186, no. 1–3, pp. 82–86, 2007.

[22] B. Andres-Toro, J. M. Giron-Sierra, P. Fernandez-Blanco, J. A. Lopez-Orozco, and E. Besada-Portas, "Multiobjective optimization and multivariable control of the beer fermentation process with the use of evolutionary algorithms," *J. Zhejiang Univ. A*, vol. 5, no. 4, pp. 378–389, 2004.

[23] A. Hawbani and X. Wang, "Zigzag Coverage Scheme Algorithm & Analysis for Wireless Sensor Networks.," *Netw. Protoc. Algorithms*, vol. 5, no. 4, pp. 19–38, 2013.

[24] S. Park, *Robust design and analysis for quality engineering*. Boom Koninklijke Uitgevers, 1996.

[25] A. Topuz and C. Hamzacebi, "Moisture ratio prediction in drying process of agricultural products: A new correlation model," *Appl. Eng. Agric.*, vol. 26, no. 6, pp. 1005–1011, 2010.

Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches

Ndiatenda Ndou
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
ndiatenda.ndou@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

Ashwini Jadhav
Faculty of Science
The University of the Witwatersrand,
Johannesburg, South Africa
ashwini.jadhav@wits.ac.za

***Abstract*—This research provides a comparative study of the genre classification performance of deep-learning and traditional machine-learning models. Furthermore, we investigate the performance of machine-learning models implemented on three-second duration features, to that of those implemented on thirty-seconds duration features.**

We present the categories of features utilized for automatic genre classification and implement Information Gain Ranking algorithm to determine the features most contributing to the correct classification of a music piece. Machine-learning models and Convolutional Neural Network (CNN) were then trained and tested on ten GTZAN dataset genres. The k-Nearest Neighbours (kNN) provided the best classification accuracy at 92.69% on three-seconds duration input features.

***Index Terms*—machine-learning, deep-learning, music genre classification, CNN, MFCC**

I. Introduction

Genre is one of the most common of factors distinguishing music pieces. Human responses to genre can be biased, however, broad genre definitions exist worldwide. Observing the shift of music to digital platforms, it becomes clear that the automating the task of music classification would be beneficial to all parties involved.

This research explores automatic music genre classification with the aim to show that machine-learning and deep-learning approaches can be utilized to classify music from only the audio signal itself, reducing search-time for music pieces within the large music databases that have emerged with digital music platforms. We compare the deep-learning approach to traditional machine-learning models, furthermore, we investigate the performance of machine-learning classifiers with three-seconds duration features, to those implemented with thirty-second duration features.

This study was conducted in three phases, namely, 'phase A', 'phase B, and 'phase C'. Phase 'A' and 'B' utilized six traditional machine learning classifiers to perform automatic music genre classification, however, the two phases experiment with different dimensions of input features. Phase 'C' provides the deep-learning approach and a machine-learning approach with more audio excerpts but shorter duration.

Related literature shows that digital music platform users are more likely to browse music by genre than artist similarity or recommendations, therefore, successful music genre classification will allow end-users to efficiently browse music within genre categories [10].

This paper continues with a brief background and review of related literature, followed by Section III with the procedures implemented. Section IV presents the automatic music genre classification results, and Section V concludes the paper.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

A. Music Features

We present four categories of features utilized for music genre classification. The correct set of features needs to be selected in order to perform correct and informed classification.

1) Magnitude-based features: these features can be described as timbral features, describing the loudness, pitch, and compactness of music [1]. Timbral features of music are essential for humans to categorize and group together different sounds coming from a single source [19]. Some examples of features belonging to this category are spectral features which are embedded in the magnitude spectrum, a spectrum obtained from the absolute value of the Fourier transform of a music chord [1], this examples include: spectral rolloff, spectral flux, spectral centroid, spectral spread, spectral decrease, spectral slope, spectral flatness, and Mel Frequency Cepstral Coefficients (MFCCs).

2) Tempo-based features: these are the features that describe the rhythmic aspects of music such as the rhythm and tempo [1]. Examples of features belonging to this category are; Tempo measured in beats per minute (BPM), Energy (audio signal intensity) measured using the root mean square (RMS), and the Beat Histogram to visualize important properties of audio signals through evaluation of the histogram peak, amplitude, and other statistical measures.

Author(s)	Dataset	Model used	Classification Accuracy
Sturm (2013) [20]	GTZAN	Sparse Representation Classification	83.00%
Bergstra *et al.* (2006) [5]	GTZAN	ADABOOST	82.50%
Li *et al.* (2003) [12]	GTZAN	Support Vector Machines	78.50%
Lidy *et al.* (2007) [13]	GTZAN	Sequential Minimal Optimization	76.80%
Benetos and Kotropoulos (2008) [4]	GTZAN	Non-negative Tensor Factorization	75.00%
Choi *et al.* (2016) [6]	Navier Music	CNN	75.00%
Bahuleyan (2018) [3]	Audio Set	CNN	65.00%
Tzanetakis and Cook (2002) [21]	GTZAN	Gaussian Mixture Mode	61.00%

Table I: Various studies that have shown capability to perform genre classification. The columns list the author(s), dataset utilized, model implemented, and the classification accuracy attained.

3) Pitch-based features: features belonging to this category describe the pitch of a music piece, this is an essential building block of the harmony, key, and melody of an audio piece [11]. This category is important to explore because pitch perception determines the frequency level of the underlying audio signal [11]. An example of a pitch-based feature is the ‘zero crossing rate’, which is the count of sign changes in consecutive blocks of an audio excerpt [1].

4) Chordal progression features: these group of features explore the pitch ‘chroma’, which is a twelve-dimensional vector with each dimension representing one pitch class [11]. Chroma can also be viewed as a distribution, where both the number of occurrences of a pitch and its energy can be deduced from the class values [11].

B. Related Work Results

A review of related literature reveals that several studies have displayed the capability to solve the problem of music genre classification. We present notable genre classification algorithms in Table I.

II. Methodology

This section outlines the method and set of experiments performed in the studies reviewed. The procedures carried out include further preprocessing of the dataset, feature selection, and a description of the machine learning classifiers employed.

A. Data Description

The dataset utilized for all three studies conducted was the GTZAN dataset [17]. This GTZAN dataset is an ensemble of 1000 excerpts of thirty second duration each. The 1000 music pieces are categorized into 10 genres with 100 music pieces for each genre.

For one of the studies conducted [9], the original dataset was duplicated and divided into 10 000 excerpts of three seconds duration each. This procedure provided more training data, however, the dataset did not have a consistent number of samples per genre, with some genres having slightly less or more than 1000 music pieces.

B. Feature Extraction and Representation

We have identified four categories of features that are generally hypothesized to contribute in the correct classification of music genre. Prior to the selection of these features for model implementation, vital preprocessing experiments have to be conducted to make the raw data suitable for the classification task. Feature extraction was performed in this study for two purposes:

- **Dimensionality reduction:** the raw data dimensions are usually too large, that is, an entire raw audio file may be too large to process efficiently. Related studies show that a feature set is used to present the data with fewer values, a single feature value may be produced for an entire audio signal [11].
- **Meaningful representation:** the raw audio fie contains all the information we can possibly extract and use, however, it is important that we represent the musical aspects in an interpretable manner by machines or humans [16].

The computation of features from a music excerpt usually gives rise to an n-dimensional vector, where the value of n is dependent on the length of the audio piece under analysis. If the value of n is large, we deal with high dimensional feature vectors which are inefficient to process, therefore, for a feature vector $V = (v_1, v_2, v_3 \ldots, v_n)$, the following feature representations were explored:

- **Mean:** the average value of feature V, computed as:

$$\mu V = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{1}$$

- **Standard deviation:** a measure of the spread of values of feature V, computed as:

$$\sigma V = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(vi - \mu v)^2} \tag{2}$$

- **The Feature Histogram:** obtained by arranging the feature’s local window intensities into bin ranges then taking a count of each bin’s contents and modelling a frequency histogram [1]. The normalized histogram bin values can be used for classification.

- **Mel Frequency Cepstral Coefficients (MFCC) Aggregation:** this representation takes the first n coefficients that form part of the short-term sound power-spectrum [8], [14]. Independently, each dimension is assessed, producing n coefficients per dimension. For this work, $n = 4$ was selected.
- **Area Moments:** this is an important concept in computer vision and image processing. This work follows a classic image moments implementation where 10 area moments were produced for image processing, treating each image as a two-dimensional vector $V(v_1 v_2)$, with v_1, v_2 indexing the underlying matrix [14]. We treat the extracted feature values from the audio signal as two-dimensional images and apply the moments algorithm in the work cited above.

C. Feature Selection

In this section, we present the various features utilized to perform automatic music genre classification in the studies this paper extends [1], [9], [16]. Feature selection is essential for the reduction of irrelevant and redundant data, the reduction of which may result in improved model learning accuracy, and reduced training time. Information gain ranking algorithm was utilized for comparison of the various features' contribution to a correct classification.

Features Maintained	**Rep.**	**Dim. 54**
Spectral Contrast	Mean	7
Spectral Rolloff	Mean + SD	2
Spectral Flux	Mean + SD	2
Spectral Crest	Mean + SD	2
Spectral Flatness	Mean + SD	2
Spectral Decrease	Mean + SD	2
Spectral Kurtosis	Mean + SD	2
Spectral Slope	Mean + SD	2
Spectral Skewness	Mean + SD	2
Spectral Centroid	Mean + SD	2
Spectral Spread	Mean + SD	2
Spectral Entropy	Mean + SD	1
Zero Crossing Rate	Mean + SD	2
Mel Frequency Cepstral Coefficients	Mean	17
Root Mean Square	Mean + SD	2
Beat Histogram	Sum + Mean + SD	3
Temporal Statistic Spread	Mean + SD	2
Features Eliminated	**Rep.**	**Dim. 51**
Spectral Crest Factor	Mean + SD	2
Spectral Tonal Power Ratio	Mean + SD	2
Mel Frequency Cepstral Coefficients	SD	35
Chroma	Mean	12

Table II: The set of features selected for training the employed machine-learning classifiers in 'phase B'. The upper shaded portion the table presents the features maintained after using Information Gain Ranking (IGR) algorithm, while the lower portion presents the eliminated features. The column heading acronym Rep. and Dim. are the feature representation and feature dimension respectively. A total of 54 features were selected in 'phase B', [16](sic).

The work presented in this paper was carried out over three studies referred to as phases, therefore, we continue by presenting three different sets of features selected. TableIII presents the features selected for 'phase A', Table IV presents the features for 'phase B', and TableII presents the feature set selected for 'phase C'.

Features Maintained	**Rep.**	**Dim. 459**
Spectral Flux	MFCC	4
Spectral Variability	MFCC	4
Compactness	Mean + SD	2
MFCCs	MFCC	52
Peak Centroid	Mean + SD	2
Peak Smoothness	SD	1
Complex Domain Onset Detection	Mean	1
Loudness + Sharpness and Spread	Mean	26
OBSI + Radio	Mean	17
Spectral Decrease	Mean	1
Spectral Flatness	Mean	20
Spectral Slope	Mean	1
Shape Statistic Spread	Mean	1
Spectral Centroid	MFCC	4
Spectral Rolloff	SD	1
Spectral Crest	Mean	19
Spectral Variation	Mean	1
Autocorrelation coefficients	Mean	49
Amplitude modulation	Mean	8
Zero Crossing + SF	MFCC	8
Envelope Statistic Spread	Mean	1
LPC and LSF	Mean	12
Root Mean Square	Mean + SD	2
Fraction of low energy	Mean	1
Beat Histogram	SD	171
Strength of Strongest Beat	Mean	1
Temporal Statistic Spread	Mean	1
Chroma	MFCC	48
Features Eliminated	**Rep.**	**Dim. 223**
Peak Fux	20-bin FH	20
Peak Smoothness	Mean	1
Shape Statistic Centroid and Skewness	Mean	1
Shape Statistic Kurtosis	Mean	2
Strongest Frequency of Centroid	MFCC	4
Spectral Rolloff	Mean	1
Strongest Frequency FFT	MFCC	4
Envelope Centroid, Skewness and Kurtosis	Mean	4
Beat Histogram	Mean	171
Strongest Beat	Mean + SD	2
Strength of Strongest Beat	SD	1
Fraction Low Energy	SD	1
Beat Sum	MFCC	4
Relative Difference Function	MFCC	4
Temporal Statistic Centroid	Mean	1
Temporal Statistic Skewness	Mean	1
Temporal Statistic Kurtosis	Mean	1

Table III: The set of features selected for training the employed machine-learning classifiers in 'phase A'. The upper shaded portion the table presents the features maintained after using Information Gain Ranking (IGR) algorithm, while the lower portion presents the eliminated features. The column heading acronym Rep. and Dim. are the feature representation and feature dimension respectively. A total of 459 features were selected in 'phase A', [1](sic).

D. Traditional Machine-Learning Models

For this research, we implemented the following off-the-shelf machine-learning models were implemented through

Features Maintained	**Rep.**	**Dim. 57**
Chroma	Mean + SD^2	2
Root Mean Square	Mean + SD^2	2
Spectral Centroid	Mean + SD^2	2
Spectral Bandwidth	Mean + SD^2	2
Spectral Rolloff	Mean + SD^2	2
Zero Crossing Rate	Mean + SD^2	2
Mel Frequency Cepstral Coefficients	Mean + SD^2	20
Harmony	Mean + SD^2	2
tempo	Mean	3
Features Eliminated	**Rep.**	**Dim. 51**
Spectral Crest Factor	Mean + SD	2
Spectral Tonal Power Ratio	Mean + SD	2
Chroma	Mean	12

Table IV: The set of features selected for training the employed machine learning classifiers in ‘phase C’. The upper shaded portion the table presents the features maintained after using Information Gain Ranking (IGR) algorithm, while the lower portion presents the eliminated features. The column heading acronym Rep. and Dim. are the feature representation and feature dimension respectively. A total of 57 features were selected in ‘phase C’, [9](sic).

the Scikit Learn library [20]: k-Nearest Neighbours, Linear Logistic Regression, Multilayer Perceptron, Random Forests trees, and Support Vector Machines. The hyperparameters for each model are provided in Section III.

E. Deep-Learning Approach

The Convolutional Neural Network (CNN) architecture in this research was constructed using Keras [7]. The CNN built here has an input layer and five convolutional blocks, with each convolutional block consisting the following: The Convolutional Neural Network (CNN) architecture in this research was constructed using Keras [7]. The CNN built here has an input layer and five convolutional blocks, with each convolutional block consisting the following:

- Convolutional layer with mirrored padding, 1x1 stride, and 3x3 filter
- The rectified linear activation function (ReLu)
- Maximum pooling with 2x2 stride and window size
- Probability of 0.2 for dropout regularization

The last layer of the CNN outputs the probabilities of ten label classes through a fully-connected layer implementing the SoftMax activation function. The class that attains the highest probability becomes the classified label for a given input. The CNNs were trained on the spectrograms, twenty Mel Frequency Cepstral Coefficients (MFCC) of the three-or-thirty-seconds feature set, and the extracted spectrograms.

F. Evaluation Metrics

To reduce bias and produce credible results, we performed 3-reapeted 10-fold cross-validation prior to the models classifying the test dataset. We utilize the classification accuracy and training time to evaluate the performance of all employed models.

III. Results and Discussion

This section outlines the results obtained when we use the features provided in Table III, IV, and II to perform genre cassification on ten GTZAN genres.

A. Traditional Machine-Learning Models

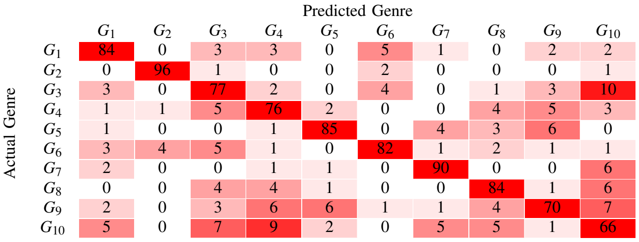
The traditional machine-learning models implemented in this research were tested on the GTZAN dataset, the results are presented in Table V, VI, and VII.

Table V presents the results obtained during ‘phase A’ of this research. The Linear Logistic Regression provided the best classification accuracy at 81%, however, with the exception of the Multilayer Perceptron, we note that the logistic regression had the longest training time. We also note that the níve Bayes classifier was outperformed by all the trained classifiers.

Table VI presents the results obtained during ‘phase B’ of this research. The Support Vector Machines (SVM) provided the best classification accuracy at 80.80%. The SVM also attained a relatively low training time taking 0.3 seconds to build. Logistic Regression displayed notable accuracy again with a classification accuracy of 75.80%, however, failing to outperform the LogitBoost implementation followed in ‘phase A’.

Table VII presents the various models’ hyperparameters and performance during ‘phase C’ of this research. The k-Nearest Neighbour (kNN) provided the best classification accuracy at 92.69%, furthermore, the kNN attained the shortest training time of 78 milliseconds. We note that ‘phase C’ utilized a three-seconds duration feature set as opposed to the thirty-seconds duration dataset utilized in ‘phase A’ and ‘B’ of this research.

Figure 1 presents the confusion matrix attained when classifying music into ten GTZAN genres using Linear Logistic Regression models during ‘phase A’ of this research. We note the significant overlap between rock and country music, where ten country music excerpts were classified as rock music. Furthermore, rock music was the most misclassified genre, with rock music excerpts classified as blues, country, disco, metal, and pop.



Actual Genre	Predicted Genre									
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	84	0	3	3	0	5	1	0	2	2
G_2	0	96	1	0	0	2	0	0	0	1
G_3	3	0	77	2	0	4	0	1	3	10
G_4	1	1	5	76	2	0	0	4	5	3
G_5	1	0	0	1	85	0	4	3	6	0
G_6	3	4	5	1	0	82	1	2	1	1
G_7	2	0	0	1	1	0	90	0	0	6
G_8	0	0	4	4	1	0	0	84	1	6
G_9	2	0	3	6	6	1	1	4	70	7
G_{10}	5	0	7	9	2	0	5	5	1	66

Figure 1: A confusion matrix obtained in the classification of music into ten GTZAN genres using Linear Logistic Regression during ‘phase A’ of this research. The row labels represent actual genre labels, while the column labels represent the predicted genre labels, where: G_1 = **Blues**, G_2 = **Classical**, G_3 = **Country**, G_4 = **Disco**, G_5 = **Hiphop**, G_6 = **Jazz**, G_7 = **Metal**, G_8 = **Pop**, G_9 = **Reggae**, and G_{10} = **Rock**.

Classifier	Accuracy	Training Time (s)	Hyperparameters
Linear Logistic Regression	81.00%	25.2500	maximum number of itterations for LogitBoost=500
Random Forests	75.70%	18.0800	number of trees = 1000
Support Vector Machines	75.40%	3.8200	kernel degree=3, tolerance=0.001, epsilon for loss function=0.1, used polynomial kernal: $\gamma u' v + coef_o$, and did not normalize
Multilayer Perceptron	75.20%	27.480	number of hidden layers= number of hidden classes, learning rate=0.3, training time=500 epochs, validation threshold=20
k-Nearest Neighbour	72.80%	0.0100	number of neighbours=1, using absolute error for cross-validation, and appied linear search algorithm
naíve Bayes	53.20%	0.5600	used normal distribution for numeric attributes and supervised discretization

Table V: Classification results and implementation details of each of the models employed during ‘phase A’ of this research. The columns list the accuracy, training time and hyperparameters related to the implementation of each classifier, [1](sic).

Classifier	Accuracy	Training Time (s)	Hyperparameters
Support Vector Machines	80.80%	0.3000	radial basis function kernel, tolerance=0.001, and regularization=0.17
Multilayer Perceptron	77.30%	0.2300	hidden layers=2, learning rate=0.02, activation=ReLu, max iterations=200, solver=adam, and tolerance=0.0001
Logistic Regression	75.80%	0.0800	solver=newton-cg and max itterations=500
Random Forests	72.40%	61.080	split function=gini, number of trees = 100, and max depth 100
k-Nearest Neighbour	69.70%	0.0110	k=7 with manhattan distance metric, weighting=distance
naíve Bayes	54.50%	0.0019	Gaussian naíve Bayes with smoothing

Table VI: Classification results and implementation details of each of the models employed during ‘phase B’ of this research. The columns list the accuracy, training time and hyperparameters related to the implementation of each classifier, [16](sic).

Classifier	Accuracy	Training Time (s)	Hyperparameters
k-Nearest Neighbours	92.69%	0.0780	nearest neighbours=1
Multilayer Perceptron	81.73%	60.620	activation=ReLu solver lbfgs
Random Forests	80.28%	52.890	number of trees=1000, max depth=10, $\alpha = e^{-5}$, and hidden layer sizes=(5000,10)
Support Vector Machines	74.72%	3.8720	decision function shape=ovo
Logistic Regression	67.52%	3.6720	penaty=12, multi class=multinomial

Table VII: Classification results and implementation details of each of the models employed during ‘phase C’ of this research. The columns list the accuracy, training time and hyperparameters related to the implementation of each classifier, [9](sic).

B. Deep-Learning Approach

In this subsection, we present the classification results of the Convolutional Neural Network (CNN) when trained on spectrograms, three-seconds features, and thirty-seconds features. Table VIII compares the accuracy attained and brief details of the implementation followed.

We see that the classification accuracy provided by the CNN is relatively lower than that provided by traditional machine learning models. The highest classification accuracy attained with the CNN is 72.40%, where the three-second feature

Classifier	Epochs	Test Loss	Accuracy
CNN (3-Sec Features)	50	0.873	72.40%
CNN (Spectrograms)	120	2.254	66.50%
CNN (30-Sec Features)	30	1.609	53.50%

Table VIII: Classification results attained from CNN implementation using the three-seconds duration, thirty-second duration, and spectrogram input feature sets, [9](sic).

set was utilized. The three-second feature set provides more training data which could explain the higher accuracy attained through it. Thirty-seconds duration features gave the CNN implementation the lowest accuracy at 53.50%. We note that the implementation of CNN with spectrograms attains higher accuracy as the number of epochs is increased, however, time and computational constraints did not allow increasing epochs further than 120 in this research.

IV. Conclusion

This work aimed at automatic music genre classification using deep-learning and traditional machine-learning models. A review of related literature revealed the capability of these classifiers and a benchmark to compare the work of this research. We note that the reliability of a learning model is dependent on the quality of its ground truth, therefore, it is essential to ensure the ground truth is well-founded and motivated.

This research was conducted in three phases, namely, 'phase A', 'phase B', and, 'phase C'. Each phase had a significance that aligns with the contribution made by this research to the current body of work. We present music genre classification via machine-learning and deep-learning approaches, furthermore, this work provides a comparison of the accuracy of machine-learning models and deep-learning models in completing the classification task.

After training several classifiers, the k-Nearest Neighbours (kNN) provided the best accuracy at 92.69%, furthermore, the kNN had a relatively low training time of 78 milliseconds. The higher accuracy attained by kNN relative to related literature can be explained by the three-seconds duration feature set which provides more training data. Backed by these findings, We conclude that three-second duration input features can provide better accuracy than thirty-second duration input features.

Further noteworthy performances were provided by the Linear Logistic Regression and Support Vector Machines (SVM), attaining 81.00% and 80.80% respectively. The Convolutional Neural Network (CNN) implementations followed in this research provided relatively low accuracy, with the most accurate CNN implementation attaining 72.40%.

This work has shown that automatic music genre classification is possible, furthermore, traditional machine learning models tend to outperform deep-learning approaches.

Acknowledgment

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835).

References

[1] R. Ajoodha, R. Klein, and B. Rosman, "Single-labelled music genre classification using content-based features," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015, pp. 66–71.

[2] R. Ajoodha, R. Klein, and M. Jakovljevic, "Using statistical models and evolutionary algorithms in algorithmic music composition," in *Encyclopedia of Information Science and Technology, Third Edition*. IGI Global, 2015, pp. 6050–6062.

[3] H. Bahuleyan, "Music genre classification using machine learning techniques," 2018.

[4] E. Benetos and C. Kotropoulos, "A tensor-based approach for automatic music genre classification," in *2008 16th European Signal Processing Conference*, 2008, pp. 1–4.

[5] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and adaboost for music classification," *Machine Learning*, vol. 65, pp. 473–484, 12 2006.

[6] K. Choi, G. Fazekas, and M. Sandler, "Explaining deep convolutional neural networks on music classification," 2016.

[7] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[8] I. Fujinaga, "Adaptive optical music recognition," Ph.D. dissertation, McGill University, CAN, 1997, aAINQ29937.

[9] D. S. Lau and R. Ajoodha, "Music genre classification: A comparative study between deep-learning and traditional machine learning approaches," in *Sixth International Congress on Information and Communication Technology (6th ICICT)*. Springer, 2021, pp. 1–8.

[10] J. H. Lee and J. S. Downie, "Survey of music information needs, uses, and seeking behaviours: preliminary findings." in *ISMIR*, vol. 2004. Citeseer, 2004, p. 5th.

[11] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley Online Library, 10 2012.

[12] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ser. SIGIR '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 282–289. [Online]. Available: https://doi.org/10.1145/860435.860487

[13] T. Lidy, A. Rauber, A. Pertusa, and J. Iñesta, "Combining audio and symbolic descriptors for music classification from audio," 2007.

[14] C. McKay, R. Fiebrink, D. McEnnis, B. Li, and I. Fujinaga, "Ace: A framework for optimizing music classification," in *ISMIR*, 2005.

[15] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?" in *ISMIR*, 2006.

[16] T. Nkambule and R. Ajoodha, "Classification of music by genre using probabilistic graphical models and deep learning models," in *Sixth International Congress on Information and Communication Technology (6th ICICT)*. Springer, 2021, pp. 1–6.

[17] A. Olteanu. Gtzan dataset - music genre classification. [Online]. Available: https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," 2018.

[19] B. L. Sturm, "Alexander lerch: An introduction to audio content analysis: Applications in signal processing and music informatics," *Computer Music Journal*, vol. 37, no. 4, pp. 90–91, 2013.

[20] B. L. Sturm, "On music genre classification via compressive sampling," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1–6.

[21] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

Decision Making Engine for Task Offloading in On-device Inference Based Mobile Applications

Vihanga Ashinsana Wijayasekara
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
vihanga.2017581@iit.ac.lk

Prathieshna Vekneswaran
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
prathieshna.v@iit.ac.lk

***Abstract*—On-device AI is one of the latest cutting-edge technologies which allows devices to run the machine learning models on the device. This provides low latency, fewer privacy concerns, and many other advantages. But even though modern mobile devices come with great performance, there are situations where these mobile devices cannot handle the resource requirements of these on-device based mobile applications. As a solution for this lack of resources issue, mobile devices can transfer their heavy computational tasks to other devices such as nearby devices to minimize the burden. Transferring all computational tasks to resource-rich devices will not yield favorable under all circumstances. Depending on various conditions such as network strength and execution time, the mobile device should be able to choose a way to execute the task, either execute locally or execute remotely with the help of a resource-rich device. This paper will discuss the implementation of a decision making engine that decides when to offload machine learning tasks in on-device inference based Android mobile applications. Main objective of the proposed solution is to minimize the execution delay. Experimental results show that our decision making engine has an accuracy of 81.33% in identifying the best decision. The proposed research is a novel approach designed for on-device inference based Android mobile applications.**

Keywords—offloading decision making, smartphones, on-device Inference

I. Introduction

Smartphones are intertwined with day-to-day activities. Over the years, smartphones have faced advances in increasing computational power, integrating the latest technologies, and so on. These latest technologies have led to the development of more and more sophisticated and complex mobile applications. As a result of that, today cutting-edge technologies such as artificial intelligence, prediction, and recommendation using machine learning, etc. can be found in many mobile applications.

A. On-device AI and the smartphone

Machine learning was able to transform the mobile applications into more person-oriented applications that can provide accurate predictions and recommendations based on user's behavior. Usually, machine learning models used in these mobile applications are hosted in the cloud, which allows mobile applications to send data to the cloud, execute and get the results back. Recently there is an increasing movement towards On-Device AI, which allows the mobile applications to run the machine learning models on the devices due to the various drawbacks of the cloud-based approach such as high latency and privacy concerns.

According to Gartner, by 2022, most of the smartphones shipped by the percentage which is 80% will have on-device AI capability [1]. Since the model is much closer to the data sources such as sensors and cameras, the latency is very much lower compared to the cloud-based paradigm. Since the data does not need to upload to anywhere, there is no need to worry about privacy concerns.

Even though On-device AI seems to be the best approach, it has some drawbacks as well. Since the machine learning model runs on the device, it needs much processing power and energy. Recent research shows that there is a visible difference between on-device inference based mobile applications and cloud-based mobile applications when it comes to energy consumption [2]. Mobile devices are made lightweight and smaller to fulfill their main objective, mobility. Thus, there could be situations when required resources for on-device AI based tasks in the mobile application cannot bridge the available resources in the mobile device. This is where task offloading comes to the scene.

There is a pervasive computing technique called Cyber Foraging, which allows resource-limited mobile devices to offload their heavy computational tasks to powerful devices in the vicinity to boost mobile devices' performance [3]. Thus, it has opened the door for resource-limited mobile devices to use unused resources in resource-rich devices.

Offload machine learning-based tasks every time is not the preferred approach due to many reasons; sometimes the smartphone may have enough resources to execute the task, current network conditions might take lots of time to send the data to the resourceful device and get the result back. Due to these reasons, there should be a mechanism that decides when and what to offload based on the environment. This paper introduces a decision making engine for task offloading in on-device AI based mobile applications to resourceful personal devices such as laptops.

Further, the major contributions of this paper can be summarized as follows.

- The proposed Decision Making Engine is a novel approach designed for on-device inference based Android mobile applications. The proposed approach minimizes the execution time of the machine learning tasks in the application, considering the Round Trip Time (RTT), historical data such as previous execution times for each task, the current situation of the smartphone as well as the connected resource-rich device. The decision making engine decides when to offload based on the given parameters.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

- Unlike decision making algorithms based on pre-trained machine learning models, the decision making algorithm is based on both real time data and historical data gathered from the application itself. Even though the algorithm is same for any offload enabled application that runs on any smartphone, it will not executes in the same way for two different applications in the same device nor the same application in two different devices. This makes the developer to worry less about the compatibility of the application in smartphones in various range since the decision making algorithm acts in own way in all these devices.
- The decision making engine has an automated offloading process terminating functionality to prevent the halting problem.
- The decision making engine resides in the mobile application, that stating all the data used for the decision making engine remain within the application, which ensures the privacy of the user's application usage.

The remainder of this paper can be structured as follows. Section II describes the recent related work on the decision making process on mobile devices. Section III proposes the decision making engine for on-device AI based task offloading. Subsequently, Section IV evaluates the proposed approach. Discussion, conclusions, and future work are discussed in Section V and VI, respectively.

II. Related Work

Existing work on offloading decision making can be divided into 3 components based on the used approach; Reinforcement learning-based decision making approaches, traditional approaches that use algorithms such as the Lyapunov function, and deep learning-based approaches.

A. Reinforcement Learning Based Decision Making

Markov Decision Process (MDP) is used for modeling and solving decision making situations under stochastic conditions [4]. Predicts of the Markov Decision Process are only based on the information provided by the current state. Yang et al. [5] used MDP model to find an optimal place to offload. Notable factor in this research is it shows how the available network bandwidth of the edge servers and the mobile device location can be used in selecting the optimal place to offload.

B. Traditional Approaches Based Decision Making

Optimal Stopping Theory (OST) is concerned with the cases of choosing a time to take a particular action for the sake of maximizing the anticipated reward or minimizing the anticipated cost [6]. Alghamdi et al. [7] proposed an offloading decision-making process based on optimal stopping theory. The objective is to minimize the execution delay.

Lyapunov function, which is a scalar function, is used to prove the stability of an equilibrium point [8]. Using the Lyapunov function to control a dynamic system in an ideal way can be defined as Lyapunov optimization. Wu et al. [9] proposed a Lyapunov optimization based energy-efficient decision algorithm to minimize energy consumption while satisfying a delay constraint.

C. Deep Learning-Based Decision Making

Ali et al. [10] proposed a deep learning-based computation offloading decision-making approach for mobile edge computing. To train the DNN model, they created a dataset that is generated using a mathematical model. Yu et al. [11] proposed a Deep Supervised Learning (DSL) method for offloading decision-making processes. The decision is based on the local execution overheads and the conditions in the network side.

Learning with trial and error by being rewarded or penalized is known as Reinforcement Learning [12]. Deep-Q-Network is one of the most popular algorithms in Deep Reinforcement Learning. Van Le and Tham [13] proposed a Deep Reinforcement Learning approach for computation offload decision-making approach. It is noteworthy that the decision-making process is based on the unreliability of both users' and cloudlet's movement and the resource availability in the cloudlet. Park et al. [14] proposed a deep reinforcement learning-based (Deep-Q-Network) framework for real time offloading to make decisions related to offloading. Results shows that the proposed algorithm was able minimize energy consumption and while providing low latency.

Yu et al. [15] proposed a Deep Imitation Learning-based offloading model which could minimize the offloading cost. The decision is based on the local execution cost and remote network resource usage consideration.

It is identified that even though there are various researches to identify the best decision for executing computational tasks in mobile devices, there is a lack of research focused on task offloading in on-device inference based mobile applications and finding solutions for the halting problem.

Each approach discussed in the literature review has its advantages and disadvantages and they are shown in table I.

TABLE I. Comparison of the Approaches

Approach	Advantages	Disadvantages
Reinforcement learning Algorithms	Does not require a prior knowledge	High Complexity Curse of dimensionality
Traditional Algorithms	Ideal for slowly changing/ stable envirenments	High Complexity Hard to implement on unknown environments
Deep Learning	Can be used in dynamic environments	Long training time Requires a lot of data

Traditional algorithms and reinforcement learning algorithms are ideal approaches if the environment is known. Due to the high complexity of the traditional algorithms and reinforcement learning algorithms, it might be hard to implement these in a practical environment. When it comes to deep learning based approaches, it requires a lot of data to learn. Further, since smartphones have a wider performance and resource range, the same trained machine learning model might not be successful for all the devices. It would be an added advantage if the decision making algorithm can use the smartphone's data itself. Thus, the error that occurred by the data will be reduced. It is identified that most of the previous work does not consider the devices' own historical data such as execution times. As stated in the previous publication, historical data can be used to prevent the halting problem in computation task offloading. Thus, it was decided to investigate how historical data can be used in machine learning related tasks offloading to prevent the halting problem and further how it can be used in the decision making process. It is identified that there is a possibility of developing a decision making algorithm by using a decision tree algorithm for this purpose. Since decision trees are less complex compared to traditional algorithms mentioned in the literature review and there is a possibility that they would fit for the practical environment as well.

Therefore, in this project, the decision tree algorithm is chosen, and the most important parameters defined in the literature review are chosen. Therefore, parameters such as Round Trip Time (RTT), surrogate device information such as CPU usage and memory usage, and mobile device usage are selected to use in the decision making system.

In order to secure the privacy of the user's and device's interactions which are used in the decision making engine, it was decided to keep the decision making engine within the application. Hence, no data movement is done. Further, personal devices such as laptops were used as resource-rich devices, thus the data is only being transferred to personal devices which minimizes privacy concerns. Since the discussed approaches such as deep learning models, reinforcement learning, and complex algorithms might take considerable energy and time cost to execute, the research was focused on decision tree algorithms.

III. Methodology

A high-level diagram of the proposed system is shown in Fig. 1. Initially, the smartphone and the resource-rich device (hereinafter called the surrogate device) are connected through Wi-Fi. In the background, the surrogate device shares some information related to the surrogate device such as its Central Processing Unit (CPU) usage, memory usage with the mobile application. Further, the mobile application periodically sends data packets and waits for the surrogate device to send back data packets to the mobile phone in order to measure Round Trip Time (RTT). Round Trip Time is used for the expected remote execution time calculations. When there is a task to be executed, the decision making engine will decide whether the task needs to be offloaded or executed locally depend on various parameters.

Architecture of the proposed solution is shown in Fig. 2. One of the key characteristics of the proposed decision making engine is that it resides in the mobile application. Thus, there is no data movement in the decision making engine.

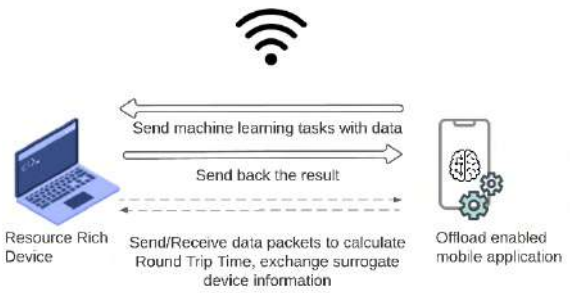

Send machine learning tasks with data
Send back the result
Resource Rich Device
Send/Receive data packets to calculate Round Trip Time, exchange surrogate device information
Offload enabled mobile application

Fig. 1 – High level diagram of the on-device inference machine learning tasks offloading system

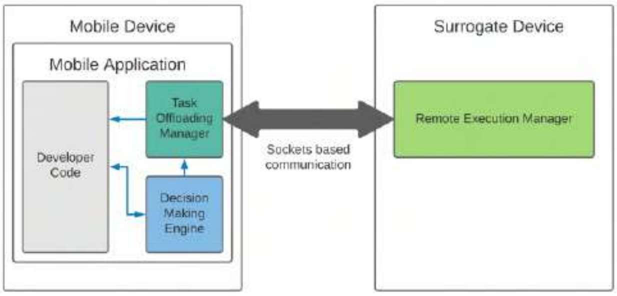


Fig. 2 - High level architecture of the offloading process

The system can be mainly divided into 5 parts: (1) Decision Maker, (2) Network Profiler, (3) Surrogate Profiler, (4) Device Profiler and (5) Database with historical data.

A high-level architecture of the decision making system is shown in Fig. 3.

A. Network Profiler

The Network profiler is used to monitor the network. It provides Round Trip Time (RTT) to the decision maker periodically. After the mobile application is connected to the surrogate device, the mobile application sends data packets to the surrogate device via the network. After the data packets are received, the surrogate device sends data packets back to the smartphone. This process is a recursive process that executes periodically. The decision making engine will measure the time spent on the above process.

B. Surrogate Profiler

The Surrogate profiler provides information related to the surrogate device to the decision maker. Occasionally the profiler gathers data from the surrogate device. It collects CPU usage, memory usage, available power, and battery status (plugged in or not connected to a power source).

C. Device Profiler

Information related to the current memory usage of the smartphone is provided by the device profiler. It provides current memory usage and whether the mobile device is currently in a low memory situation.

D. Database with Historical data

The database consists of historical data. It contains local execution times, average execution times for both local and remote executions, execution time which the remote execution takes with the task ID. Each machine learning task in the

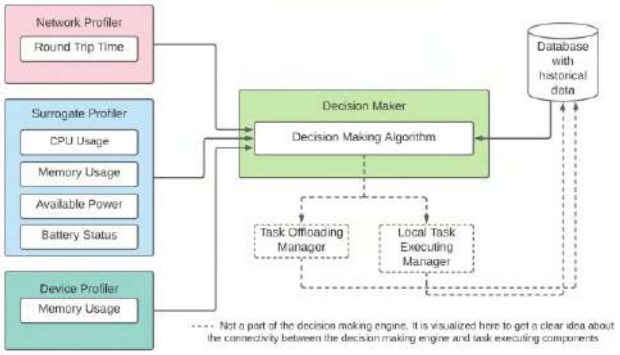

Network Profiler
Round Trip Time
Surrogate Profiler
CPU Usage
Memory Usage
Available Power
Battery Status
Device Profiler
Memory Usage
Decision Maker
Decision Making Algorithm
Database with historical data
Task Offloading Manager
Local Task Executing Manager
- - - - Not a part of the decision making engine. It is visualized here to get a clear idea about the connectivity between the decision making engine and task executing components

Fig. 3 - High level architecture of the decision making engine

mobile application has a unique ID. When the task is executed, the average executed time will send to the decision maker, and update the necessary record in the database with the execution times and calculate the new average execution time once the task is completed. There might be outliers in the execution times due to reasons such as network interruptions. These outliers will be identified before the database insertions and these data will be ignored.

E. Decision Maker

The algorithm used in the decision maker is a decision tree algorithm that interacts with network profiler, surrogate profiler, device profiler, and database with historical data every time.

Historical data plays a huge role in this algorithm. Not only it is used to decide whether the task needs to offload or not but also it is used to prevent the halting problem. Each time when the task is executed locally or remotely the average execution time (local and remote execution have separate average values) related to the task will be used in the decision making process and halting problem preventing process.

Since the execution happens in another device, the smartphone does not know whether the task runs forever (runs in an infinite loop) or it will halt. In 1936, Alan Turing proved that there is no general procedure to solve the halting problem [16]. Therefore, a mechanism to prevent halting issues has been implemented in this decision maker using historical data. When a task is executing remotely, the decision maker terminates the offloading session when the time consumed by the offloading task reaches the calculated remote execution delay for that task and then the decision maker executes the task locally. Formula to calculate remote execution delay for a task (hereinafter is referred to as T_r) is given (1). It is calculated using the Round Trip Time (T_{rtt}) and the average execution time which the surrogate device takes (T_{set}), which is fetched from the database.

$$T_r = T_{rtt} + T_{set} \tag{1}$$

The decision maker also checks the current memory usage of the smartphone. If the smartphone is in a low memory situation, the decision maker decides not to execute the task locally and offload it to the surrogate device.

The decision maker not only considers the smartphone's resource availability but also cares about the surrogate device's resource availability as well. When there is a task to be executed, the decision maker checks whether the surrogate device's CPU usage and memory usage are high, and the surrogate device is currently connected to a power source or it is running using its battery. If yes, it checks whether the battery is in its lower battery level. The decision maker decides whether the task needs to be offloaded or executed locally based on these parameters. How the decision maker uses these parameters in decision making is given in the full algorithm which is given in Algorithm 1.

TABLE II. NOTATIONS

Notations	Meaning
T_r	Remote execution delay for a task
T_{rtt}	Round Trip Time (RTT)
T_l	Estimated time spend for local execution
M_{low}	The mobile device is in low memory situation
S_{cpu}	CPU usage in the surrogate device
S_{memory}	Memory usage in the surrogate device
S_{power}	Power status in the surrogate device (whether the device is connected to a power source or in the battery)
$S_{battery}$	Battery percentage in the surrogate device
H_l	Historical data related to local execution
H_r	Historical data related to remote execution

Algorithm 1 Decision Maker

```
Input: $T_l$, $T_r$, $M_{low}$, $S_{cpu}$, $S_{memory}$, $S_{power}$, $S_{battery}$
Output: Decision of whether the task needs to be offloaded or executed locally
IF $M_{low}$ = True THEN
  Return 'REMOTE_EXECUTION'
ELSE
  IF $H_l$ not found THEN
    Return 'LOCAL_EXECUTION'
  ELSE IF $H_r$ not found THEN
    IF $S_{power}$ is in battery AND $S_{battery}$ is low THEN
      Return 'LOCAL_EXECUTION'
    ELSE IF $S_{cpu}$ is not in high usage AND $S_{memory}$ is not in high usage THEN
      Return 'REMOTE_EXECUTION'
    ELSE
      Return 'LOCAL_EXECUTION'
  ELSE
    IF $T_l < T_r$. or $T_l = T_r$ THEN
      Return 'LOCAL_EXECUTION'
    ELSE
      IF $S_{power}$ is in battery AND $S_{battery}$ is low THEN
        Return 'LOCAL_EXECUTION'
      ELSE IF $S_{cpu}$ is not in high usage AND $S_{memory}$ is not in high usage THEN
        Return 'REMOTE_EXECUTION'
      ELSE
        Return 'LOCAL_EXECUTION'
```

After checking all the parameters described above, the decision maker finally compares the estimated execution times in both local (T_l) and remote (T_r) execution for the given task. Average local execution time for the task is used as estimated local execution time (T_l). The task will be offloaded to the surrogate device for the remote execution if $T_l > T_r$. The task will be executed locally if $T_l < T_r$ or $T_l = T_r$.

There is a possibility that errors could occur when the remote execution happens due to network interruptions, issues from the surrogate device, and so on. Thus, apart from the above algorithm, it is handled by executing local execution when there is an error.

IV. Experiment Design

Since the execution time for a machine learning task gets longer when the smartphone is struggling with managing its resources due to limited resources, as stated in section I, the objective of this decision making engine is to minimize execution delay. The decision making engine will check which option either offload or local execution might take the minimum execution time based on the parameters.

The connection between the smartphone and the surrogate device is done using TCP (Transmission Control Protocol) socket connection and the connection is established within the Local Area Network (LAN) using Wi-Fi.

Application for the surrogate device is developed using Python to manage task execution and send data to the surrogate profiler. Psutil [17], which is a Python library for process and system monitoring is used to get surrogate system-related information.

Various Android libraries such as ActivityManager [18] are used to get the necessary information for the decision maker. TensorFlow Lite [19] is used to developing and executing the on-device inference machine learning model (tflite file).

A. *Evaluation*

To conduct the testing and evaluation, an on-device inference based Android application was developed and integrated offload and decision making functionalities. The application is developed using Android Studio and TensorFlow Lite.

Initially, testing was conducted to check the algorithm logic, followed by the performance testing. Performance testing was conducted in both a controlled environment and a realistic environment.

Since the objective of the decision making engine is to minimize execution delay, the accuracy of the decision making engine is based on the execution time of both tasks offloading and local execution. For the testing purpose, the machine learning task is executed both locally and remotely despite the decision makes by the decision making engine. It was done to measure the task execution time on both remote execution and local execution at that time. The decisions made by the decision making engine is compared with the time spend on offloading process and the time spent on local execution. If the decision of the decision making engine is the approach that took the lowest execution time, that decision is considered as a correct decision. The accuracy was calculated as the following formula given in (2).

$$Accuracy = \frac{\underline{Number\ of\ the\ correct\ decisions}}{Number\ of\ the\ total\ decisions} \qquad (2)$$

The evaluation was accomplished by conducting a self-evaluation, getting feedback from the related expertise. Expertise feedback was collected based on a criterion of the depth of the research and the impression of the project, architectural design, and the prototype implementation.

V. Results and Discussion

The proposed offloading decision making engine was tested under various situations. Initially, algorithm logic was tested, by conducting 192 test cases. The test was conducted to make sure that the decision making algorithm gives the correct output under the given conditions. It was able to achieve 100% accuracy for the logical testing. After the logical testing was conducted, performance testing was performed. For the testing, Nokia 1 Plus [20], Nokia 4.2 [21], and Nokia 8 [22] smartphones were used. The test was conducted both in a controlled environment and a realistic environment.

The results are summarized in Table III. Based on the results, the decision making engines were able to achieve an average accuracy of 92% in the controlled environment. In the realistic environment, the average accuracy was reduced to 81.33%. It is noted that the decision largely depends on the historical data. It is identified that the accuracy of the decision making engine is keeping high as long as the network is stable. The decision making engine gives wrong decisions when there are sudden unexpected interruptions.

TABLE III. ACCURACY

Smartphone	Environment	Accuracy
Nokia 1 Plus	Realistic Environment	76%
Nokia 1 Plus	Controlled Environment	84%
Nokia 4.2	Realistic Environment	72%
Nokia 4.2	Controlled Environment	96%
Nokia 8	Realistic Environment	96%
Nokia 8	Controlled Environment	96%

VI. Conclusion and Future Work

Most of the research analyzed did not consider much about the role of the device's historical data in decision making. In this project, the past data of execution times were taken into account in the decision making process. Thus, the decision making engine is personalized to that device up to a certain amount. The results show that despite using machine learning models and complex algorithms, high accuracy can be also achieved from a decision tree algorithm based on historical data. The decision making engine was able to achieve 92% average accuracy in the controlled environment while maintaining an average accuracy of 81.33% in the realistic environment. The limitation of this approach is that the accuracy of the decision making engine might get reduced if there is less or no historical data. The following points and the limitation mentioned above are considered as future enhancements.

- Advance this algorithm further to support multi-user environments.
- When decision making, network interruptions take into account.

This is a part of the ongoing research which is focused on offloading machine learning tasks in mobile devices to nearby devices. It was decided to integrate the decision making approach discussed in this paper into a fully-fledged on-device AI framework for machine learning tasks offloading for Android mobile applications.

Acknowledgment

Mr. Guhanathan Poravi, Mr. Sharmilan Somasundaram, and the faculty of the Informatics Institute of Technology are

being appreciated for the motivation, expertise, and insights provided continuously.

REFERENCES

[1] "Newsroom - Gartner Highlights 10 Uses for AI-Powered Smartphones," *Gartner*, Jan. 04, 2018. https://www.gartner.com/en/newsroom/press-releases/2018-03-20-gartner-highlights-10-uses-for-ai-powered-smartphones (accessed Sep. 20, 2020).

[2] T. Guo, "Cloud-based or On-device: An Empirical Study of Mobile Deep Inference," *ArXiv170704610 Cs*, Apr. 2018, Accessed: Oct. 07, 2020. [Online]. Available: http://arxiv.org/abs/1707.04610.

[3] R. Balan, J. Flinn, M. Satyanarayanan, S. Sinnamohideen, and H.-I. Yang, "The case for cyber foraging," in *Proceedings of the 10th workshop on ACM SIGOPS European workshop: beyond the PC - EW10*, Saint-Emilion, France, 2002, p. 87, doi: 10.1145/1133373.1133390.

[4] Q. Hu and W. Yue, Markov decision processes with their applications. New York: Springer, 2008.

[5] G. Yang, L. Hou, X. He, D. He, S. Chan, and M. Guizani, "Offloading Time Optimization via Markov Decision Process in Mobile Edge Computing," *IEEE Internet Things J.*, pp. 1–1, 2020, doi: 10.1109/JIOT.2020.3033285.

[6] T. S. Ferguson, Optimal Stopping and Applications. UCLA 2008.

[7] I. Alghamdi, C. Anagnostopoulos, and D. P. Pezaros, "Time-Optimized Task Offloading Decision Making in Mobile Edge Computing," in *2019 Wireless Days (WD)*, Manchester, United Kingdom, Apr. 2019, pp. 1–8, doi: 10.1109/WD.2019.8734210.

[8] "Method of Lyapunov Functions," Math24. https://www.math24.net/method-lyapunov-functions (accessed Dec. 12, 2020).

[9] H. Wu, Y. Sun, and K. Wolter, "Energy-Efficient Decision Making for Mobile Cloud Offloading," *IEEE Trans. Cloud Comput.*, vol. 8, no. 2, pp. 570–584, Apr. 2020, doi: 10.1109/TCC.2018.2789446.

[10] Z. Ali, L. Jiao, T. Baker, G. Abbas, Z. H. Abbas, and S. Khaf, "A Deep Learning Approach for Energy Efficient Computational Offloading in Mobile Edge Computing," *IEEE Access*, vol. 7, pp. 149623–149633, 2019, doi: 10.1109/ACCESS.2019.2947053.

[11] S. Yu, X. Wang, and R. Langar, "Computation offloading for mobile edge computing: A deep learning approach," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Montreal, QC, Oct. 2017, pp. 1–6, doi: 10.1109/PIMRC.2017.8292514.

[12] "Deep Reinforcement Learning | DeepMind." https://deepmind.com/blog/article/deep-reinforcement-learning (accessed Apr. 05, 2021).

[13] D. Van Le and C.-K. Tham, "A deep reinforcement learning based offloading scheme in ad-hoc mobile clouds," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Honolulu, HI, Apr. 2018, pp. 760–765, doi: 10.1109/INFCOMW.2018.8406881.

[14] S. Park, D. Kwon, J. Kim, Y. K. Lee, and S. Cho, "Adaptive Real-Time Offloading Decision-Making for Mobile Edges: Deep Reinforcement Learning Framework and Simulation Results," *Appl. Sci.*, vol. 10, no. 5, p. 1663, Mar. 2020, doi: 10.3390/app10051663.

[15] S. Yu, X. Chen, L. Yang, D. Wu, M. Bennis, and J. Zhang, "Intelligent Edge: Leveraging Deep Imitation Learning for Mobile Edge Computation Offloading," *IEEE Wirel. Commun.*, vol. 27, no. 1, pp. 92–99, Feb. 2020, doi: 10.1109/MWC.001.1900232.

[16] "Why is Turing's halting problem unsolvable? - Scientific American." https://www.scientificamerican.com/article/why-is-turings-halting-pr/ (accessed Dec. 14, 2020).

[17] G. Rodola, psutil: Cross-platform lib for process and system monitoring in Python.

[18] "ActivityManager," *Android Developers*. https://developer.android.com/reference/android/app/ActivityManager (accessed Feb. 22, 2021).

[19] "TensorFlow Lite | ML for Mobile and Edge Devices," *TensorFlow*. https://www.tensorflow.org/lite (accessed Oct. 06, 2020).

[20] "Nokia 1 Plus - Full phone specifications." https://www.gsmarena.com/nokia_1_plus-9538.php (accessed Mar. 31, 2021).

[21] "Nokia 4.2 - Full phone specifications." https://www.gsmarena.com/nokia_4_2-9603.php (accessed Mar. 31, 2021).

[22] "Nokia 8 - Full phone specifications." https://www.gsmarena.com/nokia_8-8522.php (accessed Apr. 04, 2021).

Scattering-based Quality Measures

Rasha Kashef
Electrical, Computer, and Biomedical Engineering Department
Ryerson University
Toronto, Canada
rkashef@ryerson.ca

Abstract- Various clustering algorithms use diverse settings, parameters, and initializations, generally result in different clustering solutions. Therefore, it is essential to compare and evaluate the clustering results and select the methods that best fits the "actual" data distribution. This can be achieved by using informative quality metrics that reflect the "goodness" of the resulting solutions compared to the ground truth. Different Extrinsic validation metrics have been provided in the literature, including F-measure, Entropy, Rand Index, and Purity. However, there is a gap in the literature in evaluating the level of divergence between multiple clusterings in an aggregate, especially in consensus clustering. In this paper, we propose three scattering measures that calculate the divergence level (i.e., scattering level) between two or more clustering algorithms. The proposed metrics are Scatter F-score, Scatter Entropy, and Scatter Purity. The proposed scattering measures are variants of the traditional F-measure, Entropy, and Purity quality measures. The scattering measures are used as pre-assessment criteria for deciding which clustering algorithms to combine in an aggregate. Experimental results on artificial, real, and text datasets show that the scattering measures play an important role in enhancing the clustering quality in consensus clustering and increasing the feasibility of the consensus.

***Index Terms*—Machine Learning, Clustering, Classification, Validation Measures, Divergence, Internal and External metrics**.

I. INTRODUCTION

Machine learning has received great attraction in many applications, including gene expression analysis [1][2], segmentation [3]-[5], prediction and time-series analytics [6]-[11], distributed computing [12]-[15], outlier detection [16]-[18], recommendation systems [19]-[24], text mining [25][26], and beyond. Machine learning can be classified into unsupervised and supervised methods [27][28]. In unsupervised learning, the training of the algorithm does not depend on any external knowledge, i.e., it does not use any external labeling for building the model [29]. However, in the supervised approach, to build a prediction, classification, or regression model, the data must be labeled to develop the required model. At some point during the learning process, the solutions provided must be assessed using validations metrics. There are many validation metrics in the literature used to determine the quality of the machine learning method's solutions. In clustering analysis, the validation process estimates how well a partition fits the underlying structure of the data. Cluster validation measures can be categorized into two classes, external clustering validation and internal clustering validation [30]. External measures evaluate the result based on some supervised information available, while internal ones only evaluate the result based on the information intrinsic to the raw data. F-measure, NMI measure, Entropy, and Purity are popular external clustering measures [31]. A good validation should be invariant to data size changes, cluster size, and the number of clusters [32].

The traditional external quality metrics measure the disparity between the original labeling of the dataset (i.e., class labels) and the resulting clusterings. In consensus clustering [4][18][26], two or more clustering algorithms are combined to provide better clustering solutions than individual-based methods. However, there is a research gap in assessing the quality of the aggregate before building the consensus. Various work has been done by trial and error without a solid evaluation of the feasibility or the quality of the combined collection of the clustering algorithms. This pre-assessment stage before building the consensus is mandatory to reduce the computational time and increase the chance of obtaining a high-quality aggregate with the proper set of combined algorithms. In this paper, we propose the scattering-based validation metrics that measure the diversity of the clustering solutions obtained from two or more clustering algorithms. These measures can be easily applied to the aggregate before actually building the consensus to evaluate the feasibility of the aggregate. The scattering-based validation measures are considered variants of the original extrinsic measures in which we consider the ground truth as the opposed clustering method. We proposed the scatter F-score (SF), scatter-Entropy (SE), and scatter-Purity (SP) evaluation measures. The effectiveness of using the scattering measured is shown through the application on six datasets with various levels of overlapping and sparsity. Two single-based clustering algorithms and one consensus clustering is used in the experiments. It has been shown that using the proposed metrics assesses in determining which

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

clustering aggregates should be combined to obtain a successful ensemble or consensus.

The rest of the paper is organized as follows: In section 2, extrinsic quality measures are introduced. Section 3 discussed the adopted clustering algorithms. In section 4, the proposed scattering measures are presented. Experimental analysis and results are shown in Section5. Finally, the conclusion and future directions are given in Section 6.

II. EXTRINSIC QUALITY MEASURES

The external clustering measures are based on external criteria. Such that these measures are used when the real partition of the clustered data is given a priori. A cluster validity has two features: (1) the quality of clusters using homogeneity and separation, and (2) the second feature relies on a given "ground truth. The "ground truth" is obtained from domain knowledge or other knowledge repositories. Thus, the evaluation depends on prior knowledge, i.e., class labels. This labeling compares the resulting clusters with the original labels; these measures are known as extrinsic metrics.

A. F-measure-based Measures

F-measure is based on pairing similar clusters in two clusterings solutions [31]. It relies on a pre-processing step such that each cluster is mapped to a class [33]. F-measure is broadly used in text mining based on precision and recall concepts [34]-[36].

$$\mathrm{R}(Si, Lj) = \frac{n_{ij}}{n_i} \quad (1)$$

$$\mathrm{P}(Si, Lj) = \frac{n_{ij}}{n_j} \quad (2)$$

where n_{ij} is the number of instances of class Si that are mapped to cluster j, n_j is the number of instances in cluster Sj, and n_i is the number of instances in class Li. For a cluster Si and class Lj, the F-measure is:

$$\mathrm{F(Si, Lj)} = \frac{2*R(i,j)P(i,j)}{P(i,j)+R(i,j)} \, (3)$$

The total F-Measure is formulated as the weighted sum of the *max* values of the F-Measures for each cluster.

$$F = \frac{1}{N}\sum_{i=1}^{k} n_i \max_j \frac{2n_{ij}}{n_i+n_j} (4)$$

F-measure cannot be applied for cases where nested clustering is presented [36], and it cannot handle the problem of class size imbalance properly [34]. However, F-measure is appropriate for partitional clustering since it tends to split a large and pure cluster into many smaller disjoint partitions. The **hF-measure** is commonly used for nested clustering [36]. The **L-measure** is a derivative of the F-measure to compute the quality of a computational lexicon using some clustering criterion. The L-measure achieved significant results when applied to small-sized data [37].

B. Homogeneity-based Measures

Entropy provides a solution to the "matching problem by measuring the homogeneity of a solution [33]. The entropy of a cluster is computed as:

$$E_j = \sum_i p_i \log{(p_i)} \;\; (5)$$

Pi is the percentage of instances mapped from the clustering to the defined initial classes. The overall entropy is calculated as the weighted sum of the cluster-wise entropies.

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \; (6)$$

V-measure (values lie in [0, 1]) is an entropy-based measure that measures the relative satisfaction of the criteria of homogeneity and completeness. It depends on the number of clusters, the size of the data set, and the adopted clustering algorithm [33]. The ***V-measure*** favors solutions with many clusters, and it strongly favors a clustering having many small clusters [38]. The V-measure is computed as the harmonic mean of homogeneity h and completeness c of the clustering with a tuning parameter β [33].

$$v_\beta = (1+\beta)\frac{h*c}{\beta*h+c} \, (7)$$

where h stands for homogeneity, and c stands for completeness.

VI is the variation of information with two properties: it satisfies the metric axioms, and it is convexly additive [39]. The range of scores given by VI depends on the size of the dataset [38]. The possible values of VI lie in [0, 2log *n*], where *n* is the size of the dataset. The VI has limited use when evaluating the performance on different datasets. The VI is defined as:

$$\mathrm{VI(C, K)} = \mathrm{H(C|K)} + \mathrm{H(K|C)}(8)$$

where H(C|K) is the conditional entropy of the class distribution C given the proposed clustering K, and H(K|C) is the conditional entropy of the clustering distribution given the proposed class [38].

Normalized Variation of Information (NVI) lies in [0, 1], and it keeps the convex additivity property of VI but not its metric axioms. Perfect solution means that the NVI = 0. NVI achieves better performance on a highly non-trivial NLP application with large datasets[38][39].

Purity measures the homogeneity of a solution such that a clustering solution with purity values close to 0 is considered a poor solution [31][33]. The difference between Purity, normalized Van Dongen (NVD) [40], Criterion H (CH)[41], and Centroid Similarity Index (CSI) [42], is related to their matching. If the matching result is the same, these indexes will provide equivalent results [31]. The purity of a set of clusters is calculated as below.

$$Purity = \sum_{r=1}^{k} \frac{1}{n} max_i(n_r^i) \quad (9)$$

n_r^i is the points in class *i* that are grouped in cluster *r*.

The **Normalized Mutual Information (NMI)** quantifies the shared information between the clustering and the true partition. The NMI values are in the range [0, 1] such that a value close to 1 indicates a decent clustering [32].

$$NMI(K,C) = \frac{I(K,C)}{\sqrt{H(K)H(C)}} \quad (10)$$

Where I(K,C) denotes the mutual information between the consensus clustering K and the true class C. H(K) denotes the entropy of K.

III. ADOPTED CLUSTERING ALGORITHMS

Clustering algorithms can be classified into two main categories, partitional and hierarchical methods. In partitional clustering, the clusters are obtained as partitions (crisp or overlapping), while in hierarchal clustering, the clustering solutions are obtained in the form of a hierarchy called dendrograms. In this paper, we will adopt two clustering algorithms and their variants, the k-means (KM) and fuzzy c-means clustering (FCM) [43]. We have also used the Cooperative Clustering (CC) clustering algorithm [4] as a consensus clustering to show the proposed scattering measures' efficiency in enhancing the clustering quality of a given aggregate.

A. The Crisp Clustering

The crisp clustering algorithm, k-means algorithm (KM) (Fig.1), chooses *k* cases randomly as initial cluster's centroids and then maps each instance x to the closest center c_i, i=1,2,..*k*. The centroids are then updated as the mean of the cases set assigned to each cluster. The objective function J defined in (Eq.11) is invoked as the stopping criterion of the iterative clustering process.

$$J = \sum_{j=1}^{k} \sum_{j=1}^{n} DisSimilaity\ (x_j, c_i) \quad (11)$$

DisSimilarity(x_j,c_i) is the distance between the point x_j and the centroids c_i. The Euclidian distance $||x_j\text{-}c_i||^2$ is used as a dissimilarity measure. A similarity measure explicitly used in document clustering is the cosine correlation measure (Eq.12). The J values can be updated as shown in Eq.13.

$$CosineSim(x,y) = \frac{x.y}{||x||\,||y||} \quad (12)$$

$$J = \sum_{j=1}^{k} \sum_{j=1}^{n} (1 - CosineSim\ (x_j, c_i)) \quad (13)$$

```
Algorithm: Crisp Clustering (KM)
Input: dataset X, and the number of clusters k.
Output: Set of k clusters S
S= {}, Randomly, select k initial centroids c_i
Begin
 Repeat
  Step1 (Assignment Stage): instances are mapped to the closest centroid,
  the objective function J (Eq.11) is calculated.
  Step2 (Revising Step): new centers c_i are computed as the average of the instances in each new cluster S_i
  Step3: Modify the set S with the new clusters S_i
 Until a stopping criterion is satisfied
Return S
```

Fig.1. The Crisp Clustering Algorithm

B. The Fuzzy Clustering

In the fuzzy c-means (FCM) algorithm (Fig.2), each point is assigned multiple membership values $u_{ij} \in [0,1]$. Initially, random memberships are designated for data point $\boldsymbol{x}_j$ such that $\sum_{i=1}^{k} u_{ij} \forall\, j = 1,..,n$, u_{ij} represents the membership of object $\boldsymbol{x}_j$ to cluster S_i, *i*=1,…,*k* and *j*=1,…,*n*.. For text datasets, we use a similarity function (e.g., *Cosine Similarity*), the cluster centers and membership formulas are modified as:

$$c_i = \sum_{j=1}^{n} u_{ij}^m * x_j \left[\sum_{\propto=1}^{d}\left[\sum_{j=1}^{n} u_{ij}^m * x_{i\propto}\right]\right]^{-1/2} (14)$$

$$u_{ij} = \left[\sum_{r=1}^{k}\left[\frac{DisSimilaity\ (x_j, c_i)}{DisSimilaity\ (x_i, c_r)}\right]^{1/(m-1)}\right]^{-1}$$

$$u_{ij} = \left[\sum_{r=1}^{k}\left[\frac{CosineSim\ (x_j,c_i)}{CosineSim\ (x_i,c_r)}\right]^{1/(m-1)}\right]^{-1} (15)$$

The objective function *J* (Eq. 16) is also used as a convergence criterion to be minimized, where *m* is the weighting exponent.

$$J = \sum_{j=1}^{k} \sum_{j=1}^{n} u_{ij}^m (1 - CosineSim\ (x_j, c_i)\ ) \ (16)$$

```
Algorithm: (FCM)
Input: Dataset X, weighting exponent $m$, and number of clusters $k$.
Output: Set of $k$ clusters S
S={}, randomly initialize the membership matrix $u_{ij}$.
Begin
  Repeat
    Step1: Compute $k$ centers using Eq.14.
    Step2: Calculate object's membership to the $k$ clusters as in Eq.15.
    Step3: Evaluate the objective function $J$ using Eq.16.
    Step4: update the set S with the newly obtained clusters
  Until a stopping criterion is satisfied
Return S
```

Fig. 2. The Fuzzy Clustering Algorithm

C. Cooperative Clustering (CC)

Combining clusterings invokes an aggregate of multiple clustering algorithms in the clustering process. Cooperative Clustering (CC) (Fig.3) enables concurrent execution of multiple clustering algorithms to achieve better performance synchronously. The cooperative clustering model is established in four stages, mainly (1) executing individual clusterings, 2) obtaining the set of co-occurred sub-clusters, (3) building sub-clusters histogram representation of the pair-wise similarities, and (4) merging of sub-clusters until *k* clusters are obtained.

```
Algorithm: Cooperative Clustering (CC)
Input: Dataset X, $A_i$ algorithms, $i$=1,..,c, and number of clusters $k$.
Output: Set of $k$ clusters S={$S_1$,$S_2$,..,$S_k$}
Initialization: S={}.
Begin
    Step 1: Generate $c$ clusterings each of size k
    Step 2: Find the set of subclusters Sb
    Step 3: build similarity histograms and assign $S=Sb$
Step 4: Repeat
    Find the most homogenous two clusters in Sb, A, and B.
    Merge A and B into C
    Remove A and B from the set S
    Add the cluster C
  Until the number of clusters in the set S equals k
Return S
```

Fig. 2. The CC Algorithm

IV. THE PROPOSED SCATTERING-BASED EVALUATION METRICS

In the proposed measures, assume we have two clustering algorithms A and B, with k clusters $S^A(k)$={ S_i^A, $1\leq i \leq k$}, and $S_B(k)$={ S_j^B, $1\leq j\leq k$}, respectively. Assume $|S_i^A|$ is the number of instances in cluster S_i (obtained by A), and $|S_J^B|$ is the number of instances in cluster S_j (obtained by B).

A. The Scattering F-score (SF)

The scatter F-score of a cluster S_i^A is defined as:

$$scatter\ Fscore(S_i^A) = max\ (\frac{2*n_{ij})}{|S_i^A|+|S_j^B|}\ (17)$$

Where n_{ij} is the number of instances in the cluster S_i^A that co-occurred in the cluster S_j^B, with respect to cluster S_i^A, the cluster with the maximum F-score value is considered the cluster S_j^B that is mapped to $S_i^{A.}$ That value is considered as the score for cluster S_i^A. The overall scattering F-score for the clustering result of k clusters is the weighted average of the F-score for each cluster S_i^A.

$$SF = \frac{\sum_{i=1}^{k}\left|S_i^A\right|*F-score(S_i^A)}{\sum_{i=1}^{k}|S_i^A|}\ (18)$$

The higher values of the overall Scatter F-score indicate a closer solution both A and B produce. This is mainly due to the greater accuracy of the resulting clusters of A mapping to B's clusters. In a consensus clustering model, we seek lower values of the scatter F-score between the adopted clustering approaches to obtain significant improvement in the clustering performance, otherwise combing two similarity clustering algorithms will not generate any significantly better results than the two methods.

B. The Scattering Entropy (SE)

The Scattering Entropy provides a measure of "homogeneity". The SE metric tells shows how homogenous two clustering solutions are. Lower values for the SE measure indicate that the two clustering solutions are of high similarity (low divergence) and vice versa. The SE of two exact clustering solutions (perfect homogeneity) is zero. Assume two partitioning results of clustering algorithms A and B consisting of *k* clusters. For a cluster S_i^A , we calculate p_{ij}, the probability that a member of cluster S_i^A is mapped to class S_j^B. The scatter entropy of each cluster S_i^A is calculated using Eq.19, where we compute the sum over all clusters in S^B:

$$\mathrm{SE}(S_i^A) = \sum_j p_{ij}\log(p_{ij})\ \ (19)$$

The total SE for S^A and S^B is formulated as the weighted sum of the cluster-wise scattering entropies.

$$SE = \sum_{i=1}^{k}\frac{|(S_i^A)|}{n}\mathrm{E}(S_i^A)\ (20)$$

The overall weighted scattering entropy avoids supporting smaller clusters compared to larger clusters. Typically, the

SE measure reports a complete matching if the cluster's Entropy is zero (i.e., the perfect similarity between A and B). If a cluster contains all the cases from two different clusters in B, its entropy will be zero, and the SE measure does not convey if a cluster is fully mapped to one or more classes, which is not the case in the SF.

C. The Scattering Purity (SP)

The Scattering Purity (SP) of solution A is calculated by averaging the cluster's precision wrt. their best matching clusters generated by B. For a single cluster S_i^A, the scatter purity is calculated as the ratio of the number of instances in the leading cluster to the total number of instances within the cluster:

$$SP(S_i^A) = \frac{1}{|S_i^A|} * \max\left({}^{i=1,..k}_{j=1,..,k} n_{ij}\right) \ (21)$$

n_{ij} is the number of instances from cluster S_i^A into cluster S_j^B, and $|S_i^A|$ is the number of instances in cluster S_i^A. To evaluate the total scattering purity for the entire *k* clusters, the cluster's purities are weighted by the cluster size, and the average value is calculated:

$$SP = \sum_{i=1}^{k} \frac{|S_i^A|}{n} SP(S_i^A) \ (22)$$

High values of the SP indicate closer solutions of A and B. The above formulation for the SF, SE, and SP can be generalized to any extrinsic quality measure with two clustering solutions labels S^A and S^B.

V. EXPERIMENTAL RESULTS

In this section, we used two single-based clustering methods as KM and FCM, and one aggregate-based (consensus) method, as the cooperative clustering (CC) method. The goal is to show that the scattering measures are used as an efficient assessment tool to decide which algorithms to combine (i.e., the feasibility of the aggregate). We evaluate the performance of each algorithm using the traditional F-measure, Entropy, and Purity.

A. Datasets

In this paper, we have used two artificial datasets, two real-world datasets, and two text documents datasets.

Synthetic Datasets: We have generated two artificial datasets, DataSet1 and DataSet2. DataSet1 contains two overlapped clusters in 3-dimensional feature space with the same covariance matrix. DataSet2 contains three overlapped clusters in 2-dimensional feature space with different covariance matrices. The number of data instances in DataSet1 and DataSet2 is 100,000 and 150,000 instances, respectively. DataSet1 and DataSet2 are generated from Gaussian distribution. Experiments were also performed where the synthetic datasets were randomized before feeding into the KM, BKM, and CC.

Real-World Datasets: The pageblocks dataset involves five classes of page layout blocks of documents that a segmentation process has detected. The dataset has 5473 instances with ten numeric are obtained from 54 documents. Such that each observation belongs to one block. The pageblocks dataset was used in both [45] and [46]. In the Imgseg dataset [47], 2310 instances with 19 numeric features were drawn randomly from a database of 7 outdoor images. The hand-segmented images create a classification for every pixel. Each instance is a 3x3 region.

Document Datasets: The UW dataset is a manually collected text corpus from the University of Waterloo[1] various websites. The data set also has many documents from other Canadian websites. The UW dataset has 314 documents, categorized into ten categories, with 469 average words per document [42]. The ReutersNews contains 2340 documents classified into 20 different categories, with an average of 289 words per document [29] [40]. A summary of the datasets is shown in Table 1.

Table 1: Summary of the Datasets

Dataset	***k***	***d***	***n***
DataSet1	2	3	100000
DataSet2	3	2	150000
pageblocks	5	10	5473
Imgseg	7	19	2310
UW	10	15133	314
ReutersNews	20	28298	2340

B. Experimental and Computational Analysis

As KM and FCM depend on the initial assignment of instances to clusters, we have used random variations of each method based on changing the initial conditions. As shown in Fig.3, the KM and FCM have similar performance for the DataSet1 and pageblocks datasets measures by the large values for the SF and SP and lower SP values (almost zero for the two datasets). On the other hand, the two algorithm shows great diversity in their solutions for the UW, ReutersNews, ImgSegm, and DataSet2. The diversity (i.e., scattering) level is measured by small SF and SP values and

[1] www.uwaterloo.ca

large SP values (as shown for the UW dataset as an example). The expectation is that when we combine the KM and FCM in an ensemble (i.e., consensus), there is a low chance of obtaining better solutions as they both have similar partitions, especially for the DataSet1 and pageblocks. As there is a higher level of divergence between KM and FCM in the UW, ReutersNews, ImgSegm, and DataSet2, there is an opportunity to find a better aggregate with better results than both algorithms. Thus, the scattering measures, SF, SE, and SP, indicate that KM and FCM can be combined for the four datasets, UW, ReutersNews, ImgSegm, and DataSet2. Simultaneously, it is not feasible to combine them for the DataSet1 and pageblocks as no better aggregate can be obtained.

Figures 4-5 show the quality of the single-based algorithms and the consensus CC algorithm. It can clearly be demonstrated that as KM and FCM have very similar performance for the DataSet1 and pageblocks, the percentage of improvement is low (only up to 2%, 10%, and 7% for the F-measure, Entropy, and Purity, respectively). On the other hand, due to the highly diverse solution between the two methods for the other four datasets, the consensus CC achieved an improvement of up to 63%, 34%, and 52% improvement in the F-measure, Entropy, and Purity, respectively. These results conclude that the proposed scattering measures are effective pre-assessment tools to decide on the feasibility of the ensemble prior to the design of an aggregate. These measures can be applied for supervised and unsupervised ensemble learning as a significant validation step in the preprocessing stags.

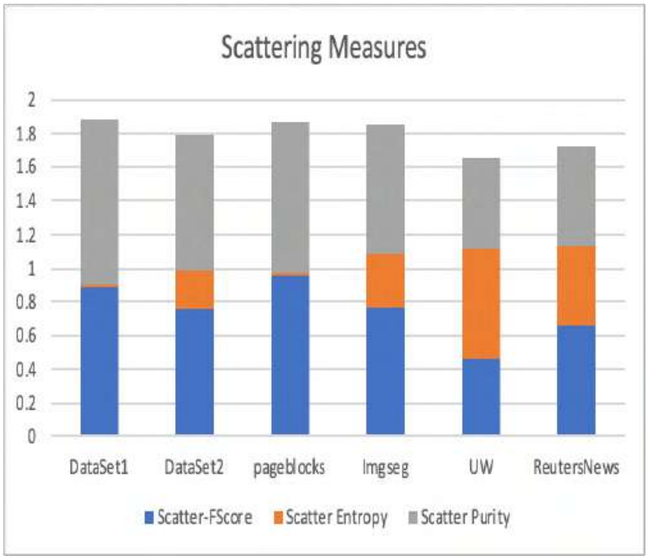


Fig. 3. The Scattering Measures

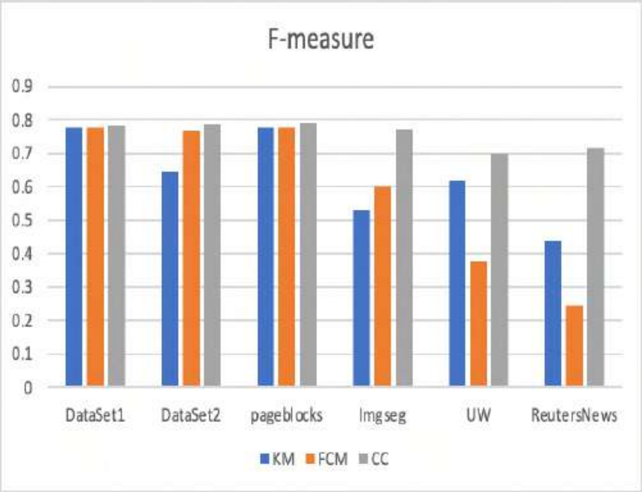


Fig. 4. The F-measure of the KM, FCM, and CC for the Six Datasets

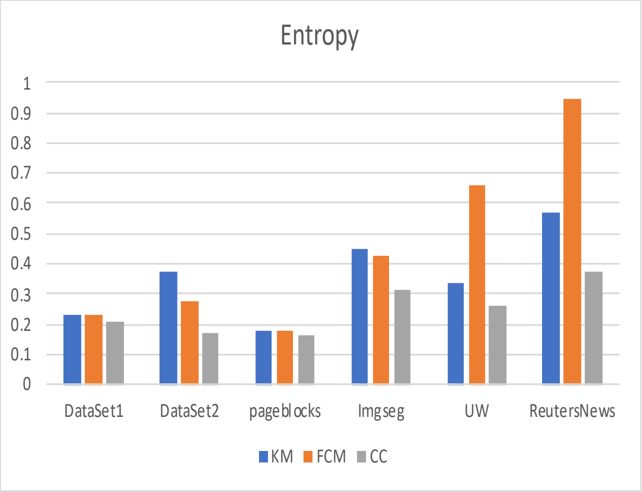


Fig. 5. The Entropy of the KM, FCM, and CC for the Six Datasets

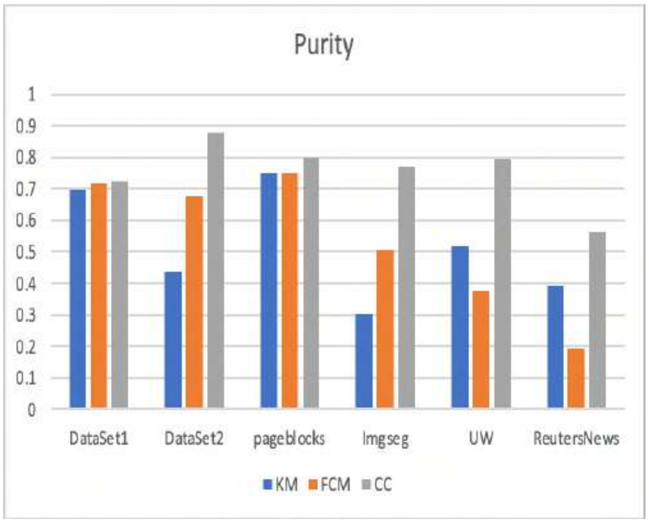


Fig. 6. The Purity of the KM, FCM, and CC for the Six Datasets

VI. CONCLUSION AND FUTURE DIRECTIONS

Validation metrics play an important factor in assessing the performance of various machine learning algorithms. The process of unsupervised learning is a more challenging task than supervised learning as no class labels are provided to evaluate the quality of the adopted algorithm, especially for building an ensemble. Thus, there is a need to provide a pre-assessment of the invoked algorithm before designing an aggregate. In this paper, we have proposed three scattering measures that effectively evaluate the divergence or the diversity in the clustering solutions to build a successful and feasible consensus. Experimental results have shown that clustering solutions with similar partitioning show no improvement when combined, while high diverse solutions indicate a high level of refinement when aggregated. These results are confirmed using the proposed three scattering measures. The proposed measures can be generalized to evaluate the performance of both supervised and unsupervised methods prior to the ensemble modeling stage. Future directions include investigating the scalability of the scattering measures if we have more than two methods, such iterative pair-wise evaluation of the scattering measures is used. We also plan to investigate the application of the proposed measures for different consensus learning approaches.

REFERENCES

[1] van IJzendoorn, David GP, et al. "Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas." PLoS computational biology 15.2 (2019): e1006826.

[2] R. Kashef and M. S. Kamel, "Towards Better Outliers Detection for Gene Expression Datasets," 2008 International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies, Bucharest, Romania, 2008, pp. 149-154, doi: 10.1109/BIOTECHNO.2008.29.

[3] Sweeney, Elizabeth M., et al. "A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI." PloS one 9.4 (2014): e95753.

[4] Kashef, R., & Kamel, M. S. (2010). Cooperative clustering. Pattern Recognition, 43(6), 2315-2329.

[5] Kashef, R., & Kamel, M. S. (2009). Enhanced bisecting k-means clustering using intermediate cooperation. Pattern Recognition, 42(11), 2557-2569.

[6] McNally, S., Roche, J., & Caton, S. (2018, March). Predicting the price of bitcoin using machine learning. In 2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP) (pp. 339-343). IEEE.

[7] Tan, X., & Kashef, R. (2019, December). Predicting the closing price of cryptocurrencies: a comparative study. In Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems (pp. 1-5).

[8] Ibrahim, A., Kashef, R., Li, M., Valencia, E., & Huang, E. (2020). Bitcoin Network Mechanics: Forecasting the BTC Closing Price Using Vector Auto-Regression Models Based on Endogenous and Exogenous Feature Variables. Journal of Risk and Financial Management, 13(9), 189.

[9] Ibrahim, A., Kashef, R., & Corrigan, L. Predicting market movement direction for bitcoin: A comparison of time series modeling methods. Computers & Electrical Engineering, 89, 106905.

[10] Tobin, T., & Kashef, R. (2020, June). Efficient Prediction of Gold Prices Using Hybrid Deep Learning. In International Conference on Image Analysis and Recognition (pp. 118-129). Springer, Cham.

[11] A. F. Ibrahim, L. Corrigan and R. Kashef, "Predicting the Demand in Bitcoin Using Data Charts: A Convolutional Neural Networks Prediction Model," 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), London, ON, Canada, 2020, pp. 1-4, doi: 10.1109/CCECE47787.2020.9255711.

[12] Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A survey on distributed machine learning. ACM Computing Surveys (CSUR), 53(2), 1-33.

[13] Kashef, R., & Niranjan, A. (2017, December). Handling Large-Scale Data Using Two-Tier Hierarchical Super-Peer P2P Network. In Proceedings of the International Conference on Big Data and Internet of Thing (pp. 52-56).

[14] Kashef, R., & Kamel, M. (2006, November). Distributed Consensus hard-fuzzy document clustering. In Proceedings of the Annual Scientific Conference of the LORNET Research Network.

[15] Yeh, T. Y., & Kashef, R. (2020). Trust-Based Collaborative Filtering Recommendation Systems on the Blockchain. Advances in Internet of Things, 10(4), 37-56.

[16] Tian, L., Fan, Y., Li, L., & Mousseau, N. (2020). Identifying flow defects in amorphous alloys using machine learning outlier detection methods. Scripta Materialia, 186, 185-189.

[17] Li M, Kashef R, Ibrahim A. Multi-Level Clustering-Based Outlier's Detection (MCOD) Using Self-Organizing Maps. Big Data and Cognitive Computing. 2020; 4(4):24. https://doi.org/10.3390/bdcc4040024

[18] Kashef, R. F. (2018, January). Ensemble-Based Anomaly Detetction using Consensus Learning. In KDD 2017 Workshop on Anomaly Detection in Finance (pp. 43-55). PMLR.

[19] Radivojević, T., Costello, Z., Workman, K., & Martin, H. G. (2020). A machine learning Automated Recommendation Tool for synthetic biology. Nature communications, 11(1), 1-14.

[20] Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. Applied Sciences, 10(21), 7748.

[21] M. Ebrahimian and R. Kashef, "Efficient Detection of Shilling's Attacks in Collaborative Filtering Recommendation Systems Using Deep Learning Models," 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore, 2020, pp. 460-464, doi: 10.1109/IEEM45057.2020.9309965.

[22] Ebrahimian M, Kashef R. Detecting Shilling Attacks Using Hybrid Deep Learning Models. Symmetry. 2020; 12(11):1805. https://doi.org/10.3390/sym12111805

[23] Kashef, R. (2020). Enhancing the Role of Large-Scale Recommendation Systems in the IoT Context. IEEE Access, 8, 178248-178257.

[24] Nawara, D., & Kashef, R. (2020, September). IoT-based Recommendation Systems–An Overview. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-7). IEEE.

[25] Pano, Toni, and Rasha Kashef. "A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19." Big Data and Cognitive Computing 4.4 (2020): 33.

[26] Kashef, R., & Kamel, M. S. (2007, October). Hard-fuzzy clustering: a Consensus approach. In 2007 IEEE International Conference on Systems, Man and Cybernetics (pp. 425-430). IEEE.

[27] Rasha Kashef, A boosted SVM classifier trained by incremental learning and decremental unlearning approach, Expert Systems with Applications, 2020, 114154, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.114154.

[28] G. Hass, P. Simon and R. Kashef, "Business Applications for Current Developments in Big Data Clustering: An Overview," 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore, 2020, pp. 195-199, doi: 10.1109/IEEM45057.2020.9309941.

[29] Close, L. and Kashef, R. (2020) Combining Artificial Immune System and Clustering Analysis: A Stock Market Anomaly Detection Model. Journal of Intelligent Learning Systems and Applications, 12, 83-108. doi: 10.4236/jilsa.2020.124005.

[30] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 17 (2001) 107–145.

[31] Rezaei, M., & Fränti, P. (2016). Set matching measures for external cluster validity. IEEE Transactions on Knowledge and Data Engineering, 28(8), 2173-2186.

[32] Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. International Journal of computers and communications, 5(1), 27-34.

[33] Rosenberg, A., & Hirschberg, J. (2007, June). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In EMNLP-CoNLL (Vol. 7, pp. 410-420).

[34] De Souto, M. C., Coelho, A. L., Faceli, K., Sakata, T. C., Bonadia, V., & Costa, I. G. (2012). A Comparison of External Clustering Evaluation Indices in the Context of Imbalanced Data Sets. brazilian symposium on neural networks.

[35] Deb, D., Fuad, M. M., & Angryk, R. A. (2006, January). Distributed hierarchical document clustering. In ACST (pp. 328-333).

[36] Draszawka, K., & Szymanski, J. (2011). External Validation Measures for Nested Clustering of Text Documents. ISMIS Industrial Session, 369, 207-225.

[37] Dalli, A. (2003, April). Adaptation of the F-measure to cluster based lexicon quality evaluation. In Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable? (pp. 51-56). Association for Computational Linguistics.

[38] Reichart, R., & Rappoport, A. (2009, June). The NVI clustering evaluation measure. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (pp. 165-173). Association for Computational Linguistics.

[39] Meilă, M. (2007). Comparing clusterings—an information based distance. Journal of multivariate analysis, 98(5), 873-895.

[40] van Dongen, Stijn: Performance Criteria for Graph Clustering and Markov Cluster Experiments. Technical Report INS–R0012, Centrum voor Wiskunde en Informatica, 2000.

[41] M. Meila and D. Heckerman, “An experimental comparison of model based clustering methods,” Machine Learning, 41(1-2), pp. 9–29, 2001.

[42] P. Fränti, M. Rezaei and Q. Zhao, “Centroid index: cluster level similarity measure,” Pattern Recognition, 47(9), pp. 3034-3045, 2014.

[43] Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4).

An Approach to design Human Assisting Prototype Robot for providing Fast and hygienically secure environment to Clinical professionals in order to fight against COVID19 in Hospitals

Kamran Hameed*
Department of Biomedical Engineering
Imam Abdulrahman Bin Faisal University
Dammam, Kingdom of Saudi Arabia
khKhawaja@iau.edu.sa

***Abstract*— In the current era robots play vital role in several industries, hospitals, and research organizations etc. In hospitals robots support and nursing staff currently may also use in this era during COVID19 Pandemic to transport such patient sample or dispose of their usage belonging to avoid spreading of this pandemic to clinical staff. From making deliveries, dispensing medication, visiting patients to aiding surgeries, robots are improving the way hospitals function. Central Sterilization Supply Department (CSSD) offers sterilization amenities to Outpatient Department (OPDs), wards and operation theatre (OT) of hospitals. It provides facilities to receive, clean, pack, disinfects, sterilizes, store and distribute instruments, in accordance with well-delineated protocols and standardizes procedures. An alarming boost in Hospital Ac-quired Infections (HAI) demonstrates the necessity of a well-organized CSSD to prevent surge in HAI. Although quality assurance procedures are followed to ensure safety and efficiency at all levels: appropriate handling of contaminated items, decontamination, proper cleansing, and instrument care, but the necessity of human presence to implement these procedure in-creases the threat of HAI. Thus, an intelligent autonomous multi-sensor wireless controlled NXT robot using MATLAB is proposed to handle contaminated items and sterilize equipments in the centralized sterilization sup-ply department. The proposed system is entirely automated and eliminates the need of human presence to implement quality assurance procedures, thus decreasing the threat of HAI. An archetype using Lego NXT robotic kit is developed to implement tasks performed in CSSD. MATLAB is used to control NXT robot instead of Lego Mindstorms software, due to its versatility, capability to perform computationally intensive tasks and wide usage by researchers. A powerful control system is designed that allows user to interrupt, control and monitor autonomous NXT robots.**

Keywords— ***COVID19 , Robotics Invention System, Central Sterilization Supply Department, Hospital Acquired Infection, Operation Theater.***

Introduction

In the current era robots play vital role in several industries, hospitals, and research organizations etc. In hospitals robots are the newest members of the hospital support and nursing staff. This project focuses on design of an intelligent autonomous multi-sensor wireless controlled NXT robot using MATLAB to handle contaminated item and sterilize equipments in the centralized sterilization supply department specially to protect clinical professional. The

system is entirely automated and quality assurance procedures, thus decreasing the threat of HAI. The choice of robotic podium for this project is the Lego Mind storms NXT. This system facility to collect, clean, pack, disinfects, sterilizes, stores, and distributes instruments, in accordance with well-delineated protocols and standardizes procedures [1-5].

Proposed System works as a single unit, which is control by MATLAB. MATLAB runs in monitoring computer which is wirelessly connected with robot via Bluetooth dongle and all other equipment in CSSD is connected with the serial port connection. We have to use advance software for controlling dongle. The dongle works as bi-direction serial port which enables us to use serial commands to communicate with Lego robot. For establishment of strong link between MATLAB and Lego NXT we have to install some program files in brick. This file helps us to easily communicate with robot by using MATLAB. First of all, we upgrade firmware of Lego NXT from 1.21 to 1.26. This was done because old firmware does not support RWTH toolbox. This Toolbox provides communication with robots with Bluetooth or USB. The toolbox includes routines for supporting interactions between robots and MATLAB. We use RWTH toolbox in MATLAB the commands were send through the MATLAB after creating the connections and robot work according to the given instructions remember as the robot have motors and sensors all were activated by via wireless connection no user or wire interface is there. as we know that the CSSD staff work manually al the process, they take contaminated material from the O.T and they enter into the department and processed it through washer and autoclaves there are some indications which already known by the staff that the equipment shows it become sterile, in short all work done by the staff according to the instructions. Same process is performing by the robot technology NXT ROBOT more accurate more smoothly and shows best performance there according to the instructions. The washer and autoclave are controlled by the microprocessor which also takes instructions through MATLAB. All the equipments of the department are interconnected with each other, there data is transfer via port to the main circuit board and then via serial port the data is transfer to the computer placed in an operating room. In this way a continues

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

feedback is provided and robot is monitored continuously so there is almost no chance of human error. The instructions of doors open, and robot enter and exit all timing are set and command by MATLAB [6-18]

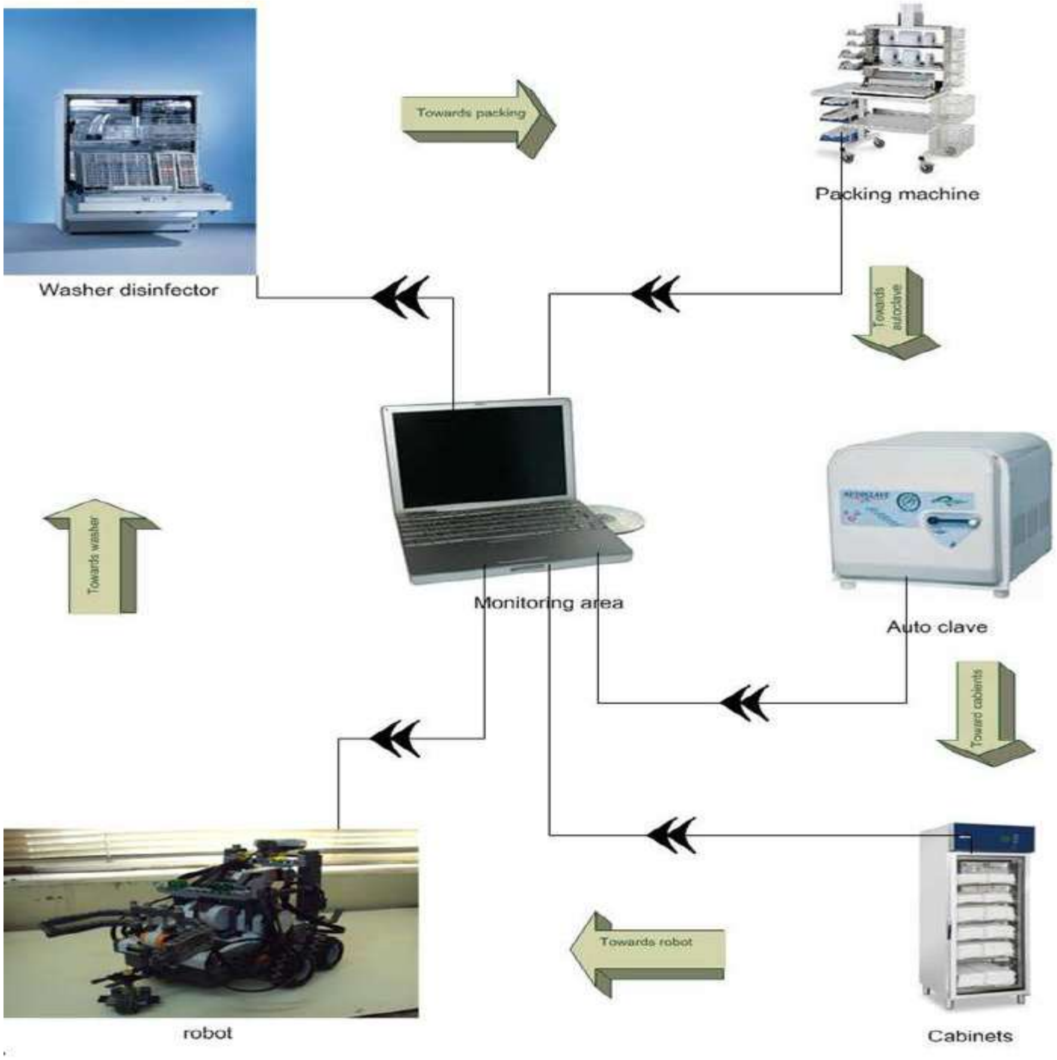


Figure 1. Shows the Working Principle of the proposed designed

I. Design Methodology

A. Material Selection in proposed designed

Lego Mindstorms kit with MATLAB software were used in this project, in addition to design whole prototype some extra materials are also used in the completion of complete prototype design as shown in Figure 2. Lego Mindstorms consists of different construction blocks, sensors and controlling brick [6], The controller brick used in the proposed system in controlling all functions of Lego robot, which includes ,4 input ports,3 output ports,1 Bluetooth connection. Light Sensors used to help the Robots to move on the desired location where the object placed, Ultrasonic Sensors are integrated with designed robot to assure not to collide with any surface or objects, Sound Sensor also used to identify if the hazardous material or object detected by the robots or not, Touch Sensor were also incorporated with robots to hold the object after feeling touch sensation. Three servo motors are also used to control the robots' gripper and arm angle. Shaft Encoder used to measure the speed of the servo motor to ensure that the designed robot manipulate its task accurately, finally Blue tooth dongle were also used, this wireless system helps robot to easily move anywhere and send back its location and status to monitoring system. This adapter is high speed and medium range which is about 100m in circle. Adapter is capable to operate in bi-direction mode. [6]. The conveyer belt is used inside the equipments to transfer the material from entrance to exit which is placed by the robot. The conveyer belt has widely used in industries for the supply chain and transfer of the material from the idea of that industries make me an example to provide a better way to hold and control the material through conveyer belt. It consists of single motor which is the center of the belt and belt is rotate on it when the motor allows to move clockwise or anti clockwise.

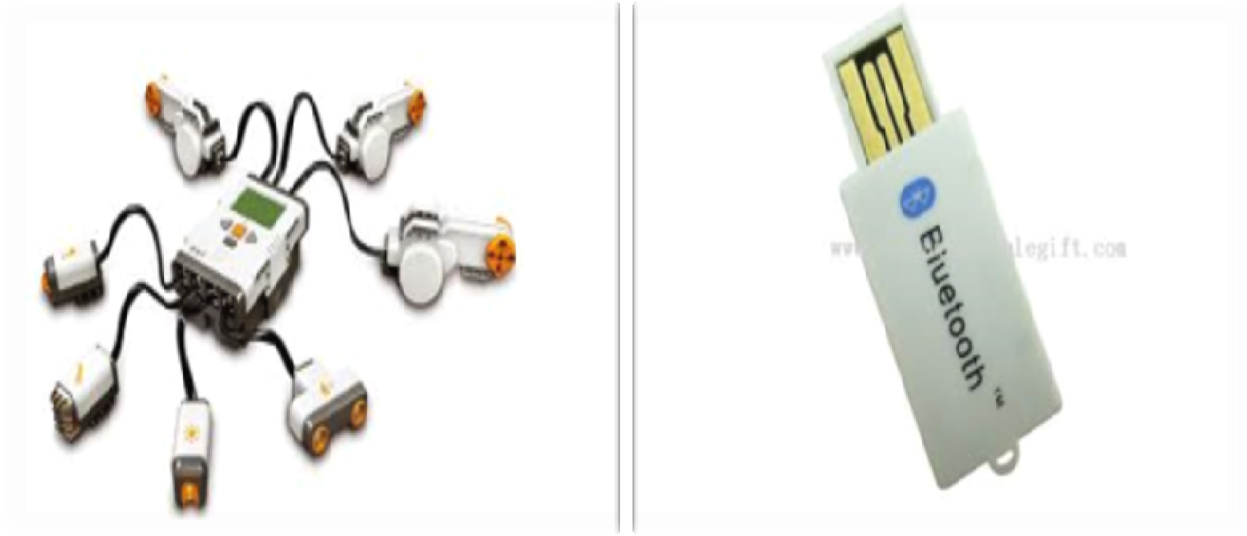


Figure 2. At Left: Lego NXT Controller units connected with different sensors, At Right shows the Bluetooth dongle used with the system for wirelessly controlling the whole system.

B. HARDWARE ARCHITECTURE

C. SOFTWARE DECSRIPTION

In Robot designing two NXT Robots' bricks were used to control severe servo motors especially four motors used among which two were used for driving and steering wheels of designed robot while one used for robotic gripper and one for robotic lifter as shown in Figure 1-2. The proposed designed system also has ability to sense different color packages with the help of color sensor to detect and distinguish visible colors based on these color sections Robot can easily pick and place of different color and shape objects, an ultrasonic sensors also used to detect obstacle during the Robot moving towards the object for handling or dispose of contaminated object like the mask or any belonging used by COVID 19 patient to protect the direct interaction of ward boy form this pandemic. ,an ultrasonic sensor may also use to retrieve the robot's distance to its staring point and light sensor was also used to rover the Robot in the straight path which may detect the lines for its path findings where user want to move the robot to move. The wireless BLUETOOTH communication used in the proposed system which provides access to control and communicate with robot from computer. 8-bit microcontroller and a 100x64 LCD monitor. This brick supports up to four sensorial inputs and can control up to three servomotors. It also has an interface displayed by the LCD System works as a single unit which is control by MATLAB. MATLAB runs in monitoring computer which is wirelessly connected with robot via Bluetooth dongle and all other equipment in CSSD relates to the serial port connection. We have to use advance software for controlling dongle. The dongle works as bi-direction serial port which enables us to use serial commands to communicate with Lego robot. For establishment of strong link between MATLAB and Lego NXT we have to install some program files in brick. This file helps us to easily communicate with robot by using MATLAB. First, we upgrade firmware of Lego NXT from 1.21 to 1.26. This was done because old firmware does not support RWTH toolbox. The toolbox includes routines for supporting interactions between robots and MATLAB. We use RWTH toolbox in MATLAB the commands were send through the MATLAB after creating the connections and robot work according to the given instructions remember as the robot have motors and sensors all were activated by via wireless connection no user or wire interface is there. as we know that the CSSD staff work manually al the process, they take contaminated material from the O.T and they enter into the department and processed it through washer and autoclaves there are some indications which already known by the staff that the equipment shows it become sterile, in short all work done by the staff according to the instructions. Same process is performing by the robot technology NXT ROBOT more accurate more smoothly and shows best performance there according to the instructions. The washer and autoclave are controlled by the microprocessor which also takes instructions through MATLAB. All the equipments of the department are interconnected with each other, there data is transfer via port to the main circuit board and then via serial port the data is transfer to the computer placed in an operating room. In this way a continues feedback is provided and robot is monitored continuously so there is almost no chance of human error. The instructions of doors open, and robot enter and exit all timing are set and command by MATLAB. The robot performs the task till packaging process after the completion of the task it will come back to its starting position.

The Robot receive command from PC from the execution of its work wirelessly for this purpose Bluetooth adapter was uses through the wireless Bluetooth adapter. The MATHLAB NXT toolbox has been used in this system [14]. The software used in this designing is a free open source subject to the GNU GENERAL PUBLIC LICENSE (GPL). Desired task commands are sent to NXT robot to achieve task by using this connection, it is also use for receiving data from sensors like distance from ultrasonic sensor of each robot. A replica of CSSD was designed as shown in Figure 1 (Working Principle Diagram) to A replica of CSSD was designed as shown in Figure 1 (Working Principle Diagram) to provide better practical approach system, we built replica of whole department with all necessary equipment use in process. Prototype includes model of washer disinfector, autoclave, supply cabinet and special chamber for sterilizing robot, all equipments are design by acrylic sheets. Models are mechanically active to perform basic functions like opening and closing of doors, receiving, and transmitting commands to main controlling computer. [15]. PIC controller is use as Communication Bridge between CSSD replica and PC server. C high level programming language is used to program PIC controller, PIC-C compiler. The main board is connecting the departments equipment with computer in monitoring area. this is done in a simple way that first data is sent via port to the main circuit board and then it is transfer serially to the computer where it is monitored as shown in Figure 1 and Figure 4 respectively.

For the safety of complete Robotic system used in proposed design , The system will be stopped if the system started malfunctioning for example if the user give the system instructions command to follow the straight path and execute their work as per command received through PC software if system started malfunctioning and does acts upon received instructions the system will be because of many sensors are being incorporated within the system to assure the system behaves properly without or with minim error during execution of any task performed by the system. ,whenever a robot is about to going outside the workspace or lost its path, the whole process is stopped, and robot come back to their initial position for this purpose an emergency button access given to operator in system control console which is shown in Figure 8 . This option may use if operator realized there is any automatic shut off needed in any emergency case.

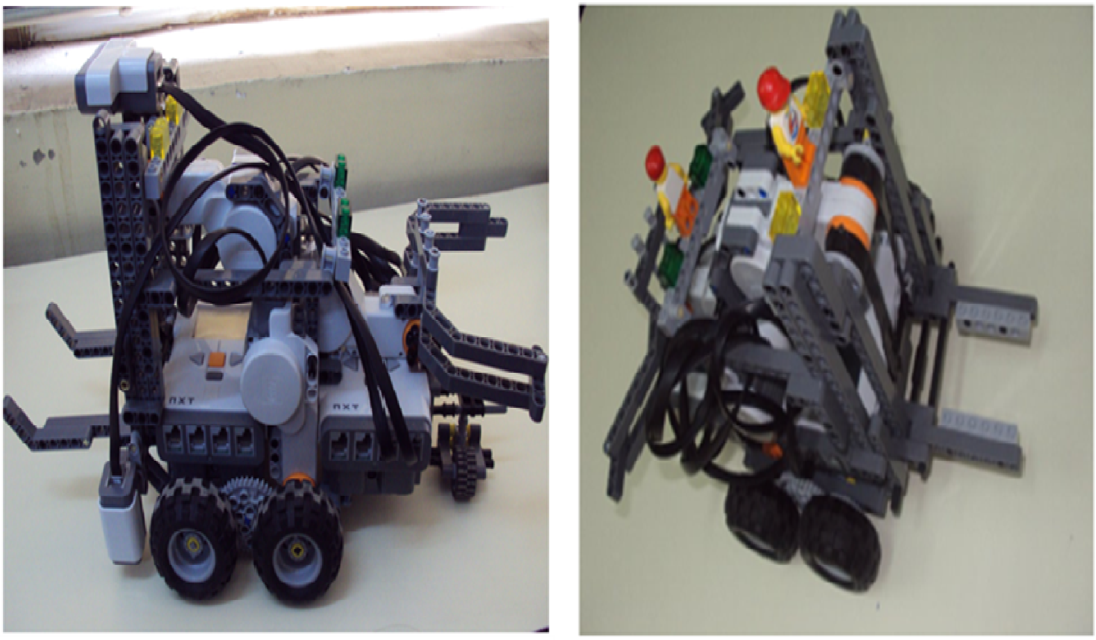


Figure 3. Proposed Robot Hardware Structure

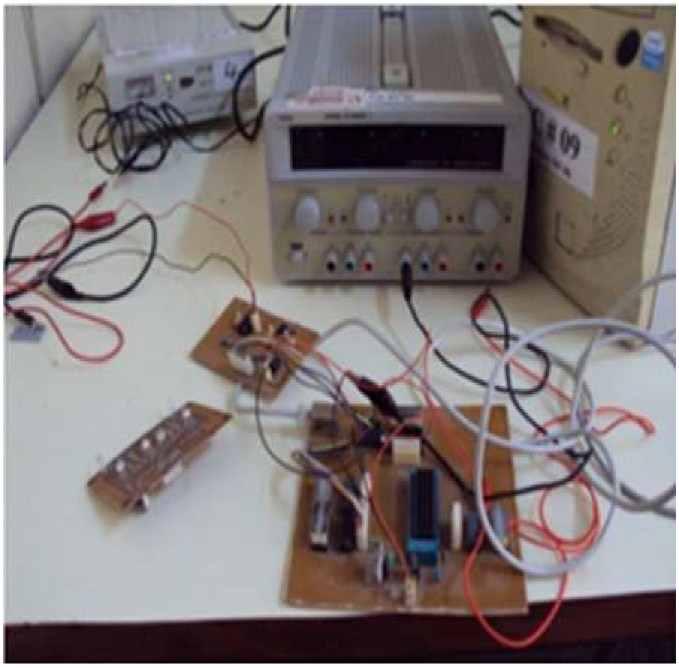


Figure 4. Project Hardware Circuit

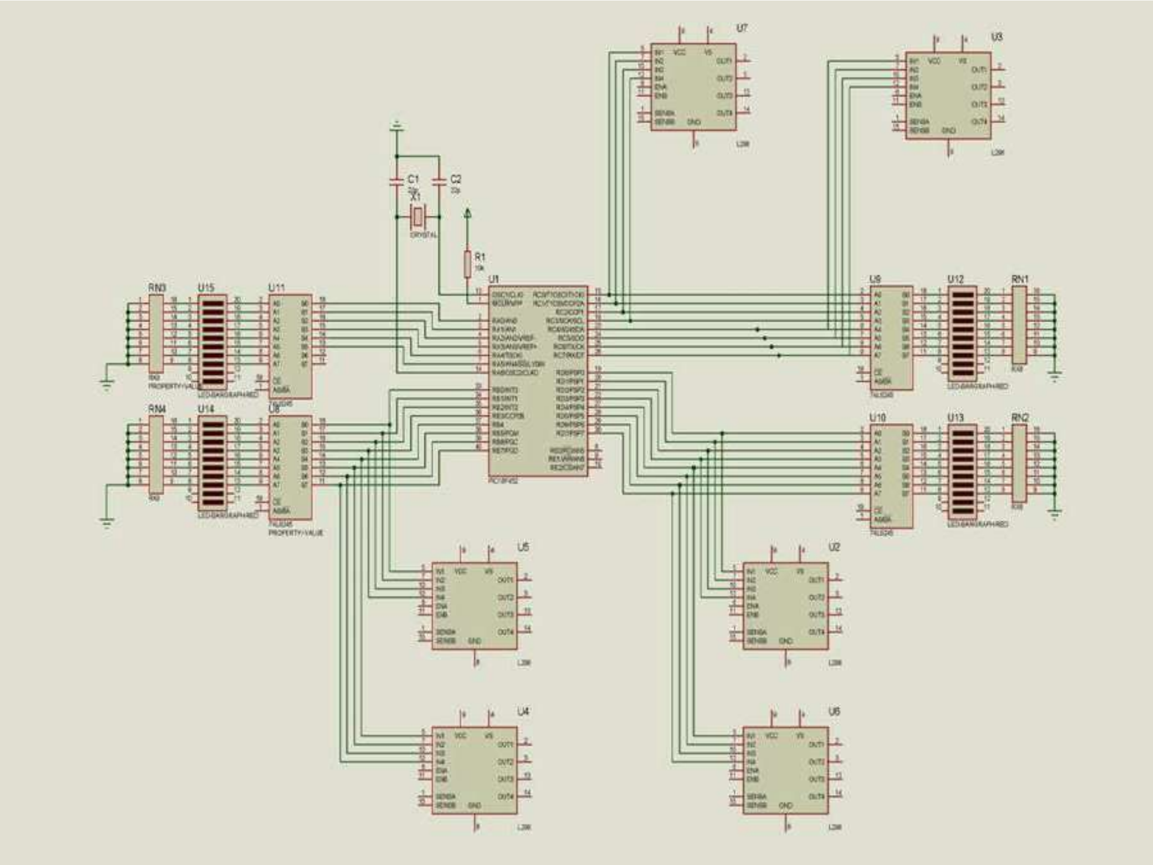


Figure 5. Hardware Circuit Schematic designed for the proposed system

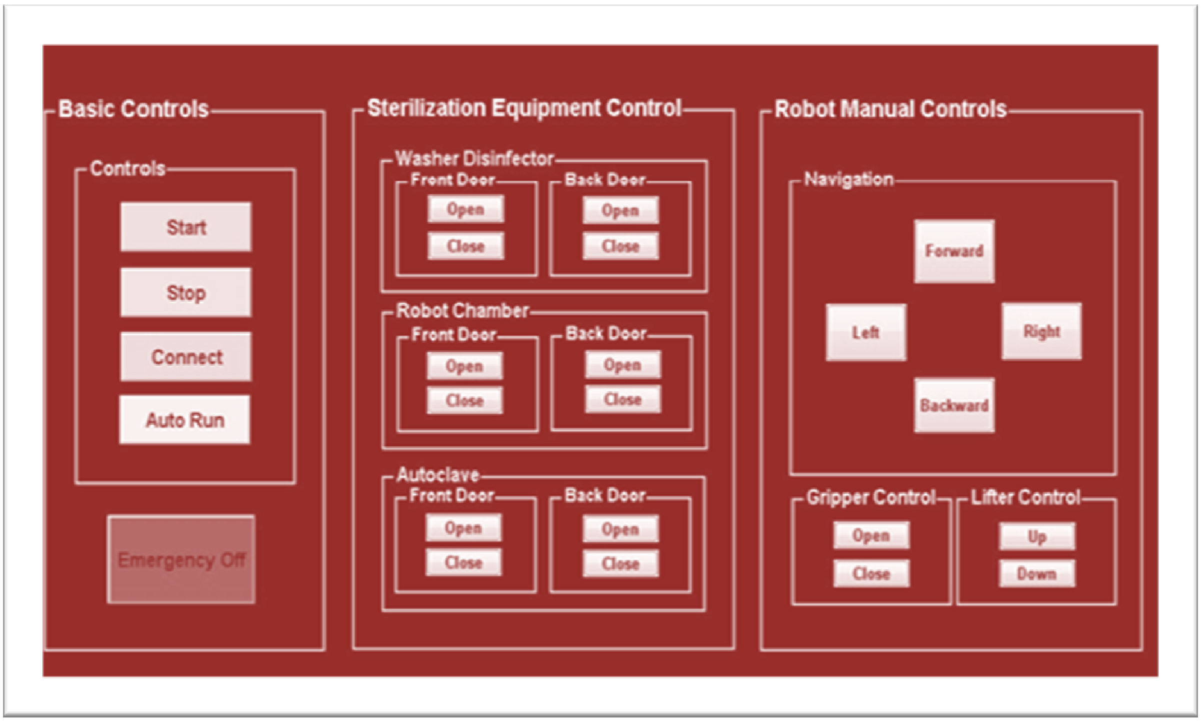


Figure 7. Basic Algorithm Flow Chart used in the Proposed Robot System Design

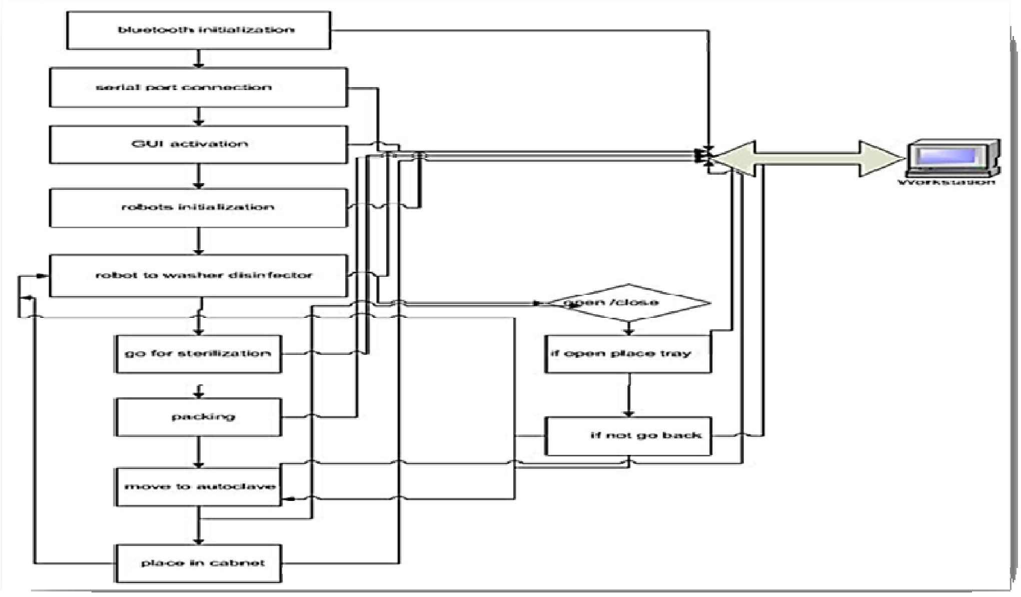


Figure 8. Proposed System operating console At Left: Basic control with star/end/auto start/emergency switches, At Center: Different operating console having provision of sterilization equipment control , ,Robot chamber control and autoclave operation control depicted, At Right: Robot Navigation control depicted with gripper control provision. with basic access control.

II. CONCLUSION

Proposed designed project may provide provides fast and hygienically secure environment to hospital Central Sterilization Supply Department especially now a days to cope the spreading of COVDID19 in clinical staff special in ward. Our system helps hospital administration to keep their hygienic quality at top. Whole system is cost effective and able to do those works which are hazardous for health of human. All hazardous material will be carrying out by robot. Whole process will automatic so very less chance of any error. In case of emergency whole system will auto, shutoff and inform operators immediately. System ensures sterilization process more reliable and accurate. We hope in future our design plays important role in maintaining hygienic quality of hospital and also make whole process faster. Robot can work freely without any risk of infections. System will able to work 24x7 without any stop. Designing of GUI for simplicity for user. System can be monitor by wireless. This phase covers all monitoring and feedback systems. In this part we try to make our robot vision intelligent so it can differentiate between objects and equipments. Image processing also helps us to find out location of robot in CSSD.

REFERENCES

[1] K. Murugan *et al*, "Medicine Distribution Robot and Human Less Intervention for Covid-19 Affected People (AKM MED ASSISTIVE BOT)," *IOP Conference Series. Materials Science and Engineering,* vol. 1049, *(1),* 2021.

[2] V. Jahmunah *et al*, "Future IoT tools for COVID-19 contact tracing and prediction: A review of the state-of-the-science," *International Journal of Imaging Systems and Technology,* 2021.

[3] Z. M. Jessop *et al*, "Personal protective equipment for surgeons during COVID-19 pandemic: systematic review of availability, usage and rationing," *British Journal of Surgery,* vol. 107, *(10),* pp. 1262-1280, 2020.

[4] M. Falahchai, Y. Babaee Hemmati and M. Hasanzade, "Dental care management during the COVID-19 outbreak," *Special Care in Dentistry,* vol. 40, *(6),* pp. 539-548, 2020.

[5] A. Kaushik, S. Patel and K. Dubey, "Digital cardiovascular care in COVID-19 pandemic: A potential alternative?" *Journal of Cardiac Surgery,* vol. 35, *(12),* pp. 3545-3550, 2020.

[6] A. S. Al-Ogaili *et al*, "IoT technologies for tackling COVID-19 in Malaysia and worldwide: Challenges, recommendations, and proposed framework," *Computers, Materials & Continua,* vol. 66, *(2),* pp. 2141-2164, 2020;2021;.

[7] C. M. Goldrick and M. Huggard, "Peer learning with lego mindstorms," in 2004.

[8] N. Mohan, *Advanced Electric Drives: Analysis, Control, and Modeling using MATLAB/Simulink.* (1st;1; ed.) 2014*9781118485484*. DOI: 10.1002/9781118910962.

[9] Anonymous "Tackling problem of non-sterile equipment," *Straits Times (Singapore : Daily),* 2018.

[10] E. Danahy *et al*, "LEGO-based Robotics in Higher Education: 15 Years of Student Creativity," *International Journal of Advanced Robotic Systems,* vol. 11, *(2),* pp. 27, 2014.

[11] Getinge, "CSSD Solutions and Equipment", July 2007

[12] M. P. Cuéllar and M. C. Pegalajar, "Design and implementation of intelligent systems with LEGO Mindstorms for undergraduate computer engineers," *Computer Applications in Engineering Education,* vol. 22, *(1),* pp. 153-166, 2014.

[13] M. B. R. Vallim, J. -. Farines and J. E. R. Cury, "Practicing engineering in a freshman introductory course," *IEEE Transactions on Education,* vol. 49, *(1),* pp. 74-79, 2006.

[14] G. A. S. Pereira and E. J. R. Freitas, "Navigation of Semi-autonomous Service Robots Using Local Information and Anytime Motion Planners," *Robotica,* vol. 38, *(11),* pp. 2080-2098, 2020.

[15] Y. A. P. Wayan Reza, C. G. I. Partha and I. W. Widhiada, "Simulation of a Differential-Drive Wheeled Mobile Lego Robot Mindstorms NXT," *Applied Mechanics and Materials,* vol. 776, pp. 319-324, 2015.

[16] M. Toz and S. Kucuk, "Dynamics simulation toolbox for industrial robot manipulators," *Computer Applications in Engineering Education,* vol. 18, *(2),* pp. 319-330, 2010;2009;.

[17] L. Fortunati, A. M. Manganelli and G. Ferrin, "Arts and crafts robots or LEGO(R)MINDSTORMS robots? A comparative study in educational robotics," *International Journal of Technology and Design Education,* 2020.

[18] Z. M. Jessop *et al*, "Personal protective equipment for surgeons during COVID-19 pandemic: systematic review of availability, usage and rationing," *British Journal of Surgery,* vol. 107, *(10),* pp. 1262-1280, 2020.

MicroPython-based Sensor Node with Asymmetric Encryption for Ubiquitous Sensor Networks

Ulrich ter Horst, Hagen Hasberg, Stephan Schulz
Faculty of Engineering and Computer Science
Hamburg University of Applied Sciences, D-20099 Hamburg, Germany
Email: {stephan.schulz, ulrich.terhorst, hagen.hasberg}@haw-hamburg.de

***Abstract*—This work introduces a new microcomputing node with long-term resistant data security, based on asymmetric and symmetric encryption combined with the modern and established scripting language Python. The presented microcomputing node integrates a MicroPython runtime environment to address a wide audience of application engineers as user base instead of a selected group of embedded engineers, who have deep knowledge in programming IoT devices using C/C++. It combines its scripting capabilities with security features of modern smartcards and secure cellular networking based on 4G.**

***Index Terms*—IoT, MicroPython, smartcard, sensor node, encryption**

I. Introduction

The acquisition and processing of real-time sensor data with real-time forecasts and downstream control is becoming a standardized tool in the industrial sector and general automation thanks to powerful networks and power-saving processors. The research area Internet of Things (IoT) lives from the connection of cyber physical systems (CPS) to data centers or other cloud instances. Over the last few years, a paradigm shift from a classic industrial control system to powerful, reliable and secure decentralized embedded systems based on wireless networks has developed. The pre-processing in fog computing corresponds to the worldwide highly available cloud systems of large providers. A central element of acceptance in the industrial sector is data security and encryption of process data and configurations of industrial systems. Data centers of industrial suppliers stand next to large companies that offer highly developed evaluation methods. The monitoring of systems for predictive maintenance by IoT is becoming increasingly easier to implement. The expansion of existing industrial plants in brownfield sites or the integration of sensor nodes in urban areas will be an important technological application in the upcoming years.

This paper presents a microcomputing node with secure 4G connectivity, which is capable to interface a wide variety of sensors, actuators and industrial networks in order to perform both sensory and control tasks simultaneously. Data streams such as process data, configurations or scripts can be securely encrypted via OpenPGP similar to RFC4880 by the use of a smartcard [1]. The microcomputing node is based on an adapted implementation of MicroPython and thus ensures easy access via a high-level programming language.

II. Applications of Sensor Nodes

Modern cloud providers offer REST-APIs and M2M protocols to integrate microcomputing nodes into solutions running on their servers. A commonly used protocol is MQTT. A simple sensor node application in the context of industrial IoT is illustrated in Fig. 1. Sensor data is received by the microcomputing node and uploaded into the AWS cloud. If necessary the data can be filtered and pre conditioned within the node itself. Afterwards the data is marshalled into a JSON string and inserted into an MQTT message. Fig. 1 shows two use cases. In the first case the data within the MQTT message is cleartext information and the messages topic belongs to an AWS Thing shadow, which can be understood as the digital twin of the microcomputing node. The message updates the shadows state and according to the attached rules, the sent data will be extracted and forwarded to any AWS cloud service. It can be displayed in a semi-realtime dashboard using AWS SiteWise, it can be processed further using AWS Analytics or it can be stored into an S3 Bucket to be processed by an A.I. application using AWS SageMaker.

All these applications are possible, because the data can be read and interpreted by AWS, the cloud provider. If this does not fit the users needs, then the cryptography features of the nodes smartcard can be applied before the data is marshalled for the transmission. In both cases the sensor data is securely transmitted via TLS, but in the second case it is already signed and encrypted using asymmetric cryptography before the transport encryption applies. The cloud provider still gets the necessary meta data to assign the message to the correct digital twin but the sensor data payload itself is unreadable ciphertext from the perspective of the cloud provider. By this procedure the user has excluded the cloud provider from his applications chain of trust. Due to asymmetric encryption it is guaranteed, that the data can only have been sent from the trusted microcomputing node, that it is unmodified and that only the user can decrypt it.

Both use cases can also be combined within one application. A logistic service provider for example can detect, upload and analyze unwanted incidents during operation without disclosing critical insights of his business to the cloud provider. If any cargo stays in one place unexpectedly long or it gets

978-1-6654-4067-7/21/$31.00 ©2021 European Union

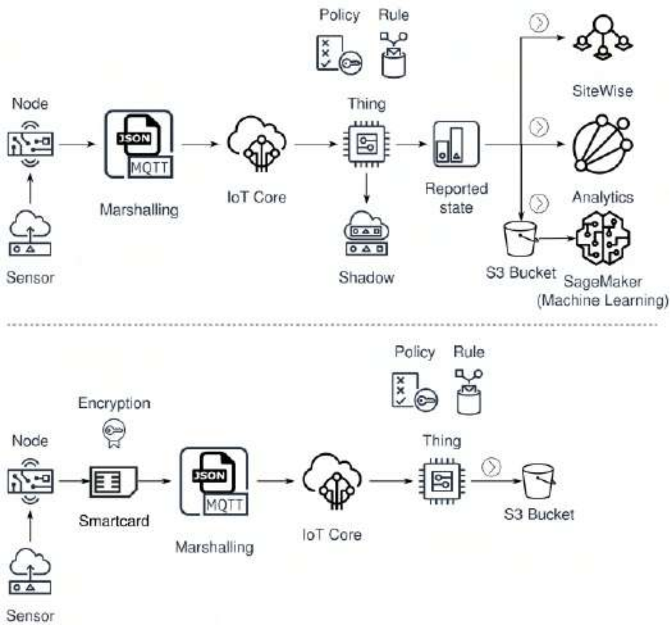


Fig. 1. Illustration of two AWS solution architectures. In the first solution the data is only encrypted during the upload into the cloud. All AWS Services can be applied to interpret the data. The solution below shows an encrypted upload. Here AWS can only act as storage hoster and has no insight into the process data.

exposed to vibrations and shocks of unacceptable amplitudes, these incidents can be logged into the cloud. The incidents date and location of occurrence can then be utilized as features for further analysis to avoid or reduce the amount of these incidents in the future. So these features are wanted to be shared with the cloud provider to use the cloud based analytic services. But other features like type of cargo or customers associated with the cargo may be not wanted to be disclosed. Hence this information is better uploaded in ciphertext form. This procedure offers the logistic service provider to examine the undisclosed information in the future outside of the cloud. Without the custom encryption inside the sensor node, the logistic service provider has to associate an incidents information to a unique identifier and store the critical information on self hosted instrastructure, which is wanted to be avoided in the first place, so that both data sets can be linked by the identifier for further inspection.

III. Platforms

The node is expected to be a realtime-capable and generic foundation, on which specialized solutions can be created. Therefore the platform, on which the MicroPython runtime environment operates, must weigh in several and partially conflicting requirements to support the scripting languages universal applicability. For example several IoT use cases do rely on solar and battery powered sensor node solutions. So energy efficiency and short response times out of deep sleep states are requirements that have a higher impact on platform decisions than data rates or bandwidths. IoT devices are also often considered as insecure and easy targets in IT security and hence as threats for a corporate networks integrity [5], [6]. So it is important to minimize the IoT devices attack surface, because often they use the same network as other corporate computers and thus have the same thread level [7]. Also the formfactor can be important in terms of size. Sometimes IoT devices are required to blend in into their surroundings, or they have to be integrated in space constrained environments, or their deployment itself must be simplified and be more cost efficient, because their destination is in a remote area.

The main decision in identifying the right platform for the microcomputing node is the choice between a Linux-capable System on Chip (SoC) and a more constrained microcontroller unit (MCU). The presented node is based on an MCU. The computing performance of MCUs can be considered as generally lower than the performance of SoCs. So bandwidth intense use cases like computer vision applications or the use of interfaces like fast ethernet are excluded from the nodes scope. But in most cases IoT applications require lower data rates. The limited computing performance is an acceptable constraint of MCUs, which outperform Linux capable SoCs in other categories. MCUs consume less energy than SoCs, because of lower clock frequencies, simplier hardware architecture and less software overhead. The Linux kernel is not realtime capable by default and must be patched to guarantee the required response times. And even if a Linux based system is patched and configured correctly, it is harder to maintain than a more minimalistic firmware of an MCU, because of the complexity of Linux operating systems in terms of their code size. Linux based SoCs have to execute several utility processes, like systemd. Every peace of software in the system does not only offer comfort for the developers and users, it also offers potential vulnerabilities to be exploited by attackers. An MCUs firmware is advantageous through its simplicity. If implemented correctly, it not only offers less attack surface, it also switches faster and therefore more often between deep sleep states and working phases, which helps to save even more energy.

The chosen MCU is an STM32F415. Its set of peripherals allows the microcomputing node to implement a sufficient set of hardware interfaces and timing capabilities. A first prototype based on a commercial off the shelf development board, that is based on the STM32F415 is shown inf fig. 2. The microcontroller board in the middle is adapted by a custom designed PCB to offer the required hardware interfaces via boxed headers. Configurable pull up/down resistors, two CAN transceivers and three load switches are also integrated on the prototypes adapter board.

Some of the firmwares internal modes of operation play an important role for data security, although all signing and encryption processes are sourced out into the trusted hardware of a smartcard. The protection against firmware readouts by unauthorized parties is a required feature. Many MCUs have insufficient readout protection capabilites and also other STM32 MCU families have shortages here, but an STM32F4 based microcontroller seems to offer sufficient protection to date [2]. Another important aspect of the choice to use an

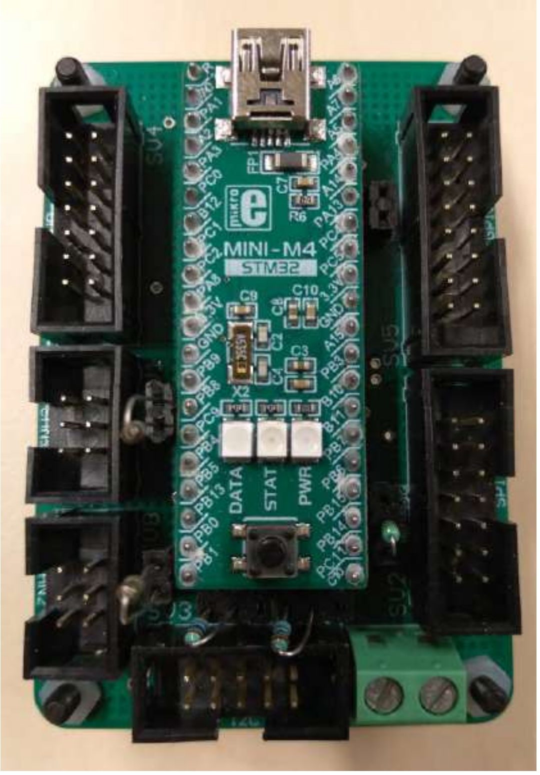


Fig. 2. A first prototype, based on a commercial off the shelf development board and a custom designed adapter PCB.

STM32 based MCU is the quality of existing libraries for the Hardware Abstraction Layer (HAL) and the quality of documentation.

IV. System Architecture

To optimize the cost efficiency of IoT sensor network deployments, it is advantageous to design the microcomputing node in such a way, that deployments do not require the installation of additional infrastructure like routers, switches or relays. Therefore the presented node integrates a cellular (4G) modem and can fully rely on existing infrastructure.

After the internet connection is established by the modem, the cloud provider can be chosen freely, because of standardized messaging protocols like MQTT, which uses TCP/IP on the transport layer and is secured with TLS. Since most of the common cloud providers like Amazon Web Services (AWS), Google Cloud, Microsoft Azure, SAP Hana and Bosch IoT Suite use MQTT, a switch between providers demands only configuration changes on the application layer. Although the presented microcomputing node supports Over-The-Air (OTA) updates of the whole firmware, these updates are rarely needed, because e.g. configurations, business logic and device drivers can be implemented in the form of MicroPython script files, that are exchangeable remotely via MQTT. Hence the sensor network solution is capable to grow and evolve along with the industrial business operation it has to conrol and monitor. It is able to adapt to new problems and challenges without demanding replacements or any manual labour on site, which results in substantial cost savings during operation.

All cryptographic and thus security-relevant functions and operations are outsourced onto a smartcard module, which offers several advantages including a less constrained selection of microprocessors as main processing unit, because it is not mandatory to have specific ciphers or other cryptographic hardware integrated in the MCU. Smartcards contain certified hardware specifically hardened against physical attacks to prevent any manipulations of the cryptographic operations and to maintain the secrecy of all private key information. Since the secret information is stored on the same trusted hardware device that executes the cryptographic operations, there is no reason for any private key to ever leave the trusted device. In contrast to permanently soldered cryptographic ASICs, a smartcard offers more flexibility. In case of compromised key information or outdated and no longer secure cryptography standards the smartcard can easily be replaced or updated and can remain in the microcomputing node. Also when the application demands new specific features like domain specific ciphers for PGP, block chain applications or MiFARE compatibility, then a smartcard offers the requested diversity without the replacement of the whole microcomputing node.

The nodes firmware is built around the MicroPython runtime environment. The layered structures of the firmware is illustrated in fig. 3. The application layer consists of MicroPython scripts. Below that is the MicroPython runtime environment and the operating system situated. Both of these subsystems are built on top of the middleware layer, that implements utilities like input/output orchestration of the MicroPython runtime environment, block devices for file systems and routines to print system information for MicroPython application developers. Below the middleware is the Hardware Abstraction Layer (HAL) located as exchangeable foundation for maximum portability.

MicroPython offers the user to implement an IoT solution in the modern and established scripting language Python. The microcomputing node is capable to execute a subset of Python 3 instructions that can be loaded from a script or entered during runtime via a Read Evaluate Print Loop (REPL). The firmware is designed in such a way, that any hardware interface can be used as input/output for the REPL. Hence MicroPython instructions can be tunneled via UART, CAN, SPI, I²C or USB-VCP and the user can interact with the node locally or remotely via a cellular connection for example. The REPL demands parsing and compiling during runtime before the instructions can be executed. Scripts on the other hand can also be uploaded in a previously cross-compiled bytecode format for immediate execution to save runtime and memory resources. The microcomputing node supports three sources of MicroPython scripts. They can be loaded over the REPL, they can be imported from the internal file system or they are statically flashed into the firmware as a so called frozen module and can be imported from ROM. The REPL and the file system are accessible remotely via the cellular network. The set of frozen modules can also be altered when

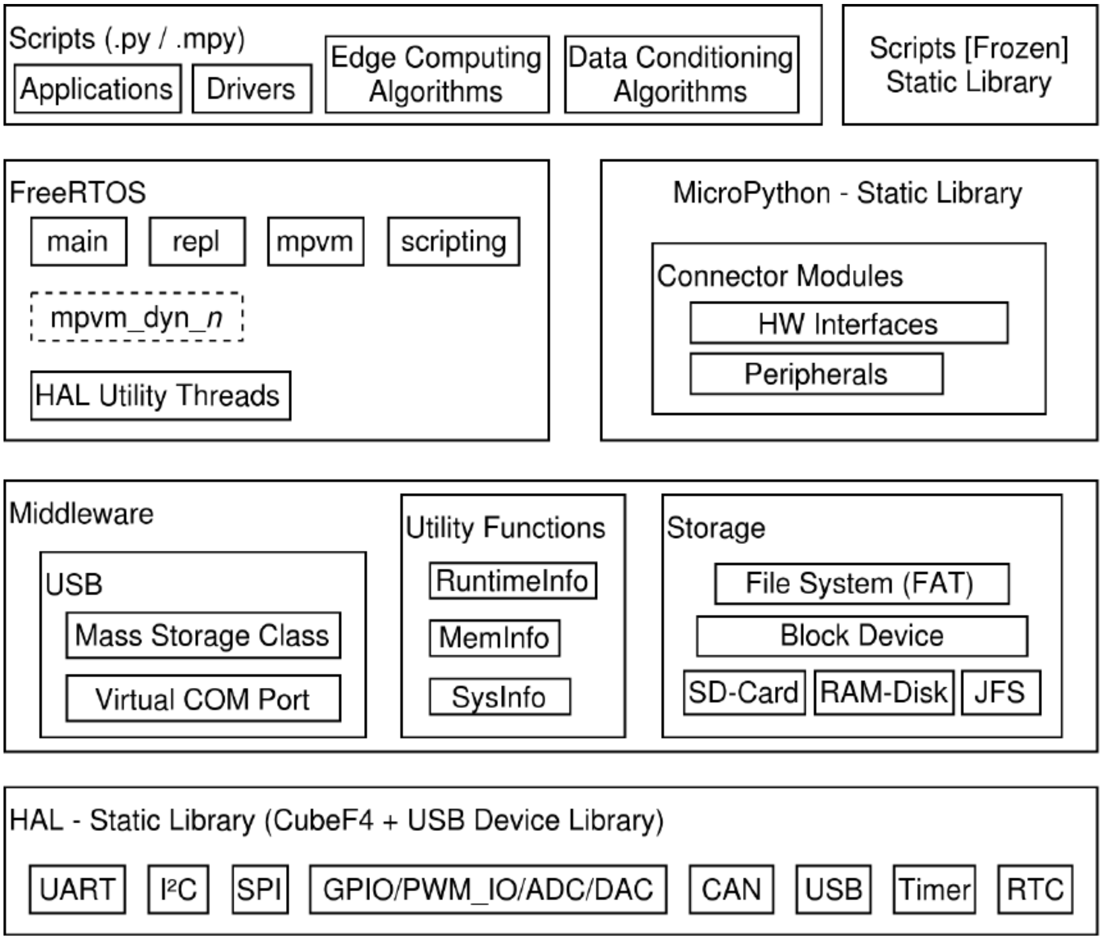


Fig. 3. Layered illustration of the microcomputing nodes firmware components.

an OTA update of the firmware is applied.

The firmwares application layer consists of MicroPython scripts, possibly also including device drivers of connected sensors or application specific algorithms for data conditioning or process state detections to generate warnings and other information, that can be uploaded to connected cloud systems.

The MicroPython runtime environment exists in the operating system layer in form of a static library so that it can be upgraded to newer MicroPython versions quickly and independently. MicroPython offers multithreading and the REPL is available anytime regardless of any scripts running in the background. To support these features and to maintain the realtime capability in parallel to responsive hardware interface handling for sensor communications the realtime operating system FreeRTOS is used. Different HAL utility threads facilitate deferred interrupt handling of the hardware interfaces. Every user generated python thread is managed through its own FreeRTOS thread. Interactions of the REPL are maintained by a seperate REPL thread. A scripting thread listens for incoming scripts via proprietary protocols on different hardware interfaces. The MicroPython runtime environment has a dedicated thread to run all incoming instructions from the other threads.

In order to make use of any microcontroller peripherals from within the MicroPython runtime environment a variety of MicroPython modules is defined in C and C++ language. The user can import these modules in REPL mode or from a script. They act as glue code to the middleware and the HAl. The HAL is based on STMicroelectronics CubeF4 HAL and like the MicroPython port it is linked into the firmware as static library. The firmware and especially its HAL follows the Ressource Acquisition Is Initialisation (RAII) ideom and avoids dynamic memory allocation from the heap at its abstraction layer. All input and output buffers are allocated from the application layer during firmware startup. Additionally the call of *malloc* is strictly avoided to offer maximum safety during runtime. The only free-storage (heap) allocation is done by the MicropPython runtime environment. therefore out of memory exceptions from heap overruns or heap fragmentation can only occur within the MicroPython heap and there overruns can be caught from the firmware, which allows for graceful reset procedures.

V. MCU Flash Memory Utilization

The MCUs flash memory serves as nonvolatile storage for three Firmware components. It holds the bootloader, that is necessary for OTA-Updates. Further it contains the actual firmware and the internal file system. Table I shows the flash distribution between these three subsystems. One can see, that the MCUs flash memory is organized in twelve sectors of different sizes. The bootloader is the smallest subsystem, it uses three of the smallest sectors. At the current state of development the bootloader is still rudimental and listens for update instructions on CAN bus only. So it is still small enough to fit in less than three sectors, but because of its internal mode of operation it must be distributed among at least three separate sectors.The firmware is located at dedicated sectors, independent from the bootloader, so that the bootloader does not overwrite itself or its fallback state information during updates, which ensures robustness against power losses. When all features are included in the compilation process and when debugging symbols are enabled, the firmware uses 58% of the flash memory. The file system needs at least two sectors to operate, so that the firmware is allowed to grow even further. But without debugging symbols it is possible to shrink the firmware down by two sectors to about 33% of flash and thus to increase the file system.

TABLE I
Flash memory sectors of STM32F415

Index	Usage	Size$[KiB]$	Percentage
0	Bootloader	16	5 %
1	Bootloader	16	
2	Bootloader	16	
3	Firmware	16	58 %
4	Firmware	64	
5	Firmware	128	
6	Firmware	128	
7	Firmware	128	
8	Firmware	128	
9	File System	128	37 %
10	File System	128	
11	File System	128	

Sector sizes vary between 16 and 128 kibibytes, which is common for MCUs. The internal file system of the microcomputing node demands the flash storage to modify its content during runtime. Flash memory is only rewritable in units of sectors. So even when the file system needs to modify a single bit, a whole sector must be rewritten and every erase cycle wears out the flash cells. The block device below the file system must compensate for the sector rewrite limitation and implement a wear leveling algorithm to enhance the flash cells life cycle. As shown in fig. 3 three different block devices are implemented for the file system. The sd-card version implements an SDIO interface, which outsources the wear leveling to a connected sd-card and makes the internal file system obsolete. The RAM-disk also avoids the usage of the internal flash, but due to its volatile nature, it is unable to hold its content persistently throughout a power outage. Persistent storage is a mandatory feature of the microcomputing node and since an sd-card can fail because of mechanical malfunction and would also increase the formfactor, the internal flash storage is used as block device. Therefore a third block device based on a journaling flash storage (JFS) is implemented. A flash sector has a size of 128 kibibytes and contains 250 blocks of 512 bytes. Flash cells can not be erased from zero to one individually but an individual conversion from one to zero is possible. The JFS uses these individual changes from one to zero to change block states in header bytes and to create new copies of blocks. So when a blocks content must be changed, first a new copy of that block is written into a free space of the sector, containing only ones. Then the new blocks state is set to valid and at last the state of the old version of the block is set to invalid, which can also mean deprecated. When a sector is full, it gets transferred to an available free sector. That is why not all sectors can be used to hold data at once and that is also why at least two separate sectors must be assigned to the JFS. The sector transfer process is implemented in a similar way as the block modification, to ensure that power outages do not corrupt the block devices content. This implementation can be the foundation of a robust journaling file system, like LittleFS [3], [4]. But in the current state the FAT file system is used to be compatible to Windows and Linux based host systems by default. By using FAT within the microcomputing node, it is possible to share the block device with the host system as USB Mass Storage Device (MSC) and to exchange files without the usage of the REPL. A journaling flash storage is necessary to maximize the microcomputing nodes lifespan, even if no journaling file system is used on top of it, because by journaling write operations between flash cells, a wear leveling is accomplished.

VI. Proof of Concept Applications

During its development the microcomputing node was used as a sensor node and as a bus converter unit (BCU) in several research projects. The BCU was e.g. the basis for several subsystems within a test-stand for small electric ducted fans for unmanned aerial vehicles.

Multiple sensors with different hardware interfaces were controlled and sampled by the microcomputing node, which was called BCU within that context. The aggregated data was broadcasted on a CAN bus by each BCU, so that all data could be read out by a workstation via a CAN-USB bridge. All BCUs were supervised and updated from that workstation via CAN bus. The bootloader mentioned above offered a way to update the BCUs firmware and MicroPython scripts were also exchanged via CAN bus. Fig. 4 gives an overview of the test-stands system design.

Four BCUs integrated the sensors using a variety of hardware interfaces. The power supply unit and the load cell were connected via UART. They delivered data for power to thrust ratio measurements. SPI and I²C busses were used for pitot tubes, microphones and three-axes accelerometers to measure air stream velocities and vibrations at different locations. New protocols or interfaces were also implemented during this use case. For example a OneWire interface was

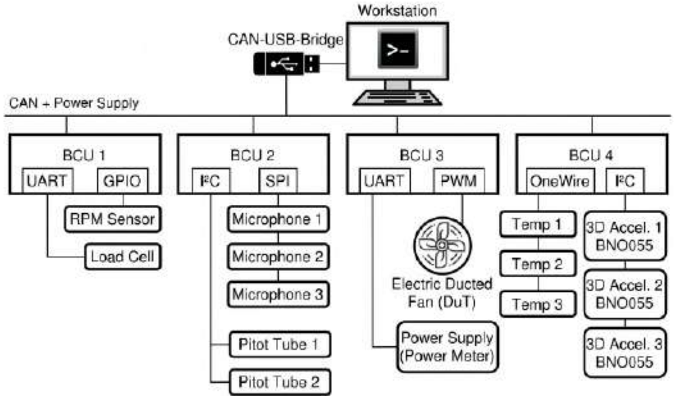


Fig. 4. System design of a test-stand for electric ducted fans for unmanned aerial vehicles. The component that aggregates different sensor data and broadcasts it via CAN bus is based on the presended microcomputing node and is called Bus Converter Unit (BCU) within this project.

added for temperature measurements and a pulse counter was implemented to measure the ducted fans rotational speed. Thus the microcomputing node has proven as convenient and generic sensor node solution for the development of the test-stand.

The microcomputing nodes reliability in terms of the IoT cloud connection was examined with another prototype. It was designed as sensor node, that is mounted on a DIN rail, which is established in industrial applications. The prototype is based on a commercial off the shelf perfboard as shown in fig. 5.

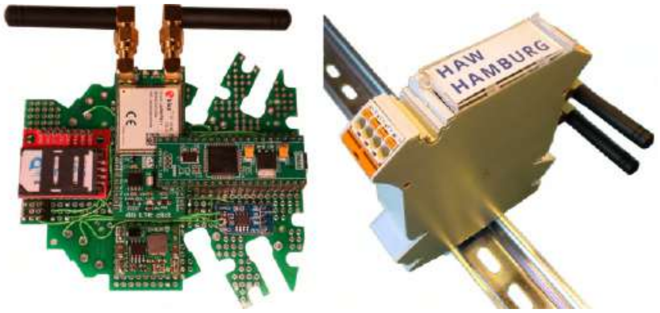


Fig. 5. A prototype focused on industrial IoT applications with 4G connectivity and a small formfactor to be mounted on a DIN rail.

In this approach the microcomputing node offers CAN bus connectivity, REPL interaction via UART and USB-VCP, as well as file system accessability over USB-MSC. A GPIO pin is configurable to interface OneWire sensors, act as 3.3V ADC/DAC or can be used as digital input/output. The focus of this prototype was to create an industry suitable sensor/actor node with 4G connectivity in a small form factor, which can be mounted onto standard DIN rails together with other industrial electronics like PLCs.

Multiple temperature sensors were connected to this prototype using the OneWire interface. It was then integrated as AWS Thing within the two AWS solutions shown in fig. 1 and described in section two. The sensor node successfully passed endurance tests in which it uploaded temperature data to its AWS Thing shadow with an update interval of one minute. It was located in a remote area with poor cellular network coverage and successfully maintained operation based on EDGE network. To harden the microcomputing node against connection losses, the MicroPython based modem driver had to be modified multiple times. This was successfully done over the air via the REPL, without requiring a complete firmware update.

VII. Conclusion

The presented microcomputing node introduces an innovative foundation for IoT sensor/actor nodes based on the established scripting language Python. With its trusted hardware in form of an intergrated smartcard, it offers sufficient security features, based on asymmetric and symmetric encryption. Thus a smartcard is a valid and flexible approach to meet the security requirements of modern IoT applications based on MCUs.

It is capable to interface a wide variety of sensors and actors and integrates these reliably in an IoT cloud solution. Because of the scripting capability the microcomputing node represents an accessible technology to a wide audience of application engineers, who do not need to have deep knowledge in programming constrained embedded systems. This is considered as key advantage of the presented microcomputing node, because experienced embedded system engineers have become an increasingly constrained resource.

Future research must show to what extent it is applicable to switch from classical C/C++ programming to script-based application development in the area of IoT. The presented microcomputing node has proven, that it is possible to solve complex measurement and monitoring tasks by the use of a scripting language on a constraint device according to current status.

References

[1] J. Callas, L. Donnerhacke, H. Finney, D. Shaw, R. Thayer “OpenPGP Message Format,” RFC 4880, DOI 10.17487/RFC4880, November 2007.

[2] J. Obermaier, S. Tatschner, “Shedding too much Light on a Microcontroller’s Firmware Protection,” Proceedings of the 11th USENIX Conference on Offensive Technologies, August 2017.

[3] J. Jongboom, “A high-integrity embedded file system,” https://os.mbed.com/blog/entry/littlefs-high-integrity-embedded-fs/, January 2018.

[4] Open Source: https://github.com/littlefs-project/littlefs, February 2021.

[5] A. Reineh, A. Martin, “Threat-Based Security Analysis for the Internet of Things,” Proceedings - 2014 International Workshop on Secure Internet of Things, SIoT 2014, 35-43, DOI 10.1109/SIoT.2014.10.

[6] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, A. Zanella, “IoT: Internet of Threats? A Survey of Practical Security Vulnerabilities in Real IoT Devices,” in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8182-8201, Oct. 2019, DOI 10.1109/JIOT.2019.2935189.

[7] C. Vorakulpipat, E. Rattanalerdnusorn, P. Thaenkaew, H. Dang Hai, “Recent challenges, trends, and concerns related to IoT security: An evolutionary study,” 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 405-410, DOI 10.23919/ICACT.2018.8323774.

An IoT Design Approach to Residential Energy Metering, Billing and Protection

Moses Oluwafemi Onibonoje, *Member, IEEE*
Department of Electrical/Electronics and Computer Engineering
Afe Babalola University
Ado-Ekiti, Nigeria
onibonojemo@abuad.edu.ng

***Abstract*—The fairness of electricity services lies in the provider being able to deliver the expected quality power, and recover the returns on investment. Energy theft, metering lapses, billing errors, and cumbersome payment procedures constitute the bulk of the non-technical power losses and contribute majorly to the incapacity of the electricity vendors to run a profitable business, and serve the customers effectively. A real-time approach has been identified as the innovative need at resolving the enumerated issues within the electricity value chain. This study is leveraging on the Internet of Things (IoT) technology to propose an extended modelled system capable of providing real-time data management, residential power system control, interactive platform for the vendors and consumers. The energy billing was modelled and developed from resourceful components. A website was developed with user-friendly interface. The unique features of the system design are the possibility of customers to load their electricity credit online, and the supplier being able to lock or disconnect any defaulting customer remotely.**

***Keywords*— electricity billing, energy theft, internet-of-things, residential power system, smart metering.**

I. Introduction

Electricity is a commodity with high-valued demand and supply, as it contributes extensively to the economy and better life for both corporate and individual entities. The monitoring and protection of the power devices in homes and facilities are very important. Meanwhile, ensuring that the actual electricity consumed is accurately tariffed is also crucial [1]. The fairness of any service relationship lies in the provider being able to deliver the expected quality service, and recover the invested capital through profitable returns on investment (ROI). The process of monitoring the track of electricity being consumed for verification is presently tedious and non-assuring, as manual meters are employed for data reading and recording. The metering strategy of the customers' facilities has experience various evolutions in recent times [2].

In Nigeria, energy metering by human data collection has been the major practice, until lately that pre-paid metering is gradually being introduced. The process involves human operators visiting the targeted facilities to collect the energy meter readings on a monthly basis, and use the collected to bill the customers. The process has been faulted by many errors associated with human interference and lapses, thereby resulting in consumers being billed on an estimated basis. Also, the pre-paid metering being introduced in batches as the new solution is being reported with cases of energy theft by connection bypass, and credit inconsistency [3]. The energy theft largely results in huge revenue loss by the providers. Energy theft is a major problem that manual, analog, digital and pre-paid meters have been unable to adequately curtail. In analog meters previously introduced, many consumers could slow down the rotation of the device by using the permanent magnet. In a similar way, digital and prepaid meters are being bypassed by consumers aiming at stealing electricity [4], [5]. The electricity theft, billing errors, and the customers payment debts constitute the bulk of the non-technical power losses, and the subsequent revenue losses. These are electricity problems that can therefore, be mitigated by real-time energy monitoring and billing over the internet, innovative technologies and integrated devices and infrastructure [6]. The growing population of electricity consumers [7], and the need to efficiently deliver their required electricity has made the adoption of Internet of Things (IoT) imperative.

The IoT is a brilliantly evolving technology which has integrated the global space with established and intelligent computing. IoT enables interconnectivity among smart devices such as mobile phones and computers, the internet, people and things. IoT has successfully formed a synergistic bond with other evolving technologies such as blockchain and machine learning to build autonomous systems that can communicate with, and control devices and objects necessary for various innovative solutions. The IoT technology has helped to propel the importance of improved electronic devices in the provision of better living condition, and enabling profitable business venturing through cost-effective and timely transactions and business executions. It has birthed attractive ideas for solution provision in various sectors including safety, protections, accounting, entertainment and security [8]–[12].

The IoT technology has been introduced in many other areas of the electricity value chain. The three cardinal sectors of power chain: generation, transmission and distribution have had benefits of integrated IoT technology [13]–[21]. A transition towards a digitized and decentralized mode of energy billing and payment will be a major milestone in ensuring profitable ROI for interested investors. The energy billing system will be efficiently and effectively managed using the IoT. The management will involve such real-time systems that can gather data on electricity connections status, metering, billing, consumer payment, and energy audit. Importantly, the stakeholders at both the supply and demand ends could be integrated into the interfacing platforms of such systems to benefit in the real-time engagements for energy solutions.

The study in [1] is herewith leveraging on IoT to extend the monitoring and protection of residential power systems capability to real-time billing, monitoring, recharge and control. This paper presents an advanced metering, billing and payment system with interactive application interface for consumers and the electricity supplier. The remaining part of the paper is structure as thus. Section II consists of the material and method of the system. Section III contains the results and discussion, while the conclusion is presented in section IV.

II. Methodology

The approach is to advance the IoT model and design to make available a system capable to provide real-time metering, billing, and recharge payment., in addition to the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

monitoring, control and protection of residential power system. The modified system is modelled, hardware and software designed, and then implemented. The system is designed based on the energy consumed on a line within a power supply section.

A. System Model

The energy consumed on a line, j over a period of time interval, t is defined by $e_{j,t}$ as indicated in equation (1).

$$e_{j,t} = P_{j,t}t = I_{j,t}V_{j,t}\cos\theta \qquad (1)$$

where $P_{j,t}$ defines the power consumed on the line within the period.

The tariff mechanism adopted is the progressive pricing technique referred to as the stepwise energy pricing (SEP) [22]. The assumption is that the energy quantity is categorized into n-steps, with each step corresponding to a unit price increase with the step. The monthly clearing price is equal to the sum of product (SOP) of electricity quantity across all the steps and the corresponding price. Considering the SEP schematic as shown in Fig 1, the energy quantity vector across all the steps is given by equation (2) while the unit pricing vector is given by equation (3).

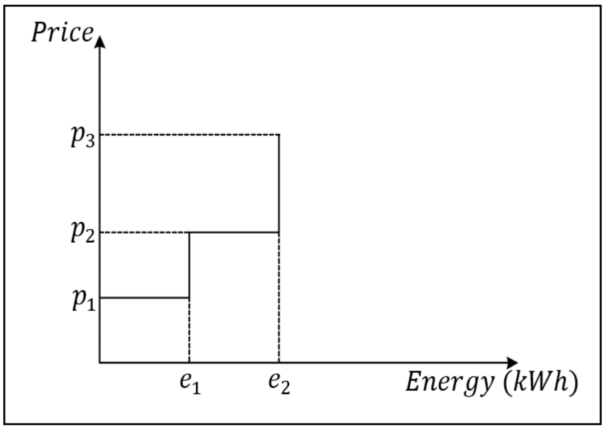


Fig. 1. The schematic of the SEP model

$$e_j = [e_{1,}, e_{2,}, \dots \dots .] \qquad (2)$$

$$p = [p_1, p_2, p_3] \qquad (3)$$

The monthly clearing price, P_c is a function of the price and the quantity of energy usage on a stepwise basis, as expressed by equation (4), where e_T is the total energy usage after the implementation of SEP.

$$P_c = p_1 e_1 + p_2(e_2 - e_1) + p_3(e_T - e_2) \qquad (4)$$

The stepwise tariff model can be optimized to encourage energy efficiency, using the genetic algorithm analysis.

B. Harware Design

The system architecture is as shown in the block diagram of Fig 2. A unique website was designed for the system to connect and interface the smart devices with the internet.

The building blocks of the system include four main units: the sensor module for current, voltage and temperature parameters measurement; the transceiver unit being integrated with the WIFI and GPRS modules for internet connection for the website and the mobile application; the relay unit which performs the make-or-break connection to the end power users; and the Arduino microcontroller which processes the received signal and output control signal to the relay unit. The ACS712 current sensor combined with the LM35 temperature sensor, an input voltage capacity measurement of about 400Vac. The radio frequency XBee 2.4 GHz by Digimesh was the transceiver module integrated with the antenna as shown in the hardware connection of Fig 3.

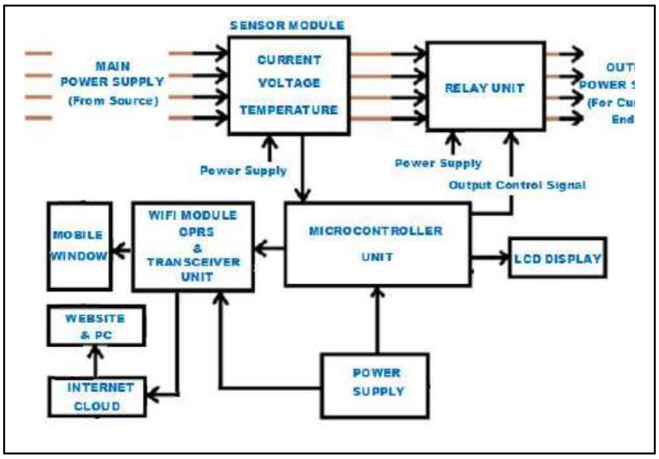


Fig. 2. The system block diagram

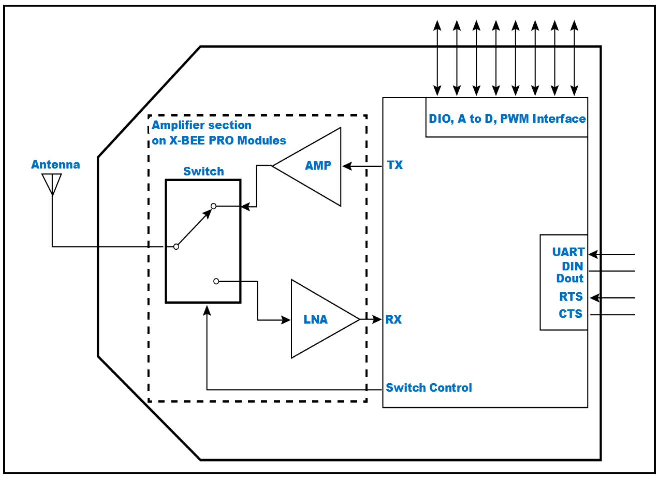


Fig. 3. Internal block diagram of the transceiver module

C. Software Design

The software unit of the system is designed for a two-way communication feature. The main functions include the initialization with the commands over the internet website and to integrate the microcontroller and adjoining electronic units for control. The data for power consumption and the corresponding the billing information are directly transferred to the cloud and available on the website in real-time. The algorithm for the operational details of the system is as shown in Fig 4.

III. Results and Discussion

The resulting system from the implementation of the design, and the operation of the developed user platforms are presented in this section. The hardware components of the IoT based metering system was developed as shown in Fig 5. It consists of an Arduino nano being connected to an integrated transceiver with the WIFI and GPRS module for internet connectivity. Also, there is a four-channel relay section for disconnecting the supply to the end user, while the LCD unit displays all the data in real-time.

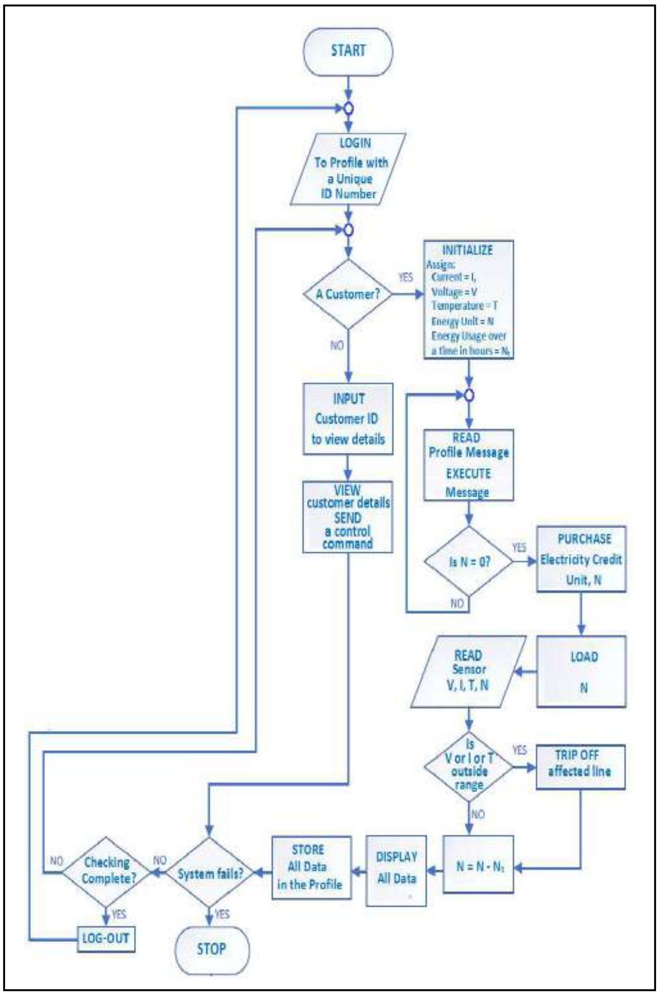


Fig. 4. Flowchart of the system operation

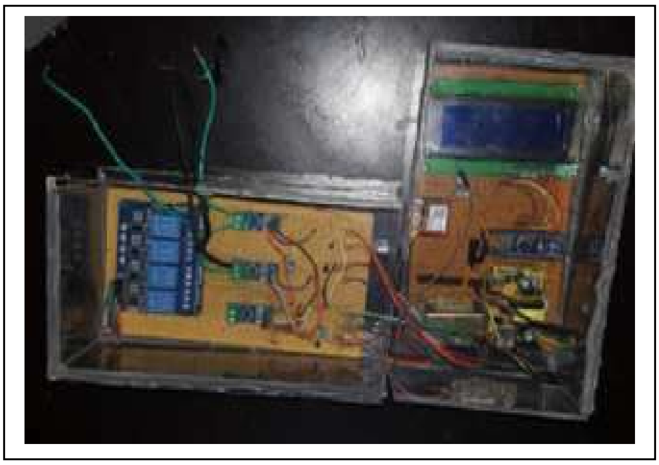
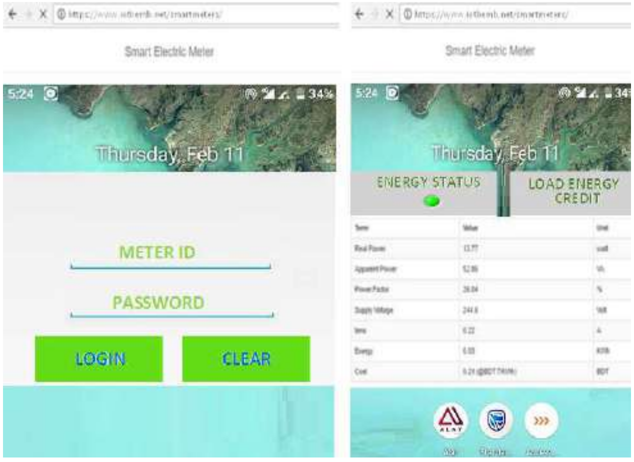


Fig. 5. The developed hardware of the IoT-based metering system

The webpage and mobile application window developed is as shown in Fig 6. The webpage has an easy-to-use interface with good outlook. A user needs a unique identity number (ID) to login and navigate through the page. After profile login, there is an option to select between a customer or supplier. A customer selects accordingly and read the message available in the inbox. If there is no electricity credit, it is indicated, and the he cannot go further unless the system is recharged with credit unit. The system is linked with a payment platform, to be activated, for automated payment.



Term	Value	Unit
Real Power	13.77	watt
Apparent Power	52.86	VA
Power Factor	26.04	%
Supply Voltage	244.8	Volt
Irms	0.22	A
Energy	0.03	KWh
Cost	0.21 (@BDT 7/kWh)	BDT

Fig. 6. The user application interface

The credit unit can be purchased and loaded manually. The sensors readings and the credit unit value are collected and stored. If any of the benchmarked parameters: current, voltage or temperature exceeds the threshold, the affected line is tripped off. All data and notifications are displayed on the LCD, and stored in the customer profile as a message. The supplier representative can also login with a unique ID, view the status of any customer power supply facility to monitor any discrepancies in the energy usage, and sends a control command to disconnect a defaulting customer. There is an automated 3 Kb/s data refreshing of the website every five seconds to allow for data update in real-time. Therefore, the data usage for the refresh is minimal.

IV. Conclusion

In this paper, an Internet of Things (IoT) based energy metering system for residential facilities has been developed. The study has extended and improved the existing work on the real-time monitoring and protection of a residential power system using IoT technology. The various units of the system were implemented on a resourceful-selection basis of the electronic devices. A user-friendly website and mobile application interface were also designed for real-time monitoring, collection of the metering and billing information for both the supplier and the customers with exclusive access codes. Previous methods and approaches adopted in energy metering and billing such as demand side load method, analog meter reading and digital prepaid metering were examined with relative flaws which have, therefore, been addressed with this IoT developed system. The developed IoT-based system is a combination of subsystems as energy meter, billing device, payment platform and residential power system controller. The customer credit is counts down according to electricity unit's usage and automatically logs off whenever the credits exhausted. The unique features of the system design are the possibility of customers to load their electricity credit online, and the supplier being able to lock or disconnect any defaulting customer remotely.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Ass. Professor N. Nnamdi and Ass. Professor P. Bokoro for their contribution to the earlier part of the work which had been previously published.

REFERENCES

[1] M. O. Onibonoje, N. I. Nwulu, and P. N. Bokoro, "An Internet-of-Things Design Approach to Real-Time Monitoring and Protection of a Residential Power System," 2019, doi: 10.1109/SEGE.2019.8859879.

[2] S. H. Mir, S. Ashruf, Y. Bhat, N. Beigh, and others, "Review on smart electric metering system based on GSM/IOT," *Asian J. Electr. Sci.*, vol. 8, no. 1, pp. 1–6, 2019.

[3] K. Ashna and S. N. George, "GSM based automatic energy meter reading system with instant billing," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 2013, pp. 65–72.

[4] M. U. Hashmi and J. G. Priolkar, "Anti-theft energy metering for smart electrical distribution system," in *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, 2015, pp. 1424–1428.

[5] S. Darby, "Smart metering: what potential for householder engagement?," *Build. Res. Inf.*, vol. 38, no. 5, pp. 442–457, 2010.

[6] R. Sathyapriya and V. Jeyalakshmi, "Hardware implementation of IOT based energy management theft detection and disconnection using smart meter," *Malaya J. Mat.*, vol. 5, no. 2, pp. 4177–4180, 2020.

[7] P. Wang, J. Y. Huang, Y. Ding, P. Loh, and L. Goel, "Demand side load management of smart grids using intelligent trading/metering/billing system," in *2011 IEEE Trondheim PowerTech*, 2011, pp. 1–6.

[8] K. Mandula, R. Parupalli, C. H. A. S. Murty, E. Magesh, and R. Lunagariya, "Mobile based home automation using Internet of Things (IoT)," in *2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2015, pp. 340–343.

[9] M. O. Onibonoje, P. N. Bokoro, N. I. Nwulu, and S. L. Gbadamosi, "An IoT-Based Approach to Real-Time Conditioning and Control in a Server Room," 2019, doi: 10.1109/IDAP.2019.8875880.

[10] S. Mahmud, S. Ahmed, and K. Shikder, "A smart home automation and metering system using internet of things (IoT)," in *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2019, pp. 451–454.

[11] M. O. Onibonoje and T. O. Olowu, "Real-time remote monitoring and automated control of granary environmental factors using wireless sensor network," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, pp. 113–118.

[12] M. O. Onibonoje and N. Nwulu, "Synergistic Technologies for Precision Agriculture," in *Artificial Intelligence and IoT-Based Technologies for Sustainable Farming and Smart Agriculture*, IGI Global, 2021, pp. 123–139.

[13] X. Fei and G. Tian, "Optimization of Communication Network Fault Identification Based on NB-IoT," *Microprocess. Microsyst.*, vol. 80, p. 103531, 2021.

[14] M. Farhoumandi, Q. Zhou, and M. Shahidehpour, "A review of machine learning applications in IoT-integrated modern power systems," *Electr. J.*, vol. 34, no. 1, p. 106879, 2021.

[15] V. Motjoadi, P. N. Bokoro, and M. O. Onibonoje, "A review of microgrid-based approach to rural electrification in South Africa: Architecture and policy framework," *Energies*, vol. 13, no. 9, May 2020, doi: 10.3390/en13092193.

[16] V. Motjoadi, P. N. Bokoro, and M. O. Onibonoje, "Review of Switching and Control Techniques of Solar Microgrids," in *2020 IEEE PES/IAS PowerAfrica*, 2020, pp. 1–5.

[17] O. E. Aluko, T. E. Fabunmi, M. O. Onibonoje, and J. O. Dada, "A Clean and Renewable Energy-Utility Solution in Nigeria," in *2020 IEEE PES/IAS PowerAfrica*, 2020, pp. 1–5.

[18] O. E. Aluko, M. O. Onibonoje, and J. O. Dada, "A Review of the Control System Roles in Integrating Renewable Energy into the National Grid," in *2020 IEEE PES/IAS PowerAfrica*, 2020, pp. 1–5.

[19] S. Tiwari, V. Agrawal, S. K. Jain, and P. K. Shrivastava, "Observation and Control of Smart Grid Using IoT and Cloud Technology," in *Intelligent Computing in Control and Communication*, Springer, 2021, pp. 559–575.

[20] S. Sambhi, S. Sambhi, and V. S. Bhadoria, "IoT-Based Optimized and Secured Ecosystem for Energy Internet: The State-of-the-Art," *Internet Things Bus. Transform. Dev. an Eng. Bus. Strateg. Ind. 5.0*, pp. 91–125, 2021.

[21] D. Sun *et al.*, "Research on IoT architecture and application scheme for smart grid," in *Proceedings of the 9th International Conference on Computer Engineering and Networks*, 2021, pp. 921–928.

[22] C. Li *et al.*, "A new stepwise power tariff model and its application for residential consumers in regulated electricity markets," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 300–308, 2012.

Wireless versus Wired Network-on-Chip to Enable the Multi-Tenant Multi-FPGAs in Cloud

Qaiser Ijaz and El-Bay Bourennane
ImViA Laboratory, University of Burgundy, Dijon 21000, France
qaiser_ijaz@etu.u-bourgogne.fr, ebourenn@u-bourgogne.fr

***Abstract*— The new era of computing is not CPU-centric but enriched with all the heterogeneous computing resources including the reconfigurable fabric. In multi-FPGA architecture, either deployed within a data center or as a standalone model, inter-FPGA communication is crucial. Network-on-chip exhibits a promising performance for the integration of one FPGA. A sustainable communication architecture requires stable performance as the number of applications or users grows. Wireless network-on-chip has the potential to be that communication architecture, as it boasts the same performance capability as wired solutions in addition to its multicast capacities. We conducted an exploratory study to investigate different performance parameters using the Noxim simulator. The preliminary results showed better average delay and network throughput for wireless network-on-chip as compared to the wired counterpart.**

Keywords—Wireless NoC, multi-FPGAs, reconfigurable computing

I. Introduction

Datacenter architectures provide for the need of computing workloads today, that is acceleration for more performance. Datacenters have heterogeneous computing resources, comprising of different processing fabrics that are suitable for a certain set of operations [1]. We have more choices than ever. GPUs offer better parallel performance, more efficient computing, and an easy-to-use programming model. On the other hand, FPGAs bring higher performance-per-watt, improved hardware acceleration, and lower inter-device latency. Assigning the right resource for the right job is mandatory for high performance and maximum utilization at the system level. Enhanced scalability and parallelism increase the possibilities for optimization at the algorithmic level, improved architecture and micro-architecture, and design methods focused on the target platform [2]. FPGAs is a competitive computing resource known for lower power consumption. The lifetime cost of data centers is majorly the cost of electrical power [3]. Therefore, data centers are looking for energy-efficient computing fabric and that opens new horizons in High-Performance Reconfigurable Computing (HRC).

FPGAs have been used to accelerate datacenter services in both academia and industry. The choice of FPGA over GPU in Catapult [4] was driven mainly by power demand. The proprietary search engine of Microsoft achieved a 90x speedup as compared to the software-only solution, at the cost of a 10% increase in power consumption. Sustainable deployment of FPGAs in data centers requires an operating system to be designed in such a way that it can serve multiple users simultaneously. Without this vital virtualization characteristic, the investment would neither be economically feasible nor sustainable. Communication architecture and an operating system to virtualize the reconfigurable fabric in heterogeneous datacenters are two prime challenges to address in this area. We published a comprehensive review [5] of the area last year to direct the scientific community towards the most pressing problems. In this study, we explored the Wireless Network on Chip (WiNoC) to solve the bottleneck communication problem in the context of multi-tenant multi-FPGAs in a datacenter. Please refer to the mentioned review for the detailed context of the problem.

Section II presents the state of the art on multi-FPGA virtualization techniques. Section III presents the theoretical foundation of WiNoC as a potential communication architecture in the given context. Section IV presents the results with an ample discussion on the limitations of this study.

II. Multi-Node Level Virtualization

The nomenclature of the node is being used in the continuity of the previous research that simply means an FPGA. The architecture depends on the organization of FPGAs to the host as illustrated in Fig. 1. In the direct model, nodes directly communicate, while the link represents the interface. In the slave model, the FPGAs are hosted on CPUs via PCIe. This means if FPGA wants to communicate with another FPGA then it must send the data through CPUs and network. In a standalone model, the FPGAs and CPUs are accessible to every node directly through the network. The hybrid models can be attained by combining two models to meet the specific requirements.

Byma and colleagues focused to minimize the virtualization overhead of standard size datacenter that provides commercial cloud services [6]. They managed to attain remarkable performance when compared to regular virtual machines while reducing the iteration time of design. Kondel and his colleagues presented the work on maximizing the utilization of high-end FPGAs by paravirtualization [7]. They provided homogeneous virtualized FPGA regions for multiple users. This accommodating multi-tenancy approach allows the individual resources to embrace the user requirements. Zhang and colleagues implemented an operating system to share the single FPGA chip among multiple users at the run-time along with an upgraded resource manager [8]. These works did not write a detailed account of communication architecture and interfaces. Weerasinghe et al. [9] illustrated a different approach where FPGAs are connected through the datacenter network. This decoupled approach is useful for hyper-scale data centers where FPGA acts as an equal and distinct computing resource. They provided the detailed architecture of the system combined with an analysis on resource estimation, from a scaling perspective. The development of tightly coupled platforms is a recent trend, for example, IMB CAPI and Intel HARP collaborated to yield. The community responded to the call for proposals and since then significant contributions have been made using these platforms.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

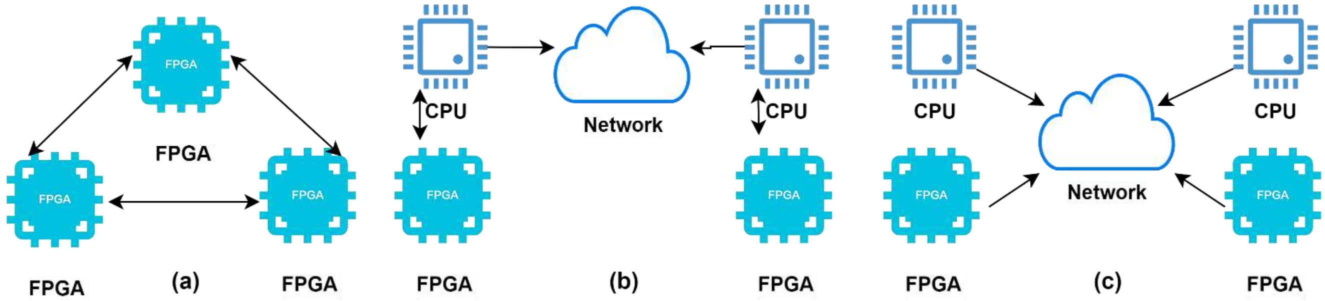


Fig. 1. Possible Arrangement and Connectivity of multiple FPGAs: a) Direct Model; b) Slave Model with PCIe interface; c) Standalone Model

Some recent examples are [10,11]. There are three sub-classes of multi-node virtualization that will be briefly revisited for the readers.

A. Custom Clusters

In parallel computing architecture, the concept of the systolic array model is based on custom clusters, where every node processes the data, and the processed data is moved from one node to another through a first-in-first-out (FIFO) buffer or network semantics. Some of these architectures [12–15] make use of fast series transceivers with FIFO buffers, Peer to Peer (P2P) connection MaxRing, and Peripheral Component Interconnect Express (PCIe) links, to transmit data across several nodes. Customized designs permit direct communication among the nodes through specific network connections. A systolic array model has been exploited by the cluster of 512 FPGAs [16] to distribute and execute the computations on multiple FPGAs. This work is a good representative of a small to medium scale datacenter.

B. Frameworks

Frameworks make use of the conventional server-client architecture, where the computational part is dedicated to one or more FPGAs and the CPU server manages the rest, including application-related data, configuration, and scheduling. The central part of this architecture is data management. Architectures specified for CPU are extendible to FPGAs. For example, the concept of the MapReduce framework has been extended where the mapping and reduction operations are accomplished by FPGA accelerators [17-19] in the same way as in CPU client-server architecture. The advantage of these frameworks is that they helped in overcoming the gap between the heterogeneity of datacenters. An example of such a cluster consisting of FPGAs and GPUs based on MapReduce is [20]. In another work [21], the researchers used the Apache Spark to accommodate the FPGAs by extending the java virtual machine (JVM) framework, however, this requires precision and increases communication overhead.

OpenCL is utilized through the Xilinx SDAccel framework to assign the data to multiple FPGAs by using an abstract layer and managing a transparent directory to virtualize the FPGAs at the lower abstraction level [22]. A few FPGA works [23] discussed an approach where multiple FPGAs can be shared by one group but configured within a setup of a matching accelerator. This approach comes with the drawback of occupying a complete FPGA that results in under-utilization, but it can be improved by the automation of the scaling algorithm. A similar approach has been proposed in [24] based on Hadoop YARN that comes with the advantage of the ease of programming. In a heterogeneous computing environment, the performance relies on a function of execution strategy. To explore the alternative execution strategies in disaggregated environments, a work based on the evaluation platform presented in [25] can be useful.

C. Cloud Service Architectures

Cloud service architectures promise computational correctness and guarantee Quality of Service (QoS) while abstracting the underlying complexity. The user is not concerned about the assigned computational node if the required acceleration or high-performance is delivered. Amazon EC2 F1 instance does not fall under this category but Microsoft's Catapult achieved a substantial speed-up in search ranking. This is a hybrid architecture as it can distribute the acceleration jobs to either of the underlying node, standalone FPGA, or a host CPU. The same performance was achieved by Baidu [26] for deep neural networks. While in [27], the FPGAs have been used as co-processors in complex problems for exploiting the multiple data streams. The architectures with network support broaden the choice of connectivity, which allows the CPU provisioning either as embedded on-chip or as a soft-core. OpenStack is known as the most common method that directly allows the user to program the FPGA through a physical or virtual address [6, 22, 28]. It offers the flexibility to the user for exploiting socket or remote routine approach to establish connectivity with an FPGA.

III. Network on Chip as Communication Backbone

In most high-performance architectures the data transfers are limited due to communication architecture and memory hierarchy, as summarized in [29,30]. The use of NoC and dedicated links in a system with a huge number of processing cores are an effective replacement for buses [31-33]. NoC comprises of multiple tunable parameters like routing algorithm, network architecture, flow control, and network topology. These endless possibilities and given the rich FPGA resources today, System on Chip (SoC) designers prefer to use it as a central communication component. It ensures the high communication bandwidth with low latency as compared to the alternative methods of communication within an FPGA.

A. Basic Architecture

The architecture of the 2D mesh of size 2x2, comprising of four routers is shown in Fig. 2, presenting exposed edge ports, that are unused ports of the corner routers. Each router has five ports, in which the local port is connected to the network interface whereas the rest of the four ports communicate with the neighboring routers or processing elements. A network interface is connected to the core,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

which represents any processing element.

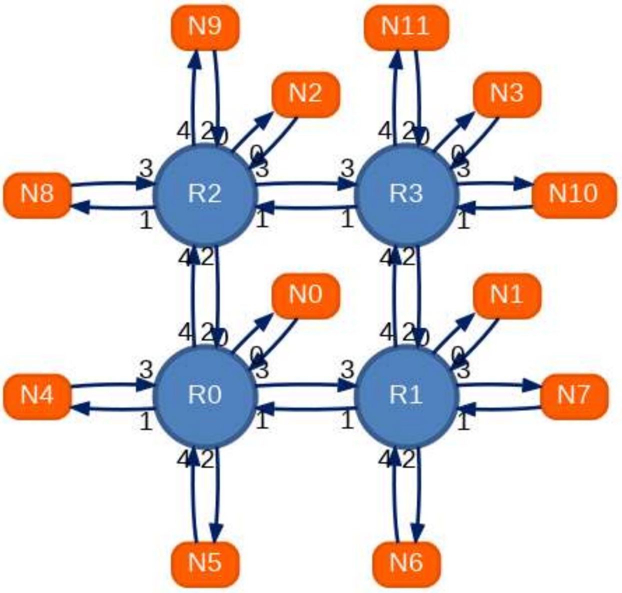


Fig. 2. A typical 2D mesh of size 2x2 with visible ports.

B. Network Layered Model

The layered model is shown in Fig. 3 along with the OSI reference model.

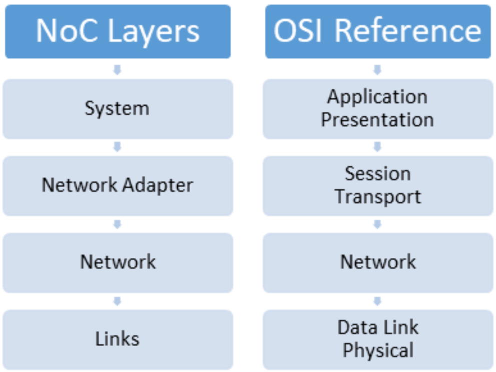


Fig. 3. Network Layered Model

Layered representation of network-on-chip contributes to the better understanding, as the top two layers of the OSI reference model are combined in the system layer. Likewise, session and transport layers correspond to the network adapter layer in NoC and the last two layers of the reference model are represented by links in NoC.

C. Wireless NoC

A case based on wireless network-on-chip was proposed by a scientist in 2012 as a potential backbone for all the multicore and manycore chips [35], another derived architecture using the same source is shown in Fig. 4. The router has a similar architecture but with an addition of a radio hub for wireless communication. The organization is illustrated in Fig. 6. The purpose of a radio hub is to govern the communication, typically single hop, between distant routers [36].

D. Objective

This exploratory study is designed to have a proof of concept of this naïve idea that wireless network-on-chip can enable inter-FPGA communication in the given context. For the comprehensive state of the art and a deeper understating of the context of the problem, readers are invited to refer to our previous published work [37-45]

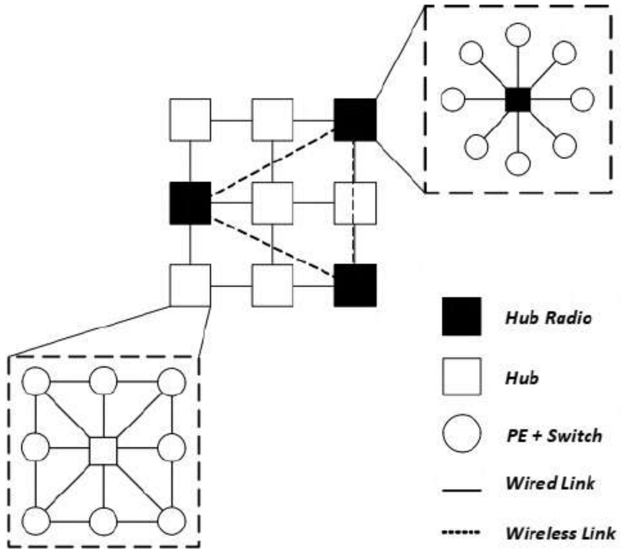


Fig. 4. 2D Mesh of Size 3x3 with Radio Hub

The inquiry-based iterative simulations were designed to establish the outcome. Fig. 5 depicts the basic idea through a concept diagram that we are aiming to achieve for multi-FPGAs in the given context.

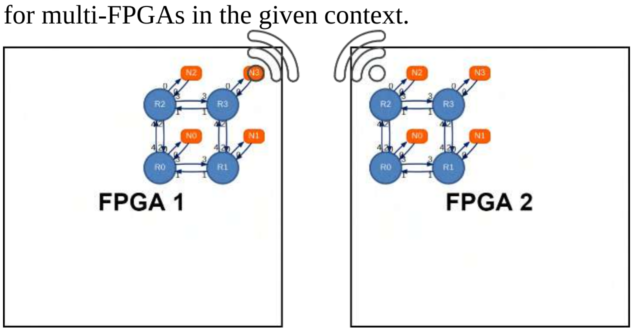


Fig. 5. Depiction of the Basic Idea

E. Choice of NoC Simulator

Simulation provides an early estimation of the decision variables that drive the physical implementation in the future. There are many simulators with a high level of abstraction that offer high code portability but lower efficiency, which we will discuss one by one.

Booksim [46] is a cycle-accurate simulator developed in C++. The first version was not intended for a specific on-chip environment but mostly a generic simulator. Booksim2 provides a wide diversity of topologies, routing algorithms, and several options to customize the micro-architecture of routers to simulate.

DARSIM [47] is a cycle-level, parallel simulator, that allows simulating both 2D and 3D mesh architectures. It provides an advanced set of NoC parameters such as different virtual channel allocation and memory models. One of the strengths of this simulator is the ability of the hardware configuration, such as bandwidth, pipeline depth, and geometry. Besides, it allows to split the tasks between cores equally and achieves cycle-accurate simulations.

Heterogeneous NoC Simulator (HNOCS) [48] is dedicated to heterogeneous NoC architectures and is based on OMNet++. OMNet++ provides C++ APIs to a

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

wide range of services to describe in detail the network topology. It also provides parallelism, various Quality-of-Service (QoS), different arbitrary technologies, and power estimation. It offers three different router types, asynchronous, synchronous, and synchronous virtual output queue, and performance statistics such as throughput, VC acquisition, and transfer latency.

Nigram [49] is a cycle-accurate and discrete event simulator developed in SystemC. It provides various network configuration commands to simulate different NoC architectures such as routing algorithms, topologies, flow control techniques. The simulation statistics include throughput and latency.

Noxim [50] is a low-level, open-source, and cycle-accurate simulator written in C++/SystemC. Noxim provides various configuration parameters such as packet and buffer sizes, packet injection rate, different routing algorithms, traffic distributions, structures, and topologies. In addition to the wired NoC simulation, Noxim also supports Wireless NoC (WiNoC) [51] evaluation and provides power consumption, throughput, and latency as performance analysis. Table 1 summarizes the NoC simulators.

TABLE 1. NOC SIMULATORS SUMMARY

Reference	Simulator	Abstraction Level	Supported Topologies
[46]	BookSim	High	Multiple
[47]	DARSIM	High	All
[48]	HNOCS	High	Mesh
[49]	Nigram	Low	All
[50]	Noxim	Low	Multiple

IV. RESULTS AND DISCUSSION

Table 2 present the evaluation environment with all the fixed and experimental configurations. The evaluation platform was based on 2D mesh platform. To ensure correctness, the simulation was repeated three times. The warmup and simulation times were 1000 and 10000 clock cycles. The routing algorithm was fixed XY as this evaluation of the routing algorithm was not the objective of the study, however, this can be studied separately. Changes in average delay, network throughput, and the total energy were observed under various traffic distributions, random, transpose, shuffle, and butterfly. The use of synthetic traffic is a limitation to address, as it does not necessarily reflect the traffic patterns of a real application. We collated the performance and power analysis for a wireless and wired network-on-chip under similar conditions. We designed the experiment for three different sizes as the network-on-chip will grow as the users or applications increase.

Results are straightforward but we would like to discuss a few trends. Better average delay and network throughput were observed for wireless network-on-chip as compared to the wired counterpart, with an exception for the butterfly traffic distribution and size 16x16, which can be an outlier. Total energy consumption in wireless network-on-chip is higher for obvious reasons, radio hub and wireless transmission that is a function of average wireless utilization, that varies in each simulation iteration. As the size of the network increases, the throughput reaches the same point, which can be seen in the second last column of the table. Although the results look favorable, yet hard to quantify the delay reduction or increase in throughput from this preliminary study due to its limitations, absence of real traffic, and complete system simulation. Recent research [52], however, has reported a 45% of delay reduction with real traffic and system simulation.

V. CONCLUSION

Sustainable deployment of FPGAs in the data center requires multi-tenancy and a dynamically scalable communication architecture. Dynamically growing network-on-chip showed significant performance. Given the context, the choice of wireless network-on-chip as an inter-FPGA communication backbone has been explored under certain configurations using Noxim simulator. The observed preliminary results showed better average delay and network throughput for wireless network-on-chip as compared to the wired counterpart. The results are however limited by simulator used, synthetic traffic patterns, absence of system-level simulation, and lack of support for heterogeneity. Further investigations and performance comparisons with reference to benchmarks are required to draw any concrete conclusion.

TABLE 2. UNCHANGED AND EXPLORATORY CONFIGURATIONS WITH RESULTS FOR DIFFERENT TRAFFIC DISTRIBUTION AND SIZES

Unchanged Configurations:
Topology = 2D Mesh
Routing Algorithm = XY
Warm Up Time = 1000 cycles
Simulation Time = 10000 cycles

Performance and Power Analysis Legend:
AD = Average delay measured in cycles
NT = Network throughput measured in flits/cycle
TE = Total energy measure in joules

		NoC Size = 4 x 4			**NoC Size = 8 x 8**			**NoC Size = 16 x 16**		
		AD	**NT**	**TE**	**AD**	**NT**	**TE**	**AD**	**NT**	**TE**
Random Distribution	Wired	576.397	0.599556	2.12e-06	1329.21	2.62967	8.89e-06	407.462	18.897	3.86e-05
	Wireless	**11.1289**	**1.27578**	**3.42e-06**	**23.804**	**5.12711**	**9.91e-06**	**377.227**	**18.9016**	**4.02e-05**
Transpose Distribution	Wired	162.765	1.09	2.11e-06	227.148	4.75933	8.87e-06	466.332	13.6593	3.55e-05
	Wireless	**11.0786**	**1.29944**	**3.48e-06**	**81.2728**	**5.02878**	**1.02e-05**	**452.839**	**13.6487**	**3.73e-05**
Shuffle Distribution	Wired	633.628	0.800111	2.09e-06	249.256	4.50322	8.70e-06	873.656	11.6328	3.51e-05
	Wireless	**8.18781**	**1.26967**	**3.43e-06**	**26.8224**	**5.15322**	**1.00e-05**	**675.111**	**11.8629**	**3.64e-05**
Butterfly Distribution	Wired	400.513	1.09867	2.06e-06	17.4902	5.215	8.45e-06	131.697	18.1358	3.44e-05
	Wireless	**7.10786**	**1.24333**	**3.43e-06**	**14.2698**	**5.06633**	**9.97e-06**	**135.339**	**18.2032**	**3.61e-05**

References

[1] F. A. Escobar, X. Chang and C. Valderrama, “Suitability Analysis of FPGAs for Heterogeneous Platforms in HPC,” IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 2, pp. 600-612, February 2016.

[2] L. Carloni, F. D. Bernardinis, C. Pinello, A. Sangiovanni-Vincentelli and M. Sgroi, “Platform-based design for embedded systems,” in Embedded Systems Hand-book, CRC Press, 2005.

[3] R. Inta, D. J. Bowman, and S. M. Scott, “The Chimera: an off-The-shelf CPU/GPGPU/FPGA hybrid computing platform,” International Journal of Reconfigurable Computing, vol. 2012, p. 2, 2012.

[4] A. Putnam et al., “A reconfigurable fabric for accelerating large-scale datacenter services,” Proc. Int'l Symp. Computer Architecture (ISCA), pp. 13-24, 2014.

[5] Q. Ijaz, E. B. Bourennane, A. K. Bashir, and H. Asghar, “Revisiting the High-Performance Reconfigurable Computing for Future Datacenters,” Future Internet, vol. 12, no. 4, April 2020.

[6] S. Byma, J. G. Steffan, H. Bannazadeh, A. L. Garcia and P. Chow, “FPGAs in the cloud: Booting virtualized hardware accelerators with OpenStack,” Proc. Annual Symp. on Field-Programmable Custom Computing Machines (FCCM), pp. 109-116, May 2014.

[7] O. Knodel, P. R. Genssler and R. G. Spallek, “Virtualizing reconfigurable hardware to provide scalability Proc.in cloud architectures,” Proc. Advances in Circuits Electronics and Micro-electronics (CENICS), 2017.

[8] J. Zhang et al., “The Feniks FPGA operating system for cloud computing,” 8th Asia–Pacific Workshop on Systems (APSys), pp. 1-7, 2017.

[9] J. Weerasinghe, F. Abel, C. Hagleitner and A. Herkersdorf, “Enabling FPGAs in hyperscale data centers,” Proc. Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), Aug. 2015.

[10] P. Caldeira et al., “From Java to FPGA: An Experience with the Intel HARP System,” Proc. Int'l Symp. on Computer Architecture and High Performance Computing (SBAC-PAD), pp. 17-24, 2018.

[11] L. Feng, J. Zhao, T. Liang, S. Sinha and W. Zhang, “LAMA: Link-Aware Hybrid Management for Memory Accesses in Emerging CPU-FPGA Platforms,” Proc. 56th Annual Design Automation Conference (DAC), 2019.

[12] O. Pell and V. Averbukh, “Maximum performance computing with dataflow engines,” Computing in Science & Engineering, vol. 14, no. 4, pp. 98-103, July 2012.

[13] K. Fleming, Y. Jung, M. Adler and J. Emer, “The LEAP FPGA operating system,” Proc. Field-Programmable Logic and Applications (FPL), pp. 1-8, 2014.

[14] M. Vesper, D. Koch, K. Vipin and S. A. Fahmy, “JetStream: An open-source high-performance PCI Express 3 streaming library for FPGA-to-Host and FPGA-to-FPGA communication,” Proc. Field-Programmable Logic and Applications (FPL), pp. 1-9, 2016.

[15] M. Jacobsen, D. Richmond, M. Hogains and R. Kastner, “RIFFA 2.1: A reusable integration framework for FPGA accelerators,” ACM Transactions on Reconfigurable Technology and Systems, vol. 8, no. 4, September 2015.

[16] M. Yoshimi et al., “A performance evaluation of CUBE: one-dimensional 512 FPGA cluster,” Reconfigurable Computing: Architectures, Tools and Applications, 2010.

[17] Y. Shan et al., “FPMR: MapReduce Framework on FPGA,” Proc. Int'l Symp. Field-Programmable Gate Arrays, pp. 93-102, 2010.

[18] Z. Wang, S. Zhang, B. He and W. Zhang, “Melia: A mapreduce framework on opencl-based fpgas,” IEEE Transactions on Parallel and Distributed Systems, no. 99, pp. 1-1, 2016.

[19] J. H. Yeung et al., “Map-Reduce as a Programming Model for Custom Computing Machines,” Proc. Int'l Symp. Field-Programmable Custom Computing Machines (FCCM), 2008.

[20] K. Tsoi and W. Luk, “Axel: A heterogeneous cluster with FPGAs and GPUs,” Proc. Int'l Symp. on Field Programmable Gate Arrays, pp. 115-124, 2010.

[21] Y. T. Chen, J. Cong, Z. Fang, J. Lei and P. Wei, “When apache spark meets FPGAs: a case study for next-generation DNA sequencing acceleration,” 8th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 16), June 2016.

[22] N. Tarafdar, T. Lin, E. Fukuda, H. Bannazadeh, A. Leon-Garcia and P. Chow, “Enabling flexible network FPGA clusters in a heterogeneous cloud data center,” Proc. Int'l Symp. Field-Programmable Gate Arrays (FPGA'17), pp. 237-246, 2017.

[23] A. Iordache et al., “High performance in the cloud with fpga groups,” Proc. Int'l Conf. on Utility & Cloud Computing, 2016.

[24] M. Huang et al., “Programming and runtime support to blaze fpga accelerator deployment at datacenter scale,” Symposium on Cloud Computing (SoCC), 2016.

[25] D. Theodoropoulos, N. Alachiotis and D. Pnevmatikatos, “Multi-FPGA Evaluation Platform for Disaggregated Computing,” Proc. 25th Annual Int'l Symp. on Field-Programmable Custom Computing Machines (FCCM), 2017.

[26] J. Ouyang et al., “SDA: Software-Defined Accelerator for Large-Scale DNN Systems,” Proc. Hot Chips Symposium (HCS), Aug. 2014.

[27] E. El-Araby, I. Gonzalez and T. El-Ghazawi, “Virtualizing and Sharing Reconfigurable Resources in High-Performance Reconfigurable Computing Systems,” Proc. HPRCTA Workshop at SC'08, Nov. 2008.

[28] J. Weerasinghe, R. Polig, F. Abel and C. Hagleitner, “Network-attached FPGAs for data center applications,” Proc. Int'l Conf. Field-Programmable Technology, 2016.

[29] R. Bittner, E. Ruf, and A. Forin, “Direct GPU/FPGA communication via PCI express,” Cluster Computing, pp. 1-10, 2013.

[30] R. Mueller, J. Teubner and G. Alonso, “Streams on wires: a query compiler for FPGAs,” VLDB Endowment, vol. 2, no. 1, August 2009.

[31] W. J. Dally and B. Towles, “Route packets not wires: on-chip interconnection networks,” Proc. 38th Annual Design Automation Conference (DAC), pp. 684-689, 2001.

[32] S. Yazdanshenas and V. Betz, “Interconnect solutions for virtualized field-programmable gate arrays,” IEEE Access, vol. 6, pp. 10 497-10 507, February 2018.

[33] A. Husnain and Q. Ijaz, “Defect Rate Analysis & Reduction of MPSOC Through Run Time Reconfigurable Computing with Multiple Caches,” Proc. Int'l. Conf. Computer Technology and Development, 2011.

[34] O. A. de Lima, W. N. Costa, V. Fresse and F. Rousseau, “A survey of NoC evaluation platforms on FPGAs,” Proc. Int'l Conf. Field-Programmable Technology, 2016.

[35] S. Deb, A. Ganguly, P. P. Pande, B. Belzer and D. Heo, “Wireless NoC as interconnection backbone for multicore chips: Promises and challenges,” IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 2, no. 2, pp. 228-239, June 2012.

[36] V. Catania, A. Mineo, S. Monteleone, M. Palesi and D. Patti, “Energy efficient transceiver in wireless network on chip architectures,” Proc. Design Automation Test in Europe Conf. Exhibition (DATE), pp. 1321-1326, Mar. 2016.

[37] Q. Ijaz et al., “Human over IP: A Vesion 6 Based Global Identification System for Humans,” International Conference on Computer Technology and Development, 3rd (ICCTD 2011). Ed. Zhou, J. ASME Press, 2011.

[38] A. u. Husnain and Q. Ijaz, “Defect Rate Analysis & Reduction of MPSOC Through Run Time Reconfigurable Computing with Multiple Caches,” International Conference on Computer Technology and Development, 3rd (ICCTD 2011). Ed. Zhou, J. ASME Press, 2011.

[39] S. Arshad, M. Ismail, U. Ahmad, A. u. Husnain and Q. Ijaz, “Optimization of Fractional-N-PLL Frequency Synthesizer for Power Effective Design,” VLSI Design, vol. 2014, p1-7, July 2014.

[40] U. Ahmad, Q. Ijaz and A. u. Husnain, “Stability analysis of coupled van der pol's oscillator,” Proc. 13th Int'l Conf. on Frontiers of Information Technology (FIT), pp. 24-29, 2015.

[41] Q. Ijaz, H. Asghar and A. Ahsan, “Exploratory study to investigate the correlation and contrast between ISO 9001 and CMMI framework: Context of software quality management,” 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 2016.

[42] H. L. Kidane, E. B. Bourennane and G. Ochoa-Ruiz, “NoC based virtualized accelerators for cloud computing,” Proc. Int'l Symp. on

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Embedded Multicore/Many-Core Syst.-Chip (MCSoC), pp. 133-137, Sep. 2016.

[43] H. L. Kidane, E. B. Bourennane and G. Ochoa-Ruiz, “Run-time scalable noc for fpga based virtualized ips,” Proc. Int'l Symp. on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), pp. 91-97, Sep. 2017.

[44] H. L. Kidane and E. B. Bourennane, “MARTE and IP-XACT Based Approach for Run-Time Scalable NoC,” Int'l Symp. on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), pp. 162-167, Sep. 2018.

[45] Q. Ijaz and E. -B. Bourennane, “Wireless NoC for Inter-FPGA Communication: Theoretical Case for Future Datacenters,” 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020.

[46] N. Jiang et al., “A detailed and flexible cycle-accurate network-on-chip simulator,” in ISPASS, IEEE Computer Society, pp. 80-96, 2013.

[47] M. Lis et al., “DARSIM: A parallel cycle-level NoC simulator,” Proc. of the Sixth Annual Workshop on Modeling, Benchmarking and Simulation (MoBS), Saint Malo, France, 20 June 2010.

[48] Y. Ben-Itzhak, E. Zahavi, I. Cidon and A. Kolodny, "Hnocs: Modular open-source simulator for heterogeneous nocs", Embedded Computer Systems (SAMOS) 2012 International Conference, pp. 51-57, 2012.

[49] L. Jain, B. Al-Hashimi, M. Gaur, V. Laxmi and A. Narayanan, “NIRGAM: A simulator for NoC interconnect routing and application modeling,” Proc. of the 2007 IEEE Design, Automation and Test in Europe Conference, Nice, France, 13–16 March 2007.

[50] V. Catania, A. Mineo, S. Monteleone, M. Palesi and D. Patti, “Noxim: An open extensible and cycle-accurate network on chip simulator,” Proc. Int'l Conf. Application-specific Systems Architectures and Processors (ASAP), pp. 162-163, Jul. 2015.

[51] V. Catania, A. Mineo, S. Monteleone, M. Palesi and D. Patti, “Cycle-accurate network on chip simulation with Noxim,” ACM Transactions on Modeling and Computer Simulation, vol. 27, no. 1, pp. 4:1-4:25, 2016.

[52] H. Lahdhiri, J. Lorandel, S. Monteleone, E. Bourdel and M. Palesi, “Framework for Design Exploration and Performance Analysis of RF-NoC Manycore Architecture,” Journal of Low Power Electronics and Applications, vol. 10, no. 4, November 2020.

Disruption and Protection of Online Synchronous Learning Environments via 802.11 Manipulation

Matthew Tigner
Department of Information Technology
Georgia Southern University
Statesboro, USA
line mt05107@georgiasouthern.edu

Hayden Wimmer*
Department of Information Technology
Georgia Southern University
Statesboro, USA
line hwimmer@georgiasouthern.edu

***Abstract*— Online learning environments are vulnerable to disruption due to the present state of Wi-Fi security on the common at-home router. Current WPA2-protected devices are susceptible to a type of attack that uses disassociation and deauthentication packets to remove all connected users and prevent new users from connecting to the target access point. This paper seeks to display an example of this attack in the context of online learning environments, showing the disruptive potential of this vulnerability and the impact it could have on online instruction. We begin with target discovery and information gathering, then move on to executing the attack while an instructor and several students are connected to various video conferencing platforms. We conclude by discussing potential mitigation and prevention strategies, such as using protected management frames and upgrading to WPA3 when possible.**

***Keywords—component, formatting, style, styling, insert** (key words)*

I. Introduction

The COVID-19 pandemic forced millions of teachers and students to quickly convert to online meetings for education, such as Zoom, Google Meet, and Microsoft Teams. As these videoconferencing platforms became more heavily used for educational meetings, the vulnerabilities of these online meeting platforms became increasingly visible. When discussing vulnerabilities of this new wave of online learning, one may consider the online conferencing application Zoom, who's user count jumped from approximately 10 million users per day in December of 2019 to over 200 million [1]. Among these users was a church group in California whose bible study session was intruded on by a user showing child pornography [1] [2]. Such attacks can play out in a variety of ways, such as a similar attack on instructional meetings for students.

While attacks can target vulnerabilities in the platforms themselves, the vulnerability discussed in this paper focuses on issues with connection to the internet itself. Specifically, modern routers that are using WPA2 encryption are vulnerable to disassociation attacks that prevent all wireless access to the internet for as long as the attack persists. Though various means of launching disassociation attacks are available, we will use the Aircrack-ng suite (a set of tools used for testing Wi-Fi security) to gather information about the target router and MDK3 (a tool used for exploring 802.11 vulnerabilities) to launch our disassociation attack. Our goal is to remove the instructor from the online meeting environment in Zoom, Google Meet, and Microsoft Teams, thus effectively disrupting the online learning experience.

II. Background

Wi-Fi Protected Access 2 (WPA2) was introduced in 2006 as a security enhancement to the pre-existing WPA. WPA was intended to be compatible with devices that used WEP so that businesses and individuals would not need to purchase new hardware to accommodate this security update; however, this also meant that some of the same vulnerabilities present in WEP also appear in WPA [3]. WPA2 utilizes pre-shared keys and implements the National Institute of Standards and Technology's Advanced Encryption Standard. Access points using WPA2 have two options available to users: WPA2-Personal and WPA2-Enterprise. WPA2-Personal uses a key exchange between the user and the access point while WPA2-Enterprise utilizes a server to grant access to users. Though considered highly secure against unauthorized users, several vulnerabilities have been discovered since WPA2's deployment, such as the KRACK exploit and disassociation/deauthentication attacks [4]. Some of these vulnerabilities have been patched; however, disassociation/deauthentication attacks remain a threat to WPA2 devices that do not use protected management frames.

When discussing wireless access points and their vulnerabilities, the terms disassociation and deauthentication are frequently found. The two antonyms, association and authentication, must be somewhat understood to grasp attacks that leverage such vulnerabilities. Authentication occurs at the beginning of the access process. This is the stage in which a device that wants to connect to the wireless network must supply valid credentials to the access point. After a device is authenticated for access to the network it is recorded to the access point for frame delivery and gains full access to the network [5]. This is known as association.

The term "hacking" has long been a buzzword in cybersecurity and generally evokes thoughts of cybercriminals, intrusions, and stolen data. While there is no shortage of malicious users on the internet, they do not comprise the entirety of the hacking community; in fact, there are those who dedicate their proficiency with computers to benefiting the system that they hack. In its purest form, this is known as ethical hacking. Three main types of hackers are generally defined in the computing world: Blackhat, Greyhat, and Whitehat hackers. The principle separation amongst these three groups is the drive and intent for hacking. Blackhat hackers are cybercriminals who employ malicious program writing and exploitation of security holes for their own benefit. They gain illegitimate access to systems and cause harm to people and property. Greyhat hackers operate in the area between Whitehat and Blackhat hacking, using their skills to find issues and security holes, but not always using this position for their own gain. This group may contain curious but non-malicious users that want to explore their capabilities and individuals who want to benefit from finding issues themselves rather than from gathering sensitive information. For example, these users may ask for payment from a target in exchange for fixing the hole that they found and threaten to publish the issue if not compensated. Finally, Whitehat hackers are motivated by improving security through hacking. They obtain permission from a network administrator (or other relevant figures) to test security and report on issues found. These hackers are also known as ethical hackers.

Penetration testing is a vital element of any serious network security plan. To protect a system from intruders, one must think and act as an intruder. A Whitehat or Greyhat hacker may be employed by a business or an individual to attempt to breach a network in some pre-agreed upon way. This gives the employer a third-party perspective on their security measures. In this scenario, Whitehat penetration testing methods will mirror those of Blackhat hackers. The attack itself may have been discussed previously to target a specific subsystem in a business or home or may have been left more general to allow for a wider range of intrusion attempts to test for preparedness. The key to ethical hacking and penetration testing is legality and consent. Though the attack may be disruptive, the impact of the attack should be controlled and should not pose a serious threat to data integrity or confidentiality. Penetration testing is a form of ethical hacking and seeks to better the security of a given system.

III. Literature Review

Aung and Thant [6] seek to educate the reader on a variety of Wi-Fi attacks, focusing on general methods and attack techniques and listing specific examples later in the paper. The authors concisely explain an array of attack types, such as deauthentication, disassociation, dictionary attacks, etc. and their requirements (some attacks only work on WEP, for example). In addition, Aung and Thant discuss defenses for these attacks and list general tactics for detecting intruders and preventing unauthorized tampering. Finally, Aung and Thant set up a testing environment and use Aircrack-ng to execute each of the attacks described in their paper. Aung and Thant conclude by restating the importance of implementing and maintaining solid security methods in wireless networks.

Kristiyanto and Ernastuti [7] aim to test wireless security and how vulnerabilities to deauthentication attacks in Wi-Fi networks can put IoT devices at risk. They note that IoT is a growing quickly in the modern world and with this comes growing security concerns for the safety of IoT devices, in particular those devices that individuals may trust with their privacy and safety (i.e. autonomous cars, home security cameras, home alarms, etc.). Kristiyanto and Ernastuti establish a testing environment and use penetration testing tools to attempt a deauthentication attack on the test network, starting with no inside knowledge of the network. Kristiyanto and Ernastuti find success with their attack and conclude that the IEEE 802.11 frame management methods need to be updated for a long-term solution. For a short-term mitigation strategy, Kristiyanto and Ernastuti suggest APs to hold more than one MAC address to advertise in the event of an attack.

Until recently, WPA2 was the standard Wi-Fi security measure in many wireless access points. Lounis and Zulkernine [8]begin by introducing WPA3 as the next generation in the line of Wi-Fi security. WPA3 features added security measures to be more secure than its predecessor WPA2. Lounis and Zulkernine, however, identify a vulnerability in WPA3 and seek to exploit it. In their work, they show that while a legitimate user attempts to access the network an attacker can flood a WPA3 access point with messages that indicate that the legitimate user is entering the wrong password, thus preventing that user from accessing the network and effectively denying service to any target users. Lounis and Zulkernine conclude by stating WPA3 access points should implement a feature that examines authentication requests with more care by not seeking to reply as soon as a message is received; rather, the access point should respond based on a group of collected requests. If one message contains the correct authentication parameters, then that user should be allowed to proceed. This would likely need to be coupled with other measures to prevent unauthorized access. Lounis and Zulkernine show that despite the increased security provided by WPA3, attacks are still feasible.

As IoT devices become more prevalent in modern society, the security of these devices becomes a greater concern. Where once a few extra Internet-enabled may be sprinkled around a home, now we begin to see important items such as security cameras and alarms become Internet and Wi-Fi dependent. Raghuprasad et al. [9] seek to educate about attack methods (such as DoS and Man-in-the-Middle) and defense strategies for common IoT devices. They establish a small testing network consisting of a temperature/humidity sensor, a personal hotspot with internet connection, and a cloud device. It is then demonstrated that the data stream from the sensor cannot reach the hotspot when the attacker launches their DoS attack, which is launched with the help of aireplay-ng. To conclude, Raghuprasad et al. implement a “MAC changer” to their proposed system, thus preventing similar attacks (that require a machine’s MAC address) in the future.

Wireless jamming has become a greater issue as wireless technologies have spread throughout our society. Grover [10] seeks to educate the reader on three main topics: types of jamming devices and techniques, how to find an active jamming device within a network, and methods for detection and response to a jamming scenario. Grover lists an impressive number of jamming tools, categories, and types within this paper and gives summaries of each jamming method covered. Continuing this trend, Grover lists several methods for localizing jamming devices within a network and states that jamming localization methods generally fall into two categories: range-based and range-free. Finally, Grover explains detection and response methods for jamming attacks by listing fifteen methods and tools for identifying and neutralizing a jamming attack.

IV. Methods

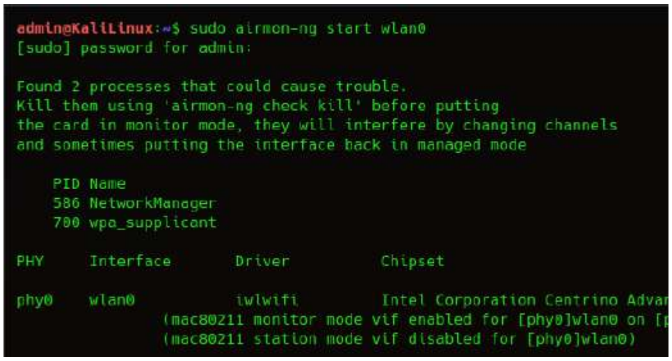
We set up a testing environment with a host connected to the target router and multiple guests operating on other networks (see Table 1 at the end of this section). The host will act as the instructor and will invite the guests to sessions on Zoom, Google Meet, and Microsoft Teams. Our goal is to disconnect the host from these meetings, thus successfully interrupting our established online learning environments. Our attack will occur in two parts. First, we discover information about the target network, specifically the router's MAC address and the channels it is broadcasting on. We then execute our attack using MDK3 on a Kali Linux machine.

A. Target Discovery

For target discovery and information gathering we will use the Aircrack-ng suite, a collection of tools used to test Wi-Fi security [11]. Once we are withing range of the target router, we place our Kali Linux machine into monitor mode using the following command:

```
sudo airmon-ng start wlan0
```

This will change our interface from *wlan0* to *wlan0mon* and allows our wireless NIC to capture data packets from all surrounding wireless access points (see Figure 1).



```
admin@KaliLinux:~$ sudo airmon-ng start wlan0
[sudo] password for admin:

Found 2 processes that could cause trouble.
Kill them using 'airmon-ng check kill' before putting
the card in monitor mode, they will interfere by changing channels
and sometimes putting the interface back in managed mode

    PID Name
    586 NetworkManager
    700 wpa_supplicant

PHY     Interface       Driver          Chipset

phy0    wlan0           iwlwifi         Intel Corporation Centrino Advar
                (mac80211 monitor mode vif enabled for [phy0]wlan0 on [p
                (mac80211 station mode vif disabled for [phy0]wlan0)
```

Figure 1 - Airmon-ng start. This command places our wireless NIC into monitor mode, allowing us to capture beacons from access points

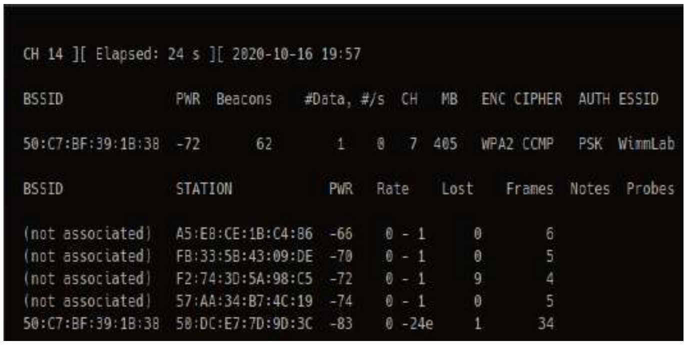
Using a separate terminal window, we activate airodump-ng (a tool provided within the Aircrack-ng suite to capture 802.11 packets from targeted access points [12]) so that all the data captured is displayed to us using the following command:

```
sudo airodump-ng wlan0mon
```

In a decently busy area this can yield a large list of available networks and will constantly be changing and reordering itself. To combat this, we will use the ESSID of the target network to narrow the search. This should make the information more manageable.

```
sudo airodump-ng --essid WimmLab wlan0mon
```

We are now able to concentrate on our target AP and gather the MAC address and the channels that the WimmLab router is broadcasting on (see Figure 2).



```
CH 14 ][ Elapsed: 24 s ][ 2020-10-16 19:57

BSSID              PWR  Beacons    #Data, #/s  CH   MB   ENC CIPHER  AUTH ESSID

50:C7:BF:39:1B:38  -72       62        1    0   7  405   WPA2 CCMP   PSK  WimmLab

BSSID              STATION            PWR   Rate    Lost    Frames  Notes  Probes

(not associated)   A5:E8:CE:1B:C4:86  -66    0 - 1      0        6
(not associated)   FB:33:5B:43:09:DE  -70    0 - 1      0        5
(not associated)   F2:74:3D:5A:98:C5  -72    0 - 1      9        4
(not associated)   57:AA:34:B7:4C:19  -74    0 - 1      0        5
50:C7:BF:39:1B:38  58:DC:E7:7D:9D:3C  -83    0 -24e     1       34
```

Figure 2 - Our airodump-ng screen, showing information for surrounding access points. In this case, we have focused our search on routers with the ESSID of WimmLab

Having recorded this information, we are ready to move on to attacking our target.

B. Jamming Attack with MDK3

Wireless jamming is a broad term used to describe intentional disruption of wireless communication [10]. This can be accomplished in a variety of ways, from utilizing a device that emits radio waves to counter the target frequency to flooding a router with packets to the degree that it cannot process the legitimate user traffic. In this usage example, we will flood a target router with deauthentication packets to remove all connected users and keep any other users from joining.

We now open another terminal window and enter the following command:

```
sudo mdk3 wlan0mon d -c 1 -b Desktop/black.txt
```

This command will utilize MDK3, a tool created to test 802.11 networks for common vulnerabilities [13]. We use MDK3's deauthentication attack (*d*) and specify channel 1 (*-c 1*) and look to the text file titled "black.txt" for the MAC address to target (*-b Desktop/black.txt*). This feature of MDK3 reads a list of MAC addresses on a text file and can either blacklist or whitelist those addresses (in the case of whitelisting, MDK3 would target all addresses except those specified). In our case, we seek to target one specific address and jam the traffic from the router.

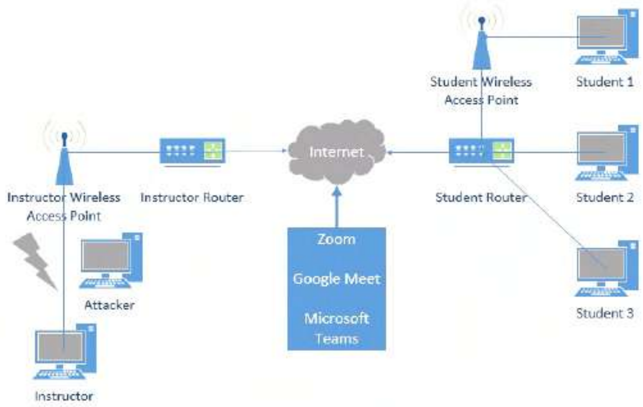


Table 1 - Test network diagram

V. Results

Within 10-20 seconds of executing the attack the host is removed from the meetings and is given some form of a "Trying to connect" message by the platform in question (see Figures 3 - 8). As the router is being jammed, all wireless access to the internet is also prevented until either the attack is cancelled by the attacker or the user switches to an unaffected router. It should be noted that this attack does not impact wired traffic from the target router.

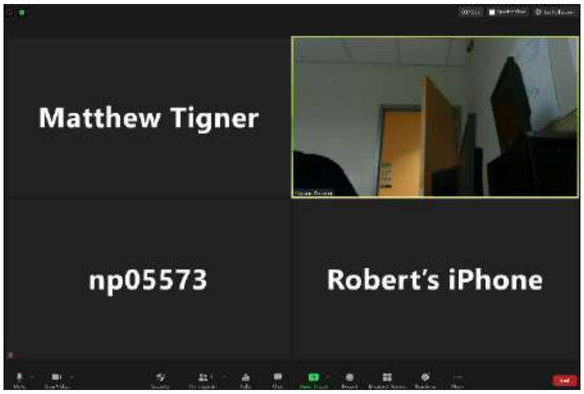
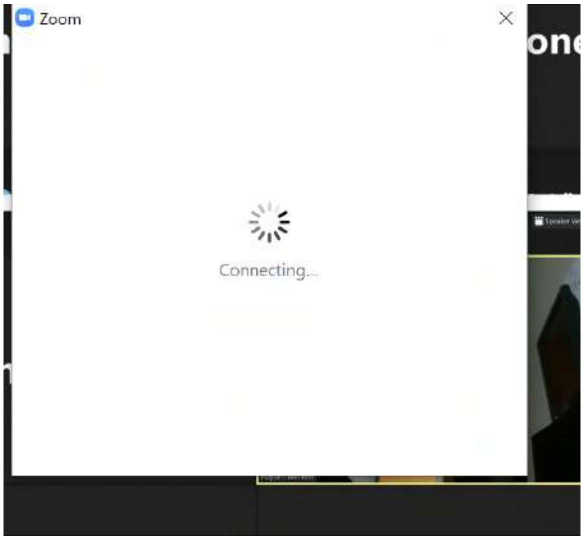


Figure 3 – Student view of a Zoom meeting with several colleagues



Connecting...

Figure 4 - The host is disconnected from the meeting

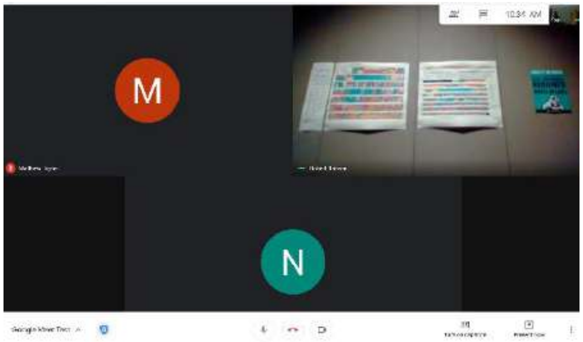
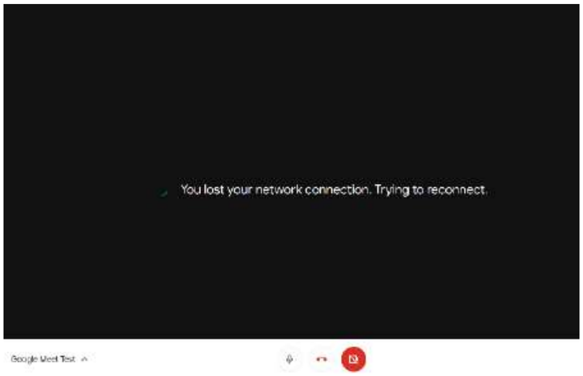


Figure 5 - Google Meet session with colleagues



You lost your network connection. Trying to reconnect.

Figure 6 – Host's connection severed after attack is launched

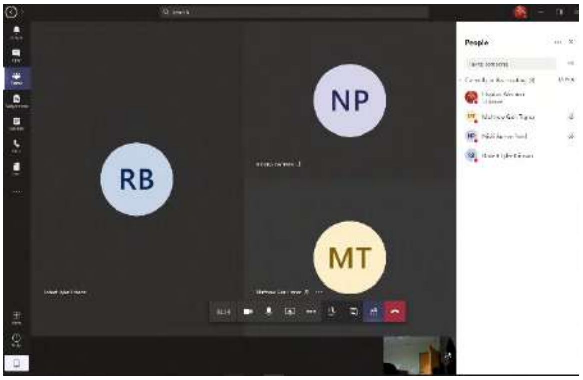
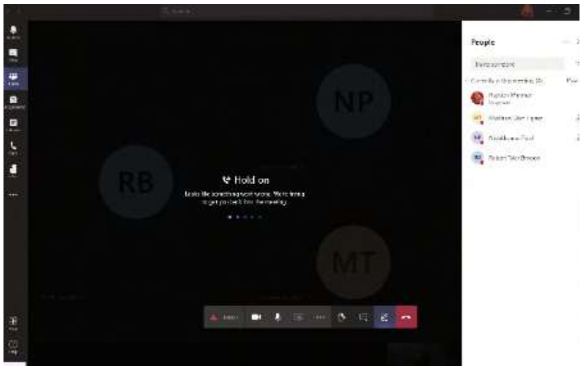


Figure 7 - A Microsoft Teams meeting is in progress with multiple colleagues



Hold on

Figure 8 - Host is successfully disconnected from MS Team meeting. In this case, the effect was immediate

VI. Implications and Defense

With online learning becoming so prevalent during the COVID-19 pandemic, attackers find themselves rich with opportunities to negatively impact online instruction. The immediate implication of this research is that malicious users could easily interfere with online learning environments if they are within range of a target, which could be a student or an instructor. Worth restating are the facts that an attacker needs no private information about the target router and only needs to be in range of the router to launch the attack. A hacker could easily park outside the home of an instructor and, using the methods described above, could completely halt all of that instructor's classes. Furthermore, it can be assumed that IoT devices in the home that rely on wireless connections are vulnerable to this attack. This creates potential for attackers to inhibit security devices around the home, thus posing a risk for harm to people and/or property.

There are several methods available to help to counteract this threat. Implementation of Intrusion Detection Systems (IDS) such as Kismet will help to mitigate an attack by notifying the victim once the system notices unusual traffic. This would be useful as a victim could immediately tell if their failed connection was due to network performance issues or if it was malicious. Beyond detection methods, the use of protected management frames in compatible devices will hinder this attack as packet forging is prevented, meaning that an attacker could no longer flood the router with disassociation packets [14]. The next generation of Wi-Fi Protected Access, WPA3, will require these management frames on its network [15]. Therefore, this attack vector should become obsolete on devices using WPA3. Additionally, the use of a wired connection would also secure this attack vector, as disassociation/deauthentication packets only apply to wireless communication.

VII. Conclusion

Using the Aircrack-ng suite and MDK3, we successfully gathered enough information to launch a disassociation attack against an instructor's router, successfully removing the host from the educational instruction sessions on Zoom, Google Meet, and Microsoft Teams. This effectively halts the educational environment as students would be left without an instructor, severely hindering their education in the event of a persistent attacker that repeatedly targets the same instructor. Familiarizing one's self with the current industry standards for Wi-Fi security and diligently upgrading all related devices can reduce the risk of victimization from this attack, resulting in safer and more secure environments for education and learning.

VIII. References

[1] "US Church Sues After Bible Study 'Zoombombed' by Child Abuse," 14 May 2020. [Online]. Available: https://www.bbc.com/news/world-us-canada-52668124. [Accessed 15 October 2020].

[2] K. O'Flaherty, "Zoom's Security Nightmare Just Got Worse: But Here's The Reality," Forbes, 5 June 2020. [Online]. Available: https://www.forbes.com/sites/kateoflahertyuk/2020/06/05/zooms-security-nightmare-just-got-worse-but-heres-the-reality/#4dc4901e2131. [Accessed 15 October 2020].

[3] "WPA Security Explained: What is Wi-Fi Protected Access?," AT&T, 29 June 2020. [Online]. Available: https://cybersecurity.att.com/blogs/security-essentials/wpa-security-explained-what-is-wi-fi-protected-access. [Accessed 29 November 2020].

[4] "What You Need to do About the WPA2 Wi-Fi Network Vulnerability," Norton, 18 January 2018. [Online]. Available: https://us.norton.com/internetsecurity-emerging-threats-what-to-do-about-krack-vulnerability.html. [Accessed 11 November 2020].

[5] "Understanding IEEE 802.11 Authentication and Association," Intel, 19 October 2020. [Online]. Available: https://www.intel.com/content/www/us/en/support/articles/000006508/network-and-i-o/wireless.html. [Accessed 11 November 2020].

[6] M. A. C. Aung and K. P. Thant, "IEEE 802.11 Attacks and Defenses," *Seventeenth International Conference on Computer Applications,* 2019.

[7] Y. Kristiyanto and Ernastuti, "Analysis of Deauthentication Attack on IEEE 802.11 Connectivity Based on IoT Technology Using External Penetration Test," *CommIT,* vol. 14, pp. 45-51, 2020.

[8] K. Lounis and M. Zulkernine, "Bad-Token: Denial of Service Attacks on WPA3," *Proceedings of the 12th International Conference on Security of Information and Networks,* pp. 1-8, 2019.

[9] A. Raghuprasad, S. Padmanabhan, A. Babu and P. Binu, "Security Analysis and Prevention of Attacks on IoT Devices," *2020 International Conference on Communication and Signal Processing (ICCSP),* pp. 876-880, 2020.

[10] K. Grover, "Jamming and Anti-Jamming Techniques in Wireless Networks: A Survey," *International Journal of Ad Hoc and Ubiquitous Computing,* 2014.

[11] "Aircrack-ng," Aircrack-ng, 26 January 2020. [Online]. Available: https://www.aircrack-ng.org/doku.php?id=Main. [Accessed 29 September 2020].

[12] "Airodump-ng," Aircrack-ng, 26 January 2020. [Online]. Available: https://www.aircrack-ng.org/doku.php?id=airodump-ng. [Accessed 29 September 2020].

[13] "MDK3 Package Description," Kali Tools, [Online]. Available: https://tools.kali.org/wireless-attacks/mdk3. [Accessed 30 September 2020].

[14] "What Are Protected Management Frames," WiFi Alliance, [Online]. Available: https://www.wi-fi.org/knowledge-center/faq/what-are-protected-management-frames. [Accessed 2 October 2020].

[15] "WPA3 Security," WiFi Alliance, [Online]. Available: https://www.wi-fi.org/discover-wi-fi/security. [Accessed 2 October 2020].

Water Level Control System using Programmable Logic Controller (PLC): Rujban Water Supply System

Yousef M. K. Ali *, Omar A. Zargelin*, Fadel Lashhab¶ and Abdulbasit Alaribi*
* *Department of Electrical and Electronic Engineering, Al Zintan University*, Al Zintan, Libya.
¶ *Department of Electrical Engineering and Computer Science, Howard University, Washington DC, USA.*
Email: *zargelin@uoz.edu.ly, yousef.ali@uoz.edu.ly,¶fadel.lashhab@howard.edu, and *1516702041@uoz.edu.ly

***Abstract*—The main challenges related to water affect human settlements' sustainability: the lack of access to safe and clean water at all times. These problems have massive consequences on human health, safety, economic growth, and development. Rujban city faces this crisis despite more than five decades trying to resolve this problem. In this paper, we describe the existing water supply system, which consists of three main components: Karthoom Automated Control System (KACS), Idref Automated Control System 1 (IACS1), and Idref Automated Control System 2 (IACS2). We propose the control system design of the KACS to meet the city's demand and keep the collecting tank with an acceptable level of water continuously. This paper presents the Karthoom Automatic Control System (KACS) design to control the water supply system using the Programmable Logic Controller (PLC). PLC controller is a central part of the proposed automatic control system due to its reliability and ease of installing and maintenance. Besides, we use PLC to carry out the water level control, sequencing, monitoring, display, and control functions, such as logic, sequence, timing, counting, and arithmetic logic. To maintain an appropriate water level in the collecting tank, we control the operation of the feeding submersible pumps ("on" or "off") by sensing the collecting tank's water levels. This is accomplished by utilizing four-stage-level sensors, seven water flow sensors, and an RF transmitter/receiver to monitor and control all pumps' operations and protect them from dry running and then accomplish the sensing stage. We introduce the control system components, the operation flow chart, and the proposed system's implementation. Finally, we implemented and conducted a simulation experiment using the Siemens S7-200 PLC to control the overall water supply system and evaluate the proposed control approach's performance.**

***Index Terms*—An automatic control system, Programmable Logic Controller (PLC), Siemens S7-200 PLC, water flow sensors, submersible pumps, Wireless communication network.**

I. Introduction

Despite a sufficient number of deep wells and a reservoir for collecting water in the Karthoom area, the lack of water availability is considered a problem that is elevating and rising, especially with the population's growth in the city. To develop root solutions for these obstacles which led to the lack of water supply in the Rujban town, we propose a design of an automatic control system to control water flow from seven wells located in the Karthoom area water level in the local collecting tank. The water level in the tank was measured using the sensors. Four different water levels were considered, we then decided which submersible pump must turn "on" accordingly by sending an "on" signal wirelessly to the target pump(s). The system also prioritizes which of the pumps must be operated based on each pump's number of working hours. The lifetime of each pump is maintained in which each of them has to run periodically. Besides, the flow of water from the operating wells was monitored using a flow sensor for each well while protecting the submersible pumps from dry running by sending an "off" signal through a wireless link to the target pump to turn it off.

PLC controllers are usually a central part of many automatic control systems in the industry due to their reliability and ease of installing and maintenance. PLCs are used for the internal storage of instructions to carry out control functions, such as logic, sequence, timing, counting, and arithmetic. They are also widely used to monitor and control plants or equipment in industries such as oil and gas refining, water and waste control, energy, transportation, etc.

The KACS system is located in harsh desert environments that can damage any cabling. The long distances between the wells' locations and the collecting tank lead to higher installation and maintenance costs due to low scalability and connectors' high failure rate. These reasons have inspired us to use advanced wireless technologies as appropriate solutions to these problems and reduce the installation cost of the KACS system, which is distributed over a vast land area. Moreover, the system will display each pump's operational status (Run or Stop) and the water level in the collecting tank.

The PLC controller will run or stop any pump by sending RF signals "on" or "off" signal" via a transmitter. The transmitter can send the RF signal to seven different receivers with Omni-directional antennae. The receiver at the other end can receive signals sent by the transmitter, which are only specified for it by a specific unique code that uses ZigBee technology.

There are two main contextual areas of focus and problem studying: wireless networks (WSN) and control systems. At the same time, the previous paper's primary focus is on WSN [19, 14], Crucial consideration needs to be given to other areas. There are many existing technologies and various approaches in WSN regarding outdoor applications, but we depend mostly on the reference (book) and radio mobile information. In [3],

978-1-6654-4067-7/21/31.00 ©2021 *IEEE*

a smart wireless technology designed and developed using a microcontroller for various crops with irrigation systems. The authors of this work use the Zigbee wireless control system to create an intelligent irrigation system that provides an efficient water supply for use by humans and agriculture, which is crucial to reduce water use and reduce the workforce needed for irrigation. The concept of water level monitoring and management introduced by [16]. This paper presented water level monitoring and management within the context of water's electrical conductivity. This paper also proposed a web and cellular-based monitoring service protocol to determine and sense water levels globally. The researchers in [17] analyzed the existing oil pumping system and discovered that it has a high power-consuming process and should entail more manual power. They proposed a sensor network-based intelligent control system for saving power and efficient monitoring of oil wells' health. If any abnormality is detected in the oil well data sensing, the maintenance manager is notified through an SMS via GSM. This system allows oil wells to be monitored from remote places. In [7], designed an approach to help people living in multi-storied buildings that face inconvenience when there is a shortage of water in the tank. This system automatically pumps water from the reservoir to the tank whenever the water level decreases. The water level is checked using a ping sensor connected to a micro stamp controller to switch on the pump as soon as it is empty. Automatic irrigation to the plants is provided by [8], which helps in saving money and water. An 8051 microcontroller was used to control the entire system. The proposed controller is programmed as giving the interrupt signal to the sprinkler. The solar-powered water pump controller was designed and deployed remotely by [12] to detect the tank's water levels and pumping the water from a remote source such as a pond or well when required. For more details on wireless communication, vision, and technical challenges, see [13], and [4] and their reference in which described the current wireless systems and emerging systems and standards. The gap between these systems and the vision for future wireless applications indicates that much work remains to be done to make this vision a reality. In [5] a low-cost solution to apply fuzzy logic control for a water level control system using an Arduino was proposed. It also gives a low-cost hardware solution and a practical procedure for system identification and control. A prototype system for artificial control and monitoring was implemented by [1] using IoTs, two control systems were introduced: classical PID (Proportional Integral Derivative) and fuzzy logic with a comparison between them. For more information about designing an implementation of water systems' control using some existing techniques raised in the literature, see [10, 18, 9, 2], and their cited references.

This paper proposes a control mechanism to control Rujban's water supply system using the Siemens S7-200 PLC. We present several distinct contributions:

- First, we described the existing water supply system and its components.
- Second, we proposed the control system's main components in more detail, including the PLC controller, the water tank, water level/flow sensors, transmitter/receiver stations, and auxiliary components.
- Third, we proposed the Karthoom Automatic Control System (KACS) design to control the water supply system using the Programmable Logic Controller (PLC).
- Finally, we implemented a simulation experiment using the Siemens S7-200 PLC to control and evaluate the proposed control system for the overall water supply system.
- This paper's primary goals:
  1) Reducing energy consumption.
  2) Eliminating water wastage.
  3) Providing continuous water flow.
  4) Maintaining the service life of the pumps.
  5) Reducing the labor power of maintenance.
  6) Ensuring the city's necessary daily water usage.

The paper's organization as follows: In Section II, we define the problem statement. Section III describes the existing water supply system, which consists of three main components: Karthoom Automated Control System (KACS), Idref Automated Control System 1 (IACS1), and Idref Automated Control System 2 (IACS2). In Section IV, we propose the control system's main components in more detail, including the PLC controller, the water tank, water level sensors, water flow sensors, transmitter/receiver stations, and auxiliary components. The implementation and experimental simulation results are conducted in Section V using the Siemens S7-200 PLC to control the overall water supply system and evaluate the proposed control approach's performance.

II. Problem Statement

To fully understand the water dilemma and its infrastructure and history in Rujban, you could review previous papers [15, 19]. The whole water providing system of the Rujban city can be divided into three subsystems, and each subsystem is divided into two parts, a wireless communication network and an automatic control system, as follows:

1) Pumping water from the producing wells into the collecting tank in the Karthoom area: There are seven wells currently producing water and connected to the karthoom collecting tank. The KACS is divided into wireless communication network part [15], and the automatic control system, which is the subject of this paper.
2) Pumping water from Karthoom collecting tank using big centrifugal pumps to the collecting tank located in the Rujban center. The wireless communication network part is studied in the previous paper [15, 19], and the automatic control system is addressed in this paper.
3) Pumping water from the Rujban center collecting tank to the upper distribution tanks and water providing management ”IACS2” will be another future research.

After conducting a thorough study on the Karthoom water operation and management dilemma, we discovered the following:

- Manual operation of submersible pumps for wells and their remoteness from the central station where the distance is sometimes greater than 3 kilometers.
- Lack of high-level water monitoring in the collecting tank.
- Lack of means of transportation and difficulty in moving from the station to wells.
- Power cuts from time to time and the consequent stop of water pumping and being obliged to operate it.
- Lack of knowledge of water interruption.
- High labor prices. The primary motivation for conducting this study and suggesting possible economic solutions is to solve this problem and promote the city, especially since the bulk of the expenses necessary to provide water for residential communities have been paid. One of those costs is the big modern water distribution network, which cost 16 million Libyan dinars for all city districts.

III. Design and Development

In this section, we divide this KACS system into two parts. The first part is related to the control process, which is the subject of this paper. The other is related to wireless communication, and it is addressed in our previous research. The block diagram of the KACS is shown in Fig. 1.

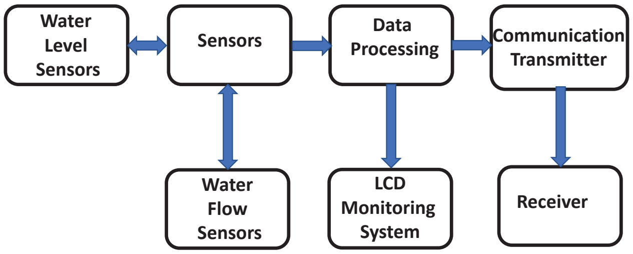


Figure 1. Block diagram of the KACS.

The water pumping process is controlled according to the tank's water level measured by the four-level sensors. Thus, the process control mode is on/off control. In this process, the on/off controller turns pumps ”on” when the water level is measured at its minimum value. And turns the pumps ”off” when the water level reaches its maximum value. The system block diagram, designed according to environmental conditions and the desired process control.

The water pumping process from the seven water wells to the collecting tank at the Karthoom area requires water level monitoring in the tank to avoid flooding and wastage of water or completely drain the water tank. Furthermore, the daily travel from and to the remote wells located away from the collecting tank, as shown in Fig. 2, could keep a worker busy without maintenance. The low efficiency of this type of system prompted Rujban city to install an automatic control system. The control system will enable water level monitoring and distribution from a central location, reduce overall operating costs, and maintain a running water distribution system. The proposed control system can gather data, monitor, and control remote pumps from a central location at the Karthoom area. The control process consists of monitoring the collecting tank's level and accordingly controlling the seven submersible pumps' operation and the two centrifugal pumps. The centrifugal pumps are used to pump water from the collecting tank to the Rujban city center's assembly tank. The PLC controller checks the tank's water level using four-level sensors (i.e., very Low, Low, High, and Very High water levels). These four-level values are inputs to the PLC controller, which starts and stops the submersible pumps based on the measured water level (this will be discussed in more details in Section IV). The digital inputs of the PLC controller will also include seven water flow sensors, one for each pipe that connected the well with the collecting tank and two push button to start/stop the two centrifugal pumps, and another push button to be used in the case of an emergency shutdown to disable the PLC control system or to change the control process to a manual mode. The PLC controller will send the control signals to remote far away, from 1.8 to 3.2 km, pumps via an RF modem. One transmitter unit is located on the collecting tank's side, connected to the central PLC controller, and shares a base code for initializing registers for the communication protocol (txrx.c) [12]. A receiver stations at each well, which will receive an ”on/off signals” sent by the PLC via the transmitter end. The site layout is shown in Fig. 2.

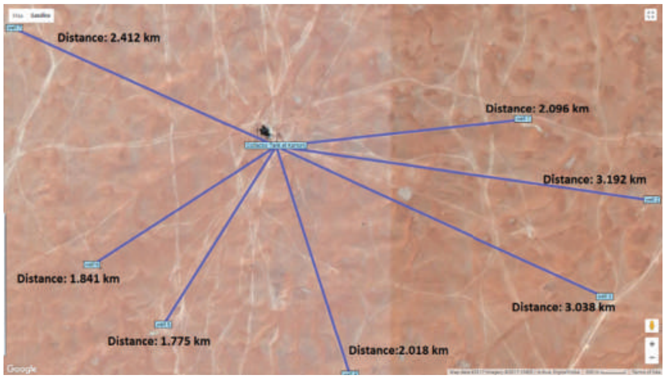


Figure 2. Wireless links layout

The main goal is to develop an automatic control system to control the collecting tank's water level using a PLC controller. The following are the objectives that need to be achieved; these objectives will act as a guide and will restrict the system to be implemented for certain situations:

1) To check the water level in the tank: the PLC controller will respond based on the detected water level and run the well pumps accordingly:
   - If the tank's water level is high high (>= 98%), an alarm will be generated, and the labor has to convert to manual process to turn off all submersible pumps and may turn ”on” centrifugal pumps.

- If the tank's water level is very high ($>= 95\%$), all submersible pumps must be turned "off" and may centrifugal pumps turned "on".
- If the water level in the tank is high ($>= 80\%$), two out of the seven submersible pumps will continue in an "on" situation.
- If the water level in the tank is low ($>= 60\%$), the PLC controller will turn "on" four pumps.
- If the tank's water level is very low ($>= 10\%$), all seven water pumps will be turned on, and the centrifugal pumps will be stopped to protect the motor from dry running.
- If the tank's water level is low low ($<= 8\%$), an alarm will be generated and the labor has to convert the control system to a manual process to turn on all submersible pumps and turn "off" centrifugal pumps.

2) If the PLC controller detects no water flow from any particular running pump, the corresponding pump will be turned off after a specific time.
3) The PLC controller code will be written using the Ladder diagram language: the code is implemented and simulated using the Delta WPLSoft 2.49 software [6], And it will include all control sequences in the process as explained in the previous steps. There are different PLC brands available in the market, such as Siemens, Allen Bradley, Mitsubishi, ABB, Schneider, etc.

IV. Control System Components

This section will describe our proposed control system's main components in more detail, including the PLC controller, the water tank, water level sensors, water flow sensors, transmitter/receiver stations, and auxiliary components.

A. The PLC Controller

The PLC controller is a computer programmed to perform almost any control, sequencing, monitoring, and display function. The primary function of the PLC is to send wireless signals to the water pump to turn it on or off, based on the water level in the tank, using floating limit switches that indicate the current water level in the tank. This work will introduce a convenient solution to the process plants where cabling is expensive or not possible. It also has lower installation and maintenance costs, provides reliable operation, and robust and flexible construction. Usually, a PLC system's essential functional components are a processor unit, memory, power supply unit, input/output interface section, communications interface, and programming device, as shown in Fig. 3. The input and output sections are where the processor receives external devices' information and sends control commands to external devices. In our case, the PLC's inputs will include four digital level sensors and seven flow sensors. The outputs might be connected to motor starter coils or other visual and audible alarms. All input and output signals in this project are digital, and no existence analog signals. The pumps are connected to the PLC controller's digital output wireless via specific relay circuits connected to the receiver station at each well. For this paper, the Siemens S7-200 PLC was proposed to automatically control the overall system and reduce the design and control complexity. Siemens provides different S7-200 CPU models with various features and capabilities to create sufficient water pumping process solutions. The CPU 222 [16] is a perfect candidate for the presented control system with eight inputs/6 outputs and two expansion modules to increase the number of inputs/outputs up to 94 maximum, see Fig. 4, [11]. Typically, PCs are used to write the control program using the PLC Ladder logic. The control program is then downloaded to the PLC memory through a physical connection such as the RS 485 cable. The first step of the system implementation is to download the Ladder logic program into the PLC. The S7-200 CPU Models contain all instructions and logic circuits required to control the whole system, as shown in Fig. 5.

B. Water Level Sensors: Limit Switches

The level sensor's assembly consists of four float switches, these float switches are placed inside the collecting tank, and they are arranged at very low water level (10%), low water level (60%), high water level (80%), and very high water level (the tank is full). These limit switches are connected to the digital input of the PLC controller. The basic operation is as follows: when the water level reaches one of the four limit switches, the contact of that switch will close, and a +24 V signal is set on the PLC's digital input, which makes the control system respond by opening or closing the relay contacts to turn the pump on/off according to the detected water level. A very high alarm switch would be positioned at a level of 98% to turn off all submersible pumps, prevent flooding and wastage of water from the tank, and activate a warning audibly and a flashlight. A very low alarm switch of 10% can be utilized to detect a low water level near the bottom of the tank and activate audibly and flashing light alarms; accordingly, any operating discharge pump of the two pumps will be turned off.

C. Water Flow sensors

The flow sensors are generally open contacts that close if water flows from a specific well's pipeline connected to the collecting tank. There will be one flow sensor for each submersible pump. The flow sensors are used to protect pumps from dry running or leakage in the piping system. Therefore, the control system will help prevent unwanted failures and costly repairs.

D. Transmitter Unit

Data are transferred from the PLC to the transmitter, and the USART module with Atmega 32 controls it. When initialization occurs, tasks are timed on a $1ms$ time scale given by timer 0 Afterward, it interrupts the timer 0 compare, match, and (USARTUDREvect), triggered by an empty transmitter buffer. The serial model's baud rate is set to 4000 bps. The MCU sends one packet of information every $200ms$, and

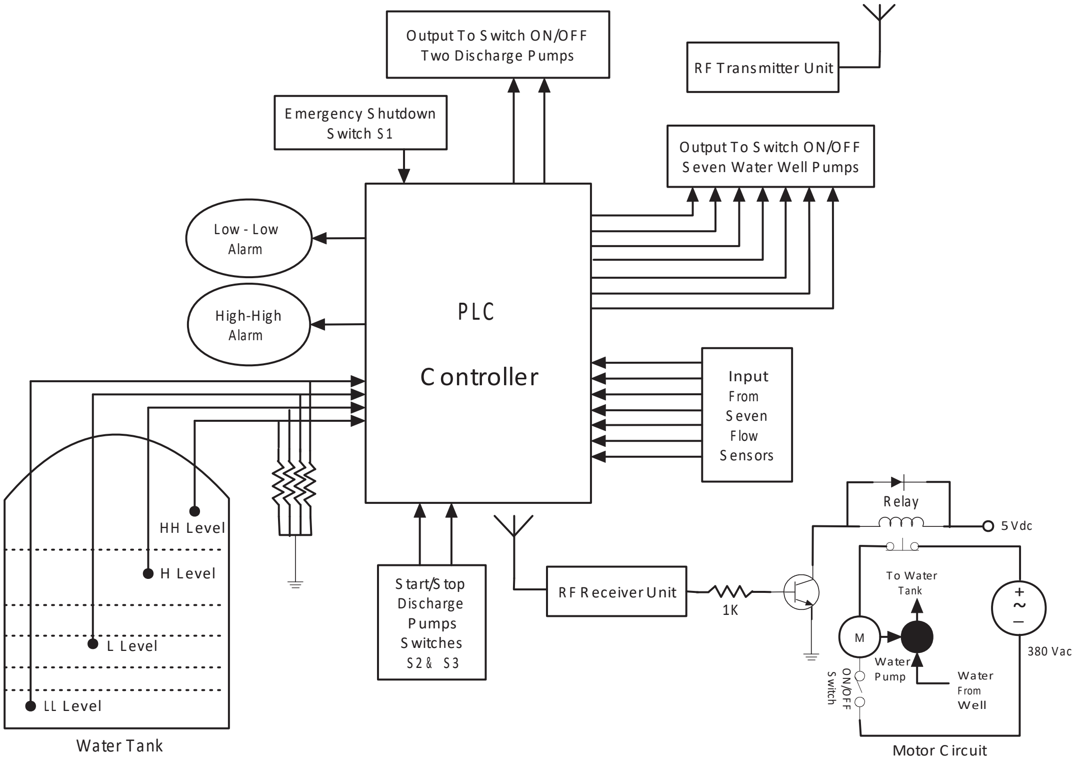


Figure 3. Block diagram of the control system.

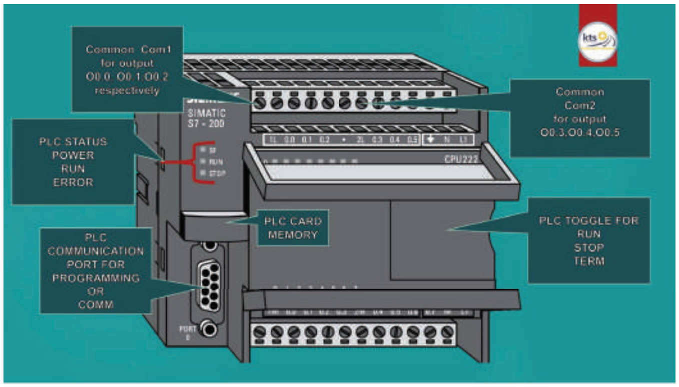


Figure 4. The Siemens S7-200 PLC CPU 222, [11].

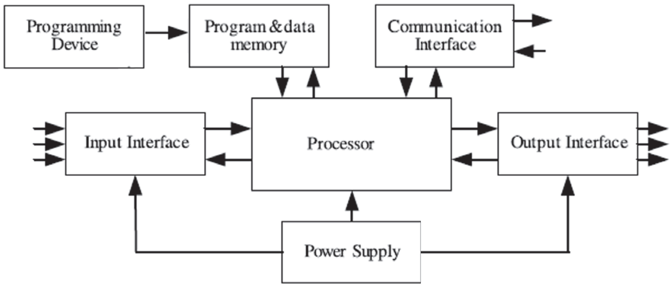


Figure 5. Basic Structure of PLC.

at the end of each $200ms$ unit, the float switch's State is checked, where the character symbol "on" is transmitted when the button is "on" and the character sign "off" is sent when it isn't. A modified version of the packet structure available by Desai et al. is used. It consists of three synchronization bytes: a start character, a single data byte, and an end character. Each byte that is transmitted is governed by DC balancing, meaning that it consists of an equal number of $0s$ and $1s$; this ensures that the receiver's calibration is not lost during the first transmission. Symbol "off" (149) and Symbol "on" (180) are the data symbols that are transmitted one at a time when the empty transmit buffer occurs (USARTUDREvect). Afterward, the bits are outputted by the USART module out of PIN $D.1$, which happens to be sent connected to an NPN transistor base to turn the transmitter on and off. At least two-bit flips are required before one symbol is falsely decoded as another, which benefits the characters used above.

E. Receiver Unit

The receiver unit must detect its data and turn the voltage on or off to control the pump. Here, the timer "0" is used once again to generate a 1 ms timer to compare match interrupts, and the baud rate is also once again set to 4000 bps because it must see that the receiver and the transmitter operate at the same baud rate. When the USART buffer (UDR) receives a byte of data, the (USARTRXCvectISR) gets triggered. At this time, the program checks for the synchronization character, and if it is not there, the program ignores the received byte, but if it is there, then the program finds the start character by searching the next three bytes, and if it is found., the program would receive the data byte. The reception of data bytes stops if the stop character is found or 32 bytes are received, which is the maximum number of received bytes. Afterward, the data's validity is checked by the program (if the 2nd element of the message array is the start character). The 3rd element of the message array is checked if the data is valid to see if its symbol is in the ”on” sign. Pin B.0 is written as a logical high to turn the pump on if the 3rd character is the symbol ”on” Still, if it is not, then the pin B.0 is written as a logic low to turn it on if and only if the Symbol ”On” is received, and the pump gets turned off if after 3 minutes no other symbol ”on” is obtained, which is to account for any problems at the transmitter end.

F. Auxiliary Components

Two bush buttons will be installed to start and stop the two discharge pumps, and another bush button will be used in the case of an emergency to shut down the PLC control system or if the operator wants to set the whole system in manual mode. In addition, to control some high-power devices such as flashing lights, speakers, and high voltage electric motors using a low volt PLC controller, interface devices between the PLC and the high power devices are required. Usually, mechanical relays (contractors) are available to switch currents from milli-amperes to higher amperes. We should use a relay circuit in the water pump system to adapt to a high voltage AC. The relay circuit's output is connected to the motor's negative side of the cable—the positive side of the line combined with 380V AC [3].

V. System Implementation and Experimental Simulation Results

A. Software Description

The ladder logic diagram for the PLC controller was written and simulated using the Delta WPLSoft 2.49 software [6].

When the system is powered up, the PLC controller senses the digital inputs connected to the level limit switch and flow sensors; if the PLC digital inputs receive any changes, then the controller sends signals to the digital outputs of the PLC accordingly. The whole operation makes a cycle or repeats itself concerning the input signals according to the flow chart shown in Fig. 6.

B. Operation Description

The designed control system aims to meet the city's demand for water by using an automatic control system to operate the pumping water process as a cycle and keep the collecting tank full of water at all times. The control process will be achieved by monitoring the input signals' level and then sending control output signals according to turn on/off water pumps or sending alert signals when the tank's water level is almost full (HH alarm) or nearly empty (LL alarm). As shown in Fig. 6, the control process can be described in the following points:

1) Suppose the water level is very low (LL Level $< 10\%$), which means the tank almost empty. In this case, the PLC turns “off” any operating discharging pump, and ”on” signals will be sent wirelessly to all submersible pumps via the RF transmitter to restore the low water level in the tank at 60%. Moreover, a ”LL alarm” signal will be activated. Audible alarms and flashing lights will be activated to alert the collecting tank operator's near emptiness and protect the two discharging pump motors from dry running.
2) If the water level reached the low level's limit switch (60%). In this case, the PLC input connected to the L limit switch will send a +24v signal, which means the water level is at 60%; accordingly, the PLC will turn ”off” three pumps, and the other four pumps will continue operating. Until the tank's water level reaches a high level at 80%, the PLC input connected to the H limit switch will receive a +24v signal. Therefore, the PLC will turn ”off” two pumps out of the four operating pumps. Table I shows the description of the tank filling and draining control procedure.

Table I
Tank Filing and Draining Control Procedure

Tank Level	Water Level	Pumps Status	Draining
”HH”	97%	Alarm	”on”
”VH”	95%	All Pumps ”off”	”on”
”H”	80%	Two Pumps ”on”	”on”
”L”	40%	Four Pumps ”on”	”on”
”VL”	10%	All Pumps ”on”	”off”
”LL”	8%	Alarm	”off”

3) If the tank is almost full and the water level becomes High High (”HH”) at 98%, that means the PLC system does not work. An audible / flashing alarm signal will be activated to let the labor convert the system to manual operation and turn all submersible pumps off.
4) If the tank is almost full and the water level becomes Very High (”VH”) at 95%, the PLC will send turn ”off” signals to all operating submersible pumps. And at the same time, the PLC will send a ”VH” alarm” signal note.
5) Furthermore, the pump priority based on the number of working hours will be used as an indicator by the PLC controller to determine the next pump(s) to be turned ”on”. Each pump will run cumulative 24 hours before the next pump takes over priority in this control system.
6) Finally, after sending an ”on” signal to any pump, the PLC will wait 5 minutes before checking the flow sensor for that particular pump; if the PLC digital input connected to that flow sensor receives a positive signal (+24v), then this means the submersible pump is producing water. The pump's motor will continue running. Otherwise, the PLC will send an ”off” signal to stop that pump and protecting the pump's motor from dry running.

C. Control Mechanism

During operation and to start any pump, the PLC controller will send an ”on” signal to a particular submersible pump

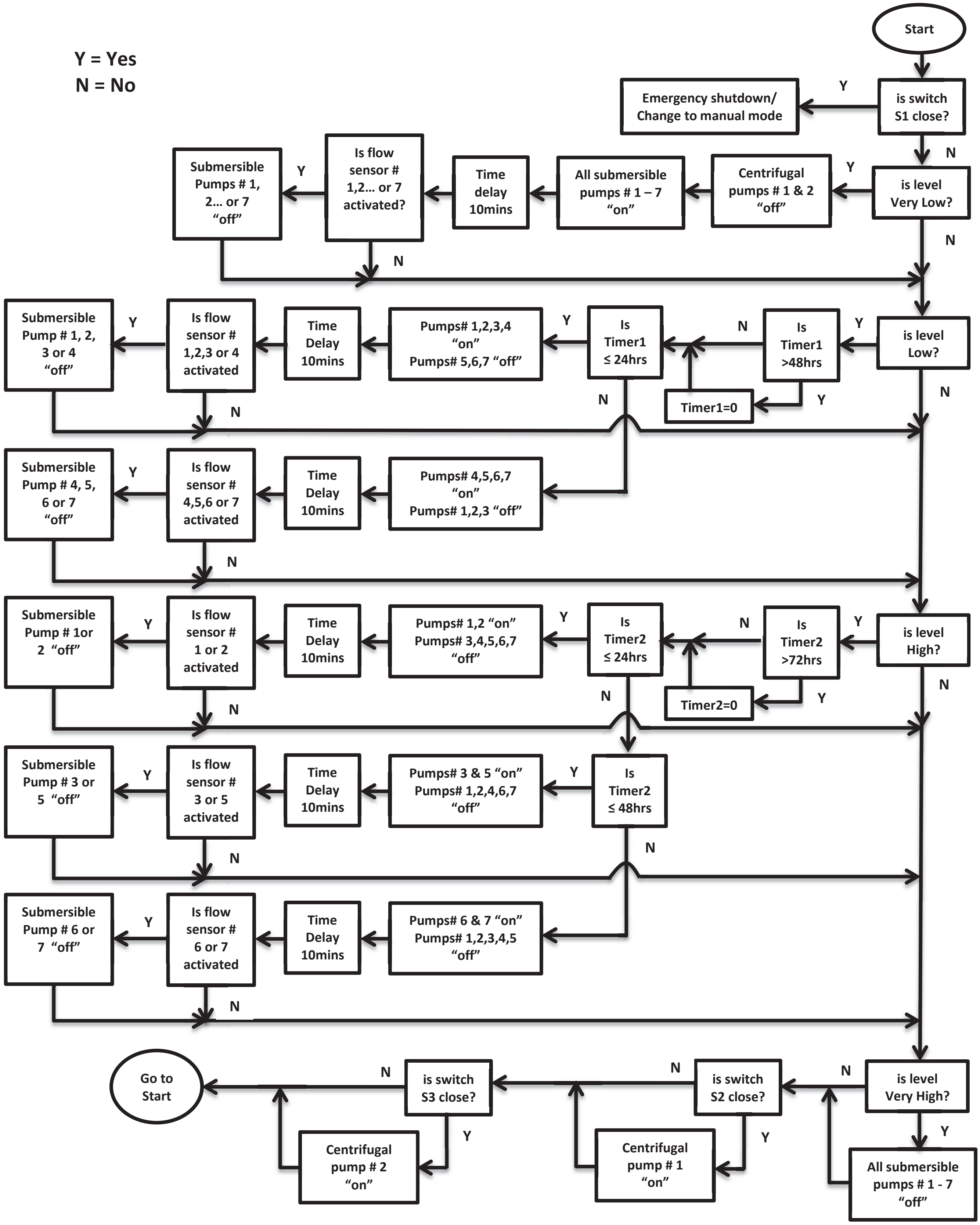

Y = Yes
N = No
Start
is switch S1 close?
Y
Emergency shutdown/ Change to manual mode
N
is level Very Low?
Y
Centrifugal pumps # 1 & 2 "off"
All submersible pumps # 1 – 7 "on"
Time delay 10mins
Is flow sensor # 1,2… or 7 activated?
Y
Submersible Pumps # 1, 2… or 7 "off"
N
N
is level Low?
Y
Is Timer1 >48hrs
Y
Timer1=0
N
Is Timer1 ≤ 24hrs
Y
Pumps# 1,2,3,4 "on" Pumps# 5,6,7 "off"
Time Delay 10mins
Is flow sensor # 1,2,3 or 4 activated
Y
Submersible Pump # 1, 2, 3 or 4 "off"
N
N
N
Pumps# 4,5,6,7 "on" Pumps# 1,2,3 "off"
Time Delay 10mins
Is flow sensor # 4,5,6 or 7 activated
Y
Submersible Pump # 4, 5, 6 or 7 "off"
N
is level High?
Y
Is Timer2 >72hrs
Y
Timer2=0
N
Is Timer2 ≤ 24hrs
Y
Pumps# 1,2 "on" Pumps# 3,4,5,6,7 "off"
Time Delay 10mins
Is flow sensor # 1 or 2 activated
Y
Submersible Pump # 1or 2 "off"
N
N
N
Is Timer2 ≤ 48hrs
Y
Pumps# 3 & 5 "on" Pumps# 1,2,4,6,7 "off"
Time Delay 10mins
Is flow sensor # 3 or 5 activated
Y
Submersible Pump # 3 or 5 "off"
N
N
Pumps# 6 & 7 "on" Pumps# 1,2,3,4,5 "off"
Time Delay 10mins
Is flow sensor # 6 or 7 activated
Y
Submersible Pump # 6 or 7 "off"
N
is level Very High?
Y
All submersible pumps # 1 - 7 "off"
N
is switch S2 close?
Y
Centrifugal pump # 1 "on"
N
is switch S3 close?
Y
Centrifugal pump # 2 "on"
N
Go to Start

Figure 6. Flow chart of system design.

through the RF transmitter using 5700 MHz wireless links. The receiver unit on that water well will catch that signal. Upon receiving this signal, the receiver unit sends a logic high signal (+24V) to the motor's relay circuit. The output of the relay circuit is connected to the motor pump's cable as a neutral. The other side of the motor's line is connected to AC 380 v. This will start the motor of that particular pump, and the water is pumped from the well to the collecting tank. If the PLC controller sends an ”off” signal to a specific submersible pump, the receiver unit on that well will catch that signal. Upon receiving this signal, the receiver unit sends a logic low signal (0V) to the motor's relay circuit. This will disconnect the pump's engine from the AC 380V power source, and the pump will be turned off. The motor circuit is shown in Fig. 3 is used to control the motor on the receiver side [12]. An ”on/off” switch is used to manage the motor driver circuit manually. The transmitter and receiver units share a base code for initializing registers for the communication protocol (txrx.c). We will describe the program of the transmitter end later in this paper.

To sum our results up, we implemented the PLC with the S7-200 model to convert manual operation and management into a fully automatic operation for KACS. Because of the limited number of pages of this conference, we omitted the PLC ladder program and some simulation results. We hope to add them to future journals. Our results showed many rewards, including reducing energy consumption, eliminating water wastage, providing continuous water flow, protecting pump damage, maintaining the service life of the equipment, reducing the labor intensity of maintenance, avoiding stealing water from the water distribution system, and ensuring the city's necessary daily water usage.

VI. Conclusion and Future Work

Most Middle East and African countries suffer from water shortages due to insufficient and old water supply systems. This paper presented an efficient, low-cost, flexible, economical, and easily configurable approach for water level control using a wireless solution; we proposed an Automatic control of Rujban's water supply system based on PLC. The water level control system combines two parts; wireless sensors and PLCs. This paper focused on the control system design for efficient automated water pumping through water level/water flow sensors, which turn on/off a specific number of pumps, and Siemens S7-200 PLC controller. First, we described the existing water supply system and its main elements. Second, we proposed the control system's main features in more detail, including the Siemens S7-200 PLC controller, water level/flow sensors, transmitter/receiver units, and auxiliary components. Third, we provided the overall operation procedure to convert it from fully manual to 100% automatic with pumps operation time recycling. Finally, we implemented a simulation experiment using the PLC to control, manage, and evaluate the proposed control system. Our results showed many rewards, including reducing energy consumption, eliminating water wastage, providing continuous water flow, protecting pump damage, maintaining the service life of the equipment, reducing the labor intensity of maintenance, avoiding stealing water from the water distribution system, and ensuring the city's necessary daily water usage. As future work, to solve the city's complete water problem, we will design and implement two other automatic control systems; the first system for IACS1 and the second system for IACS2. Furthermore, we will design and implement the overall water level control system in real-time, including wireless communication and control using PLC.

References

[1] Methaq A Ali, Abbas Hussein Miry, and Tariq M Salman. “IoT Based Water Tank Level Control System Using PLC”. In: *2020 International Conference on Computer Science and Software Engineering (CSASE)*. IEEE. 2020, pp. 7–12.

[2] Zafer Aydogmus. “Implementation of a fuzzy-based level control using SCADA”. In: *Expert Systems with Applications* 36.3 (2009), pp. 6593–6597.

[3] Ritesh Boda. “A Design and Development of Smart Wireless Sensor Network for Farming Water Supply System”. In: ().

[4] Jane Butler. “Wireless Networking in the Developing World, Third Edition”. In: *International Journal of Scientific and Research Publications* (2013).

[5] Fayçal Chabni et al. “The application of fuzzy control in water tank level using Arduino”. In: *International Journal of Advanced Computer Science and Applications* 7.4 (2016), pp. 261–265.

[6] INC. DELTA ELECTRONICS. *Delta WPLSoft 2.49 software*. https://wplsoft.software.informer.com/. Feb. 2021.

[7] Sudip Dogra. “Design and Development of Automatic Water Level Controller Using Stamp Processor and Ping Sensor”. In: *International Conference on Futuristic Trends in Computing and Communication* (2015).

[8] Venkata Naga Rohit Gunturi. “Micro controller based automatic plant irrigation system”. In: *International Journal of Advancements in Research & Technology* 2.4 (2013), pp. 194–198.

[9] Suppachai Howimanporn, Sasithorn Chookaew, and Chaiyaporn Silawatchananai. “Comparison between PID and Sliding Mode Controllers for Rotary Inverted Pendulum Using PLC”. In: *2020 4th International Conference on Automation, Control and Robots (ICACR)*. IEEE. 2020, pp. 122–126.

[10] Cosmina Illes, Gabriel Nicolae Popa, and Ioan Filip. “Water level control system using PLC and wireless sensors”. In: *2013 IEEE 9th International Conference on Computational Cybernetics (ICCC)*. IEEE. 2013, pp. 195–199.

[11] Kavisatechsolutions.com. *SIEMENS PLC WIRING S7-200 PLC WIRING DIAGRAM*. https://www.kavisatechsolutions.com/2020/08/siemens-plc-wiring.html. Feb. 2021.

[12] Roland Krieger. "Remote solar-powered water pump controller". PhD thesis. Cornell University, 2014.

[13] Andreas F Molisch. "Wireless communications". In: 34 (2012).

[14] Mahmud Mohamed Nagasa et al. "Designing Wireless Network for Water Issue in the City of Zintan". In: *International Science and Technology Journal (ISTJ), Libya* (2019).

[15] Yousef M. K. Ali Omar A. Zargelin and Fadel Lashhab. "Wireless Communication and Control Systems: Case Study: Karthoom Water System". In: *Submitted to IEEE 1st International Maghreb Meeting of the conference on Sciences and Techniques of Automatic control and computer engineering MI-STA* (2021).

[16] SM Khaled Reza, Shah Ahsanuzzaman Md Tariq, SM Mohsin Reza, et al. "Microcontroller based automated water level sensing and controlling: design and implementation issue". In: *Proceedings of the world congress on engineering and computer science*. Vol. 1. 2010, pp. 20–22.

[17] C Rojiha. "Sensor network based automatic control system for oil pumping unit management". In: *International Journal of Scientific and Research Publications* 3.3 (2013), pp. 1–4.

[18] Marios Tyrovolas and Tibor Hajnal. "Inter-communication between Programmable Logic Controllers using IoT technologies: A Modbus RTU/MQTT Approach". In: *arXiv preprint arXiv:2102.05988* (2021).

[19] Omer Zergalin and Walid KA Hasan. "Post Study Design of Wireless Network for the Water Dilemma in the city of Rujban". In: *The International Journal of Engineering and Information Technology (IJEIT)* 4.2 (2018).

Performance Analysis of Microstrip Patch Antenna for the Diagnosis of Brain Cancer & Tumor using the Fifth-Generation Frequency Band

Sayed Abdul Kadir Al-Nahiun
Department of Electrical and Electronic Engineering (EEE) American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
nahiunkf42@gmail.com

Fardeen Mahbub
Department of Electrical and Electronic Engineering (EEE) American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
mahbubfardeen1998@gmail.com

Rashedul Islam
Department of Electrical and Electronic Engineering (EEE) American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
saydulbabar147570@gmail.com

Shouherdho Banerjee Akash
Department of Electrical and Electronic Engineering (EEE) American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
akashbanerjee906@gmail.com

Raja Rashidul Hasan
Department of Electrical and Electronic Engineering (EEE) American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
hemal@aiub.edu

Md. Abdur Rahman
Department of Electrical and Electronic Engineering (EEE) American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
arahman@aiub.edu

***Abstract*—Brain Cancer and Tumors are common death factors over the world. Determining the location of a brain tumor at an early stage is difficult due to its minimal size and some disadvantages of the mechanisms used for the diagnosis of the brain tumor. In this paper, a Rectangular Microstrip Patch Antenna has been designed for Microwave Imaging (MI) with a frequency range of 1.5 GHz to 3 GHz at a resonant frequency of 2.3 GHz (5G-Band) in the CST Studio Suite Software to identify brain tumors. FR-4 Substrate material has been used to design the Antenna. The Antenna dimension that has been designed in this paper is 60.46*78.73*1.7 mm^3 and the radiating patch of the Antenna was fed by a feedline, which is rectangular in size. The human brain phantom has been created in the CST software with six different homogenous layers of skin, fat, skull, dura, CSF (Cerebrospinal Fluid), and the Brain. Besides, a 5mm tumor was also placed inside that human brain. The Antenna was applied in the brain phantom both with and without the tumor to analyze the Antenna's performance. A Reflection Factor ($S_{1,1}$) of -30.76 dB and -30.88 dB were also achieved respectively after applying the Antenna in the brain phantom with and without the tumor. Other obtained performance parameter values were also provided in this paper, such as Directivity (2D & 3D), Radiation Efficiency, Polar Radiation, Specific Absorption Radiation (SAR), etc. the Antenna will be a safer choice for the detection of brain tumor. 5G frequency band has been used here because the free space antenna can be used in communication (5G mobile communication, WLAN, Wi-Fi), and as well as for body applications.**

Keywords*—*Brain Cancer, Microwave Imaging, Patch Antenna, 5G-Band, SAR, Reflection Factor, VSWR, Directivity.

I. Introduction

Cancer has recently become the second-largest mortality cause globally, and it is responsible for the deaths of 9.6 million people every year all over the world, according to the World Health Organization (WHO). There are numerous types of Cancer according to their symptoms, and Brain Cancer is one of them. The cancers of the central nervous system and Brain form a group of tumors that are rare and heterogeneous. These kinds of tumors are liable for 3% of cancer cases and are more prominent in males than females. Though the incidence rate is low, brain cancer ranks in the top 10 in cancer deaths [1]. The variables that define the tumor state are its scale, location, and development status [2]. There are two forms of a tumor which include benign and malignant. A benign tumor is a type of tumor without any cancerous cells and is not very much harmful. Moreover, its tissue development is prolonged [3]. On the other hand, a malignant tumor is the formation of cancerous cells and is considered deadly, risky, and unpredictable. Moreover, it spreads very quickly to the surrounding tissue and hampers its respective functions to a significant extent [2]. Earlier, various approaches such as X-ray Mammography, Computed Tomography (CT scan), Magnetic Resonance Imaging (MRI), Ultrasound, Radiography, and many other techniques could be used for detecting brain tumors or brain cancer cells [4]. But all these approaches have some drawbacks, such as their high ionizing radiations, and sometimes, they are unable to make the difference between the tumors that are benign and malignant [5].

One of the recent successful invented approaches for detecting the brain tumor is Microwave Imaging (MI) using a Microstrip patch antenna due to its low price, non-ionizing radiation, and faster working process over the previous approaches. The working mechanism of the Microwave Imaging (MI) system depends on the dielectric properties (such as permittivity, conductivity) of the stable cells and the tumor-affected cells of the human brain [6]. A microstrip patch antenna is an integral factor of the Microwave Imaging systems that irradiates the human head under examination with microwaves

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

that pass around the head and then detect another antenna that acts as a receiver. The receiver antenna produces the tumor-mirrored information that has been analyzed with an appropriate signal processing technique for receiving 3D images of the brain under tests [7]. However, for better diagnosis, specific microstrip patch antennas' performance parameters, such as SAR (Specific Absorption Rate), Reflection Factor ($S_{1,1}$), Radiation Pattern, Directivity, and Polarization, requires to be adequately analyzed [8]. A Microstrip Patch Antenna aims to design in this paper to diagnose brain cancer and tumor by using as a part of its improvement from the past. The operating frequency of the Antenna is 2.3GHz which is under Sub-6 GHz 5th Generation spans (450 MHz - 6 GHz) [9].

II. Modelling of the Microstrip Patch Antenna in Free Space

At first, an operating frequency of 2.3 GHz (5G-Band) has been selected, which varies from 1.5-3 GHz. The Width of the patch was calculated from Equation-1 using the Velocity of Light (C_o), Operating frequency (f_r), and the Dielectric Constant (ε_r) value of 4.3.

$$W = \frac{c_0}{2f_r\sqrt{\frac{\varepsilon_r+1}{2}}} \quad (1)$$

The Effective Dielectric Constant, ε_{reff} has also been calculated using Equation (2) [10].

$$\varepsilon_{reff} = \frac{\varepsilon_r+1}{2} + \frac{\varepsilon_r-1}{2}\left(1+12\frac{h}{w}\right)^{-0.5} \quad (2)$$

Here *h* represents the height of the substrate, and *w* represents the width of the patch.

After that, the Effective Length, $L_{eff,}$ is calculated using Equation (3) [10].

$$L_{eff} = \frac{c_0}{2f_r\sqrt{\varepsilon_{reff}}} \quad (3)$$

The actual length of the patch has been determined. To determine it, the fringing factor is subtracted from the effective length, which was shown in the following Equation (4) [9].

$$\Delta L = 0.412\frac{\left(\frac{w}{h}+0.264\right)(\varepsilon_{reff}+0.3)}{(\varepsilon_{reff}-0.258)\left(\frac{w}{h}+0.813\right)} \quad (4)$$

The patch's actual length has been determined using the following Equation (5) [10].

$$L = L_{eff} - 2\Delta L \quad (5)$$

Following that, the Width of the Ground Plane's corresponding values, $W_g=2*W$, and the Ground Plane's Length, $L_g=2*L$, have been calculated [5].

TABLE I. Parameters of the Proposed Model

Parameter with Symbols	Dimensions (mm)
Patch Length, *L*	30.23
Ground Plane Length, L_g	60.46
Patch Width, *W*	38.5
Ground Plane Width, W_g	77
Ground Thickness, h_t	0.035
Substrate Height, h_s	1.7
Feedline Width, W_f	3.2
Feedline Insertion, F_i	8.9295
The Gap between Patch and the Feedline, G_{pf}	1

An inset feed transmission feedline which is 50-ohm were also connected to the proposed antenna model. The free space parameter of the Antenna have been represented in Table I.

The 3D model and the proposed Antenna model's geometry are shown in Figures 1 and 2, respectively.

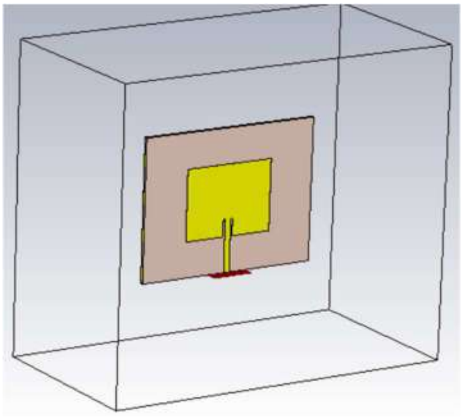


Fig. 1. 3D model of the Microstrip Patch Antenna

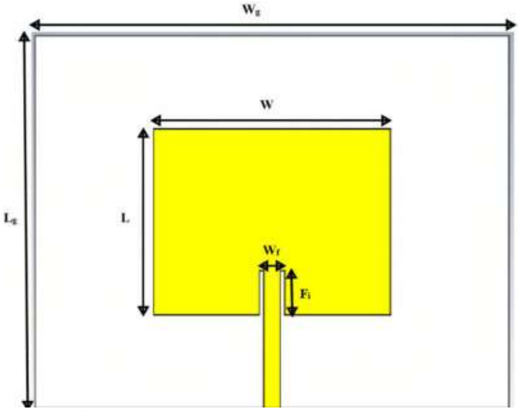


Fig. 2. Geometry of the Designed Microstrip Patch Antenna

III. Modeling of the Microstrip Patch Antenna in Brain Phantom

A human head model consisting of six different layers has been designed and simulated using the CST Studio Suite 2019 software. These layers include Skin, Fat, Skull, Dura, Cerebrospinal Fluid (CSF), and the Brain (white matter). While preparing the human head design, a radius of 81 mm was considered, and the dimensions & electrical properties of each layer have been listed in Table II. Figure 3 shows the six-layered healthy Human Head Phantom, and Figure 4 shows the entire Head Phantom.

TABLE II. Parameter Values of the Brain Phantom

Tissue	*Permittivity*	*Conductivity (S/m)*	*Density (Kg/m^3)*	*Thickness Mm*	*Thermal Conductance KJ/Kg/K*	*Heat Capacity W/m/K*
Skin	42.85	1.59	1090	1.50	0.50	3.662
Fat	10.82	0.26	910	0.15	0.24	2.973
Skull	14.96	0.59	1850	0.60	0.36	2.524
Dura	42.03	1.66	1130	0.40	0.44	3.364
CSF	66.24	3.45	1005.9	0.50	0.62	4.200
Brain	42.53	1.51	1030	82.25	0.56	3.682
Tumor	55.8	4.1	1070	5	-	-

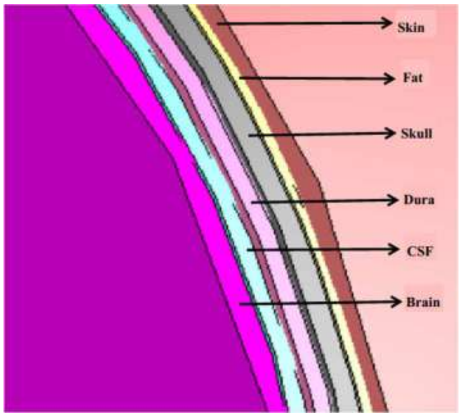


Fig. 3. Six-Layered healthy Human Head Model

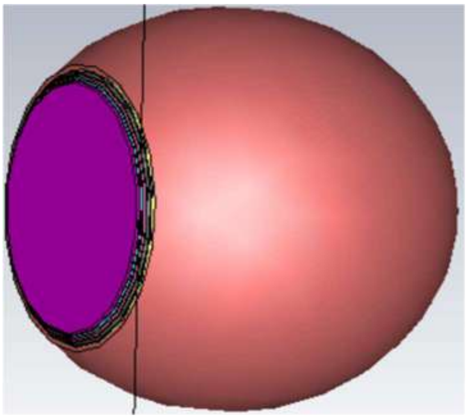


Fig. 4. The Full Head Phantom

Since there are variations in each layer's dimensions, each layer possesses different electrical properties such as Heat Capacity, Conductivity, Permittivity, etc. After designing the human brain phantom, a 5 mm tumor was placed on the Human Brain. Then, the Microstrip Patch Antenna was applied to the brain phantom both with and without the tumor. For avoiding the negative effect of the electromagnetic field, the ground plane of the Antenna has been attached to the head phantom's surface [9]. The coordinates of the proposed Antenna's position were (x=0, y=0, z=85). Figure 5 shows the tumor-affected brain Phantom with the Microstrip Patch Antenna, and Figure 6 represents the Location of the Tumor after applying the Microstrip Patch Antenna.

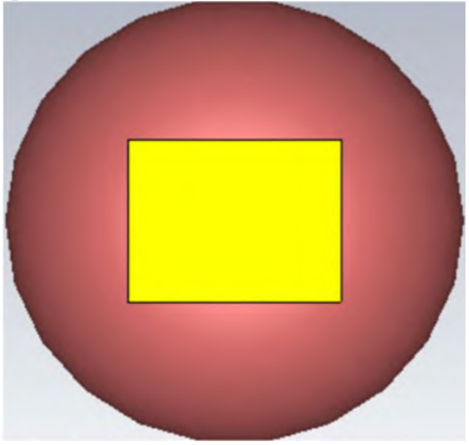


Fig. 5. Tumor affected Brain Phantom with Microstrip Patch Antenna

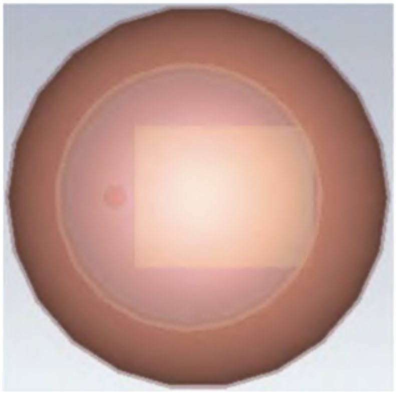


Fig. 6. Location of the Tumor after applying the Microstrip Patch Antenna

IV. Result Analysis

The designed Antenna has been applied in the Brain Phantom consisting of without the tumor and with a tumor, respectively. All the determined results from the simulation have been discussed below.

A. *Reflection Factor*

Reflection Factor is the measure of the RF energy that an Antenna can accept, and it is measured in terms of dB [11]. However, a smaller magnitude of Reflection Factor means less energy is being passed to the proposed Antenna, reducing its effectiveness. Therefore, as the proposed Antenna in this research work is for body applications, therefore from that perspective, a higher magnitude of Reflection Factor is always preferable. For the tumor-less Brain phantom implemented in this research work, a Reflection Factor of -30.87 dB and a Bandwidth of 82.8 MHz have been successfully obtained. Moreover, the Brain Phantom consisting of the tumor has also been implemented from which a Reflection Factor of -30.76 dB and a Bandwidth of 82.8 MHz has also been successfully obtained. All the results have been obtained at an operating frequency of 2.3 GHz. Figures 7 & 8 shows the Reflection Factor ($S_{1,1}$) value of the Brain Phantom without the tumor and as well as with the tumor, respectively.

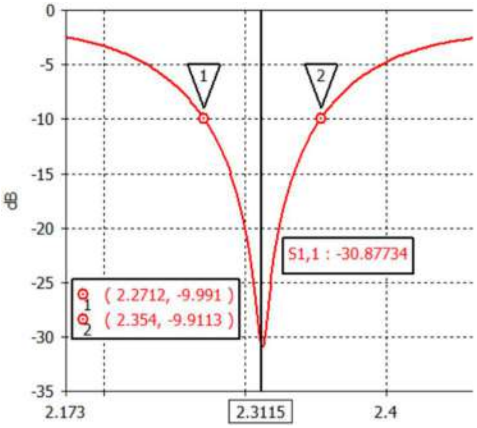


Fig. 7. Reflection Factor of the Brain Phantom without Tumor

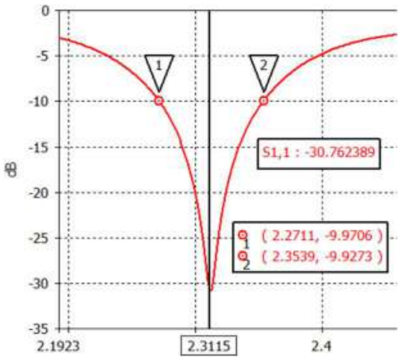


Fig. 8. Reflection Factor of the Brain Phantom with Tumor

B. Directivity

It is the measure of the radiation pattern concentration of an antenna in a particular direction, and it is expressed in terms of dB [12]. An antenna radiates more concentrated beams when the magnitude of the Directivity of that Antenna is higher. However, a higher directivity also resembles that the beam will travel to a significant distance. Here, after applying the proposed antenna model on the Brain Phantom consisting of the tumor, a 2D and 3D Directivity of dB has been successfully achieved. Figures 9 & 10 shows the 2D and 3D Radiation Pattern (Directivity) of the Brain Phantom Efficiency Pattern of the Brain Phantom with Tumor, respectively.

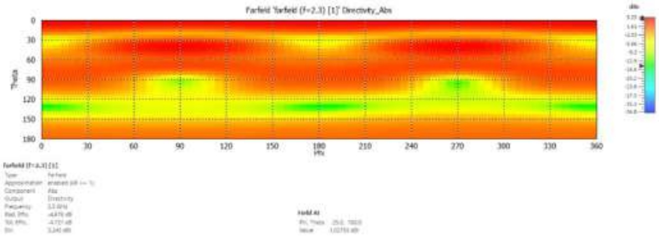


Fig. 9. 2D Radiation Pattern (Directivity) of the Brain Phantom with Tumor

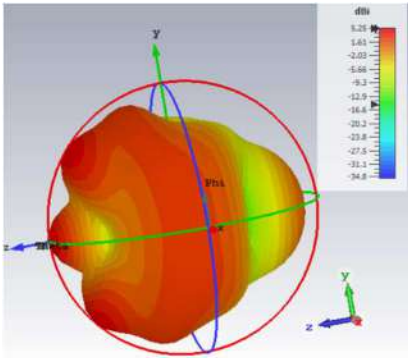
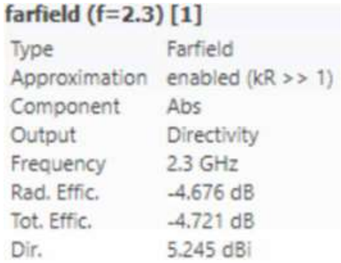


Fig. 10. 3D Radiation Pattern (Directivity) of the Brain Phantom with Tumor



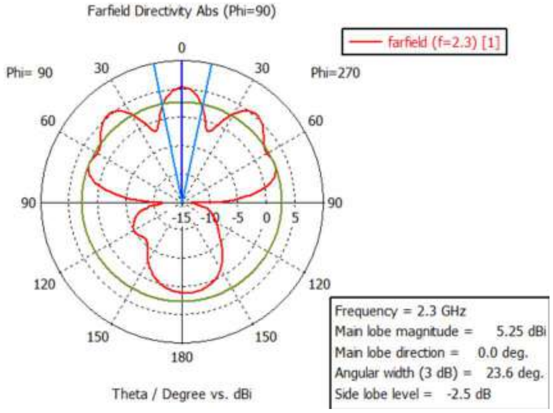
farfield (f=2.3) [1]

Type	Farfield
Approximation	enabled (kR >> 1)
Component	Abs
Output	Directivity
Frequency	2.3 GHz
Rad. Effic.	-4.676 dB
Tot. Effic.	-4.721 dB
Dir.	5.245 dBi

Fig. 11. 3D Radiation Pattern (Directivity) value of the Brain Phantom with Tumor

C. Farfield Radiation (Polar Form)

In this work, the proposed antenna model has been placed on the Brain Phantom consisting of a tumor. From the simulation, the Farfield Radiation has been determined (Polar form, Phi=90°). From the simulation, the main lobe magnitude of 5.25 dBi and a side lobe level of -2.5 dB has been successfully achieved for the Operating Frequency of 2.3 GHz. Figure 12 shows the Farfield Radiation (Polar Form) of the Brain Phantom with Tumor.



Frequency = 2.3 GHz
Main lobe magnitude = 5.25 dBi
Main lobe direction = 0.0 deg.
Angular width (3 dB) = 23.6 deg.
Side lobe level = -2.5 dB

Fig. 12. Farfield Radiation (Polar Form) of the Brain Phantom with Tumor

D. Surface Current

The applied electromagnetic field induces an actual electric current, known as Surface Current, especially in metallic antennas [13]. After applying the proposed antenna model in the Brain Phantom consisting of the tumor, a Surface Current of 34.6 A/m has been successfully achieved from the simulation. Figure 13 shows the Surface Current of the Brain Phantom with Tumor.

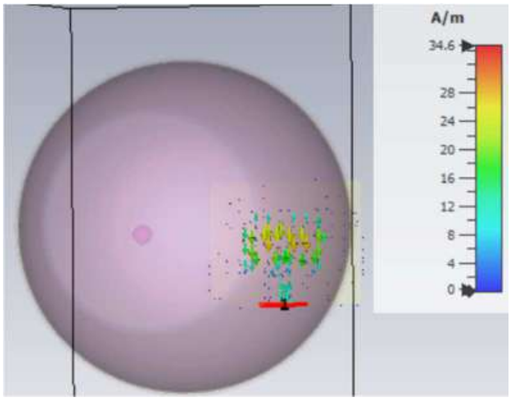


Fig. 13. Surface Current of the Brain Phantom with Tumor

E. Specific Absorption Rate (SAR)

SAR is the measure of the maximum allowable level of Electromagnetic Radiation produced by a Communication Antenna in numerous wireless devices. It is an important parameter, and it requires to be maintained within a standard level. The IEEE C95.3-2002 standard governs the standard level of SAR, which specifies that for one gram of average body tissue, the level of SAR does not exceed 1.6 W/kg, as per the ICNIRP and FCC guidelines [2].

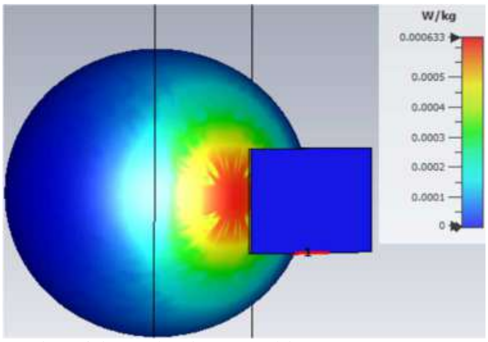


Fig. 14. SAR value of the Brain Phantom without Tumor

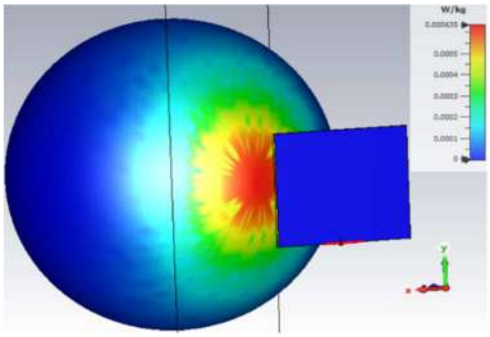


Fig. 15. SAR value of the Brain Phantom with Tumor

Here, after applying the Antenna's model on the Brain Phantom consisting of no-tumor, a SAR value of 0.000633 W/kg has been achieved. Similarly, after using the Antenna on the Brain Phantom consisting of the tumor, a SAR value of 0.000635 W/kg has also been successfully determined. Figures 14 & 15 shows the SAR value of the Brain Phantom without the tumor and as well as with the tumor, respectively.

V. Comparative Study

As a part of the comparative study, numerous past research works have been thoroughly analyzed and compared with the current works. After the brief comparison, it can be determined that the antenna model that has been proposed in this research work is much more effective and advantageous, which includes its ability to detect low sized tumors precisely. Table III portrays the brief comparison of the proposed antenna model with the past research works. Since the distance of the proposed antenna from the human brain phantom is bigger than the compared research, as a result, the SAR value is much lower than the compared research works. Due to the antenna’s long distance from the human brain phantom, the proposed antenna is also safer due to its low SAR and low ionizing radiation.

TABLE III. Comparison of the Antenna model with previous Research Paper

Reflection Factor ($S_{1,1}$), dB	*Directivity, (in dBi)*	*SAR, (in W/kg)*	*Tumor Size (in mm)*	*Distance of the Antenna from Brain Phantom*	*Ref.*
-18.2	-	0.9501	10	3 mm	[4]
-9.15	-	0.4610	35	10 mm	[14]
-26.69	6.469	1.95	5	10 mm	[8]
-26	6.60	-	5	-	[6]
-22.30	4.731	0.03101	5	83.79 mm	[15]
-30.76	5.425	0.000635	5	85 mm	Proposed Model

VI. Conclusion

In this research work, a Rectangular Microstrip Patch Antenna has been designed to diagnose brain cancer in the early stage. This paper exhibits a precise pathway for characterizing numerous tissues in various simulated scenarios after applying the Antenna in the healthy human brain model and the tumor-affected brain model. At first, a human head phantom model has been prepared in the CST Studio Suite Software 2019. The designed Antenna operates at a frequency of 2.3 GHz was placed on the head's phantom. Therefore, a better performance was achieved by using the 2.3 GHz operating frequency after applying the antenna on the tumor-affected brain phantom. Thus, 2.3 GHz was used as the operating frequency. The Antenna's Reflection Factor was determined -30.87 dB after applying the Antenna on the tumor-free brain phantom and -30.76 dB after applying the Antenna on the tumor-affected brain phantom. A Radiation Efficiency was found to be -4.6820749 dB and -4.6753968 dB after using the Antenna on the Cancer free brain phantom, and the Cancer affected brain phantom respectively. A Surface Current of 34.6 A/m and a Directivity of 5.245dBi was received with Cancer-affected brain phantom. A SAR value of 0.000633 W/kg and 0.000635W/kg were also obtained after applying the Antenna on the tumor-free brain phantom and the tumor-affected brain phantom. From all the determined data, it has been learned that the proposed antenna model has been susceptible to Diagnose Brain Cancer from the cancer-affected brain. After analyzing the SAR and other performance parameters such as Reflection Factor, VSWR, Directivity, Radiation Pattern, Radiation Efficiency, etc. of the Antenna in the presence and absence of tumor, it can be concluded that the designed Antenna is very much feasible and capable enough for detecting the tumor in the Human Brain. However, near the edge of the human head model's brain, the tumor was being placed. Therefore, considering the tumor's location in another area of the human head or the center of the brain, further testing and analysis are required.

References

[1] Cardoso, V., Dinis, H. and Mendes, P.M., 2019, November. Electric Field Array Detector for Millimeter Wave Assistance on Brain Tumor Resection. In *2019 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)* (pp. 1-4). IEEE.

[2] Raihan, R., Bhuiyan, M.S.A., Hasan, R.R., Chowdhury, T. and Farhin, R., 2017, August. A wearable microstrip patch antenna for detecting brain cancer. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)* (pp. 432-436). IEEE.

[3] Chowdhury, T., Farhin, R., Hassan, R.R., Bhuiyan, M.S.A. and Raihan, R., 2017, September. Design of a patch antenna operating at ISM band for brain tumor detection. In *2017 4th International Conference on Advances in Electrical Engineering (ICAEE)* (pp. 94-98). IEEE.

[4] Singh, T., Singh, S., Singh, M. and Kaur, R., 2019, September. Design of Patch Antenna to Detect Brain Tumor. In *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (Vol. 1, pp. 1-6). IEEE.

[5] Kahwaji, A., Arshad, H., Sahran, S., Garba, AG and Hussain, R.I., 2016, March. Hexagonal microstrip antenna simulation for breast cancer detection. In 2016 International Conference on Industrial Informatics and Computer Systems (CIICS) (pp. 1-4). IEEE.

[6] Rokunuzzaman, M., Samsuzzaman, M. and Islam, M.T., 2016. Unidirectional wideband 3-D antenna for human head-imaging application. IEEE Antennas and Wireless Propagation Letters, 16, pp.169-172.

[7] Shekhawat, S., Sharma, V., Jain, P.K., Sharma, B.R., Saxena, V.K. and Bhatnagar, D., 2019, December. An Off-diagonal Feed Elliptical Patch Antenna with Ring Shaped Slot in Ground Plane for Microwave Imaging of Breast. In 2019 IEEE Indian Conference on Antennas and Propogation (InCAP) (pp. 1-4). IEEE.

[8] Paul, L.C., Hossain, M.N., Mowla, M.M., Mahmud, M.Z., Azim, R. and Islam, M.T., 2019, November. Human Brain Tumor Detection Using CPW Fed UWB Vivaldi Antenna. In *2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)* (pp. 1-6). IEEE.

[9] Danneberg, M., Datta, R., Festag, A. and Fettweis, G., 2014, June. Experimental testbed for 5G cognitive radio access in 4G LTE cellular systems. In *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)* (pp. 321-324). IEEE.

[10] Balanis, C.A., 2016. Antenna theory: analysis and design. John wiley & sons.

[11] Kang, J.S., Kim, J.H., Kang, N.W. and Kim, D.C., 2012, July. Antenna measurement using S-parameters. In 2012 Conference on Precision electromagnetic Measurements (pp. 658-659). IEEE.

[12] Mahbub, F., Islam, R., Al-Nahiun, S.A.K., Akash, S.B., Hasan, R.R. and Rahman, M.A., 2021, January. A Single-Band 28.5 GHz Rectangular Microstrip Patch Antenna for 5G Communications Technology. In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1151-1156). IEEE.

[13] Alzahed, A.M., Mikki, S.M., Antar, Y.M., Clénet, M. and Jovic, S., 2016, October. Characterization of a rectangular patch antenna using ACGFSEM approach. In 2016 IEEE Conference on Antenna Measurements & Applications (CAMA) (pp. 1-3). IEEE.

[14] Elkorany, A.S., Helmy, R.M., Saleeb, A.A. and Areed, N.F., 2019, December. Microstrip Patch Antenna Linear Arrays for Brain Tumor Detection. In *2019 14th International Conference on Computer Engineering and Systems (ICCES)* (pp. 425-431). IEEE.

[15] Sinha, S., Niloy, T.S.R., Hasan, R.R., Rahman, M.A. and Rahman, S., 2020, January. A Wearable Microstrip Patch Antenna for Detecting Brain Tumor. In *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)* (pp. 85-89). IEEE.

Resilience of Autonomous Vehicle Object Category Detection to Universal Adversarial Perturbations

Mohammad Nayeem Teli
Department of Computer Science
University of Maryland
College Park, Maryland 20740

Seungwon Oh
Department of Electrical and Computer Engineering
University of Maryland
College Park, MD 20740

***Abstract*—Due to the vulnerability of deep neural networks to adversarial examples, numerous works on adversarial attacks and defenses have been burgeoning over the past several years. However, there seem to be some conventional views regarding adversarial attacks and object detection approaches that most researchers take for granted. In this work, we bring a fresh perspective on those procedures by evaluating the impact of universal perturbations on object detection at a class-level. We apply it to a carefully curated data set related to autonomous driving. We use Faster-RCNN object detector on images of five different categories: person, car, truck, stop sign and traffic light from the COCO data set, while carefully perturbing the images using Universal Dense Object Suppression algorithm. Our results indicate that person, car, traffic light, truck and stop sign are resilient in that order (most to least) to universal perturbations. To the best of our knowledge, this is the first time such a ranking has been established which is significant for the security of the data sets pertaining to autonomous vehicles and object detection in general.**

I. Introduction

In an autonomous vehicle set up we can categorize the interactions between various components based on the interactions of a vehicle and one or more of the following: landscape (e.g., roads, barriers, buildings, trees etc.), other automobiles, people, and road signs. Autonomous vehicles often recognize each of these categories through the use of Machine Learning models for object detection. While objection detection techniques have made a huge progress, so have the various adversarial attacks, as explained in the next few sections. Most of the studies have heavily concentrated on the techniques of adversarial attacks and the means of making them more and more efficient in breaking the object detector. A lot of research has also concentrated on defense approaches. However, there is very limited or no research done to empirically understand robustness of each individual object category. In this work, we mostly concentrate on specific object categories and their ability to withstand adversarial attacks.

More specifically, we answer questions like, is it more difficult to recognize a person in comparison to a car when an object detector is subjected to universal perturbations? It is very important to understand the impact of such adversarial attacks at a category level so that targeted efforts could be made to address such weaknesses in object detection. Also, different object categories have different features and understanding the affects on them individually would enable us to build systems that can encompass methods to overcome the weaknesses of an object detector.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Due to the vulnerability of deep neural networks to adversarial examples, adversarial attacks have been receiving huge attention over the past several years [1]. In the context of the adversarial attacks, universal perturbations research has also received a lot of interest [2] since it is much more efficient and transferable than image-specific attacks. However, most of it is directed toward image classification and not much toward object detection [3]. More importantly, we observed that there are two conventions that most work on adversarial attacks seem to be holding on: evaluation of adversarial attacks on all (1) **classes** and (2) **images**.

In this paper, we address issues pertaining to the impact of dataset curation and its need, and, the effect of universal perturbations on individual categories of objects. First of all, we question whether the experiments on adversarial attacks have been measuring their impact rigorously, for instance, quantified by the fooling rate (the proportion of images that change labels when perturbed by our universal perturbation). Many experiments have been naively using the whole data set which may lead to assessing the effect of an attack in a more optimistic way. It is owing to the fact that the object detector may already have some incorrect predictions on some of the images in the first place, regardless of the perturbation. The number of those images could be quite high. This is supported by the fact that the mAPs on many data sets are still not high enough to ensure this even on clean images, for example, the COCO AP for the state-of-the-art object detector is still less than 60. Thus, we argue that those examples should not be counted as an example of a successful attack and also be removed from the data set to prevent data set bias from the evaluation.

Secondly, we also observed that there is less research toward understanding the impact of universal perturbations on specific categories related to autonomous driving. Most of the experiments and evaluations are conducted while including all the classes in a data set. This may widen the gap between understanding the effect of universal perturbations in general. In more specific applications a successful adversarial attack may perform differently in the self-driving car domain. Given that deep learning has various applications beyond just self-driving cars, we need to also consider evaluating adversarial attacks at a class-specific level and compare which attack is more deadly to a specific problem, and not just in general. To those who are closely related to the domain, this could have more meaningful implications rather than a general evaluation on a general-purpose data set. In the context of autonomous vehicles, we would like our system to be more reliable, and robust to attacks on the person class than may be those on stop signs. This is due to the fact that the consequences for a vehicle to crash into a person in comparison to a stop sign is vastly different.

In this work, we incorporate data set curation and class-level evaluation of robustness into experimental framework. We present a rigorous evaluation on the impact of universal perturbations on detecting specifically five object categories related to autonomous vehicles: person, stop sign, car, truck and traffic light. We run a series of controlled experiments by perturbing images of these categories using universal perturbations. Finally, we rank the vulnerability of detecting each object category by calculating the metrics for each object category individually and compare them.

We summarize the main contributions of this paper as follows: (1) firstly, we demonstrate the vulnerability of object detectors in detecting autonomous vehicle-related categories to universal perturbations. This is achieved using a simple iterative method of using the gradient of soft-max output with respect to input image and perturbation, (2) secondly, we establish a class-level vulnerability ranking with regards to adversarial attacks, and (3) thirdly, we isolate the impact of adversarial perturbations. This is done by including only those images wherein the detector correctly predicted at least one of the designated classes.

II. Related Works

A. Adversarial Attacks on Object Detection

Deep Neural Networks are ubiquitous in today's world within and outside the lab that have been built to alter the way applications are run. Its impact can be seen all around us. However, it does not come without its pitfalls. They are susceptible to adversarial attacks [4] [5][6][7]. This is a result of many noise injection techniques that cause failure in object detection and classification

Object detection in images, specifically related to autonomous vehicles, that is the focus of this work, come under adversarial attacks as well [8]. Adversarial examples are generated to be similar to the actual data, $\mathbb{D}$, with the goal of a mis-classification. Consider a classification function, f that correctly classifies an image, x, with a label y, and x' gets incorrectly classified with a label $\hat{y}$: $\arg\min_{\eta} f(x+\eta) = \hat{y}$. According to Szegedy et al. [4] we can solve the following optimization problem to find the perturbation

$$\min_{x'} ||x' - x||_p, \qquad s.t. \quad f(x') = \hat{y}$$

where $||.||_p$ is a distance defined on the metric space X. The most commonly used distance metric for object recognition in adversarial attacks is the p-norm:

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

where $p \in \{0, 2, \infty\}$. According to Serban et al. [8] the choice of p is picked as follows:

- $p = 0$ measures the dissimilarity in coordinates between the perturbed and original images with regards to the number of pixels altered in the original image.
- $p = 2$ measures the Euclidean distance between the original and the adversarial image.
- $p = \infty$ measures the maximum bound for altering each pixel in an image without restricting the number of changed pixels.

Deng et al. [9] presented an analysis of adversarial attacks on autonomous driving models using five adversarial attacks and four defense methods on three driving models and found them to be vulnerable. Sharma et al. [10] demonstrated the vulnerability of the machine learning models of autonomous vehicles in their failure of detecting adversarial attacks.

B. Universal Adversarial Perturbation

The attacks explained above are mostly image-specific, However, a much stronger class of adversarial attacks, called universal attack generates a single adversarial example to misclassify all images in the data set [3]. Benz et al. [11] introduced a double targeted universal adversarial perturbations (DT-UAPs) to bridge the gap between the instance discriminative image-dependent perturbations and the generic universal perturbations and its potential as a physical attack. Sitawarin et al. [12] expand the scope of adversarial attacks using Out-of-distribution attacks by enabling the adversary to generate these starting from an arbitrary point in the image space. Prior attacks are restricted to existing training/test data (In-Distribution). They demonstrated that Out-of-Distribution attacks can outperform In-Distribution attacks on classifiers and exposed a new attack vector for these defenses.

Metzen et al. [13] proposed attacks against semantic image segmentation which removes a target class, such as, pedestrians, from the segmentation while leaving the segmentation otherwise unchanged. Li et al. [14] propose, for the first time, an iterative method to generate universal adversarial perturbations against object detection. Yingwei et al. [15] propose that regionally homogeneous perturbations can well transfer across different vision tasks (attacking with the semantic segmentation task and testing on the object detection task) and the defenses are less robust to regionally homogeneous perturbations.

As is noticed in the previous works the emphasis has been on the different adversarial attacks and defenses against such attacks. However, to the best of our knowledge, very little to no work has been performed to understand the robustness of the image categories. Our work focuses on the area of generating

TABLE I: Autonomous vehicle object categories in COCO dataset

category	Number of training samples
person	64115
bicycle	3252
car	12251
truck	6127
bus	3952
motorcycle	3502
traffic light	4139
stop sign	1734

universal adversarial perturbations against object detection, but proposes a new evaluation framework.

III. Dataset curation

In this research, we curate image categories from the popular and challenging COCO dataset by focusing on the categories specific to autonomous vehicles. Since it is an object detection dataset, most of the images would contain more than one category in a single image. This is done by design, since in a real world scenario a given scene would contain more than a single object category. Table I contains the number of such instances for each class in the COCO training set.

Among the vehicle categories, we only look at the top two by the number of instances and leave out bicycle, bus, and motorcycle and finalized with 5 categories: person, car, truck, stop sign, and traffic light for this experiment. Also, we only include the training set in COCO because the number of instances of each category is large enough, and most importantly, we meticulously study the independent impact of perturbations and leave out any incorrect classifications due to the object detection failure. To this end, using the training set on which an object detector has been trained would give us higher probability of collecting images with true positives. Some of the sample images from the dataset are shown in Figure 1.

In order to rule out missed detections due to the failure of object detector, we curate the dataset based on the results of the object detector prior to perturbing an image. As long as the object detector is able to detect at least one instance of the object category, we would retain that image in our dataset. In case the detector missed all the object instances of a target class in an image, that image would not be used further.

IV. Methods

A. Adversarial perturbation

Universal perturbation is an image-independent perturbation that would cause a label change for a number of images [3]. Our goal is to test the impact of such perturbations in object detection. Li et al. [14] introduced Universal Dense Object Suppression (U-DOS) algorithm to find universal perturbations that would blind an object detector on training set, I, so that it may fail to find objects in most of the images in I and at the same time remain imperceptible to a human eye. We use this approach since it is for the first time that the existence of universal perturbations on object detection tasks is performed. We made two modifications to the approach, though: 1) to simplify the loss function, we do not include the background probability in the objective function , and, 2) we use a normalized L1-norm by dividing it by the number of pixels.

U-DOS is an iterative algorithm that updates the universal perturbation, $\mathbf{v}$, by generating a vector, v_i, for each image, $I_i \in I$, so that the detector, D, fails to find any objects in the perturbed image, $I_i + \mathbf{v} + v_i$. We modify the original objective function in Li et al. [14] to only include the sum of class confidence probabilities of detection instances. The modified objective function is given as,

$$\mathcal{F}(I_i, \mathbf{v}, v_i) = \sum_{r_n \in \mathcal{R}_{D, I_i, \mathbf{v}+v_i}} P_{obj}(r_n | I_i + \mathbf{v} + v_i) \tag{1}$$

where, $P_{obj}(r_n|I_i + \mathbf{v} + v_i)$ is the probability of the object class with the highest confidence, and $\mathcal{R}_{D,I_i,\mathbf{v}+v_i}$, is the set of detection results on the perturbed image. Equation 1 is minimized while limiting the magnitude of the perturbation, v_i, as

$$v_i \leftarrow \underset{v_i'}{\text{argmin}} \left[\mathcal{F}(I_i, \mathbf{v}, v_i') + ||\mathbf{v} + v_i'|| \right] \tag{2}$$

To solve for v_i, the algorithm starts with a zero-vector and optimizes it with gradient descent.

$$\begin{aligned} v_i^{(0)} &= \mathbf{0} \\ v_i^{(n+1)} &= v_i^{(n)} - \alpha \nabla_{v_i^{(n)}} (\mathcal{F}(I_i, \mathbf{v}, v_i^{(n)}) + ||\mathbf{v} + v_i^{(n)}||) \end{aligned} \tag{3}$$

where, α is the learning rate and the gradient, $\nabla_{v_i^{(n)}}(*)$, is computed with back-propagation.

We modified Li et al.'s objective function to make it more applicable to object detectors in general: originally the paper included the term of increasing the background class probability and decreasing the ones for specific/all object class, but we removed the term increasing the probability of the background class because not all object detectors output a soft max probability that includes the background class.

B. Metrics

In order to evaluate the performance of the universal perturbations in blinding the object detector, we use two metrics proposed by [14]: **image-level blind degree** and **instance-level blind degree**.

1) Image-level blind degree: Image-level blind degree is defined as the ratio of the images in which the object detector, D, can find at least one object with confidence above a threshold, θ, to the total number of images, N. It is expressed as

$$B_{img}(D, \mathbf{v}, \theta) = \frac{\sum_{i=1}^{N} ind(I_i, D, \mathbf{v}, \theta)}{N} \tag{4}$$

where, N is the total number of images, and I_i denotes the i-th image, $\mathbf{v}$ the universal perturbation added to these images, $ind(I_i, D, \mathbf{v}, \theta)$ is an indicator function represented as

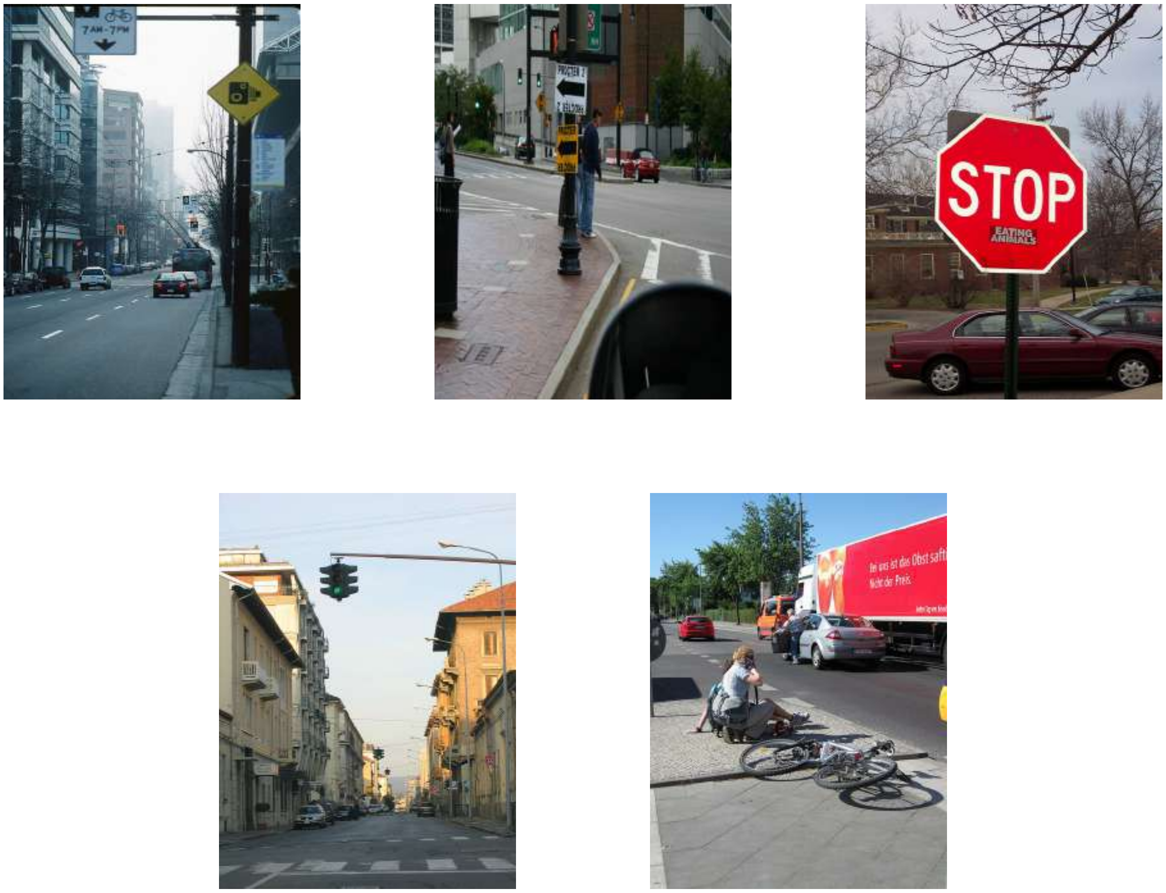


Fig. 1: Sample images of the selected object categories

$$ind(I_i, D, \mathbf{v}, \theta) = \begin{cases} 1, & \text{if } \exists r \in \Re_{D,I_i,\mathbf{v}}, P_{obj}(r|I_i + \mathbf{v}) > \theta \\ 0, & \text{otherwise} \end{cases}$$

where the set of detection results on the perturbed image is $D(I_i+\mathbf{v})$ given by $\Re_{D,I_i,\mathbf{v}} = D(I_i+\mathbf{v})$, with $\mathbf{v} \in \mathbb{R}^d$ being the perturbation. $P_{obj}(r_n|I_i+\mathbf{v}+v_i)$ is the probability of the object class with the highest confidence and the image-independent perturbation, $\mathbf{v}$, must satisfy the following two constraints:

1) $\Re_{D,I_i,\mathbf{v}} = \emptyset$
2) $||\mathbf{v}||_\infty < \xi$

2) *Instance-level blind degree:* Instance-level blind degree calculates the average number of instances that the detector, D finds in each image with confidence beyond the given threshold θ and is expressed as

$$\mathcal{B}_{ins}(D, \mathbf{v}, \theta) = \frac{\sum_{i=1}^{N} |\{r|r \in \mathcal{R}_{D,I_i,\mathbf{v}} \text{ and } P_{obj}(r|I_i + \mathbf{v}) > \theta\}|}{N} \tag{5}$$

where $|*|$ denotes the number of elements in the set.

V. Experiments

In order to test the impact of perturbations on our selected categories, we use Faster-RCNN-FPN-Resnet50 (trained on COCO2017 train) model in Detectron2 [16] [17] library because of its high mean Average Precision (mAP) on the COCO2017 test. We also fine-tuned three hyperparameters: number of epochs(n_epoch), step size (α) and ξ for the max inf-norm of the perturbations. After a series of pilot experiments, we settled on the following values for the parameters: $n\_epoch = 250, \xi$ (max l_p norm) = 10, $\alpha = 20, max\_imgs = 500$, and $score\_threshold, \theta = 0.7$. In our experiments, we keep track of the instance-level and image-level blind degree as we vary perturbation norm for each of the five categories of objects. This is done so that the object detector fails to find objects of a given class at a certain norm. Figure 2 visually shows two examples of the object detector blindness before and after the perturbations.

To evaluate universal perturbations on each category, we conduct a targeted attack on each category. We compute an universal perturbation that attacks only the targeted category by considering only the object proposal classified that category

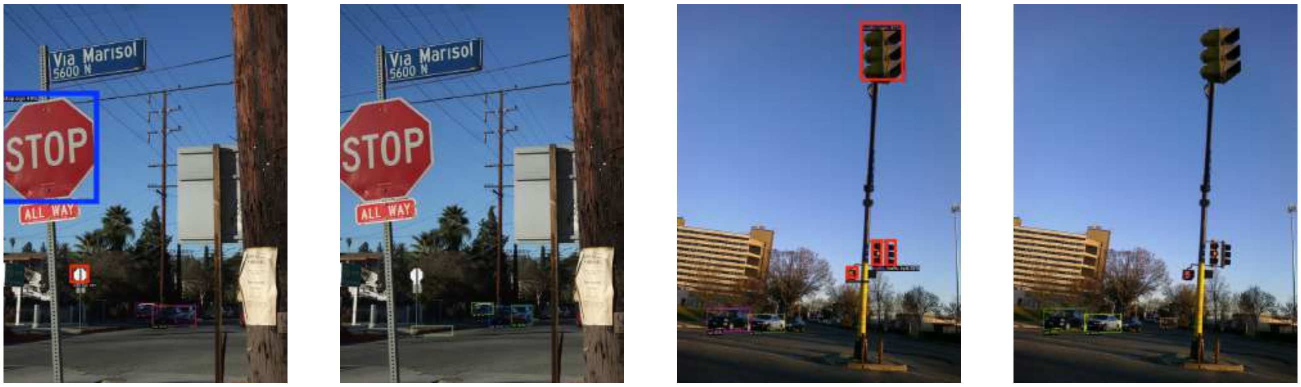


Fig. 2: Samples detection results with and without perturbation. 1st and 3rd columns: detection with clean images. 2nd and 4th columns: detection with perturbed images.

TABLE II: Rank of image level and instance level blind degree resilience of different categories to perturbations as epochs and norm increase

Epochs		Norm	
Image	**Instance**	**Image**	**Instance**
person	person	person	person
stop sign	car	car	car
car	traffic light	traffic light	traffic light
traffic light	stop sign	truck	truck
truck	truck	stop sign	stop sign

and optimizing on the loss based on those object proposals to remove final detections of that class. Figure 3 presents the effect of perturbations on each category as the epochs and the norm are increased.

Based on Figure 3, Table II presents the ranks of resilience of each of our categories. It is evident that the person is the most resilient category at image-level as well as instance-level, followed by car, traffic light, truck and stop sign as the norm is increased. It is also clear that person category is again most resilient to perturbations as you increase the number of steps in projected gradient descent to find a stronger perturbation. However, norm is a more significant metric as it is related to the visibility of the perturbations. It is also notable that a single universal perturbation is able to make the detector fail in detecting objects of all categories, except for, the person category in most of the images, at a norm which is imperceptible to humans.

VI. Conclusion

In this paper, we explore the vulnerability of object detection on five autonomous vehicle object categories to universal perturbations. In diligently designed experiments, we carefully curated the realistic and diverse COCO data set and isolate the failure of object detection without perturbations, from the failure with perturbations. This work lies at the core of the security of data sets in the realm of autonomous vehicle driving. To the best of our knowledge, this is the first study of its kind that ranks the resilience of these five categories to universal perturbations. These results are significant since the perturbations can be imperceptible to humans while causing real harm in an adversarial attack. We believe a more fine-grained and detailed view on the experiments can be beneficial to the rigorous and evaluation of adversarial attacks.

VII. Future Work

Having established a categorical order based on the resilience of image categories to adversarial attacks, this research could take many different directions. One of the possible future effort would be to build and train object detectors while incorporating special features of each of the categories. Another possible direction would be to study the reasons behind the resilience of each of these categories and employ that in training robust object detectors. A different approach of future work would involve studying various new perturbation techniques to further impact the resilience of the object categories. We could also explore multiple new vehicle data sets and explore the effect of perturbations across data sets and algorithms.

References

[1] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[2] A. Chaubey, N. Agrawal, K. Barnwal, K. Guliani, and P. Mehta, “Universal adversarial perturbations: A survey,” 05 2020.

[3] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

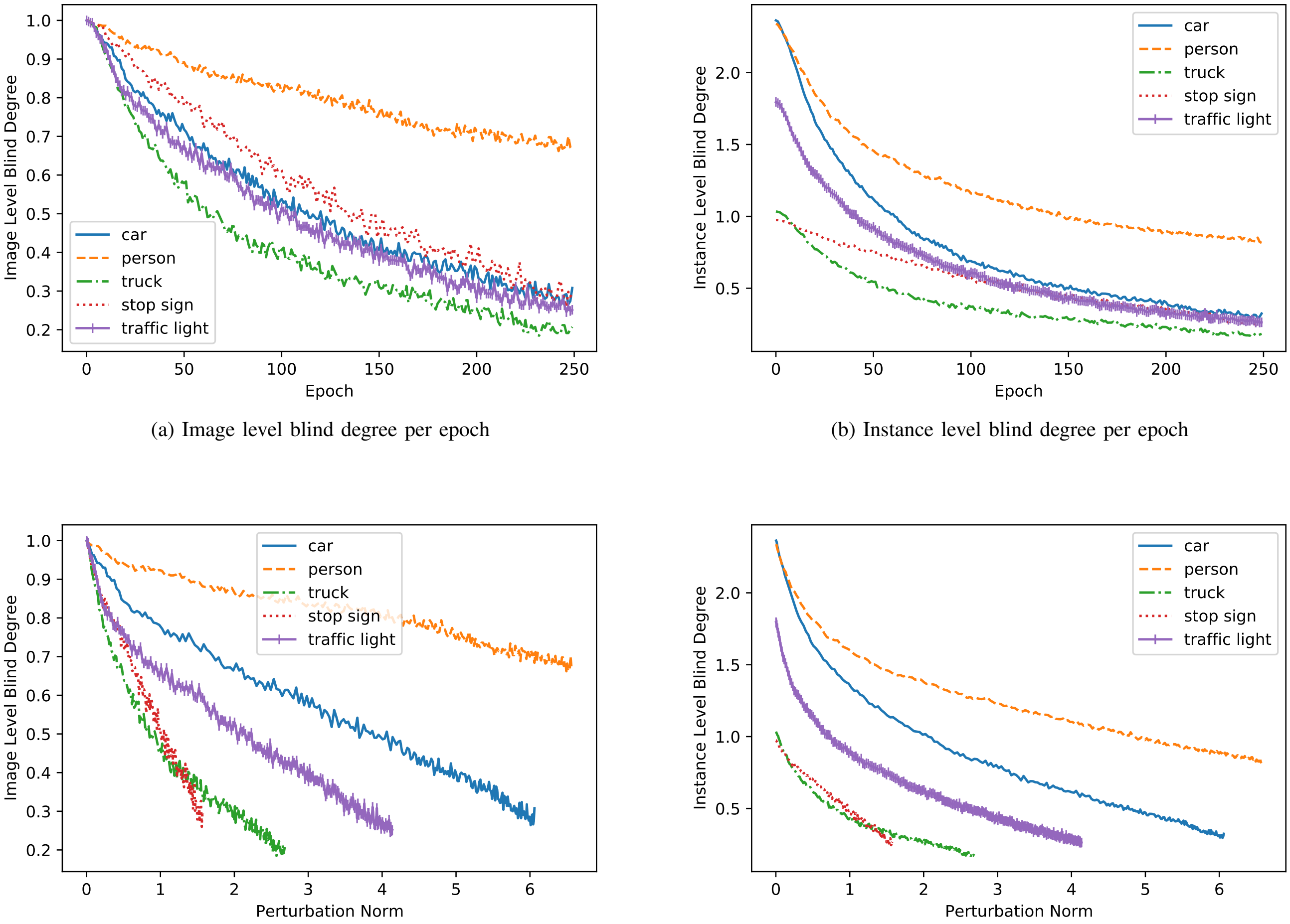
[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014.

[5] M. Ozdag, “Adversarial attacks and defenses against deep neural networks: A survey,” *Procedia Computer Science*, vol. 140, pp. 152–161, 2018, cyber Physical Systems and Deep Learning Chicago, Illinois November 5-7, 2018.

[6] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[7] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “Adversarial attacks and defences: A survey,” *CoRR*, vol. abs/1810.00069, 2018.

[8] A. Serban, E. Poll, and J. Visser, “Adversarial Examples on Object Recognition: A Comprehensive Survey,” *arXiv e-prints*, p. arXiv:2008.04094, Aug. 2020.



(a) Image level blind degree per epoch

(b) Instance level blind degree per epoch

(c) Image level blind degree vs. norm

(d) Instance level blind degree vs. norm

Fig. 3: Image-level and instance-level blind degree vs. epochs and norms. Both blind degrees are decreased as we increase the norm of perturbation.

[9] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim, "An analysis of adversarial attacks and defenses on autonomous driving models," in *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2020, pp. 1–10.

[10] P. Sharma, D. Austin, and H. Liu, "Attacks on machine learning: Adversarial examples in connected and autonomous vehicles," in *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, 2019, pp. 1–7.

[11] P. Benz, C. Zhang, T. Imtiaz, and I. S. Kweon, "Double Targeted Universal Adversarial Perturbations," *arXiv e-prints*, p. arXiv:2010.03288, Oct. 2020.

[12] C. Sitawarin, A. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "Darts: Deceiving autonomous cars with toxic signs," *ArXiv*, vol. abs/1802.06430, 2018.

[13] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2774–2783.

[14] D. Li, J. Zhang, and K. Huang, "Universal adversarial perturbations against object detection," *Pattern Recognition*, p. 107584, 2020.

[15] Y. Li, S. Bai, C. Xie, Z. Liao, X. Shen, and A. L. Yuille, "Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12356. Springer, 2020, pp. 795–813.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[17] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019.

Ontology Creation Model based on Attention Mechanism for a Specific Business Domain

1st Maryam Heidari
George Mason University
mheidari@gmu.edu

2nd Samira Zad
Florida International university
Szad001@fiu.edu

3nd Brett Berlin
George Mason University
fberln@gmu.edu

4nd Setareh Rafatirad
George Mason University
srafati@gmu.edu

***Abstract*—Ontology can describe real-world objects and relationships. It can be used for searching, web mining, and semantic analysis. However, creating the ontology for a different domain is a labor-intensive and computationally expensive task. This work introduces a new model to automate ontology creation for the real estate domain based on a natural language processing approach. The new model can save significant time and human labor in ontology creation in several domains. Also, the new model can enhance retrieval information systems in advanced semantic web search engines.**

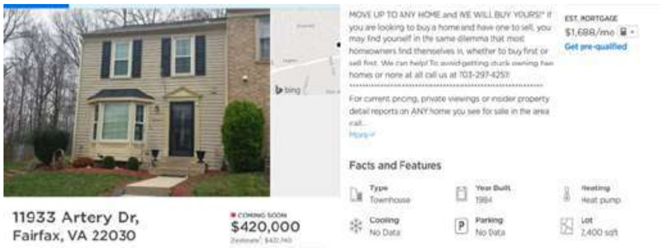
I. Introduction

In this work, we apply attention model in natural language processing to create real estate ontology. The real estate market has two main categories, commercial and residential. several studies have been done in the creation of ontology for the commercial real estate [1] [2] [3] [4]. Real estate website provides a wide range of housing data, which includes: property address, number of bedrooms, bathroom and all information related to the house [5] [6], like Zillow [7] and Trulia [8]. Figure 1 shows the example of information that is available on the Zillow website for each house.

Machine learning models can be used in different aspect of scientific research such as business [9]health [10], [11] [12] [13] [14], cyber security [15] [16], [16], environment [17], IOT [18] [19], Data maintenance systems [20]–[22], computer science applications [23], [24] [25] and social science [26].

Natural language processing models [27] [28] and computer vision models [29], [30] [31] can be used in advanced semantic web applications. Model optimization for graph neural networks [32] [33] [34] algorithms play an important role to make artificial intelligence applications more powerful. Natural language processing applications in social media [35] [36] [37] [38] can be used to enhance semantic web search and ontology creation.

In this work, we create an ontology for residential real estate. Each house on the Zillow website has the " overview section," which is a rich source of textual information for creating real estate ontology. This section provides a textual description of the house, which can help extract different contexts related to the real estate domain. Building the ontology for the entities and their relationships in the house description can help analyze and mine interesting knowledge, such as finding the relationship between house price and house properties, the influence of the neighborhood, and cluster similar real estate properties.



MOVE UP TO ANY HOME and WE WILL BUY YOURS!" If you are looking to buy a home and have one to sell, you may find yourself in the same dilemma that most homeowners find themselves in, whether to buy first or sell first. We can help! To avoid getting stuck owning two homes or none at all call us at 703-297-4257!

****

For current pricing, private viewings or insider property detail reports on ANY home you see for sale in the area call...

More

EST. MORTGAGE
$1,688/mo
Get pre-qualified

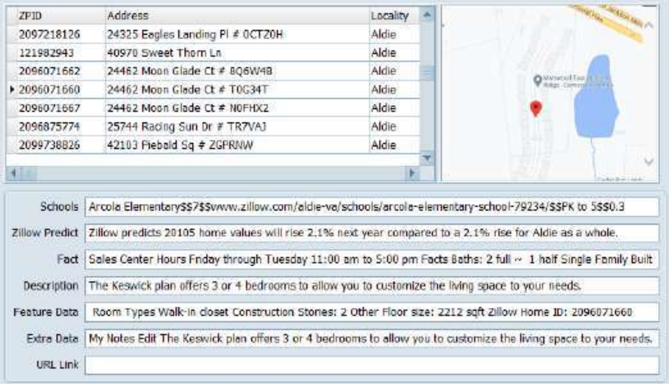
Facts and Features

Type Townhouse	Year Built 1984	Heating Heat pump
Cooling No Data	Parking No Data	Lot 2,400 sqft

11933 Artery Dr,
Fairfax, VA 22030

COMING SOON
$420,000
Zestimate®: [illegible]

Fig. 1. Zillow real estate website



ZPID	Address	Locality
2097218126	24325 Eagles Landing Pl # 0CTZ0H	Aldie
121982943	40970 Sweet Thorn Ln	Aldie
2096071662	24462 Moon Glade Ct # 8Q6W4B	Aldie
2096071660	24462 Moon Glade Ct # T0G34T	Aldie
2096071667	24462 Moon Glade Ct # N0FHX2	Aldie
2096875774	25744 Racing Sun Dr # TR7VAJ	Aldie
2099738826	42103 Piebald Sq # ZGPRNW	Aldie

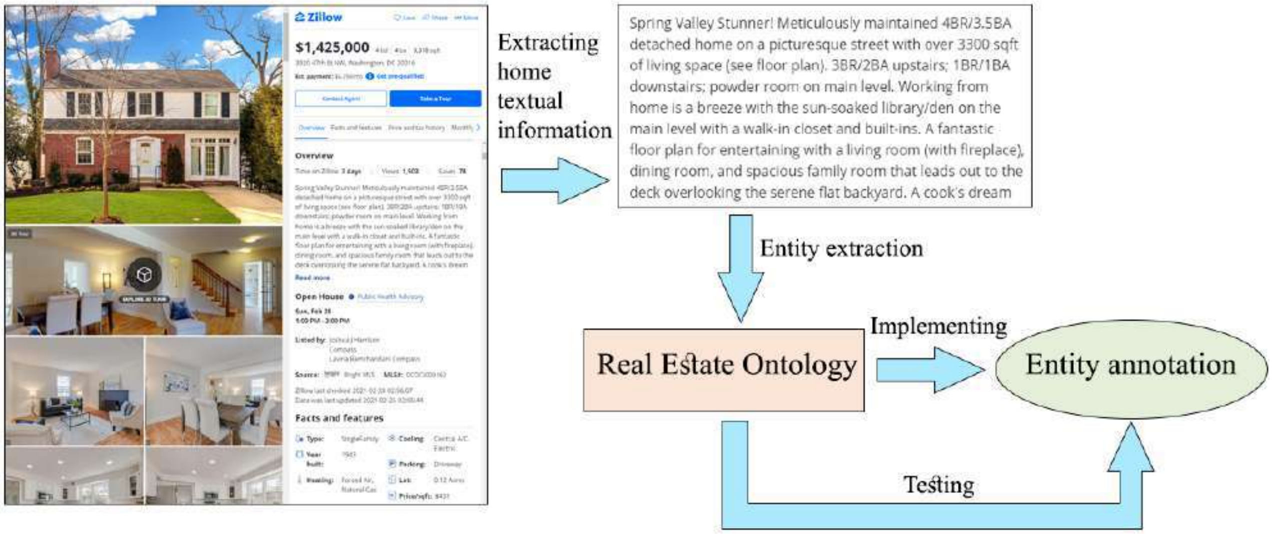
Schools: Arcola Elementary$$7$$www.zillow.com/aldie-va/schools/arcola-elementary-school-79234/$$PK to 5$$0.3
Zillow Predict: Zillow predicts 20105 home values will rise 2.1% next year compared to a 2.1% rise for Aldie as a whole.
Fact: Sales Center Hours Friday through Tuesday 11:00 am to 5:00 pm Facts Baths: 2 full ~ 1 half Single Family Built
Description: The Keswick plan offers 3 or 4 bedrooms to allow you to customize the living space to your needs.
Feature Data: Room Types Walk-in closet Construction Stories: 2 Other Floor size: 2212 sqft Zillow Home ID: 2096071660
Extra Data: My Notes Edit The Keswick plan offers 3 or 4 bedrooms to allow you to customize the living space to your needs.
URL Link:

Fig. 2. Real estate data example

This project can enhance information retrieval systems and increase customer satisfaction in commercial and residential real estate. For example, when someone wants to sell a house, the system can estimate the house's price based on statistical and textual information and find a potential buyer. Another benefit of real estate ontology is that it improves clustering models for a real estate property and makes it more convenient to search for a specific house. In this way, our project has two major contributions, house price estimation and advanced semantic search for real estate properties. First, we build a domain-specified ontology based on a semi-automated approach since we use human annotators to extract entities and make a ground truth to test the new proposed model. Second, we automate the ontology creation process by using a natural language processing approach, attention model in transformers [39]–[41].

In this project, the main goal is to extract internal and external house features from the textual information to create

978-1-6654-4067-7/21/$31.00 ©2021 IEEE



Spring Valley Stunner! Meticulously maintained 4BR/3.5BA detached home on a picturesque street with over 3300 sqft of living space (see floor plan). 3BR/2BA upstairs; 1BR/1BA downstairs; powder room on main level. Working from home is a breeze with the sun-soaked library/den on the main level with a walk-in closet and built-ins. A fantastic floor plan for entertaining with a living room (with fireplace), dining room, and spacious family room that leads out to the deck overlooking the serene flat backyard. A cook's dream

Fig. 3. Ontology creation process by human annotators

real estate ontology. We have three milestones for doing this project: data collection, building ontology, and testing the model.

1) For each house on the Zillow website, we collect both house descriptions and house features. Data includes house description(textual information), address, zip code, number of bedrooms and bathrooms, rent price, and city.
2) In the semi-automated step, human annotators extract information from real estate descriptions and find real state entities and their relationships. The human-created ontology is the ground truth for our new model. Also, this ontology is checked by a human expert in the real estate domain. Moreover, we modify this ontology based on the expert's idea.
3) Design an NLP model which can automate ontology creation. Test the final model with a ground truth ontology which is created in the first phase.

II. Data Collection

We write a customized software application to collect real estate data by using Zillow API. Using Zillow APIs, we collect about one million real estate records from the website based on three different house types: single-family, townhouse, and condo. The data will be separated into two parts. One part is for ontology creation, and the other part is for testing the ontology. Figure 2 shows our application and the example of data provided by this application.

In this data set, we fully utilize the text part, which describes the real estate information. The texts are unstructured datasets, and they need to be clean before doing text mining. Each record in this work is a text which property owner provides. For example, in the house description, the textual information includes bedroom, kitchen, living room, floor description. These words are the entity that we need to provide more annotation for them. All records in the data set need to have semantic annotation. The adjective words are also important entities for our ontology.

III. Method

A. Text mining and Building ontology by human expert

Figure 3 shows the process of building ontology by human annotators. All textual information is extracted and divided into two groups—one for ontology implementation and the other for testing.

In the First step, We use amazon sage maker ground truth [42] to annotate all entities and create the original real estate ontology by a human expert.

In the second step, to test the created ontology in the first step, human annotators use the ontology in step one to annotate unseen real estate text data set to test the entity annotation method: real estate entities includes bedrooms, bathrooms, area space from the description section on the Zillow website. The entities extracted in the second step show that how much the first step should be repeated to create a more precise and reliable ontology as a ground truth. We continue this iterative process until the model achieves 94% accuracy.

B. Building ontology with OntoGen

OntoGen is software for ontology creation. We use OntoGen to extract the real estate entities and then compare the results with the ground truth ontology for this research. OntoGen assigns the words into several clusters. Also, These clusters give a conceptual idea about how many clusters we should expect in the ground truth ontology. The word frequency in the home description's texts should be similar to the word frequency of the annotated texts by ground truth ontology. Each word is considered as an element or entity. We compare the most frequent elements with our ontology entities to check the correct entity extraction by ontology.

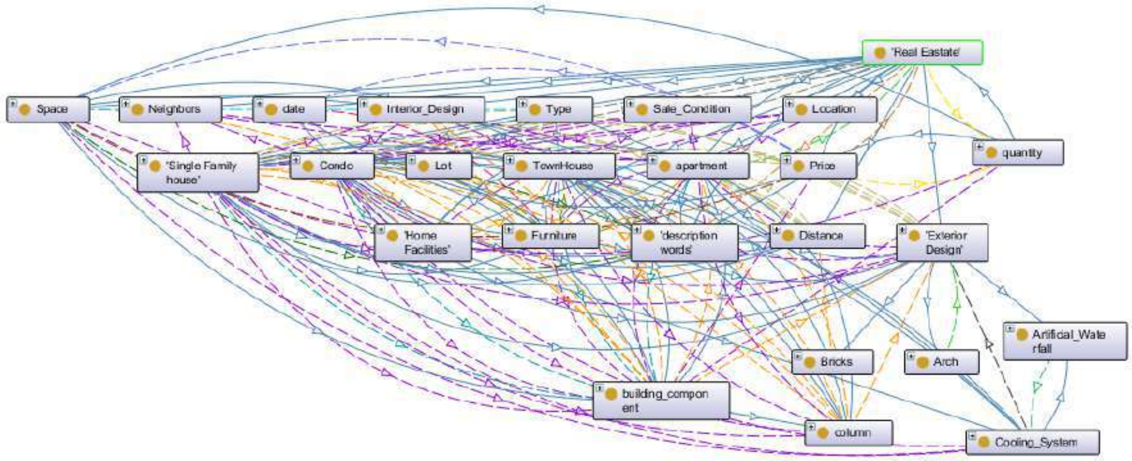


Fig. 4. Ontology based on attention mechanism

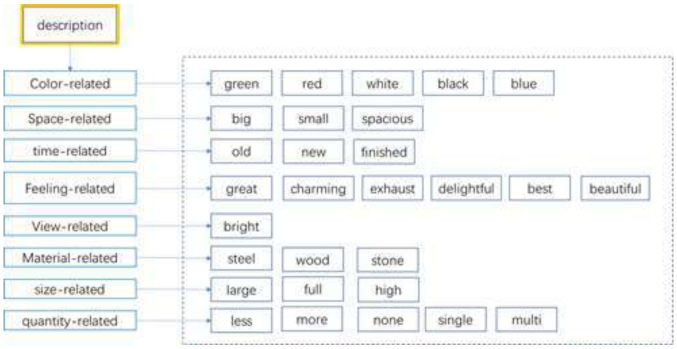


Fig. 5. Descriptive words

C. Building ontology with Protege

We build the ontology with the Protégé. There are four classes: date, description words, quantity, and real estate words. Date class has a subclass of day, month, and year. Description words include all descriptive words in the text. Figure 5 shows descriptive words. The real estate class includes many elements of house features, including room, exteriors, indoor places. The ground truth ontology has the same elements and classes as the Protege ontology.

D. Building ontology with attention model as a natural language processing approach

In this work, to automate the ontology creation, we use the self-attention mechanism in the transformer [43] [44]. We create embedding for each sentence in the real estate data set. Each transformer contains multi-head attention layers that can create multi attention mechanism about each specific real estate word and learn how to get several aspects about one single word in the house description [45]. Figure4 shows the ontology which is created based on the attention mechanism. Self-attention mechanism can be explained as embedding vector for token input, the query (Q), key (K) and value (V) are created from each three parameter matrices where $W^Q \in \mathbb{R}^{d_{model}\times d_k}$,$W^K \in \mathbb{R}^{d_{model}\times d_k}$ and $W^V \in \mathbb{R}^{d_{model}\times d_v}$. So this mechanism can be shown as [46]:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

$d_k = d_v = d_{model} = 768$ in BERT base version which is used in this research.

The class label's occurrence in the ontology entity is shown with $C \in \mathbb{R}^H$, and H is the embedding size of each sentence. So a dense neural layer is treated as a classification layer, which consists of parameters $W \in \mathbb{R}^{K\times H}$, where K is the number of possible class labels that we have in our data set. The class label here means each entity will be paired by multiple entities in ontology, but we choose the one with the highest probability. The prediction probabilities $P \in \mathbb{R}^K$ are calculated by a softmax activation function is:

$$P = softmax(CW^T + b) \tag{2}$$

In this research, all the parameters optimized to minimize the loss function [46].

$$\mathcal{L} = -\frac{1}{N}\sum log\left(\hat{p_c}^{(i)}\right) \tag{3}$$

IV. Test the Ontology Based on Attention Mechanism

To test the ontology with the attention mechanism, we compare the ground truth ontology results with our new ontology. Here are two examples of the results.

In name annotation, we consider the words of the same meaning but in different formats. For example, the words ”high” and ”High” should denote the same word. To distinguish these similar words, we assign each class some neighbor's words. The neighbors help to correct mapping from an instance to an entity. For example, the word ”garden” can be recognized as a garden within the house or park. We give park the neighbors “go” “to” “near” “around”. We give the garden within the house the neighbors of ”have,” ”within.” For the

TABLE I
EXAMPLE OF ENTITY ANNOTATION BY NEW ONTOLOGY BASED ON ATTENTION

Raw text	The Remington Place II model is a gorgeous luxury home that opens into a two-story foyer with a full window wall. The main floor features an expansive kitchen and family room area with an open layout that is perfect for entertaining or dining in. The Remington Place II is designed with a flexible floor plan layout ∼with a choice for the laundry room to located on either the first or second floor. Additionally ∼residents can choose to upgrade with another bedroom and full bathroom on the main level. The upstairs is built with four full bedrooms and a landing space that overlooks the family room. The Remington Place . . .
Annotation by machine	Full, window, kitchen, bedroom, full, More, Less
Mapping by machine	Less ['is type of quantity_realted', 'is subclass of description words'] More ['is type of quantity_realted', 'is subclass of description words'] bedroom ['is type of space', 'is subclass of realestate'] full ['is type of size_realted', 'is subclass of description words'] kitchen ['is type of space', 'is subclass of realestate'] window ['is type of exterior_design', 'is subclass of realestate']
Annotation by human	Story, foyer, window, wall, floor, kitchen, family room, area, plan, laundry, bedroom, bathroom, door, ceiling, closet, vanity mirrors, shower, soaking tube,
Mapping by human	Story [' is type of building component ', 'is subclass of space', 'has a level'] Foyer [' is type of space', 'is subclass of real estate', 'has a size'] Window [' is type of building component', 'is subclass of space', 'has a size'] Wall [' is type of building component', 'is subclass of space', 'has a size'] Floor [' is type of building component', 'is subclass of space', 'has a floor covering'] Kitchen [' is type of space ', 'is subclass of real estate', 'has a tubing system] Family room [' is type of room ', 'is subclass of space', 'has some wall'] Area [' is type of size', 'is subclass of description words', 'has a value'] Plan [' is type of design', 'is subclass of architecture', 'has a location'] Laundry [' is type of cleaning space', 'is subclass of space', 'has a watering system'] Door [' is type of building component', 'is subclass of space', 'has a color'] Ceiling [' is type of building component ', 'is subclass of space', 'has an area']

recognized word "Garden," we find its won neighbors. Then we can compute the probability of these words belong to the two promising entities based on the matching neighbors.

Table I are examples based on our new ontology. The new ontology can extract the real estate entities with high accuracy. Each table contains raw data, annotated data, and mapped data.

V. CONCLUSION

This work created a human-annotated ontology by a human expert in the real estate domain, which can be used as ground truth for ontology creation in this domain. We also introduce a new model for the automated creation of ontology based on the transformer model's attention mechanism to save computational resources and time. In future work, The proposed ontology and annotation function can help to store the information of a house in a structured format for advanced semantic web search engines for the real estate market.

REFERENCES

[1] Qiu Xin, "Business solution for luxury housing market based on e-catalog ontology," in *2011 International Conference on Business Management and Electronic Information*, vol. 4, pp. 146–150, 2011.

[2] B. Di Martino, C. Mirarchi, F. F. D'Abrunzo, L. Fiorillo, N. Iovine, C. Trebbi, and A. Pavan, "A semantic and rule based technique and inference engine for discovering real estate units in building information models," in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 81–88, 2019.

[3] A. Alekseev, E. Galiaskarov, and K. Koskova, "Application of the matrix rating mechanisms and system cognitive analysis methods at the task of residential real estate conceptual designing," in *2019 IEEE 21st Conference on Business Informatics (CBI)*, vol. 02, pp. 111–116, 2019.

[4] N. Rastogi, P. Verma, and P. Kumar, "Evaluation of information retrieval performance metrics using real estate ontology," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 102–106, 2020.

[5] E. Stubkjaer, *The ontology and modelling of real estate transactions*. Routledge, 2107.

[6] K. Pancerz and P. Grochowalski, "From unstructured data included in real-estate listings to information systems over ontological graphs," in *2017 International Conference on Information and Digital Technologies (IDT)*, pp. 298–303, 2017.

[7] "Zillow realestate website." https://www.zillow.com/, 2020/03/02.

[8] "Trulia realestate website." https://www.trulia.com/, 2019/02/02.

[9] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[10] D. G. A. A. M. M. S. J. M. S. J. K. J. B. M.-C. H. Golnoush Asaeikheybari, Cory Hughart, "Precision hiv health app, positive peers, powered by data harnessing, ai, and learning," in *IEEE 2020 Second International Conference on Transdisciplinary AI (TransAI), TransAI 2020*, 2020.

[11] G. Asaeikheybari, J. Green, X. Qian, H. Jiang, and M.-C. Huang, "Medical image learning from a few/few training samples: Melanoma segmentation study," *Smart Health*, vol. 14, p. 100088, 2019.

[12] M. Saadati, J. Nelson, and H. Ayaz, "Mental workload classification from spatial representation of fnirs recordings using convolutional neural networks," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2019.

[13] N. Nazari, S. A. Mirsalari, S. Sinaei, M. E. Salehi, and M. Daneshtalab, "Multi-level binarized lstm in eeg classification for wearable devices," in *2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pp. 175–181, IEEE, 2020.

[14] N. Nazari, M. Loni, M. E. Salehi, M. Daneshtalab, and M. Sjodin, "Totnet: An endeavor toward optimizing ternary neural networks," in *2019 22nd Euromicro Conference on Digital System Design (DSD)*, pp. 305–312, IEEE, 2019.

[15] M. R. Izadi, R. Stevenson, and L. N. Kloepper, "Separation of over-

lapping sources in bioacoustic mixtures," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1688–1696, 2020.

[16] N. Etemadyrad and J. K. Nelson, "A sequential detection approach to indoor positioning using rss-based fingerprinting," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1127–1131, IEEE, 2016.

[17] H. Soltani-Jigheh and S. T. Yaghoubi, "Effect of liquid polymer on properties of fine-grained soils," 2019.

[18] N. H. Tonekaboni, S. Kulkarni, and L. Ramaswamy, "Edge-based anomalous sensor placement detection for participatory sensing of urban heat islands," in *2018 IEEE International Smart Cities Conference (ISC2)*, pp. 1–8, IEEE, 2018.

[19] N. H. Tonekaboni, L. Ramaswamy, D. Mishra, A. Grundstein, S. Kulkarni, and Y. Yin, "Scouts: A smart community centric urban heat monitoring framework," in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Advances on Resilient and Intelligent Cities*, pp. 27–30, 2018.

[20] G. Mohsenian, S. Khalili, and B. Sammakia, "A design methodology for controlling local airflow delivery in data centers using air dampers," in *2019 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 905–911, 2019.

[21] S. Khalili, G. Mohsenian, A. Desu, K. Ghose, and B. Sammakia, "Airflow management using active air dampers in presence of a dynamic workload in data centers," in *2019 35th Semiconductor Thermal Measurement, Modeling and Management Symposium (SEMI-THERM)*, pp. 101–110, 2019.

[22] M. Tradat, G. Mohsenian, Y. Manaserh, B. Sammakia, D. Mendo, and H. A. Alissa, "Experimental analysis of different measurement techniques of server-rack airflow predictions towards proper dc airflow management," in *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 366–373, 2020.

[23] O. Saremi, M. S. Panahi, and A. Sabzehzar, "An improved continuous-action extended classifier systems for function approximation," *Procedia Computer Science*, vol. 61, pp. 361–366, 2015.

[24] A. Sabzehzar, W. Shan, M. S. Panahi, and O. Saremi, "An improved extended classifier system for the real-time-input real-time-output (xcsrr) stability control of a biped robot," *Procedia Computer Science*, vol. 61, pp. 492–499, 2015.

[25] N. H. Tonekaboni, L. Ramaswamy, D. Mishra, O. Setayeshfar, and S. Omidvar, "Spatio-temporal coverage enhancement in drive-by sensing through utility-aware mobile agent selection," in *International Conference on Internet of Things*, pp. 108–124, Springer, 2020.

[26] S. Voghoei, N. Hashemi Tonekaboni, D. Yazdansepas, and H. R. Arabnia, "University online courses: Correlation between students' participation rate and academic performance," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 772–777, 2019.

[27] G. Beigi, A. Mosallanezhad, R. Guo, H. Alvari, A. Nou, and H. Liu, "Privacy-aware recommendation with private-attribute protection using adversarial learning," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, (New York, NY, USA), p. 34–42, Association for Computing Machinery, 2020.

[28] A. Mosallanezhad, G. Beigi, and H. Liu, "Deep reinforcement learning-based text anonymization against private-attribute inference," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 2360–2369, Association for Computational Linguistics, Nov. 2019.

[29] S. Amirian, Z. Wang, T. R. Taha, and H. R. Arabnia, "Dissection of deep learning with applications in image recognition," in *Computational Science and Computational Intelligence; "Artificial Intelligence" (CSCI-ISAI); 2018 International Conference on. IEEE*, pp. 1132–1138, 2018.

[30] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic generation of descriptive titles for video clips using deep learning," in *Springer Nature - Research Book Series:Transactions on Computational Science Computational Intelligence*, p. Springer ID: 89066307, 2020.

[31] M. R. Izadi, "Feature level fusion from facial attributes for face recognition," *arXiv preprint arXiv:1909.13126*, 2019.

[32] M. U. Nisar, S. Voghoei, and L. Ramaswamy, "Caching for pattern matching queries in time evolving graphs: Challenges and approaches," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2352–2357, 2017.

[33] M. R. Izadi, Y. Fang, R. Stevenson, and L. Lin, "Optimization of graph neural networks with natural gradient descent," *arXiv preprint arXiv:2008.09624*, 2020.

[34] S. Voghoei, N. Hashemi Tonekaboni, J. G. Wallace, and H. R. Arabnia, "Deep learning at the edge," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 895–901, 2018.

[35] M. Heidari, J. H. J. Jones, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[36] S. Zad and M. Finlayson, "Systematic evaluation of a framework for unsupervised emotion recognition for narrative text," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pp. 26–37, 2020.

[37] H. Karbasian and A. Johri, "Insights for curriculum development: Identifying emerging data science topics through analysis of q&a communities," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pp. 192–198, 2020.

[38] L. Madahali, L. Najjar, and M. Hall, "Exploratory factor analysis of graphical features for link prediction in social networks," in *International Workshop on Complex Networks*, pp. 17–31, Springer, 2019.

[39] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network to recommend airline tickets by using online reviews," in *IEEE 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020*, 2020.

[40] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[41] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[42] "Amazon sagemaker ground truth." https://aws.amazon.com/sagemaker/groundtruth/.

[43] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[44] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[46] Y. Huang, S. Lee, M. Ma, Y. Chen, Y. Yu, and Y. Chen, "Emotionx-idea: Emotion BERT - an affectional model for conversation," *CoRR*, vol. abs/1908.06264, 2019.

An Empirical Study of Machine learning Algorithms for Social Media Bot Detection

1st Maryam Heidari
George Mason University
mheidari@gmu.edu

2nd James H Jr Jones
George Mason University
jjonesu@gmu.edu

3nd Ozlem Uzuner
George Mason University
ouzuner@gmu.edu

***Abstract*—Social media bots can change society's perspective in different aspects of life. This paper analyzes sentiment features and their effect on the accuracy of machine learning models for social media bot detection. Social bots can use tweet sentiment to create a backfire effect and confirmation bias to create a fake trend or change public opinion. We analyze bot detection problems based on sentiment features inspired by the work by Micheal Workman [1] and create new features based on textual information of online comments. We offer a quantitative approach to create new features and compare machine learning models for bot detection. This work is based on psychological and social effects inherent in tweets' text content based on the work by [1]. The new set of sentiment features are extracted from a tweet's text and used to train bot detection models. Also, we implement the new model for the Dutch language and achieve more than 87% accuracy for the Dutch tweets based on new sentiment features. Considering new sentiment features based on psychological and social factors for a tweet's text will open a potential research area for social media bot detection.**

***Index Terms*—Social media bot, Deep learning, Bot detection, Social factors**

I. Introduction

Bots spread disinformation which can be hard to detect just based on content [2] [3] [4] [5], but advanced methods can detect bot with high accuracy [6] [7] [8]. Social media bot detection can use the offensive language detection [9] [10] or other text characteristic to detect bots. Social media bots can target various audiences by creating fake trends [11], [12]. In the study conducted by Cresci et al. [13], the authors identified multiple types of spam bots, including promoter bots, URL spam bots, and fake followers. URL spam bots spread scam URL links by embedding these malicious links in retweets of legitimate users [14]. Based on the studies by Howard et al. [15], URL sharing bots are used for constant tweet duplication of legitimate users in a certain time to spread malicious URLs. One popular bot detection service is "Botmeter", which is a supervised learning approach to detect social bots [16]. Botmeter uses metadata related to each Twitter account, such as network features, user features, and temporal features, to feed a Random Forest classifier algorithm. Network features show how information diffusion happens among multiple groups of users. User features include a user name, screen name, the creation time of account and geographic location, and temporal features show patterns in a tweet's timeline. Graph-based approaches for bot detection uses fully connected nodes among a group of Twitter accounts to detect bots [17].Machine learning models have different applications in health, cyber security and business [18]. Transfer learning models also can be used in bot detection models [19]–[22].

The idea of extraction of new sentiment features for bot detection in this work inspired by the work [1] and the effect on confirmation bias and backfire effect on online users. Micheal et at [1] in his work said, "People have different ways of cognitively formulating concepts and processing information, referred to as cognitive styles. When considering a problem or an issue, some people are more self-reflective and self-reliant in terms of these cognitive processes, known as internal focused, compared to others who tend to rely on the formulation using group interaction, known as external-focused . The effects of this can be observed in the differences between people who need quiet solitude and concentration for idea generation and those who find group processes (such as group brainstorming) a means of cognitively priming ideas ." So in this work also we consider a sentiment analysis related to confirmation bias and backfire effect from the External focus based on his work [1]. This work is a preliminary report on using sentiment features for social media bots from a psychological perspective based on theoretical formulation provided by [1] about these factors. We next discuss the social effects used in this work. One famous social effect is the "backfire effect" [23] which is a widespread phenomenon in social media. The backfire effect occurs when one side of an argument provides many facts to change the audience's mind, but the effort has the reverse effect, creating more resistance in the opposing person's mind [9]. For example, when a Twitter user encounters comments that support more opposing views for an extended period, a person decides not to change his or her original beliefs, and thus the reverse effect occurs. Providing more facts creates more resistance to accept the facts rather than transferring the truth behind events. So the target of the argument decides not to change their mind regardless of what the truth is. Another crucial social effect is "confirmation bias." This occurs when people try to interpret every event based on their existing beliefs. In social media and Twitter, tweets that support people's ideas are more popular than tweets from an opposing view. Some trends and positions apparent in social media are created using the backfire effect and confirmation bias.

Social media bots can use confirmation bias or backfire effect to create fake trends, fake news, and sell products. These types of bots try to infuse fake events by creating psychological effects on users' emotions. In this research, we

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

TABLE I
CRESCI 2017 DATASET DESCRIPTION

user type	tweet count	user count
genuine	2839361	3474
social spam bot #1	1610034	991
social spam bot #2	428542	3457
social spam bot #3	1418557	464
traditional spam bot	145094	999
fake followers	196027	3351

do not analyze the backfire effect and confirmation bias for social Media bot detection directly on a tweet's text; instead, we use the sentiment features introduced by [1] and make a new feature set to improve machine learning models for bot detection. These extracted features are then used to find the best machine learning model for social bot detection in Twitter. All methods in this work are evaluated based on accuracy, F1 score, and MCC for bot detection. Also, Our method provides high accuracy for bot detection in Dutch tweets as well as English tweets. We address three main questions in this work:

1) What are new sentiment features that discriminate between human accounts and social spambots? (These features are inspired by the studies conducted by Workman [1] about Confirmation bias and Backfire effect)
2) Does the new sentiment analysis improve the accuracy of machine learning models for bot detection?
3) Is the new method applicable just to English Tweets?

II. DATASET

In this work, we use Cresci 2017 data set [13], a labeled data set of bots and human users on Twitter. There are five categories of bots: Social Spam bot #1, #2, #3, fake followers, and traditional spambots. Social Spam bot #1 are automated accounts from a 2014 election in Rome. Social spam bot #2 are promoter bots that spend several months promoting specific hashtags on Twitter. Social spam bot #3 is Amazon accounts that share spam URLs on Amazon pointing to their products. Traditional spam bots are Twitter spammer accounts that are collected from the Cresci 2013 data set [24]. The Cresci authors bought fake followers from different websites [13]. Table I shows the statistics of the Cresci 2017 data set. In this data set, each account includes the following attributes: follower-count, friends-count, retweet-count, reply-count, number of hashtags, number of shared URL, tweet's text, screen name, and user ID.

III. METHOD

In this section we identify new features and compare their effect on the machine learning models for social media bot detection. At first we answer the following questions in this experiment.

A. New features

1.What are new sentiment features that discriminate between human accounts and social spambots?

Each Twitter account has different features, such as network features, user features, and user's text content information, as discussed earlier. In this research, based on the work Michael et al. [1] we believe that one way to detect confirmation bias or backfire effect characteristics in the user account is to find a very biased opinion in the sentiment of tweet's text which is posted by a specific Twitter account. Extraction of sentiment from the tweet's text can detect potential confirmation bias and backfire effect in a user's online posting behavior [1]. We first examine tweets which are posted by an individual Twitter account. If the number of tweets posted by an account shows a significant concentration in the number of positive, negative, or neutral tweets, it could indicate a biased opinion or intensity in idea [1] that has a potential create confirmation bias or a backfire effect. So, our model flags these types of accounts for further investigation.

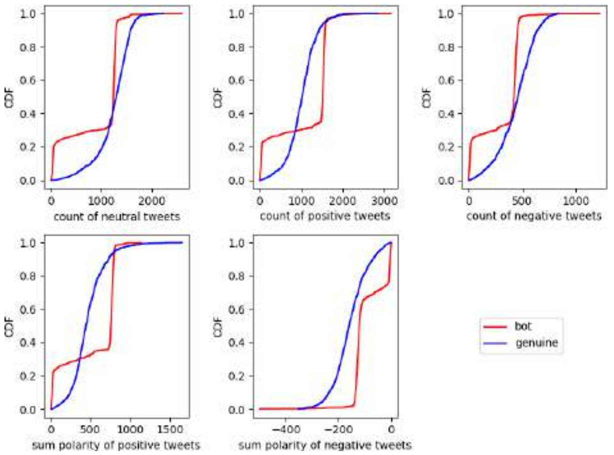


Fig. 1. Cumulative distribution function(CDF)

We provide quantitative analysis, but first, we bring a real-life example to show a clearer picture of the problem. For example, a human account on social media posts her idea about different topics in various ways. Her idea about a movie she saw last night is positive, but her opinion could be negative about the product she bought today. A human user would not post only positive comments, negative comments, or only neutral comments about all aspects of her life. When we examine tweets posted by bot accounts, the variety in the number of positive, negative, and neutral comments is very low, the user posts tweets only about a specific event, and the posts are typically skewed to extreme opinions [1]. In the following section, we apply quantitative analysis to show that the variation of tweet sentiment in each positive, negative, and neutral category is significantly different between a human and a bot account on Twitter.

Figure 1 shows a cumulative distribution function (CDF) for distinguishable characteristics of positive and negative tweets between a human account and a bot account. It can be seen that there is a significant difference in the number of positive, negative, and neutral tweets between a human and a bot account. So sentiment is a discriminating feature for a bot detection model.

We consider creating a new attribute set for each Twitter

TABLE II
NEW ATTRIBUTES

New Features	Description
Count-Neutral	Count of neutral tweets for a user
Count-Positive	Count of positive tweets for a user
Count-Negative	Count of negative tweets for a user
Sum-Positive	Sum polarity scores of positive tweets which are posted by a user
Sum-Negative	Sum polarity scores of negative tweets which are posted by a user
Average-Positive	*Sum-Positive/Count-Positive*
Average-Negative	*Sum-Negative/Count-Negative*

account based on the CDF values for positive and negative tweets. The first new attribute calculated based on the sentiment of a specific user's tweets is a polarity score for each tweet. Polarity score classifies each tweet's text into three different categories, positive, negative, and neutral. Polarity is assigned to each tweet's text by using TextBlob[1]. The polarity assigns a score X for each tweet's text such that $1 < X < 1$. The polarity score of zero means the tweet's text is a neutral comment, $0 < X$ means a positive comment, and $X < 0$ means a negative comment. Table II shows how we apply our new sentiment analysis method to extract new attributes from a tweet's text to differentiate a human account from bot accounts based on Figure 1.

B. The effect of new features in Machine learning models for bot detection

Does the new sentiment analysis improve the accuracy of machine learning models for bot detection?

In this section, we examine machine learning models based on new sentiment features. We evaluate machine learning models such as Random Forest, Support Vector Machine, Logistic Regression, and feed-forward neural network. We run bot detection classifiers with and without new features to show their effect on algorithm performance. First, the original data set attributes(previous mode of operation), and second, with the vector of new sentiment features extracted for each user(New mode). Original features for each account in Cresci 2017 include follower-count, friends-count, retweet-count, reply-count, number of hashtags, and shared URLs.

C. Random Forest

In this section, we use the Random Forest algorithm used for bot detection. This method was previously used in the study by Battur and Yaligar [25]. First, we apply the random forest on the original feature set for each Twitter account provided by Cresci 2017 [13]. Secondly, we apply the new features for each account. In both experiments, 10 fold cross-validation is used, and 20 trees are created. We determined the number of trees based on Grid search. The only difference between the two experiments is the attribute set. Table III shows that the accuracy, F1 score, and MCC are improved based on news sentiment features.

TABLE III
RANDOM FOREST PERFORMANCE

algorithm	accuracy	f1 score	mcc
random forest previous mode	0.887	0.874	0.843
random forest new mode	0.923	0.912	0.887

TABLE IV
ALGORITHM PERFORMANCE

algorithm	accuracy	f1 score	mcc
Random Forest	0.923	0.912	0.887
ANN	0.910	0.927	0.874
SVM(SVC)	0.914	0.922	0.889
SVM(SVR)	0.899	0.930	0.860
Logistic Regression	0.874	0.888	0.685

D. SVM and Logistic Regression

We first run the SVM based on original attributes, then with the new extracted features from a tweet's text for each Twitter account. Two variations of the SVM algorithm are used from the SciKit-learn library in python[2]. For the SVM kernel, we use nonlinear hyper-plane 'poly' with a degree of 7, which provides us with the best results. The data is split into 70% for the training set, 10% devset, and 20% for the test set. The input space is a set of $(x_1..x_i)$ where x_i is the attribute related to each Twitter account. . In the new sentiment analysis, we use the new extracted sentiment features. The result for SVM classification with new parameters is shown in Table IV.Based on Table IV, the F1 score and MCC value show that how the new set of features can detect bots with an accuracy of 90%, F1score nearly 91%, and MMC more than 85%. The result for SVC is better than the SVR with an MCC of 89% and accuracy of 92%.

In this research, we run logistic regression based on the new extracted attributes from a tweet's text. Table IV shows the result for logistic regression. The result shows F1score and prediction accuracy but very low MCC and so this is not a strong algorithm for bot detection based on tweet's text.

E. Neural Network

In this section, we examine the Neural Network model based on the new proposed sentiment analysis. A study by Chen et al. [26] shows how features as input space play an important role in classification results. In this research, sentiment analysis on tweet text has been done to create a new vector of features for a feed-forward neural network. As before, the model is evaluated based on two different attribute sets, original data set features and then based on new sentiment features. The neural network for proposed bot detection in this work is shown in figure 2. The Neural Network model has two hidden layers in this experiment.

[1] https://textblob.readthedocs.io/en/dev/

[2] https://scikit-learn.org/stable/modules/svm.html

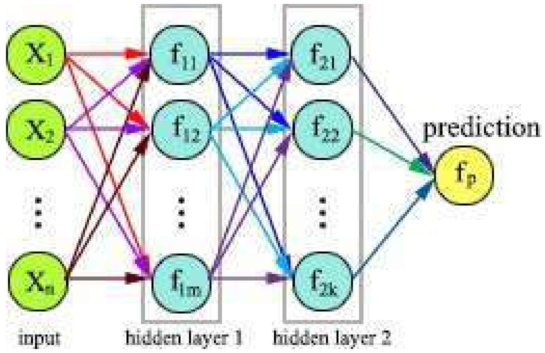


Fig. 2. ANN model for tweet bot detection

TABLE V
NEURAL NETWORK PERFORMANCE

algorithm	accuracy	f1 score	mcc
Neural network previous attribute set	0.899	0.906	0.793
Neural network new attribute set(new sentiment features)	0.910	0.927	0.874

$$y_i = \sigma(w_1 x_1 + ... + w_m x_m)$$
$$y = \begin{pmatrix} \sigma(w_{1,1}x_1 + ... + w_{1,m}x_m) \\ \vdots \\ \sigma(w_{n,1}x_1 + ... + w_{n,m}x_m) \end{pmatrix} \qquad (1)$$

In this research, based on formula 1, Twitter account feature xi where $xi \in (X_1..X_n)$ is one characteristic of user account such as retweet-count, number of followers. X shows the input to the fully connected layer where the output space is $Y \in \{Bot, \neg Bot\}$ and w_i is the weight parameter related to each feature of the Twitter account. Adding a new drop-out layer improves the prediction accuracy in comparison with a fully connected FFNN model in this experiment. The final results for two modes of operation of the fully connected neural network model are shown in TableV.

IV. Results and Algorithm Evaluation

This section provides a comparative analysis of machine learning models for bot detection. Table IV shows the results for all models. As it can be seen, new sentiment features improve Accuracy, F1 score, and MCC. The Logistic regression model has a low value for MCC in comparison with other classifiers. Regarding the prediction accuracy and F1 score, Random Forest and SVM (SVC) provide the best values. However, logistic regression has the least value for F1 score, Accuracy, and MCC. The Random Forest provides the highest value for MCC. SVM and FFNN have similar values for MCC. Based on Accuracy, F1 score, and MCC metrics and with respect to adding new sentiment features, this experiment identifies Random Forest and FFNN as powerful models for social media bot detection.

TABLE VI
DUTCH PERFORMANCE

algorithm	accuracy	f1 score	mcc
ANN	0.861	0.857	0.767
Random Forest	0.891	0.901	0.847

V. Language Dependency

In this section, we answer the third question: Can we use new sentiment features to detect bots for tweets that are not in the English language? We evaluate the effect of new sentiment features in Dutch tweets from Cresci 2017. In Table VI it can be seen that the new sentiment features provide good results for bot detection in Dutch as well. Based on new sentiment analysis, FFNN and Random Forest can detect bots in Dutch tweets with a prediction accuracy and F1 score of nearly 90%. Random Forest provides the best performance to detect bots in Dutch tweets as well as English tweets.

VI. Conclusion and future works

In this research, we evaluate the role of sentiment features on social media bot detection models and their potential role on confirmation bias and backfire effect based on a study by Michael [1]. A new set of sentiment features based on online user's posts improves model accuracy in detecting social bots. Also, applying a new set of attributes is not limited to English tweets but also shows impressive results for Dutch tweets. In our method, we consider both an account level(new set of sentiment features)and a group level approach(previous network features) for Twitter bot detection. For future work, the new method of feature engineering will be improved to extract more features from the online user comments.

References

[1] M. Workman, "An empirical study of social media exchanges about a controversial topic: Confirmation bias and participant characteristics," *Social Media in Society*, pp. 381–400, 2018.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explorations*, vol. 19, no. 1, pp. 22–36, 2017.

[3] Z. Rajabi, A. Shehu, and O. Uzuner, "A multi-channel bilstm-cnn model for multilabel emotion classification of informal text," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pp. 303–306, 2020.

[4] Z. Rajabi, A. Shehu, and H. Purohit, "User behavior modelling for fake information mitigation on social web," in *Social, Cultural, and Behavioral Modeling* (R. Thomson, H. Bisgin, C. Dancy, and A. Hyder, eds.), (Cham), pp. 234–244, Springer International Publishing, 2019.

[5] H. Karbasian, H. Purohit, R. Handa, A. Malik, and A. Johri, "Real-time inference of user types to assist with more inclusive and diverse social media activism campaigns," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 171–177, 2018.

[6] A. Rajabi, C. Gunaratne, A. V. Mantzaris, and I. Garibay, "Modeling disinformation and the effort to counter it: A cautionary tale of when the treatment can be worse than the disease," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1975–1977, 2020.

[7] L. Madahali and M. Hall, "Application of the benford's law to social bots and information operations activities," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–8, CyberSA, 2020.

[8] M. Heidari, J. H. J. Jones, and O. Uzuner, “Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter,” in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[9] F. Husain, J. Lee, S. Henry, and O. Uzuner, “Salamnet at semeval-2020 task12: Deep learning approach for arabic offensive language detection,” in *International Workshop on Semantic Evaluation (SemEval) 2020*, 2020.

[10] F. Husain and O. Uzuner, “Transfer learning approach for arabic offensive language detection system – bert-based model,” in *2021 4th International Conference on Computer Applications Information Security (ICCAIS) - Contemporary Computer Technologies and Applications*, 2020.

[11] A. Sabzehzar, G. Burtch, Y. Hong, and T. Raghu, “The role of religion in online pro-social lending,” in *40th International Conference on Information Systems, ICIS 2019*, Association for Information Systems, 2019.

[12] A. Sabzehzar, Y. Hong, and T. Raghu, “People don’t change, their priorities do: Evidence of value homophily for disaster relief,” in *41st International Conference on Information Systems, ICIS 2020*, Association for Information Systems, 2020.

[13] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pp. 963–972, 2017.

[14] Z. Chen and D. Subramanian, “An unsupervised approach to detect spam campaigns that use botnets on twitter,” *CoRR*, vol. abs/1804.05232, 2018.

[15] P. N. Howard, S. Woolley, and R. Calo, “Algorithms, bots, and political communication in the us 2016 election: The challenge of automated political communication for election law and administration,” *Journal of Information Technology & Politics*, 2018.

[16] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, “Botornot: A system to evaluate social bots,” in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pp. 273–274, 2016.

[17] W. M. Campbell, C. K. Dagli, and C. J. Weinstein, “Social network analysis with content and graphs,” *Lincoln Laboratory Journal*, 2013.

[18] M. Heidari, S. Zad, and S. Rafatirad, “Ensemble of supervised and unsupervised learning models to predict a profitable business decision,” in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[19] M. Heidari and S. Rafatirad, “Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews,” in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[20] M. Heidari and S. Rafatirad, “Using transfer learning approach to implement convolutional neural network to recommend airline tickets by using online reviews,” in *IEEE 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020*, 2020.

[21] M. Heidari and S. Rafatirad, “Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment,” in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[22] M. Heidari and S. Rafatirad, “Semantic convolutional neural network model for safe business investment by using bert,” in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[23] T. Wood and E. Porter, “The elusive backfire effect: Mass attitudes’ steadfast factual adherence,” *Springer Science+Business Media, LLC, part of Springer Nature 2018*, 2018.

[24] C. Yang, R. C. Harkreader, and G. Gu, “Empirical evaluation and new design for fighting evolving twitter spammers,” *IEEE Trans. Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.

[25] R. Battur and N. Yaligar, “Twitter bot detection using machine learning algorithms,” *International Journal of Science and Research (IJSR)*, 2018.

[26] S.-C. Chen, Y.-R. Chen, and W.-G. Tzeng, “Effective botnet detection through neural networks on convolutional features,” in *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018*, pp. 372–378, 2018.

Mechatronic Exoskeleton Systems for Supporting the Biomechanics of Shoulder-Elbow-Wrist: An Innovative Review

José Cornejo
Space Physics and Engineering Division, Bioastronautics and Space Mechatronics Research Group
Lima, PERU
jose.cornejo@ieee.org

Deyby Huamanchahua
Facultad de Ingeniería, Universidad ESAN
Lima, PERU
dhuamanchahuac@esan.edu.pe

Sofía Huamán-Vizconde
Facultad de Ingeniería, Universidad Tecnológica del Perú
Lima, PERU
s.huaman@ieee.org

Dana Terrazas-Rodas
Facultad de Ingeniería, Universidad Tecnológica del Perú
Lima, PERU
dterrazas@ieee.org

Jorge Sierra-Huertas
Facultad de Ingeniería, Universidad Tecnológica del Perú
Lima, PERU
jsierra@ieee.org

Alexander Janampa-Espinoza
Facultad de Ingeniería, Universidad Tecnológica del Perú
Lima, PERU
jesusjanampa@ieee.org

Jorge Gonzáles
Facultad de Ingeniería, Universidad Tecnológica del Perú
Lima, PERU
jorgegonzales@ieee.org

Manuel Cardona, *Senior Member IEEE*
Facultad de Ingeniería, Universidad Don Bosco
Soyapango, EL SALVADOR
manuel.cardona@ieee.org

***Abstract*— Disability is defined as a condition of the human body that limits the execution of a task or activity. According to the World Health Organization (WHO), 15% of the world's population suffers from it, recent studies indicate that the growing prevalence are a significant problem and that, consequently, the demand for rehabilitation services is rising considerably. For this reason, different authors propose the use of exoskeletons in rehabilitation therapies as an alternative solution. Thanks to the progress of this kind of technology, it is possible to create robotic systems that help people with disabilities to recover, totally or partially, the original movement of their affected limbs. This Innovative Review Article presents an exhaustive review of the main features of upper-limb exoskeletons such as Degrees of Freedom (DoF), mechanism type, rehabilitation mode, movements allowed, applications and Technology Readiness Level (TRL). Firstly, the study provides a brief description of the biomechanics of the upper limbs of the human body. Next, the material, the rehabilitation modes and the Technology Readiness Level (TRL) of each of these devices are analyzed. As a result, it was observed that aluminum and PLA are the most used materials for exoskeletons' manufacturing. In addition, it was noticed that most of these exoskeletons perform passive rehabilitation. Besides, it was identified that the most common applications are the assistance and rehabilitation in the affected limb of patients who have suffered a stroke. Finally, using TLR scale, it was determined that these mechatronic systems are between TRL5 and TRL8.**

Keywords—Biomechanics, biomechatronic, elbow, exoskeletons, rehabilitation, shoulder, upper-limbs, wrist

I. Introduction

Over the years, the incidence of disabilities in the world population has increased significantly, becoming an alarming social problem. According to the World Health Organization (WHO), a person with a disability is someone who has “any problem with the function or the structure of the body, a limitation of activity, a difficulty in the execution of a task or action, with a participation restriction” [1]. It should be noted that a considerable percentage of people suffer from this condition around the world.

In 2011, the World Health Organization (WHO) reported 1,636,800 people with disabilities [2]. That is, approximately 15% of the world's population [3]. For example, in Canada, 13.7% of the population over 15 years of age suffered this reality and in 2018, the percentage attained 22.3% [4]. Furthermore, in Latin America and the Caribbean, from 2000 to 2010, more than 66 million people lived with a type of disease; representing 12.4% and 5.4% of the of Latin American and Caribbean population, respectively [5]. From 2010 to 2012, there was an increase of more than 4.5 million, raising the number from 66 121 596 to 70 666 206 people [6]. In 2011, the first report on disability in Peru was published. The study determined that 5.2% of the Peruvian population suffered from some kind of disability [7]. It is important to mention that several countries do not have updated statistical information.

The demand for physiotherapy and rehabilitation services is increasing. Therefore, meeting this need properly becomes a challenge. Rehabilitation exoskeletons emerge as a possible solution to this problem [8], [9]; awakening the interest of the scientific community in proposing improved designs for these robotic systems.

One of the most prominent applications in exoskeletons is robot-assisted rehabilitation, due to the promising advantages they offer to the patient [10]. Since the device must imitate the natural movement of the patient’s affected limb, without causing complications that threaten the user’s safety [11], these mechatronic systems are very complex: imitating the movement of the patient's internal joints while retaining the same number of degrees of freedom (DoF) is a complex endeavor.

On the other hand, understanding the biomechanics of the upper limb allows researchers to easily analyze the processes and forces that affect the human body. The kinematics and the additional weight, generated by the patient's upper limb, are additional features that are considered in its design [12]. Finally, exoskeletons are classified according to the NASA Technology Readiness Level (TRL) scale.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

A. *Applications*

Nowadays, exoskeletons make it possible to perfect the techniques used in rehabilitation; achieving better results compared to the traditional techniques. In this section, the application of the upper limb exoskeletons detailed in Table I are presented.

Some exoskeletons in Table I were designed for the treatment of patients that had suffered strokes, which the WHO describes as a phenomenon that interrupts blood flow through the arteries [13]. An example of these devices is AGB-Exo [14], a gravity balanced motorized mechatronic system focused on shoulder and elbow rehabilitation. This device features four active DoF and one passive DoF that offers an active rehabilitation training mode. This device performs movements such as extension-flexion, pronation-supination, abduction-adduction and internal-external rotation in the patient's upper limb. Other examples are ETS-MARSE [15], CLEVER ARM [16], among others.

However, most of the exoskeletons in Table I. were designed to rehabilitate and / or assist the upper limb, that is, the shoulder, the elbow and the wrist, without considering the hands. They are oriented to the treatment of patients with disabilities who have difficulties performing joint movements, without specifying the pathology that caused it. An example is ChARMin [17], a robotic device with 6 DoF. This system provides active and passive rehabilitation modes. It performs movements such as extension-flexion, abduction-adduction, pronation-supination, and internal-external rotation in the user's upper limb. Other examples of this type of devices are ANYexo [18] and CAREX-7 [19].

As well, most of the exoskeletons in Table I were designed to rehabilitate and / or assist a specific part of the upper limb, focused on pathologies such as myopathy, Duchenne Muscular Dystrophy (DMD), among other injuries. External agents such as exoskeletons support rehabilitation of the patients through therapeutic techniques that stimulate motor skills and functions. For example, Armeo®Power [20] is an exoskeleton of 6 DoF with the purpose of rehabilitating patients who have suffered a stroke, Parkinson's disease, cerebral palsy, among others. Likewise, InMotion ARM [20] is an exoskeleton system developed by the Massachusetts Institute of Technology (MIT), whose structure provides 3 DoF. Other examples are Higuma *et.al* [17], ETS-MARSE [21], ReHab-Arm [22], seen in Fig 1.

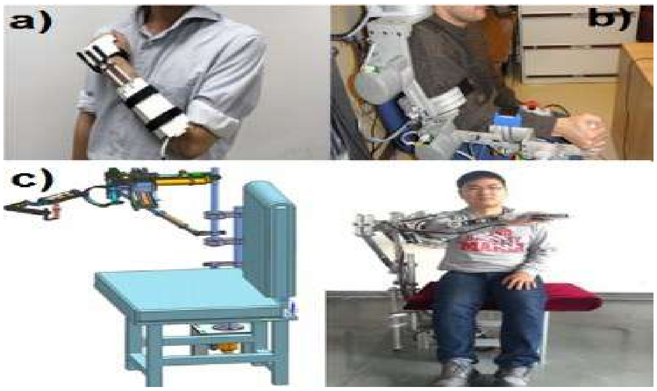


Fig. 1. Upper-Limb Exoskeletons. a) ETS-MARSE [15]. b) Rehab-Arm [22] c) Higuma *et.al* [40].

II. Biomechanics of Upper Limbs

Upper extremity exoskeletons are designed to accurately mimic the natural movement of the human body. So, biomechanical concepts are important in their design.

A. *Biomechanics of the Shoulder*

The human shoulder is a structure made up of joints with the most significant range of motion (ROM). It has three bones which are the clavicle, the scapula and the humerus. In addition, it has three independent joints. The shoulder includes the sternoclavicular, acromioclavicular, and glenohumeral joints as seen in Fig 2. The three joints form the closed kinematic chain of the shoulder girdle [23]. When exploring its range of motion (ROM), the scapulothoracic mobility (movement of the scapula over the chest wall) and the shoulder mobility should be considered [23]. Although the initial 30° of abduction are performed without movement of the scapula, the joint movement for full arm lift occurs at a ratio of 2:1 for every 3° of lift. The shoulder joint confers approximately a second physiologic thoracic scapular junction [23]. That is, when the limb has been raised, and the arm remains in a vertical position next to the head (180° of abduction or flexion of the arm), the shoulder joint has participated in 120° and the scapulothoracic junction in 60° [23]. The important movements of the shoulder girdle are those of the scapula: elevation and descent, prolongation, retraction and rotation.

The sternoclavicular joint is a saddle synovial joint, but it works like a spheroid joint. It is divided into two compartments by an articular disc. This disc is firmly attached to the anterior and posterior sternoclavicular ligaments and the interclavicular ligament. The sternoclavicular joint is stable and provides significant mobility to the shoulder girdle and upper limb. In full limb elevation, the clavicle is elevated to an angle of approximately 60° [23]. When lifting is performed by flexion, the rotation of the clavicle is followed around its longitudinal axis. The acromioclavicular joint is a flat synovial joint located 2-3 cm from the highest point of the shoulder formed by the lateral part of the acromion. The acromion of the scapula rotates on the acromial extremity of the clavicle [23]. There is no muscle that connects the bones that make the joint and move them, but it is the axioappendicular muscles that are inserted into the scapula and move it displacing the acromion over the clavicle [23].

The shoulder joint (glenohumeral) is synovial and spheroid. It allows a wide variety of movements, making it relatively unstable. The proximal area of the humerus articulates with the glenoid cavity of the scapula, and although it is relatively shallow, it is slightly enlarged by the ring of the glenoid border [23]. The glenoid cavity houses approximately one third of the proximal area of the humerus, which is held in position by the musculotendinous rotator cuff (supraspinatus, infraspinatus, teres minor, and subscapularis muscles) [23]. The shoulder joint allows movements in the three axes: flexion-extension, abduction-adduction, medial and lateral rotation of the humerus, and circumduction. Lateral rotation of the humerus increases the range of abduction. When the arm is abducted without rotation, the available articular surface is exhausted and the greater tubercle meets the coracoacromial arch, preventing further abduction. If the arm rotates laterally

180°, the tubercles rotate backwards and more articular surface is available for continued elevation [23].

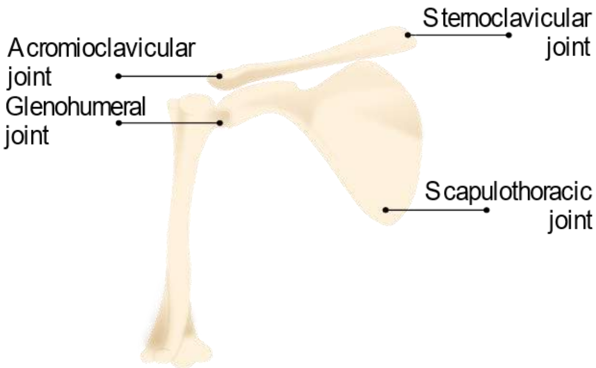


Fig. 2. Joints located at the shoulder

B. *Biomechanics of the Elbow*

The elbow joint, as seen in Fig 3., is a hinge-type synovial joint located 2–3 cm below the epicondyles of the humerus [23]. The pulley-shaped trochlea and the spheroidal condyle of the humerus articulate the trochlear incision of the ulna and the slightly concave upper face of the head of the radius. Consequently, there is a humeroulnar and humeroradial joint. The articular surfaces, covered with hyaline cartilage, are almost completely congruent when the forearm is positioned midway between pronation and supination, and the elbow is flexed at a right angle [23]. The elbow joint allows flexion and extension movements. The long axis of the ulna in full extension forms an angle of about 170° with the long axis of the humerus [23]. This angle is called the angle of transport, because of how the forearm moves away from the body. The obliquity of the ulna, and consequently the angle of transport, are more pronounced in women than in men [23]. It is believed that this allows the upper extremities to avoid hitting the wide female pelvis when swaying while walking. In anatomical position, the elbow is in front of the waist, the angle of transport disappears when the forearm is in pronation [23].

The proximal radioulnar joint is a trochoid-type synovial joint that allows movement of the head of the radius over the ulna. The proximal end of the radius articulates with the radial incision of the ulna and is held in position by the annular ligament of the radius [23]. During pronation and supination of the forearm, the head of the radius rotates within the annulus formed by the annular ligament and the radial incisions of the ulna. Supination rotates the palm anteriorly, or superiorly if the forearm is flexed. Pronation turns the palm posteriorly, or inferiorly if the forearm is flexed. The axis of these movements passes proximally through the center of the radial head and distally through the insertion of the vertex of the articular disc of the ulna [23]. During pronation and supination, it is the radius that rotates: its head rotates within the cup-shaped ring formed by the annular ligament and the radial incision of the ulna. Supination and pronation are almost always accompanied by synergistic movements of the shoulder joint and the elbow that produce simultaneous movements of the ulna, except when the elbow is flexed [23].

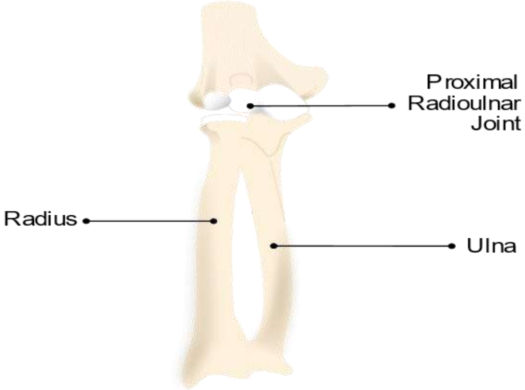


Fig. 3. Joints located at the elbow

C. *Biomechanics of the Wrist*

The radiocarpal joint is an ellipsoid-type synovial joint. The position of this joint is indicated approximately by a line joining the styloid process of the radius with the one of the ulna or by the proximal groove of the carpus [23]. The carpus (wrist seen in Fig 4.) is the proximal segment of the hand. It is made up of a group of eight carpal bones, it articulates proximally with the forearm through the radiocarpal joint and distally with the five metacarpals. The radiocarpal joint can increase its range of motion (ROM) by additional small displacements of the intercarpal and midcarpal joints. The movements that the wrist joint allows to perform them and are flexion-extension, abduction-adduction and circumduction. The degree of flexion of the hand on the forearm is greater than the one of extension; these movements are followed by similar movements in the midcarpal joint [23]. The degree of adduction of the hand is greater than that of abduction. Most of the adduction occurs at the radiocarpal joint. Circumduction of the hand consists of a successive series of flexion, extension, adduction, and abduction movements.

As mentioned in this article, biomechanics concepts are applied in the development of exoskeletons oriented to the rehabilitation of the upper limb. Within the biomechanical analysis, the researchers proceed to kinematic and dynamic modeling. Kinematic modeling is the geometric formulation of a system to obtain the values of the joints and the location of the final element (position and orientation) with respect to a reference system. Dynamic modeling consists of the physical equations of motion in a three-dimensional space and considering the forces acting on the joints. Destarac *et al.* [24], for example, performs the kinematic and dynamic modeling of the ORTE robot, obtaining the values that determine the workspace and the forces that affect it.

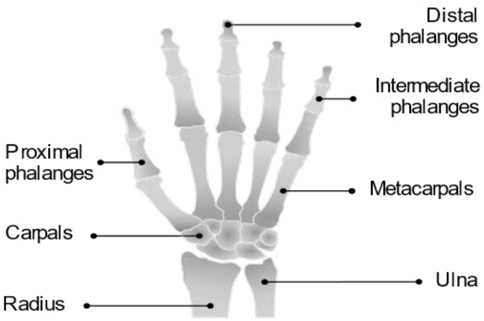


Fig. 4. Joints and bones located at the wrist

III. Mechatronic Exoskeletons

There is a wide variety of rehabilitation exoskeletons, each at a different development stage. The analyzed upper limb exoskeletons were classified according to NASA's Technology Readiness Level (TRL) scale, which is used to indicate the development status of technology projects, where TRL 1 is the lowest and TRL 9 is the highest [25], [87 – 91].

The upper limb exoskeletons shown in Table I were classified using the TRL scale. The analyzed exoskeletons were found to be between TRL5 and TRL8. Each level has a different meaning: an exoskeleton in TRL 5 is considered a full-scale development, while in TRL 6, it has already been validated in a simulated environment. In TRL 7, the system/prototype has been validated in a real environment, in TRL 8, the first prototype/system is available for commercialization, and in TRL 9, the exoskeleton is officially commercial [26].

In Table I, the most common levels observed for the exoskeletons for the upper limbs were TRL 6, TRL 7, and TRL 8, while TRL 5 was the least common. It shows the exoskeleton's names or references and their classification by degrees of freedom (DoF), mechanism, rehabilitation mode, movements, application and TRL [92-94].

TABLE I. Upper-limb Exoskeletons

Name / Ref.	DoF	Mechanism	Rehab. Mode	Movements	Application	TRL
RETRAINER-ARM [27]	4	E,S	P	E(EF),S(EL,R)	Stroke	8
EXOWRIST [28]	2	W	P	W(EF,RU)	Wrist rehabilitation	7
AGB-Exo [14]	4a,1p	E,S	AC	E(EF),FA(PS),S(AA,EF,IE)	Stroke	5
EASoftM [29]	4	E,S	P	E(EF,PS),S(EF,IE)	Different injuries	6
ORTE [24],[30]	5	E,S	-	E(EF),FA(PS),S(AA,EF,IE)	Different injuries	7
Wu *et al.* [31]	7a,2p	E,S,W	AC,P	E(EF),FA(PS),S(AA,EF,IE),W(EF,RU)	Right arm training	7
Wu *et al.* [32]	7a,2p	E,S,W	AC,P	E(EF), FA(PS),S(AA,EF,IE), W(EF,RU)	Hemiplegia	6
ULEL [33],[34]	3	E,S,W	P	(E,S,W)(EF)	Post-stroke patients	7
Jung *et al.* [35]	5	E,S	P	E(EF),S(AA,EF,MR,EV)	Shoulder - Elbow rehabilitation	6/7
NESM [36]	4a,8p	E,S	AC,P	E(EF),S(AA,EF,IE)	Post-stroke rehabilitation	7
EXOWRIST [28]	2	W	P	W(EF,RU)	Wrist rehabilitation	7
ANYexo [18]	5	E,S,W	A	-	Shoulder-Elbow-Wrist Rehabilitation	7
Chen *et al.*[37]	4	E,S	P	E(EF)S,(AA,EF,IE)	Supported shoulder–elbow motion	6/7
Copaci *et al.* [38]	2	E	AC,P	E(EF,PS)	Evaluation and therapy	8
NTUH-ARM [39]	7	E,S	AC,P	E(EF),S(AA,ED,EF,IE,PR)	Support and Arm rehabilitation	6
Higuma *et al.* [40]	2	W	P	W(AA,EF)	Rehabilitation and assistance in ADL	7
Rehab-Arm [22]	7	E,S,W	AC,AS,P	E(EF,LM),S(AA,EF,LM),W(EF,RU)	Stroke and hemiplegia	7/8
RehabRoby [41]	6	E,S,W	AC,AS,P	E(EF,PS),S(AA,EF,IE),W(EF)	Stroke	8
ETS-MARSE [15]	7	E,S,W	AC,AS,P	S(vEF,hEF,IE),E(EF),FA(PS),W(EF,RU)	Stroke	7
Zhang *et al.* [42]	3	E,S	AC,P	E(EF),S(AA,EF)	Hemiplegia	7
Tanaka *et al.* [43]	6	E,S,SG	AC,P	E(EF),S(AA,EF,IE),SG(ED,EF)	myopathy and hemiplegia	7
Ye *et al.* [44],[45]	5	E,S,W	AC,AS,P	E(EF),PS,S(EF),W(EF)	Shoulder-Elbow-Wrist Rehabilitation	7
Wu *et al.* [46]	7	E,S,W	AC,AS,P	S(AA,EF,IE),E(EF),W(EF,PS,RU)	Shouldeer-Elbow-Wrist Rehabilitation	6/7
ChARMin [17]	6	E,S,W	AC,P	S(AA,EF,IE),E(EF,PS),W(EF)	Shoulder-Elbow-Wrist rehabilitation	8
Nam *et al.* [47]	2	E,W	P	E(EF),W(DV)	Elbow-Wrist rehabilitation	6
4-soft-A [48]	2	E	P	E(EF)	Elbow rehabilitation	6
Bai *et al.* [49]	3	S	AC,P	E(AA),S(EF)	Shoulder-Elbow rehabilitation	7
Crea *et al.* [50]	2	E,S	AC,P	S(EF,IE),E(EF),FA(PS)	Elbow-Shoulder rehabilitation	6
Hasegawa *et al.* [21]	2	S	AC,P	S(EF,IE)	Shoulder rehabilitation	7
Liu *et al.* [51]	2	S	P	S(AA,EF)	Shoulder rehabilitation	6
Islam *et al.* [52]	5	S	AC,P	S(AA,EF,IE,PR)	Shoulder rehabilitation	6
SCRIPT [53]	5	W	P	W(AA,EF),T(EF,Radial AB); F(F/E)	Hand rehabilitation	7

Lobo-Prat *et al.* [54]	1	E	AC	E(EF)	Duchenne Muscular Dystrophy	6
ARMin III [55]	3	S	AC	S(AA,EF,IE,PR)	Shoulder rehabilitation	8
ArmeoPower [55]	3	S	AC	S(AA,EF,IE,PR)	Shoulder rehabilitation	8
Pneu-WREX [55]	4	S	AC	S(AA,EF,IE,PR)	Shoulder rehabilitation	8
BONES [55]	3	S	AC	S(AA,EF,IE,PR)	Shoulder rehabilitation	8
EXO-UL7 [55]	3	S	AC	S(AA,EF,IE,PR)	Shoulder rehabilitation	8
L-EXOS [55]	3	S	AC	S(AA,EF,IE,PR)	Shoulder rehabilitation	8
Huang *et al.* [56]	4	E,S,W	AC	A(EF)	Shoulder-Elbow-Wrist rehabilitation	8
Hsieh *et al*. [57]	2	S	P	S(AA,EF)	Shoulder rehabilitation	7
CAREX-7 [19]	7	E,S,W	P	S(AA,EF,IE),E(EF,PS),W(AA,EF)	Shoulder-Elbow-Wrist rehabilitation	8
Li *et al*. [58]	2	W	P	W(AA,EF)	Wrist rehabilitation	6
Ianoşi *et al.* [59]	4	S	P	A(EF)	Shoulder rehabilitation	6
Wang *et al.* [60]	3	S	P	A(EF)	Shoulder rehabilitation	7
Zhang *et al.* [61]	3	S	P	S(AA)	Shoulder rehabilitation	6
Kopke *et al.* [62]	3	S	P	A(EF)	Shoulder rehabilitation	7
Ghobj *et al.* [63]	3	S	P	E(EF),S(EF)	Shoulder-Elbow rehabilitation	7
Goltapeh *et al.* [64]	3	S	P	A(EF)	Shoulder rehabilitation	5
González-Mendoza *et al*. [65]	2	S	P	S(EF),E(EF)	Shoulder rehabilitation	6
Hsieh *et al*.[66]	3	S	P	A(EF)	Shoulder rehabilitation	7
Huang *et al.* [67]	4	S	P	A(EF)	Sensor assisted arm rehabilitation	7
Jiang *et al.* [68]	2	S	P	A(EF)	Shoulder rehabilitation	6
Nycz *et al.* [69]	3	S	P	E(EF),S(AA,EF,IE)	Shoulder-Elbow rehabilitation	7
Ali *et al.* [70]	3	S	P	E(EF),S(AA,EF,IE)	Shoulder-Elbow rehabilitation	7
Aljallad *et al*. [71]	5	S	P	S(AA,EF,IE),E(EF)	Shoulder-Elbow rehabilitation	6
Atia *et al*. [72]	4	S	P	EF	Shoulder exoskeleton	6
Casas *et al*. [73]	3	S	P	EF	Shoulder exoskeleton	6
De la Iglesia *et al.* [74]	3	E	AC	EF	Elbow rehabilitation	6
Rebelo *et al.* [75]	3	S	P	EF	Shoulder rehabilitation	6
Weiler *et al.* [76]	3	S	P	E(EF),S(AA,EF,IE)	Shoulder-Elbow rehabilitation	7
Chen *et al*. [77]	4	E,S	P	E(EF),S(AA,EF,IE)	Shoulder-Elbow rehabilitation	7
Lee *et al*. [78]	4	E,S	P	(E,S)(EF)	Shoulder-Elbow rehabilitation	6
CADEL [79]	2	A,E	-	EF	Elbow assist	6
Liu *et al*. [80]	6	E,S	-	(E,S)(EF)	Exoskeleton prototype	6
Thalagala *et al*. [81]	4a,3p	E,S	-	E(EF),S(AA,EF,IE,PS)	Rehab and Assitance in ADL	7
Vilalpando *et al.* [82]	2	S	P	S(AA,EF)	-	6
Niyetkaliyev *et al.* [83]	5	S,SG	-	SJ(AA,EF,IE,PS),SG(ED,PR)	Neurological disability	5
Noda *et al*. [84]	1	S	P	S(EF)	Sroke	7
Chien *et al*. [85]	2	S	AC,P	AA,EF	After-stroke rehabilitation	6
Zhang *et al.* [86]	2	E	-	EF,PS	Stroke	6
CLEVER ARM [16]	6a,2p	E,S,W	-	E(EF),W(EF,PS)	Stroke	6

Note Abbreviations: EF: Extension-Flexion, RU: Urnal-Radial deviation, PS: Pronation-Supination, AA: Abduction-Adduction, IE: Internal-External rotation, MR: Medial Rotation, EV: Elevation as a Vertical Movement, EL: Elevation, R: Rotation, LM: Medial-Lateral rotation, ED: Elevation-Depression, PR: Protraction-Retraction, v: Vertical, h: Horizontal, SG: Shoulder girdle, DV: Dorsiflexion-Volarflexion, A: Arm, AC: Active rehabilitation, P: Passive rehabilitation, AS Assistive rehabilitation, a: active degree, p: passive degree and SJ: Shoulder joint, Rehab. Mode: Rehabilitation Mode

A. *Materials*

The exoskeletons used for rehabilitation attempt to resemble the natural anthropology of patients. That is, they try to imitate the natural movements of the users. When designing an exoskeleton, it is essential to take into account characteristics such as mechanical strength, resistance to fatigue, weight, cost, availability, etc. Thus, this section analyzes the selection of materials used for the mechanical structure of the robotic systems shown in Table I.

a) *Metallic materials*

Metallic materials are used in exoskeletal devices due to their high resistance, hardness and rigidity. Aluminum and metallic alloys are the most used materials, used by researchers for prototype development. In this section, the exoskeletons that have metallic materials in their mechanical structure are mentioned. Some examples are RETRAINER-ARM [27], EXOWRIST [28], EASoftM [29] and 4-soft-A [48], among others; seen in Fig 5.

For EXOWRIST [28], its mechanical structure is made of aluminum alloy. This upper limb exoskeleton is a prototype for patients with cerebrovascular diseases. This system has two degrees of freedom (DoF) and is oriented to the treatment of wrist. This device performs movements such as extension-flexion and urnal-radial. It has actuators integrated into the metal mechanism that are manipulated during the physiotherapy sessions. This system is useful for the evaluation of patients with the support of a therapist, who configures the type of rehabilitation exercise to be performed.

In a different manner, RETRAINER-ARM [27] is an upper extremity exoskeleton that features four degrees of freedom (DoF). It is oriented to the treatment of stroke patients. Its mechanical structure uses aluminum metal links. The ease for manufacturing and the geometric shape of the pieces stands out. This device constitutes a non-invasive exoskeleton that assists the patient during their treatment.

For example, EASoftM [29] is an upper limb exoskeleton that features four degrees of freedom (DoF). This device performs movements such as extension-flexion, pronation-supination, and internal-external rotation in the patient's upper limb, which is oriented to the treatment of patients with different injuries. This mechatronic system aims to rehabilitate the shoulder and elbow. On the other side, the researchers have selected aluminum alloys, since it is flexible, non-invasive, it has an adequate mechanical resistance and a proportional weight. These characteristics were observed during the follow-up of the rehabilitation process of a patient.

The aluminum is one of the materials that stands out for its high availability, low cost, high manufacturability (cutting process) and preferred for the mechanical structure of exoskeletons.

b) *Plastic materials*

The use of plastic materials in robotic devices is common mainly due to their lightness. These materials are used as hybrid elements that make up the structures of robotic systems. That is, they are used together with other materials. Their elastic properties and tensile stress make them the most suitable material for deformation. In this section, some of the exoskeletons, listed in Table I, whose mechanical structure is built based on plastic materials are described. Some examples are ChARMin [17], ARMin III [55] and RETRAINER - ARM [27], among others.

For ARMin III [55], materials are selected considering design complexity. It performs movements such as extension-flexion, abduction-adduction, internal-external rotation and protraction-retraction. In this way, designers ensure that they are capable of resisting weight, resistance and fatigue. In order to design complex mechanisms, 3D printing is also used. This exoskeleton has 10% lactic polyacid to facilitate the attachment with actuators and their fixation.

On the other hand, RETRAINER-ARM [27] is a system that aims to rehabilitate the shoulder and elbow. This apparatus performs movements such as extension-flexion, elevation and rotation in the patient's upper limb. It uses flexible plastics to adapt to the anthropology of the patient, considering that the patient will use this hybrid mechanism with actuators.

For example, ChARMin [17] uses Acrylonitrile Butadiene Styrene (ABS) for manufactured parts whose geometry is complicated and requires a too complex design to support structures such as actuators. This device is made of aluminum and has joints, fixing elements, and accessories with lactic polyacid. Given the demands of the design, it is a hybrid mechanism. 3D printing is used to manufacture parts such as joints, screws, fasteners or connecting pins in order to obtain a hybrid exoskeleton that meets the fatigue and deformation requirements. This technology provides the easiest way to produce the parts, since it allows the manufacturing of highly complex designs. Acrylonitrile Butadiene Styrene (ABS) is one of the most used plastic materials for the mechanical structure of the exoskeletons listed in Table I.

Plastic materials are selected because their characteristics allow to solve several design challenges. They are resistant, they support fatigue stress, they are non-corrodible, easy to maintain and resistant to running or testing.

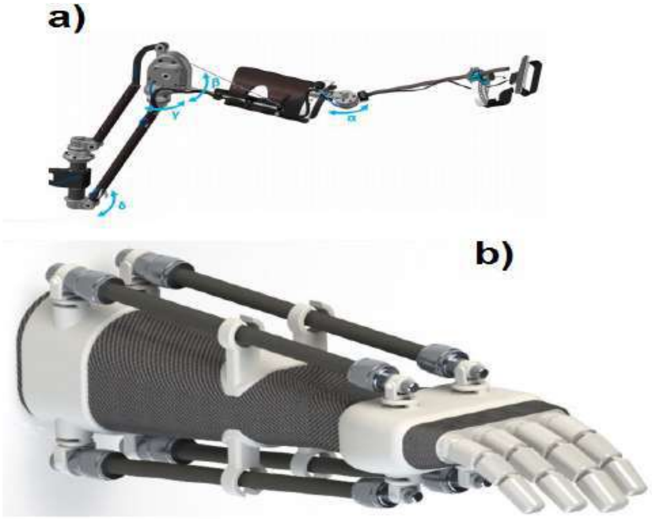


Fig. 5. Upper-limb exoskeleton robots. a)RETRAINER-ARM [27] . b)EXOWRIST [28]

B. Rehabilitation mode

Currently, using exoskeletons for rehabilitation provide a high-quality option to regain movement thanks to their precision in repetitive movements. Rehabilitation is the treatment that a patient receives with the aim of recovering, maintaining or improving the physical or mental skills necessary for daily life, and thus improve their quality of life. In this case, we will focus on physical abilities, which can be totally or partially reduced due to illnesses or injuries. Rehabilitation modes can be passive, active, and assisted [87] In this section, the rehabilitation modes offered by the exoskeletons in Table I are mentioned.

a) Passive rehabilitation

The use of exoskeletons with passive rehabilitation modes allows treating patients effortlessly: the robotic system helps the patient move the muscles and joints throughout his/her ROM. This is commonly used when the user suffers from a disability caused by hemiplegia and other common injuries. Although passive exercise does not require any effort from the patient, following this kind of rehabilitation has many benefits, such as helping them to prevent joint stiffness. These benefits are most effective when used consistently over a long period of time.

An example of this rehabilitation mode is provided by EASoftM [29]. Patients during rehabilitation can perform exercises without additional effort. The succession of exercises that the patient must follow is monitored by the therapist. Passive rehabilitation is a great way to improve blood flow to affected areas and provide sensory stimulation to the affected limb. Another benefit is shown by EXOWRIST [28], where the use of this rehabilitation mode helps prevent the worsening of spasticity of the upper limbs in patients.

In ULEL [33],[34] this mode helps patients interact with the rehabilitation professional to perform exercises. This exoskeleton focuses on passive movements to provide the patient with routines that include massages, chiropractic, and physical therapy. Similarly, the RETRAINER-ARM [27] exoskeleton aims to rehabilitate the patient through the passive mode making the shoulder and elbow move using an external force which is generated by the exoskeletal device.

b) Active rehabilitation

The active rehabilitation mode with exoskeletons implies that patients make a physical effort to move, that is, they perform the exercises and the exoskeletons offer resistance to movement in a way that helps them to strengthen the muscles. In general, physical rehabilitation stimulates the brain to reconnect through neuroplasticity, improving the ability to send signals to the muscles. Neuroplasticity occurs with both passive and active exercise but is enhanced with active exercise. One of the benefits of active rehabilitation is that it helps to strengthen muscles, being beneficial if muscle atrophy has occurred due to a reduced daily movement.

ARMin III exoskeleton [55] focuses on the rehabilitation of patients with cerebrovascular accidents (CVA). The active mode is used in order to help them regain mobility of the shoulder. Using this mode is excellent for patients suffering from hemiparesis and strokes. Another example is the AGB-Exo exoskeleton [14], since the patient makes a succession of physical effort exerted on the muscular activity of the upper limbs, active movements such as self-stretching or other movements of the muscles are performed through repetitive tasks.

NESM [36], on the other hand, has a structure focused on passive rehabilitation because it is incorporated into the spine, which generates a passive self-aligning kinematic chain. The NTUH-ARM exoskeleton [39], uses the Lyapunov theory to improve stability in the controller design, allowing gravity compensation in assisted training. It presents a ROM considered in its 7 DoF, which allows a complete rehabilitation in each section of the arm. On the other hand, the ChARMin [17] exoskeleton is focused on the rehabilitation of children with defects in the motor function of the arm, cerebral palsy, and others. In addition, the rehabilitation process has a computational interface (Unity Technology) that allows the child to train through a video game.

c) Assisted Rehabilitation

In this rehabilitation mode, robotic support is provided only when the patient cannot follow the training proposed by the therapist. This means, it helps patients to exercise their limbs to the best of their ability. This type of rehabilitation is based on swinging against an object; turning the body back until the patient feels that the shoulders are stretched avoiding spasticity. Assisted rehabilitation using exoskeletons is advantageous since it provides important benefits in relation to muscle strength, it reduces motion recovery time, achieving functional independence faster.

Many authors prefer to build exoskeletons that offer the three rehabilitation modes, as seen in Fig 6. For example, Rehab-Arm [22] performs passive, active and assisted rehabilitation. This mechatronic system with 7 DoF is aimed at rehabilitation in patients that have suffered strokes and hemiplegia. RehabRoby [41], Ye *et al.* [44],[45], ETS-MARSE [15] and Wu *et al.* [46] devices also offer these modes for stroke patients, focusing on shoulder-elbow-wrist rehabilitation, with 7, 5, 7 and 6 DoF, respectively.

Finally, the robotic systems used for the rehabilitation of upper limbs shown in Table I do not offer only one type of assisted rehabilitation, but they offer assisted, active and passive modes or a combination of these.

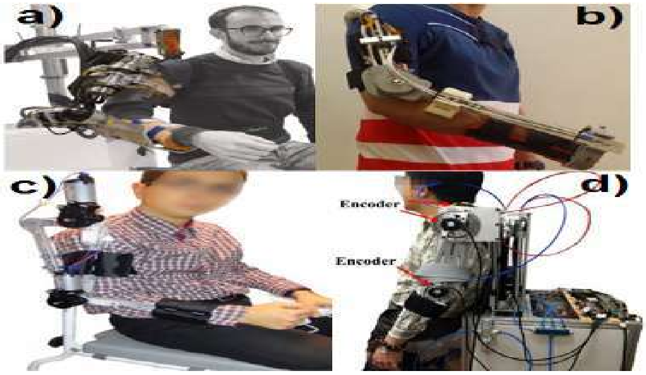


Fig. 6. Upper-limb exoskeleton robots. a)NESM [36]. b) Copaci *et al.* [38]. c) Ianosi *et al.* [59] d) Chen *et al.* [77]

IV. Conclusion

In conclusion, a brief review of the most recent exoskeletons for the rehabilitation of upper limbs damaged by diseases such as strokes, hemiplegia, Duchenne muscular dystrophy, among others, has been carried out. The latter are the most common disabilities for which exoskeletons are designed. Exoskeletons focus on passive and active rehabilitation or a combination of both. The passive rehabilitation mode was the most observed. This mode is used when the patient does not offer resistance to the effort of muscle movement.

In addition, the concept of TRL was defined and the indexed exoskeletons were categorized based on these. As a result, it was observed that the various exoskeleton projects analyzed are between TRL 5 and TRL 8. The most common category was TRL 6, which means that most of the exoskeletons shown are prototypes that have not yet passed the validation process in a group of patients. However, these devices show important advances which result in further progress for the development and manufacturing of exoskeletons for rehabilitation.

It was also mentioned that the choice of material is key for exoskeletons, since the objective of these robotic systems is to imitate the natural movements of the users. It is important to note that only a few research articles mention the materials used in the design of the exoskeletons. Nevertheless, it was observed that the most used materials were metals and plastic. Aluminum and ABS stand out among them for their interesting properties such as weight, strength, low cost, and commercial availability, while metallic titanium and PETG plastic are the least used due to their higher costs.

Acknowledgment

This research was supported by the IEEE Engineering in Medicine and Biology Society at the Universidad Tecnológica del Perú (UTP), the IEEE Student Branch at UTP and the Research Group on Biomedical Engineering and Related Careers (GIIBCA). In addition, the authors would like to thank the Bioastronautics and Space Mechatronics Research Group (https://orcid.org/0000-0002-7173-3960) for supporting Latin American exponential innovation carried out by entrepreneurs and new technologies. Finally, thanks to Research Direction of Universidad Tecnológica del Perú – UTP, filial Arequipa for the support and translation.

References

[1] "OMS | Discapacidades." https://www.who.int/topics/disabilities/es/ (accessed Feb. 23, 2021).

[2] "World report on disability." https://www.who.int/publications/i/item/9789241564182 (accessed Feb. 23, 2021).

[3] "World Report on Disability." https://www.who.int/teams/noncommunicable-diseases/disability-and-rehabilitation/world-report-on-disability (accessed Feb. 23, 2021).

[4] S. C. Government of Canada, "The evolution of disability data in Canada: Keeping in step with a more inclusive Canada," Nov. 28, 2018. https://www150.statcan.gc.ca/n1/pub/89-654-x/89-654-x2018003-eng.htm (accessed Feb. 23, 2021).

[5] C. E. para A. L. y el Caribe, *Panorama Social de América Latina 2012*. CEPAL, 2013.

[6] C. E. para A. L. y el Caribe, *Informe regional sobre la medición de la discapacidad. Una mirada a los procedimientos de medición de la discapacidad en América Latina y el Caribe. Grupo de tareas sobre medición de la discapacidad Conferencia Estadística de las Américas (CEA)*. CEPAL, 2014.

[7] "PERÚ Instituto Nacional de Estadística e Informática." https://webinei.inei.gob.pe/anda_inei/index.php/catalog/495/related_materials (accessed Feb. 23, 2021).

[8] G. Turchetti, N. Vitiello, L. Trieste, S. Romiti, E. Geisler, and S. Micera, "Why effectiveness of robot-mediated neurorehabilitation does not necessarily influence its adoption," *IEEE Rev Biomed Eng*, vol. 7, pp. 143–153, 2014, doi: 10.1109/RBME.2014.2300234.

[9] A. Sawers and L. H. Ting, "Perspectives on human-human sensorimotor interactions for the design of rehabilitation robots," *J Neuroeng Rehabil*, vol. 11, p. 142, Oct. 2014, doi: 10.1186/1743-0003-11-142.

[10] M. P. Dijkers, K. G. Akers, S. Dieffenbach, and S. S. Galen, "Systematic Reviews of Clinical Benefits of Exoskeleton Use for Gait and Mobility in Neurologic Disorders: A Tertiary Study," *Arch Phys Med Rehabil*, vol. 102, no. 2, pp. 300–313, Feb. 2021, doi: 10.1016/j.apmr.2019.01.025.

[11] Z. Li, C. Yang, and E. Burdet, "Guest Editorial An Overview of Biomedical Robotics and Bio-Mechatronics Systems and Applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 7, pp. 869–874, Jul. 2016, doi: 10.1109/TSMC.2016.2571786.

[12] M. Sarac, M. Solazzi, and A. Frisoli, "Design Requirements of Generic Hand Exoskeletons and Survey of Hand Exoskeletons for Rehabilitation, Assistive, or Haptic Use," *IEEE Transactions on Haptics*, vol. 12, no. 4, pp. 400–413, Oct. 2019, doi: 10.1109/TOH.2019.2924881.

[13] "OMS | Accidente cerebrovascular." https://www.who.int/topics/cerebrovascular_accident/es/ (accessed Feb. 23, 2021).

[14] Q. Wu, X. Wang, and F. Du, "Development and analysis of a gravity-balanced exoskeleton for active rehabilitation training of upper limb," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 230, no. 20, pp. 3777–3790, Dec. 2016, doi: 10.1177/0954406215616415.

[15] C. Ochoa-Luna, M. Rahman, M. Saad, and P. Archambault, "Robotic assisted trajectory tracking for human arm rehabilitation," *undefined*, 2014. /paper/Robotic-assisted-trajectory-tracking-for-human-arm-Ochoa-Luna-Rahman/be52699edf2b6d19d0b80f8cb3496f4929ecd77a (accessed Feb. 24, 2021).

[16] A. Zeiaee, R. Soltani-Zarrin, R. Langari and R. Tafreshi, "Design and kinematic analysis of a novel upper limb exoskeleton for rehabilitation of stroke patients," *2017 International Conference on Rehabilitation Robotics (ICORR)*, London, 2017, pp. 759-764, doi: 10.1109/ICORR.2017.8009339.

[17] U. Keller, H. J. A. van Hedel, V. Klamroth-Marganska, and R. Riener, "ChARMin: The First Actuated Exoskeleton Robot for Pediatric Arm Rehabilitation," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 5, pp. 2201–2213, Oct. 2016, doi: 10.1109/TMECH.2016.2559799.

[18] Y. Zimmermann, A. Forino, R. Riener and M. Hutter, "ANYexo: A Versatile and Dynamic Upper-Limb Rehabilitation Robot," *in IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3649-3656, Oct. 2019, doi: 10.1109/LRA.2019.2926958.).

[19] X. Cui, W. Chen, X. Jin and S. K. Agrawal, "Design of a 7-DOF Cable-Driven Arm Exoskeleton (CAREX-7) and a Controller for Dexterous Motion Training or Assistance," *in IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 1, pp. 161-172, Feb. 2017, doi: 10.1109/TMECH.2016.2618888.

[20] M. Cardona, M. A. Destarac and C. E. García, "Exoskeleton robots for rehabilitation: State of the art and future trends," *2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII)*, Managua, Nicaragua, 2017, pp. 1-6, doi: 10.1109/CONCAPAN.2017.8278480.

[21] Y. Hasegawa, T. Kitamura, S. Sakaino, and T. Tsuji, "Bilateral Control of Elbow and Shoulder Joints Using Functional Electrical Stimulation Between Humans and Robots," *IEEE Access*, vol. 8, pp. 15792–15799, 2020, doi: 10.1109/ACCESS.2020.2967466.

[22] L. Liu, Y.-Y. Shi, and L. Xie, "A novel multi-dof exoskeleton robot for upper limb rehabilitation," *J. Mech. Med. Biol.*, vol. 16, no. 08, p. 1640023, Oct. 2016, doi: 10.1142/S0219519416400236.

[23] K. L. M. Ms. P. H. Ds. FIAC, A. F. D. I. P. FAAA, and A. M. R. A. Bs. Ms. PhD, *Anatomía con orientación clínica*. 2018.

[24] M. A. Destarac, J. G. Montaño, M. Cardona, R. E. Gómez, L. J. Puglisi, and C. E. G. Cena, "ORTE Exoskeleton: Kinematic Analysis and Dynamic Modeling," in *2018 IEEE 38th Central America and Panama Convention (CONCAPAN XXXVIII)*, Nov. 2018, pp. 1–6, doi: 10.1109/CONCAPAN.2018.8596581.

[25] T. Mai, "Technology Readiness Level," *NASA*, May 06, 2015. http://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html (accessed Feb. 24, 2021).

[26] J. M. I. de A. Quintana, "Niveles de madurez tecnológica: Technology readiness levels: TRLS : una introducción," *Economía industrial*, no. 393, pp. 165–171, 2014.

[27] E. Ambrosini et al., "A Hybrid Robotic System for Arm Training of Stroke Survivors: Concept and First Evaluation," *in IEEE Transactions on Biomedical Engineering*, vol. 66, no. 12, pp. 3290-3300, Dec. 2019, doi: 10.1109/TBME.2019.2900525.

[28] G. Andrikopoulos, G. Nikolakopoulos, y S. Manesis, "Design and development of an exoskeletal wrist prototype via pneumatic artificial muscles", *Meccanica*, vol. 50, núm. 11, pp. 2709–2730, 2015

[29] V. W. Oguntosin, Y. Mori, H. Kim, S. J. Nasuto, S. Kawamura, and Y. Hayashi, "Design and Validation of Exoskeleton Actuated by Soft Modules toward Neurorehabilitation—Vision-Based Control for Precise Reaching Motion of Upper Limb," *Frontiers in Neuroscience*, vol. 11, 2017, doi: 10.3389/fnins.2017.00352.

[30] M. A. Destarac, C. E. Garcia Cena, J. Garcia, R. Espinoza and R. J. Saltaren, "ORTE: Robot for Upper Limb Rehabilitation. Biomechanical Analysis of Human Movements.," *in IEEE Latin America Transactions*, vol. 16, no. 6, pp. 1638-1643, June 2018, doi: 10.1109/TLA.2018.8444160.

[31] Q. Wu, X. Wang, B. Chen, and H. Wu, "Patient-Active Control of a Powered Exoskeleton Targeting Upper Limb Rehabilitation Training," *Front Neurol*, vol. 9, Oct. 2018, doi: 10.3389/fneur.2018.00817.

[32] Q. Wu and H. Wu, "Development, Dynamic Modeling, and Multi-Modal Control of a Therapeutic Exoskeleton for Upper Limb Rehabilitation Training," *Sensors (Basel)*, vol. 18, no. 11, Oct. 2018, doi: 10.3390/s18113611.

[33] A. Riani, T. Madani, A. Benallegue, and K. Djouani, "Adaptive integral terminal sliding mode control for upper-limb rehabilitation exoskeleton," *Control Engineering Practice*, vol. 75, pp. 108–117, Jun. 2018, doi: 10.1016/j.conengprac.2018.02.013.

[34] T. Madani, B. Daachi, and K. Djouani, "Modular-Controller-Design-Based Fast Terminal Sliding Mode for Articulated Exoskeleton Systems," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 3, pp. 1133–1140, May 2017, doi: 10.1109/TCST.2016.2579603.

[35] Y. Jung and J. Bae, "Kinematic Analysis of a 5-DOF Upper-Limb Exoskeleton With a Tilted and Vertically Translating Shoulder Joint," *in IEEE/ASME Transactions on Mechatronics*, vol. 20, no. 3, pp. 1428-1439, June 2015, doi: 10.1109/TMECH.2014.2346767.

[36] E. Trigili *et al.*, "Design and Experimental Characterization of a Shoulder-Elbow Exoskeleton With Compliant Joints for Post-Stroke Rehabilitation," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 4, pp. 1485–1496, Aug. 2019, doi: 10.1109/TMECH.2019.2907465.

[37] C.-T. Chen, W.-Y. Lien, C.-T. Chen, M.-J. Twu, and Y.-C. Wu, "Dynamic Modeling and Motion Control of a Cable-Driven Robotic Exoskeleton With Pneumatic Artificial Muscle Actuators," *IEEE Access*, vol. 8, pp. 149796–149807, 2020, doi: 10.1109/ACCESS.2020.3016726.

[38] D. Copaci, F. Martín, L. Moreno, and D. Blanco, "SMA Based Elbow Exoskeleton for Rehabilitation Therapy and Patient Evaluation," *IEEE Access*, vol. 7, pp. 31473–31484, 2019, doi: 10.1109/ACCESS.2019.2902939.

[39] S. Chen et al., "Assistive Control System for Upper Limb Rehabilitation Robot," *in IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1199-1209, Nov. 2016, doi: 10.1109/TNSRE.2016.2532478.

[40] T. Higuma, K. Kiguchi and J. Arata, "Low-Profile Two-Degree-of-Freedom Wrist Exoskeleton Device Using Multiple Spring Blades," *in IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 305-311, Jan. 2018, doi: 10.1109/LRA.2017.2739802.

[41] F. Ozkul and D. E. Barkana, "Upper-Extremity Rehabilitation Robot RehabRoby: Methodology, Design, Usability and Validation:," *International Journal of Advanced Robotic Systems*, Jan. 2013, doi: 10.5772/57261.

[42] X.-F. Zhang *et al.*, "The design of a hemiplegic upper limb rehabilitation training system based on surface EMG signals," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 12, no. 1, pp. JAMDSM0031–JAMDSM0031, 2018, doi: 10.1299/jamdsm.2018jamdsm0031.

[43] E. Tanaka, S. Saegusa, Y. Iwasaki, and L. Yuge, "Development of an ADL assistance apparatus for upper limbs and evaluation of muscle and cerebral activity," presented at the ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, IDETC/CIE 2014, 2014, doi: 10.1115/DETC2014-34914.

[44] W. Ye, Z. Li, C. Yang, F. Chen, and C.-Y. Su, "Motion Detection Enhanced Control of an Upper Limb Exoskeleton Robot for Rehabilitation Training," *Int. J. Human. Robot.*, vol. 14, no. 01, p. 1650031, Feb. 2017, doi: 10.1142/S0219843616500316.

[45] W. Ye, Z. Li and C. Su, "Development and human-like control of an upper limb rehabilitation exoskeleton using sEMG bio-feedback," *2012 IEEE International Conference on Mechatronics and Automation*, Chengdu, China, 2012, pp. 2077-2082, doi: 10.1109/ICMA.2012.6285142.

[46] Q.-C. Wu, X.-S. Wang, and F.-P. Du, "Analytical Inverse Kinematic Resolution of a Redundant Exoskeleton for Upper-Limb Rehabilitation," *Int. J. Human. Robot.*, vol. 13, no. 03, p. 1550042, Nov. 2015, doi: 10.1142/S0219843615500425.

[47] H. S. Nam *et al.*, "Biomechanical Reactions of Exoskeleton Neurorehabilitation Robots in Spastic Elbows and Wrists," *IEEE Trans Neural Syst Rehabil Eng*, vol. 25, no. 11, pp. 2196–2203, Nov. 2017, doi: 10.1109/TNSRE.2017.2714203.

[48] B. W. K. Ang and C. -H. Yeow, "Design and Modeling of a High Force Soft Actuator for Assisted Elbow Flexion," *in IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3731-3736, April 2020, doi: 10.1109/LRA.2020.2980990.

[49] S. Bai, S. Christensen, and M. R. U. Islam, "An upper-body exoskeleton with a novel shoulder mechanism for assistive applications," in *2017 IEEE International Conference on Advanced Intelligent Mechatronics, AIM 2017*, Aug. 2017, pp. 1041–1046, doi: 10.1109/AIM.2017.8014156.

[50] S. Crea *et al.*, "A novel shoulder-elbow exoskeleton with series elastic actuators," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, Jun. 2016, pp. 1248–1253, doi: 10.1109/BIOROB.2016.7523802.

[51] J. Liu, Y. Ren, D. Xu, S. H. Kang, and L.-Q. Zhang, "EMG-Based Real-Time Linear-Nonlinear Cascade Regression Decoding of Shoulder, Elbow, and Wrist Movements in Able-Bodied Persons and Stroke Survivors," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 5, pp. 1272–1281, May 2020, doi: 10.1109/TBME.2019.2935182.

[52] M. R. Islam, M. Assad-Uz-Zaman, and M. H. Rahman, "Design and control of an ergonomic robotic shoulder for wearable exoskeleton robot for rehabilitation," *Int. J. Dynam. Control*, vol. 8, no. 1, pp. 312–325, Mar. 2020, doi: 10.1007/s40435-019-00548-3.

[53] S. Ates, C. J. W. Haarman, and A. H. A. Stienen, "SCRIPT passive orthosis: design of interactive hand and wrist exoskeleton for rehabilitation at home after stroke," *Auton Robot*, vol. 41, no. 3, pp. 711–723, Mar. 2017, doi: 10.1007/s10514-016-9589-6.

[54] J. Lobo-Prat *et al.*, "Implementation of EMG- and Force-Based Control Interfaces in Active Elbow Supports for Men With Duchenne Muscular Dystrophy: A Feasibility Study," *IEEE Trans Neural Syst Rehabil Eng*, vol. 24, no. 11, pp. 1179–1190, Nov. 2016, doi: 10.1109/TNSRE.2016.2530762.

[55] A. S. Niyetkaliyev, S. Hussain, M. H. Ghayesh, and G. Alici, "Review on Design and Control Aspects of Robotic Shoulder Rehabilitation Orthoses," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 1134–1145, Dec. 2017, doi: 10.1109/THMS.2017.2700634.

[56] B. Huang, Z. Li, X. Wu, A. Ajoudani, A. Bicchi, and J. Liu, "Coordination Control of a Dual-Arm Exoskeleton Robot Using Human Impedance Transfer Skills," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 5, pp. 954–963, May 2019, doi: 10.1109/TSMC.2017.2706694.

[57] H. Hsieh, D. Chen, L. Chien, and C. Lan, "Design of a Parallel Actuated Exoskeleton for Adaptive and Safe Robotic Shoulder Rehabilitation," *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 5, pp. 2034–2045, Oct. 2017, doi: 10.1109/TMECH.2017.2717874.

[58] N. Li *et al.*, "Bioinspired Musculoskeletal Model-based Soft Wrist Exoskeleton for Stroke Rehabilitation," *J Bionic Eng*, vol. 17, no. 6, pp. 1163–1174, Nov. 2020, doi: 10.1007/s42235-020-0101-9.

[59] A. Ianoşi, A. Dimitrova, S. Noveanu, O. M. Tătar, and D. S. Mândru, "Shoulder-elbow exoskeleton as rehabilitation exerciser," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 147, p. 012048, Aug. 2016, doi: 10.1088/1757-899X/147/1/012048.

[60] X. Wang, Q. Song, X. Wang, and P. Liu, "Kinematics and Dynamics Analysis of a 3-DOF Upper-Limb Exoskeleton with an Internally Rotated Elbow Joint," *Applied Sciences*, vol. 8, no. 3, Art. no. 3, Mar. 2018, doi: 10.3390/app8030464.

[61] Q. Zhang, R. Liu, W. Chen and C. Xiong. "Frontiers | Simultaneous and Continuous Estimation of Shoulder and Elbow Kinematics from Surface EMG Signals | Neuroscience." https://www.frontiersin.org/articles/10.3389/fnins.2017.00280/full (accessed Feb. 24, 2021).

[62] J. V. Kopke, L. J. Hargrove, and M. D. Ellis, "Applying LDA-based pattern recognition to predict isometric shoulder and elbow torque generation in individuals with chronic stroke with moderate to severe motor impairment," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, no. 1, p. 35, Mar. 2019, doi: 10.1186/s12984-019-0504-1.

[63] S. Ghobj, A. Akl, A. El-Farr, M. Ayyash, and J. Abu-Khalaf, "Mechanical design for a cable driven upper limb exoskeleton prototype actuated by pneumatic rubber muscles," in *2017 International Conference on Research and Education in Mechatronics (REM)*, Sep. 2017, pp. 1–7, doi: 10.1109/REM.2017.8075232.

[64] A. N. Goltapeh, S. Behzadipour, and M. Hajihosseinali, "Design and Construction of a Planar Robotic Exoskeleton for Assessment of Upper Limb Movements," in *2019 7th International Conference on Robotics and Mechatronics (ICRoM)*, Nov. 2019, pp. 99–104, doi: 10.1109/ICRoM48714.2019.9071807.

[65] A. González-Mendoza *et al.*, "Upper Limb Musculoskeletal Modeling for Human-Exoskeleton Interaction," in *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Sep. 2019, pp. 1–5, doi: 10.1109/ICEEE.2019.8884537.

[66] H. Hsieh, L. Chien, and C. Lan, "Mechanical design of a gravity-balancing wearable exoskeleton for the motion enhancement of human upper limb," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 4992–4997, doi: 10.1109/ICRA.2015.7139893.

[67] J. Huang, J. Hong, K. Young, and C. Ko, "Development of upper-limb exoskeleton simulator for passive rehabilitation," in *2014 CACS International Automatic Control Conference (CACS 2014)*, Nov. 2014, pp. 335–339, doi: 10.1109/CACS.2014.7097212.

[68] S. Jiang, K. Young, and C. Ko, "Real-time control for an upper-limb exoskeleton robot using ANFIS," in *2017 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, Sep. 2017, pp. 25–25, doi: 10.1109/ARIS.2017.8297176.

[69] C. J. Nycz, M. A. Delph, and G. S. Fischer, "Modeling and design of a tendon actuated soft robotic exoskeleton for hemiparetic upper limb rehabilitation," *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2015, pp. 3889–3892, 2015, doi: 10.1109/EMBC.2015.7319243.

[70] S. K. Ali and M. O. Tokhi, "Control design of a de-weighting upper-limb exoskeleton: extended-based fuzzy," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 7, no. 1, Art. no. 1, Mar. 2019, doi: 10.11591/ijeei.v7i1.938.

[71] G. Aljallad, Y. Qafisheh, and J. Abu-Khalaf, "Modeling and Master-Slave Control of an Upper-Limb Exoskeleton Suit Driven by Pneumatic Muscles," in *2019 20th International Conference on Research and Education in Mechatronics (REM)*, May 2019, pp. 1–6, doi: 10.1109/REM.2019.8744106.

[72] M. G. B. Atia and O. Salah, "Fuzzy logic with load compensation for upper limb exoskeleton control based on IMU data fusion," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec. 2018, pp. 2147–2152, doi: 10.1109/ROBIO.2018.8664849.

[73] R. Casas, T. Chen, and P. S. Lum, "Comparison of Two Series Elastic Actuator Designs Incorporated into a Shoulder Exoskeleton," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, Jun. 2019, pp. 317–322, doi: 10.1109/ICORR.2019.8779448.

[74] D. H. de la Iglesia, A. S. Mendes, G. V. González, D. M. Jiménez-Bravo, and J. F. de Paz Santana, "Connected Elbow Exoskeleton System for Rehabilitation Training Based on Virtual Reality and Context-Aware," *Sensors*, vol. 20, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/s20030858.

[75] J. Rebelo, T. Sednaoui, E. B. den Exter, T. Krueger, and A. Schiele, "Bilateral Robot Teleoperation: A Wearable Arm Exoskeleton Featuring an Intuitive User Interface," *IEEE Robotics Automation Magazine*, vol. 21, no. 4, pp. 62–69, Dec. 2014, doi: 10.1109/MRA.2014.2360308.

[76] J. Weiler, J. Saravanamuttu, P. L. Gribble, and J. A. Pruszynski, "Coordinating long-latency stretch responses across the shoulder, elbow, and wrist during goal-directed reaching," *Journal of Neurophysiology*, vol. 116, no. 5, pp. 2236–2249, Aug. 2016, doi: 10.1152/jn.00524.2016.

[77] C.-T. Chen, W.-Y. Lien, C.-T. Chen, and Y.-C. Wu, "Implementation of an Upper-Limb Exoskeleton Robot Driven by Pneumatic Muscle Actuators for Rehabilitation," *Actuators*, vol. 9, no. 4, Art. no. 4, Dec. 2020, doi: 10.3390/act9040106.

[78] S. H. Lee *et al.*, "Comparisons between end-effector and exoskeleton rehabilitation robots regarding upper extremity function among chronic stroke patients with moderate-to-severe upper limb impairment," *Scientific Reports*, vol. 10, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41598-020-58630-2.

[79] G. Zuccon, M. Bottin, M. Ceccarelli, and G. Rosati, "Design and Performance of an Elbow Assisting Mechanism," *Machines*, vol. 8, no. 4, Art. no. 4, Dec. 2020, doi: 10.3390/machines8040068.

[80] C. Liu, H. Liang, N. Ueda, P. Li, Y. Fujimoto, and C. Zhu, "Functional Evaluation of a Force Sensor-Controlled Upper-Limb Power-Assisted Exoskeleton with High Backdrivability," *Sensors*, vol. 20, no. 21, Art. no. 21, Jan. 2020, doi: 10.3390/s20216379.

[81] T. D. R. G. Thalagala, S. D. K. C. Silva, L. K. A. H. Maduwantha, R. K. P. S. Ranaweera, and R. A. R. C. Gopura, "A 4 DOF exoskeleton robot with a novel shoulder joint mechanism," in *2016 IEEE/SICE International Symposium on System Integration (SII)*, Dec. 2016, pp. 132–137, doi: 10.1109/SII.2016.7843987.

[82] C. R. P. Villalpando and W. Y. Liu, "Design and modeling of a exoskeleton for the human shoulder simulating a ball and socket joint," in *2017 14th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Oct. 2017, pp. 1–5, doi: 10.1109/ICEEE.2017.8108876.

[83] A. S. Niyetkaliyev, E. Sariyildiz, and G. Alici, "A Hybrid Multi-Joint Robotic Shoulder Exoskeleton for Stroke Rehabilitation," in *2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, Jul. 2018, pp. 857–862, doi: 10.1109/AIM.2018.8452681.

[84] T. Noda *et al.*, "Development of Shoulder Exoskeleton Toward BMI Triggered Rehabilitation Robot Therapy," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2018, pp. 1105–1109, doi: 10.1109/SMC.2018.00195.

[85] L. Chien, D. Chen, and C. Lan, "Design of an adaptive exoskeleton for safe robotic shoulder rehabilitation," in *2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, Jul. 2016, pp. 282–287, doi: 10.1109/AIM.2016.7576780.

[86] G. Zhang and M. Lin, "Design of a Soft Robot Using Pneumatic Muscles for Elbow Rehabilitation," in *2018 3rd International Conference on Robotics and Automation Engineering (ICRAE)*, Nov. 2018, pp. 14–18, doi: 10.1109/ICRAE.2018.8586677.

[87] Y. Shen, P. W. Ferguson, y J. Rosen, "Upper limb exoskeleton systems—overview", *Wearable Robotics*, Elsevier, 2020, pp. 1–22.

[88] J. Cornejo, J. A. Cornejo Aguilar, and J. P. Perales Villarroel, "Innovaciones Internacionales En Robótica Médica Para Mejorar El Manejo Del Paciente En Perú," Rev. Fac. Med. Humana, vol. 19, no. 4, pp. 105–113, 2019.

[89] M. Vargas, J. Cornejo, and L. E. Correa-López, "Ingeniería Biomédica: La Revolución Tecnológica Para El Futuro Del Sistema De Salud Peruano," Rev. Fac. Med. Humana, vol. 16, no. 3, 2016.

[90] M. V. Rivera, J. Cornejo, K. Huallpayunca, A. B. Diaz, Z. N. Ortiz-Benique, A. D. Reina, G. Jamanca Lino and V. Ticllacuri, "Medicina humana espacial: Performance fisiológico y contramedidas para mejorar la salud del astronauta," Rev. Fac. Med. Humana, vol. 20, no. 2, pp. 131-142, 2020.

[91] P. Palacios, W. Castillo, M. V. Rivera and J. Cornejo, "Design of T-EVA: Wearable Temperature Monitoring System for Upper Limbs during Extravehicular Activities on Mars," *2020 IEEE Engineering International Research Conference (EIRCON)*, Lima, Peru, 2020, pp. 1-4, doi: 10.1109/EIRCON51178.2020.9254027.

[92] M. A. Gull, S. Bai, y T. Bak, "A review on design of upper limb exoskeletons", *Robotics*, vol. 9, núm. 1, p. 16, 2020.

[93] Cardona M., Destarac M., Cena C. (2020) Robotics for Rehabilitation: A State of the Art.In: Exoskeleton Robots for Rehabilitation and Healthcare Devices. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-15-4732-4_1

[94] D. A. Rozas Llontop, J. Cornejo, R. Palomares and J. A. Cornejo-Aguilar, "Mechatronics Design and Simulation of Anthropomorphic Robotic Arm mounted on Wheelchair for Supporting Patients with Spastic Cerebral Palsy," *2020 IEEE International Conference on* Engineering *Veracruz (ICEV)*, Boca del Rio, Mexico, 2020, pp. 1-5, doi: 10.1109/ICEV50249.2020.9289665.

My-Covid-Safe-Town: A mobile application to support post-Covid recovery of small local businesses

Jorge Torres
Dept. of Computer Science
College of Science and Math
Montclair State University
1 Normal Ave
Montclair, NJ, USA
torresj44@montclair.edu

Vaibhav Anu
Dept. of Computer Science
College of Science and Math
Montclair State University
1 Normal Ave
Montclair, NJ, USA
anuv@montclair.edu

Aparna S. Varde
Dept. of Computer Science
College of Science and Math
Montclair State University
1 Normal Ave
Montclair, NJ, USA
vardea@montclair.edu

Christopher Duran
Dept. of Computer Science
College of Science and Math
Montclair State University
1 Normal Ave
Montclair, NJ, USA
duranc7@montclair.edu

***Abstract*—The Covid-19 pandemic growth has led to a large desire for safety restrictions among citizens near or in Covid-19 affected areas. This includes requiring the use of masks when outdoors and an occupancy limit being placed when indoors. Some of these restrictions have been enforced by the government and can lead to infraction charges on those who choose to ignore them, but some restrictions are up to the decision of the respective individuals. This has led to varying levels of safety being applied when outside. This is especially concerning when some businesses may not be taking proper precautions to avoid the spread of Covid-19. To counterbalance this issue while also spreading awareness of the businesses that are careful enough to follow such precautions, the app My-Covid-Safe-Town (MCST) is created. MCST allows for individuals (i.e., patrons) to find businesses that fit the standard of safety and to specifically point out the steps being taken by each business to avoid the spread of the pandemic.**

Keywords—Covid-19, IoT applications, software engineering, HCI, mobile computing, multimedia technology

I. Introduction

With the end of the Covid-19 pandemic still unclear, businesses and patrons are being forced to adapt to the new environment. For patrons specifically, this includes worrying about safety precautions that are being taken by local businesses and ascertaining whether these businesses are open at hours desirable to the patron. Businesses, especially small ones, are also facing challenges such as operating with less staff and working for fewer hours. These businesses are not simply limited to small "mom and pop" shops since mid-sized businesses such as hotels that do not belong to a chain brand have struggled to stay afloat as well [1]. Large businesses would have enough stored income to continue operating while searching for methods to adapt. Small businesses, however, face problems. In a study conducted by Bartik, it has been reported that 75% of businesses participating in a usability test only had enough income stored to last a maximum of two months [2]. It is important that these businesses continue to get a steady flow of patrons because the failure to do so would imply closure.

Patrons, having been in lockdown for months, would like to frequent businesses again. This does not mean that they are willing to risk their lives or the lives of their families simply so that they can visit a shop. If a patron enters a business that does not take anti-COVID-19 precautions (e.g., requiring face masks inside the shop, sterilizing all frequently touched items, and dining outside to allow proper airflow) then that business is putting the lives of their patrons at risk. In addition, if patrons do not know the policies a business is using to protect its patrons, then it is the equivalent of blindly placing their lives in the hands of the business owners. Fear plays an important role in deciding whether visiting a business is worth the risk. Thus, informing patrons about the businesses that are including anti-COVID-19 practices and thereby reducing apprehension is important. This raises the question of how such information will be spread.

In response to this, a mobile application (i.e., app) called My-Covid-Safe-Town (MCST) is developed with the intention of informing the patron of Covid-19 related information and the manner in which small businesses are changing to accommodate to the new normal. If a patron wants to check whether a business is open or observe the precautions the concerned business is taking, MCST can present that information to the patron in a ubiquitous manner with easy access. In addition, should a patron want to ask questions (e.g., Asking if a certain item is in stock, or if the business takes reservations). MCST allows for this through email and phone calls. MCST provides patrons with the information they require by connecting them with the small businesses that lack the advertisement funds needed to convey that information. Ultimately the goal of the app is to display the shops that are open near the patron and showcase the manner in which it is keeping its patrons safe from the pandemic.

II. Related Works

Using apps to solve pandemic related problems is a common practice in recent times. A paper discussing an app tracking influenza infection mentions using similar methods as app developers are using today [3]. This particular app has used a data donation method which involves app users manually adding their information and location into the app based upon which the app analyzes that data for future use. The app has been designed around obtaining important information rapidly by using modern technology rather than using traditional methods that some government bodies might use. Ultimately the app has received about 10,000 volunteers who have collected data such as the age and date of last infection in real time. In addition, the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

data has correlated with existing basic information on influenza thereby making the concerned app a very interesting study [3].

Apps regarding Covid-19 are steadily surfacing on the Apple App Store and Google Play Store while more are being developed and released through online access. Apps for social distancing and contact tracing are highly significant today. For example, *Covid-Alert-NY* [4] and *Aarogya Setu* [5] are popular contact tracing apps in New York and India respectively, the latter having surpassed over 100 million installations in 40 days. A recent food donation app called *SeVa* [6] addresses hunger and food waste by connecting food suppliers and consumers with up-to-date information about good quality leftover food in restaurants, grocery stores and other such places in their vicinity. This is particularly helpful during Covid-19 though it has a generic usefulness as well. An app called *MDLive* [7] assists with telemedicine that is much needed during recent times since many patients prefer to make doctor visits from home. Another app called *Covid Control* [8] developed at Johns Hopkins University, a premier medical institution in the United States, is beneficial in terms of guiding the research related to this pandemic in order to address aspects such as prevention and cure. Many apps are being used by individuals on a daily basis to help prevent the spread of the disease, however there are drawbacks. According to paper by Kaspar [9], though people are taking steps to avoid Covid-19 by installing these types of apps, the overall ability to avoid Covid-19 is determined by the individual's motivation rather than the app's support. In addition, even though people want to avoid spreading the pandemic does not imply that they will always be motivated to follow proper safety procedures.

While several apps exist to help combat problems associated with Covid-19 and its aftermath, to the best of our knowledge none of these specifically provide a platform for local small business users to showcase their functionality and provide patrons ubiquitous, convenient access to adequate safety related information on these businesses for fostering better interaction. This is precisely where MCST makes a contribution. It goes beyond the "yellow pages" for information since it is more interactive. Yellow pages are static while MCST entails user-generated content and is more dynamic. Moreover, using MCST during Covid-19 recovery, stores can quickly populate additional fields and add new entries that could occur in a very short time span, e.g. "all employees are Covid-19 vaccinated". Information such as this could be highly relevant to safety precautions. Our work is thus orthogonal to the literature and makes specific contributions.

III. Methods

Prior to the creation of the app, its major requirements and design must be outlined, as sound requirements engineering is vital for software quality [24-28]. In our work, this involves developing a wireframe for the app. A wireframe is a mock-up design of the appearance and functionality of the app once it is developed [10][11]. This allows us to see how the pages will look and reduces the need to make drastic changes to the design later at the implementation stage of development. It is important to ascertain that the users have the visual clarity to see the app's functions while also having the tools they might need for using the app.

A. Wireframe Design Outline

The design of the wireframes constitutes the initial step of the app development. An example of the planning of visual clarity is presented in Fig. 1 where the "landing page" of the app is observed. The landing page is the first page the users would see when they open the app. This is also the page with the maximum paths available for navigation. It is thus important to ensure that the users can see all of these paths to understand the role of each button easily, i.e. within just a few clicks. If the users do not understand how to operate any page of this wireframe, then the design has not truly succeeded in assisting the user. However, that is because this is a wireframe on which further improvements can be made.

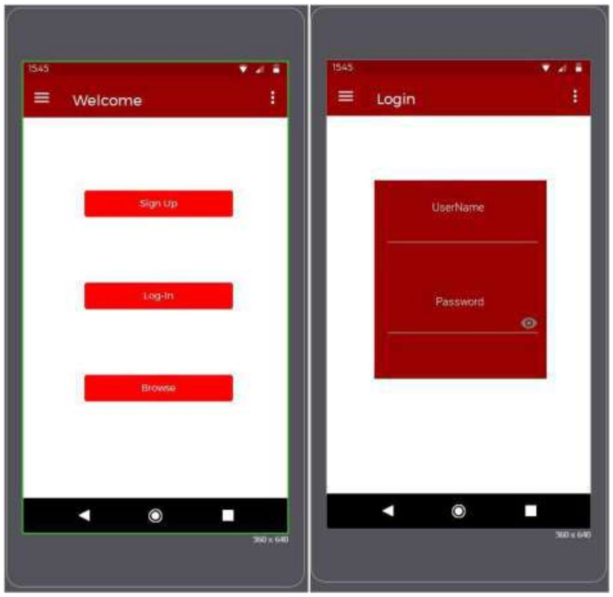


Fig. 1. Wireframe of the welcome page and login page for the MCST app

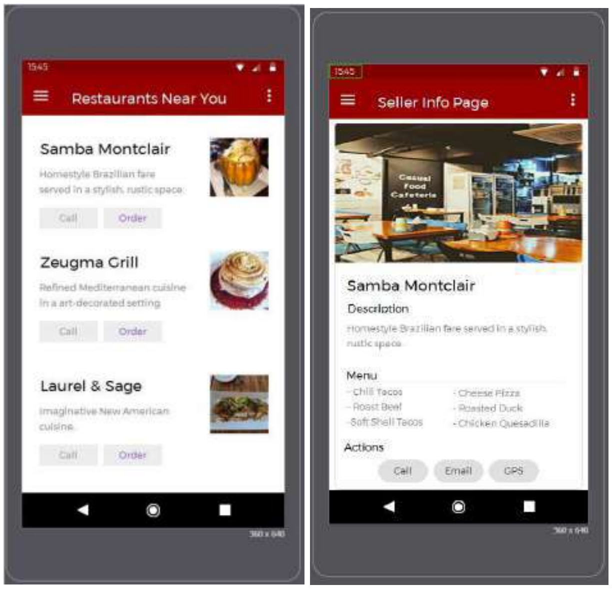


Fig. 2. Wireframe versions of pages for browsing (left) and displaying shops of small businesses (right)

Based on the figures above, we see Fig. 1 (left panel) which is the starting page through which the user can login, register, and browse the app. Fig. 1 (right panel) is an at-a-glance view of the display that appears in the pages of the small businesses. All these pages have a menu bar near the top of the screen which allows for additional options. Fig. 2 (left panel) showcases how users would find a shop that interests them. Once they click on a shop of their choice, they are directed to Fig. 2 (right panel). This is a business page that includes hours of operation and Covid-19 precautions. Likewise, after having constructed the wireframe, it is imperative to conduct usability tests in order to gauge how the targeted audience would react to the app.

B. *Wireframe Usability Testing*

Usability tests by stakeholders are an important part of developing an app. In HCI (human computer interaction) with respect to app development, stakeholders are people, groups or organizations impacted by the success or failure of the app, i.e. those who have a stake in the app [12]. In other words the stakeholders would represent the targeted users, by being the users themselves or having a clear idea of the users' needs. Without usability tests there would not be a reliable way to gauge stakeholder interest in the app. The app might seem perfect from the developers' end and have absolutely no bugs, but if there is no interest in it from the stakeholder perspective, it will not be used. It is for this reason that usability tests are conducted for this app after the outline of the wireframe that constitutes the initial app design. These are analogous to such tests conducted in the development of mid to large scale software systems in Software Engineering.

Since our study in this paper targets mainly patrons, i.e., the prospective customers of small businesses, *our sample population for usability tests focuses on patrons*, not business owners. In other words, our fundamental stakeholders for the current study's purpose are patrons. The following figures, Figs. 3 and 4, are the usability testing results achieved after showcasing the wireframe of the app to the patrons.

The results from Fig. 3 showed that our wireframes got a positive response overall. Overall, 80% responders said that they found our user interface designs easy to use (i.e., responders agree that our designs will ensure high-quality user-experience).

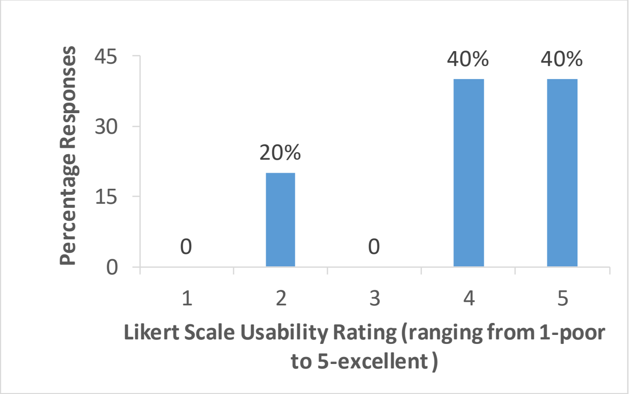


Fig. 3. Wireframe usability testing (quality): App design quality question ratings indicate that that it is well-received as evident from overall scores

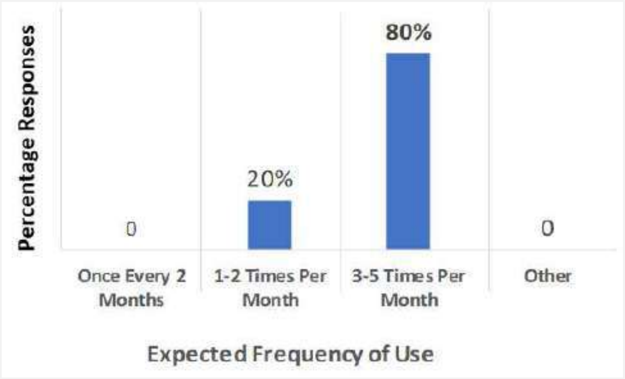


Fig. 4. Wireframe usability testing (frequency): The usage frequency question shows average use a few times a month by patrons which seems quite reasonable

The analysis presented in Fig. 4 indicates that a majority of the responders feel they will use the MCST app at least 3 to 5 times per month, and indicate this accordingly.

Observing the results of wirefame usability testing as displayed in these figures, we notice that there is interest in the app, however its design needs some improvements. We would not have been able to improve upon the app design without the help of these wireframe testing participants. Having discussed these figures it now becomes apparent that being able to navigate through each of the pages is just as important as the appearance of the pages individually. We incorporate such feedback including subjective comments from the wireframe usability tests in order to guide the final user interface design of the MCST app. A main piece of feedback is in regards to not fully understanding the app functionality; and hence this is taken into account based on which several menus are then upgraded. Another piece of feedback is in regards to the app's simplicity; and therefore we avoid adding too many complex buttons and functions in the app.

IV. Models

While wireframes are suitable for testing the appearance and visual clarity of the app, they are not feasible for perceiving the actual structure of the app. In order to accomplish this, we rely on models such as sitemaps and other Unified Modeling Language (UML) Diagrams [13]. UML Diagrams are general

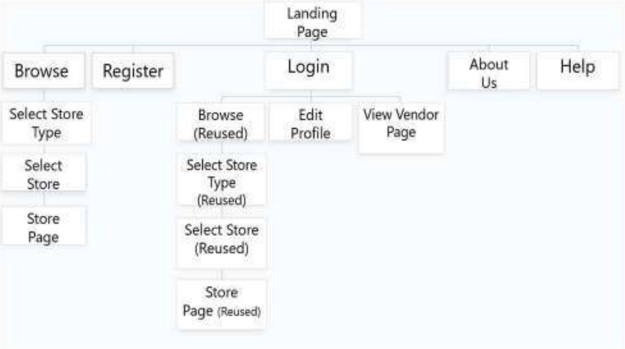


Fig. 5. Sitemap depicting the blueprint of the MCST app

purpose diagrams with the intention of allowing the developers and users of the app to easily see the app's workflow.

Considering the sitemap presented in Fig. 5, the flow of the app can be seen. This is indicative of the app structure and workflow while also allowing for changes to be made prior to the final product. For example, in Fig. 5 we can see a user starting on the landing page and then potentially clicking on login. Assuming the user then logs in, he/she will thereafter will have access to the "edit profile", "view vendor page", and "browse pages" in the app.

A sitemap shows the planned routes for accessing the site. It enables us to know where each page leads. In other words, this constitutes the blueprint of the app. However, this blueprint diagram alone is not sufficent, In the case of this specific app, there is a requirement for the data to be stored and then called upon frequently. In order to achieve this, database design is needed to map out the content and flow of the data. This is illustrated as an E-R (Entity-Relationship) diagram in Fig. 6. Here, we can see that the database is split into two tables with one being the "user account" for the vendors and another one for the businesses, Both of these contain several details such as the names and contact information needed when viewing the business pages in the latter portions of the app. The E-R diagram in Fig. 6 shows the appearance of the database from the back end. This enables us to know what should be stored in the database; in other words the absence of this could potentially lead to wasted space on redundant information. The relationships between the different entities in this app along with their data flow is visible in this diagram. Once the overall blueprint as well as the specific data flow has been accomplished, we can proceed further wirh the actual implementation of the app.

V. Implementation

```
Algorithm 1 Algorithm to determine filter
Input: n = 0 ≤ n ≤ 3
Output: intent
  Initialization:
first statement
  Switch Case
if n = 0
   intent = n
end if
if n = 1
   intent = n
end if
…
return P
```

```
Algorithm 2 Algorithm to apply filtering
Input: intent
Output: out
  Initialization:
first statement
  LOOP Process
for i = i++ do
   if i = intent
      add intent to P
   end if
end for
return P
```

Fig. 7. Algorithms 1 and 2: Pseudocode for the "browse page" in the MCST app

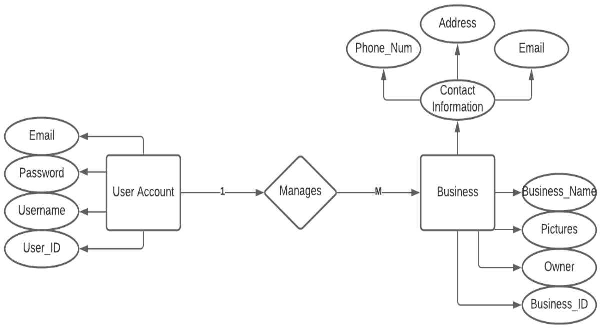


Fig. 6. E-R Diagram providing an overview of the database structure

In order to implement an app once the design is finalized, the platform and coding language must be established along with any existing programs being used. The MCST app is designed for the Android platform and therefore uses the Android Studio [14] for implementation. Android Studio encompasses both Java and eXtensible Markup Language (XML). In addition, our app development entails other software. For example, SQLite is used as a database; also calling, email, and GPS features are added. These require allowing specific permissions such as Google's API for GPS. Fortunately, since Android Studio is the main program used for developing apps it seamlessly allows for these permissions to be implemented through its program with the exception of a few API codes that needed to be transferred. The next section discusses different parts of the app with respect to their implementation details.

A. Browse Page

The "browse page", for which we present the pseudocode in Fig. 7 via Algorithms 1 and 2, portrays the toolbar used in addition to the different buttons to pass filters depending on the selection. If the user selects certain specific buttons, the search results will filter for that type of business specifically, e.g., "restaurants". These results can be seen in Fig. 10 (right panel).

B. Search Page

The "search page" is implemented with respect to the pseudocode seen in Algorithm 3 in Fig. 8. Here, we can see how the data gets sorted into a "results page" seen in Fig. 11 (left panel). Important takeaways from this pseudocode are the

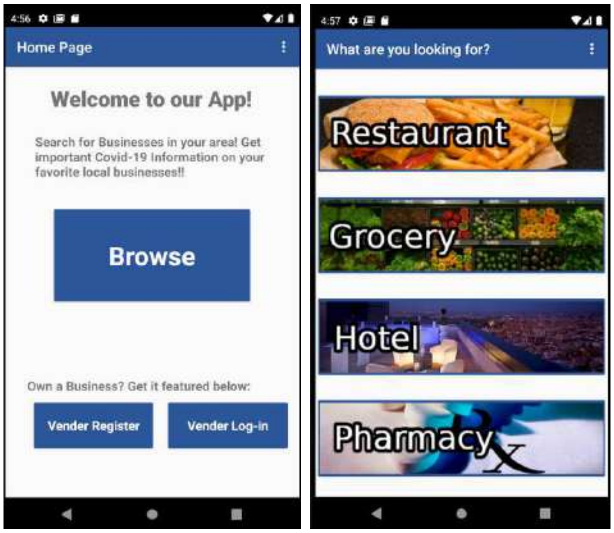
```
Algorithm 3 Display filtered results from database
Input: database results (i.e. businessName, phonenumber, etc.)
Output: intent
  Initialization:
first statement
  Declare Cursor
cursor = database results
if cursor isEmpty
   return "No data found."
end if
else while cursor moveToNext
   cursor getString (columnIndex 1)
   cursor getString (columnIndex 2)
   …
end else while
return P
```

Fig. 8. Pseudocode for the "search page"

```
Algorithm 4 Establish Email, GPS, and Phone Number
Input: intent (i.e. phonenumber, address, email)
Output: out
  Initialization:
first statement
  Call intent
intent = intent.ACTION(DIAL/VIEW/SENDTO)
third statement
  Format intent (i.e. Phone number = tel: xxx-xxx-xxxx)
startActivity (intent)
```

Fig. 9. Pseudocode for the "seller info page": Shows how the GPS, email, and call buttons operate

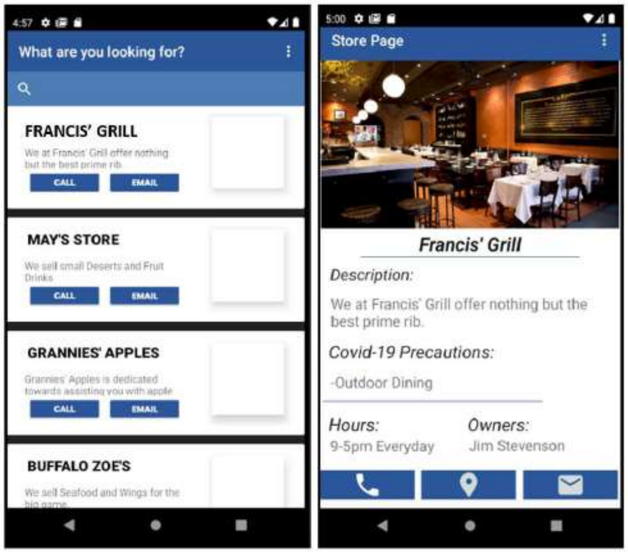
retrieval of the data from the database and the storage it into multiple arrays. These arrays are called at Index 1, Index 2… (entered as columnIndex1, columnIndex2 in the pseudocode shown in Algorithm 3). Hence, the piece of pseudocode shown in Algorithm 3 summarizes the manner in which the information from the database is stored and accessed.



Search for Businesses in your area! Get important Covid-19 Information on your favorite local businesses!!

Own a Business? Get it featured below:

Fig. 10. Example of the landing page and browse page of the MCST app



FRANCIS' GRILL

We at Francis' Grill offer nothing but the best prime rib

MAY'S STORE

We sell small Deserts and Fruit Drinks

GRANNIES' APPLES

Grannies' Apples is dedicated towards assisting you with apple

BUFFALO ZOE'S

We sell Seafood and Wings for the big game

Francis' Grill

Description:

We at Francis' Grill offer nothing but the best prime rib.

Covid-19 Precautions:

-Outdoor Dining

Hours: 9-5pm Everyday

Owners: Jim Stevenson

Fig. 11. Example of search page and the seller-info page

C. Seller Info Page

The final "seller info page", i.e., the information display page for small businesses, is primarily a collection of different buttons and text. Its functionality is summarized herewith in Algorithm 4 in Fig. 9. Note that in Fig. 10 the landing page and browse page of the app are presented. The right panel of Fig. 11 shows the seller info page. This page is certainly important since it has the numerous features useful for the patron. This is because it can call, email, and issue a GPS command to the business of a patron's choosing. In addition, it also contains any information a patron might potentially need.

VI. Experimental Evaluation

A. App Results

In the figures herewith, i.e. Figs. 10 and 11, we can visualize the front-end patron side of the app. The patron side does not have access to editing any businesses; the patrons can only search and view the shops that might interest them. They can, however, call the business from the app, email the business, and connect the GPS to the business they select.

Overall, via this implementation the aim is that the user should be able to acquire much of the information they might require through this MCST app. Any additional information can be found by contacting the business per se. Likewise, the app results have been displayed to our targeted users for their evaluations for user acceptance testing as described next.

B. User Acceptance Testing

The current version of the MCST app has been subjected to user acceptance testing in order to assess its effectiveness. Questionnaires were provided to a sample set of potential app users along with Likert scale responses [15]. This helps us to gauge how the users perceived the app. We have circulated the questionnaires among 31 users and received the following responses as illustrated in the figures next.

From Fig. 12 it seems that though most people would only use the MCST app once or twice a month while some people

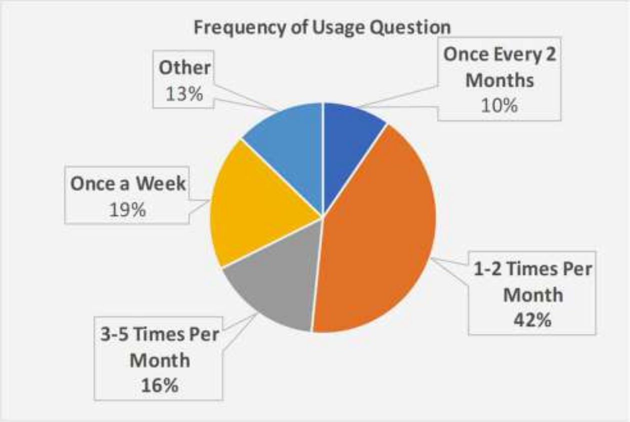


Fig. 12. Frequency of app usage estimated by the responders

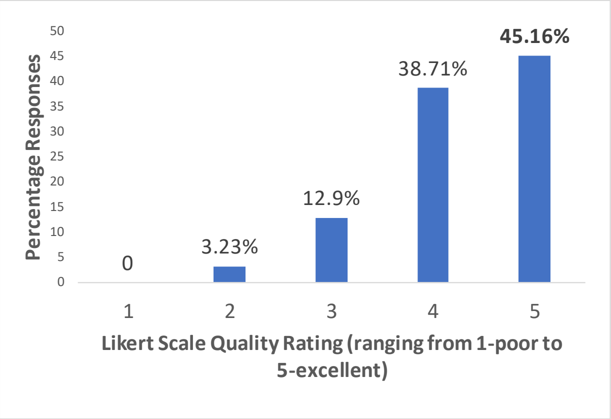


Fig. 13. User ratings for the overall quality of the app

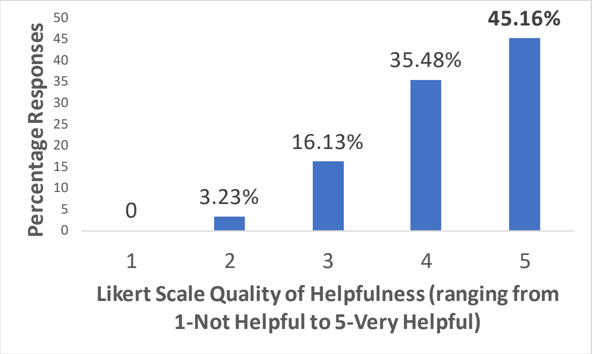


Fig. 14. Users' perceived usefulness of the app in a Covid-19 world

would use it weekly. On the whole, the results indicate that the app attracts interest. As per Fig. 13 we can infer that the app quality is well-received and it thus can be useful in a real environment. Based on Fig. 14 we can observe that the app is considered helpful by most people in Covid-19 environment. In addition to Likert scale evaluation, our user acceptance testing has also leveraged subjective assessment in order to gauge the general response to the MCST app. A snapshot of the comments from subjective assessment appears in Fig. 15. Based on this, we find that the users have well comprehended the intention of the app and have appreciated it on the whole. Some of them have provided recommendations on improving the app for an even better user experience in the future.

Hence, we can conclude that the targeted users (i.e., patrons), in general, have assessed the app as helpful and worthy of usage. It also appears as though in the current Covid-19 world, there is no such specific app identical to MCST that has already been implemented (to the best of our knowledge) and thus our MCST app would be used quite frequently. The user acceptance testing results seem promising. They also offer encouragement for any potential further enhancements to be made within future versions of the app.

VII. Conclusions and Future Works

The MCST (My-Covid-Safe-Town) app has been designed with the intention of informing users of how small businesses are taking steps to protect their patrons while also informing the patrons of any changes made to small businesses that have been impacted by Covid-19. This work falls in the general paradigm of AI in Software Engineering, in the broad realm of other such works, e.g. [16] [17] [18] addressing aspects such as software quality and requirements. The MCST app includes aspects such as: being able to find out whether a small business is closed or whether its operating hours have changed (instead of a patron walking to the shop and viewing its hours posted on the front door only to realize that it is closing early and will open the following day or worse still finding that it is temporarily shut down). Although the MCST app is still missing advanced features that require additional time investment, the fundamental features are provided and hence the core intention of the app is achieved. After having viewed the usability tests, it can be inferred that the participants comprising the potential clientele recognize the intention of the app and its goal. The app per se is designed to start on a small-scale basis and has the backbone to grow further as it becomes more popular. This app primarily covers the Montclair region in NJ, USA and it can also be used on a nationwide level; any browsing limitations can easily be implemented and updated. This app truly has the potential to grow further. It can help several people who fear leaving their homes during this Covid-19 pandemic and its aftermath. Our work on the MCST app development here is supplementary to other apps contributing to aspects of smart cities such as smart environment [19] [20], smart government [21] and smart living [4] [5] [6]. This MCST app thus makes broader impacts on smart cities, analogous to other works in the literature [22] [23] [29]. This indicates a broader perspective of the work herein besides its immediate impact on helping to combat the adverse effects Covid-19 pandemic and its aftermath.

- *"Password recovery would be useful. This is a well-designed app and a great idea."*
- *"I would very much like to see this app put to use"*
- *"Can you add something for other healthcare businesses besides pharmacies? What about therapy, rehab, counseling etc. Much needed in COVID."*
- *"It's very good work. Such apps should be designed for topics in environmental studies too."*
- *"Nice work! You should keep track of the number of real users of your app if you release this."*
- *"The app is a great idea. It has the potential to connect small business owners and patrons. This will greatly help in the cause of COVID recovery."*

Fig. 15. Comments of usability test responders (responders are asked to give general comments about the MCST app in an open-ended question)

As regards further work, after reviewing the usability tests and reflecting back on additional features that have not been implemented thus far, the MCST app does have room for improvement. There are participants in the usability test who have noticed issues such as the categories being quite limited, however those features can be added quite easily. A few mentions include minor UI (User Interface) improvements to enhance navigation on the vendor side as well as adding analytics to track the type of users that visit the app and the shops they view. A big concern that requires additional improvements is checking whether the vendor is truthful when entering his or her business information on the app. This is because false vendors could pose fake shops or enter existing shops with false or malicious information. A potential solution to this would be to either have moderators to verify each individual shop and check for validity or have the app be community-operated. In the case of the latter, the community itself would need to check whether the individual shop's claims are valid. Accordingly, a "report function" will be needed to acknowledge these claims. This constitutes some aspects of future work.

In addition to this veracity of information that can be addressed in further enhancements, other future work includes: augmentation of the functionality with respect to adding more categories of businesses such as healthcare related services; enhancement of the UI features; and publicity aspects to gain more visibility for the app. Some of these issues would be addressed as we envisage releasing the MCST app that is on our roadmap. *The MCST app can be found at the following link* [30]*:* https://dx.doi.org/10.13140/RG.2.2.12191.48802/2*.* The details of this app along with its code, data and other relevant files are available on the GitHub pages of the developers and can be provided to some interested users upon request. The ownership of the MCST app rests with its designers and developers who are the authors of this paper.

Acknowledgments

Dr. Vaibhav Anu and Dr. Aparna Varde have a joint internal grant SGPD from the College of Science and Math at Montclair State University (MSU). Also, Dr. Varde is partly supported by a grant from the National Science Foundation, USA. This is NSF MRI: Acquisition of a High-Performance GPU Cluster for Research and Education. Award Number 2018575. In addition, both Dr. Anu and Dr. Varde have support from the Faculty Scholarship Program at MSU.

References

[1] H. Shin and J. Kang, "Reducing perceived health risk to attract hotel customers in the COVID-19 pandemic era: Focused on technology innovation for social distancing and cleanliness," *International Journal of Hospital Mgmt*. 2020, doi: 10.1016/j.ijhm.2020.102664.

[2] A. W. Bartik, M. Bertrand, Z. Cullen, E. L. Glaeser, M. Luca, and C. Stanton, "The impact of COVID-19 on small business outcomes and expectations," *Proceedings of the National Academy of Sciences of the United States of America*. 2020, doi: 10.1073/pnas.2006991117.

[3] K. Fujibayashi, H. Takahashi, M. Tanei, et al, "A new influenza-tracking smartphone app (Flu-report) based on a self-administered questionnaire: Cross-sectional study," *JMIR mHealth uHealth*, 2018.

[4] NY Department of Health, "Covid Alert NY", 2020, https://coronavirus.health.ny.gov/covid-alert-ny-what-you-need-know

[5] A. Jhunjhunwala "Role of Telecom Network to Manage COVID-19 in India: Aarogya Setu" *Transactions of the Indian National Academy of Engineering*, Jun 2020, pp. 1-5, PMC7264964.

[6] C. Varghese, D. Pathak, and A. Varde, "SeVa: A foor donation app for smart living", I*EEE CCWC,* Jan 2021, pp. 408-413.

[7] MD Live Mobile App 4.0, https://www.mdlive.com/mobileapp/

[8] Johns Hopkins, "Covid Control", 2020, https://covidcontrol.jhu.edu/

[9] K. Kaspar, "Motivations for social distancing and app use as complementary measures to combat the COVID-19 pandemic: Quantitative survey study," *Journal of Medical Internet Research*. 2020, doi: 10.2196/21613.

[10] D. Rees, "What is wireframing," *Experience UX*. https://www.experienceux.co.uk/faqs/what-is-wireframing/

[11] "All-in-one prototyping tool for web and mobile apps," *Justinmind*. https://www.justinmind.com/ (accessed Dec. 13, 2020).

[12] Y. Rogers, H. Sharp, and J. Preece, *Interaction Design: Beyond Human-Computer Interaction*, 4th ed. Hoboken: Wiley, 2015.

[13] "Unified Modeling Language," *Wikipedia*. https://en.wikipedia.org/wiki/Unified_Modeling_Language.

[14] "Download Android Studio and SDK tools : Android Developers," *Android Developers*. https://developer.android.com/studio

[15] L. Liedke, "Likert Scale Definition (How to Use It, With Examples)," 2020. https://wpforms.com/beginners-guide-what-is-a-likert-scale-and-how-to-use-it/ (accessed Feb. 07, 2021).

[16] K. Z. Sultana, V. Anu and T.Y. Chong, "Using software metrics for predicting vulnerable classes and methids in Java projects: A machine learning approach", *Journal of Software Evolution and Process,* 2021, 33(3):e2303.

[17] E. Onyeka, A. Varde, V. Anu, N. Tandon and O. Daramola, "Using Commonsense Knowledge and Text Mining for Implicit Requirements Locatization", *IEEE ICTAI*, Nov 2020, pp. 935-940.

[18] M. Felderer and R. Ramler, "Quality assurance for AI-based systems: Overview and challenges" arXiv 2102.05351v1, Feb 2021.

[19] D. Pathak, A. Varde, C. Alo, F. Oteng, "Ubiquitous access for local water management through HCI based app development", *IEEE UEMCON*, Oct 2019, pp. 226-232.

[20] D. Karthikeyan, S. Shah, A. Varde, C. Alo, "Interactive visualization and app development for precipiation data in Sub-Saharan Africa", I*EEE IEMTRONICS,* Sep 2020, pp. 302-308.

[21] C. Varghese, A. Varde and X. Du, "An ordinance-tweet mining app to disseminate urban policy knowledge for smart governance", *Conf on e-Business, e-sService and e-Scoiety* (*I3E)*, Vol. 2, pp. 389-401.

[22] M. Alamaniotis, I. Ktistakis, "Fuzzy leaky bucket with application to coordinating smart appliances in smart homes", *IEEE ICTAI*, Nov 2018, pp. 878-883.

[23] R. Giffinger, H. Kramar, G. Haindlmaier, F. Strohmayer, *European Smart Cities,* Technical Report, TU Wien (Vienna Institute of Technology), Dept. of Spatial Planning, Vienna, Austria, 2015.

[24] V. Anu, G. S. Walia et al., "Usefulness of a Human Error Identification Tool for Requirements Inspection: An Experience Report", *In: Grünbacher P., Perini A. (eds) Requirements Engineering: Foundation for Software Quality. REFSQ 2017. Lecture Notes in Computer Science*, vol 10153. Springer, Cham, 2017.

[25] V. Anu, W. Hu et al., "Development of a Human Error Taxonomy for Software Requirements: A Systematic Literature Review", *Information and Software Technology Journal*, vol. 103, pp. 112-124, Nov 2018.

[26] V. Anu, G. S. Walia, W. Hu, J. C. Carver and G. Bradshaw, "Using A Cognitive Psychology Perspective on Errors to Improve Requirements Quality: An Empirical Investigation", *27th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 65-76, 2016.

[27] V. Anu, G. Walia, W. Hu, J. C. Carver and G. Bradshaw, "Issues and Opportunities for Human Error-Based Requirements Inspections: An Exploratory Study," *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Toronto, ON, Canada, 2017, pp. 460-465.

[28] M. Singh, V. Anu, G. S. Walia and A. Goswami, "Validating Requirements Reviews by Introducing Fault-Type Level Granularity: A Machine Learning Approach", *In Proceedings of the 11th Innovations in Software Engineering Conference (ISEC)*, pp. 1-11, 2018.

[29] V. Chang, "An ethical framework for big data and smart cities", *J. of Technological Forecasting and Social Change*, 2021, 165: 120559.

[30] J. Torres, C. Duran, V. Anu and A. Varde, "The MCST App: My-Covid-Safe-Town", March 2021, https://dx.doi.org/10.13140/RG.2.2.12191.48802/2

The impact of Social Media Marketing on the Travel Intention of Z Travelers

Bui Thanh Khoa
Faculty of Commerce and Tourism
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Viet Nam
khoadhcn@gmail.com
buithanhkhoa@iuh.edu.vn

Nguyen Minh Ly
Faculty of Commerce and Tourism
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Viet Nam
ly.nguyenminh2811@gmail.com

Vo Thi Thao Uyen
Faculty of Commerce and Tourism
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Viet Nam
uyen.vo.gtd@gmail.com

Nguyen Thi Trang Oanh
Faculty of Commerce and Tourism
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Viet Nam
ntto0516@gmail.com

Bui Thanh Long
School of Science and Technology
RMIT University
Ho Chi Minh City, Viet Nam
s3748575@rmit.edu.vn

Abstract— **The Covid-19 pandemic has dramatically affected the tourism industry's development; however, the demand for traveling increases in the young people, significantly, the Generation Z, who was born in 1996. A bright outlook awaits the tourism industry in the post-Covid-19 period. The study aims to determine the relationship between the economic perspective, social perspective, user-generated content trust, and inbound travel intention of Generation Z travelers. The mixed-method research was done to archive the research objectives. The research results pointed out that there were positive impacts of economic and social perspectives on user-generated content trust; moreover, user-generated content trust positively affected the inbound travel intention. Some managerial implications were proposed for the tourism business managers to increase their business performance.**

Keywords— *economic perspective, social perspective, user-generated content trust, and inbound travel intention, Z traveler.*

I. Introduction

In global integration, tourism is an economic sector that brings significant benefits to the country. In recent years, the trends of travel have noticeably emerged on the Internet. Social media marketing of websites, representing various forms of consumer-generated content such as blogs, virtual communities, wikis, social networks, collaborative tagging, and media files shared on sites like YouTube and Flickr, have gained substantial popularity in online travelers' use of the Internet [1]. For example, users, who use social media, can share their travel experience on TripAdvisor or Booking.com. Social media users (e.g., influencers) can also serve as endorsers to promote travel destinations on social media such as Facebook, TikTok, YouTube, Instagram [2]. Simultaneously, when visitors use social networking platforms to portray, recreate, and relive the journeys, the Internet rapidly mediates tourism experiences [3]. Young people today are also more interested in advertised and well-rated places on social media.

Given the tourism product's intangible service nature, the social network is an excellent resource for travelers seeking knowledge [4, 5]. Social advertisement is more and more popular to enhance the customer experience [6]. It is widely accepted in both academia and business that social media has become an important advertising venue for marketers [7]. In line with the industry's enthusiasm for social media advertisements, the number of academic, social media advertising studies has risen dramatically in recent years [8]. Along with the influence of the media, customer confidence in the tourist destination is also crucial. Many factors that help with the tourism industry's growth increase trust, such as creating good services and taking care of tourism security create a safe and friendly environment for travelers guests when receiving related services while going travel. Tourists often travel to a familiar place and a country with crime rates and political unrest low [9]. Tourists will have a right impression of a tourist destination when that place is not reflected by social media and guaranteed by the local government's calendar [3].

Furthermore, it discussed the economic perspective of travel time, i.e., the idea that travel time is wasted time, by examining how journeys are encountered, especially on trips that are often repeated as commuting; or along specific routes with highway driving [10]. Opportunities for co-present meetings with friends and family are minimal due to the geographic distribution of social networks, the availability of a vehicle, and free time constraints. As a result, the possibility of 'teleportation' for some people suggested that there could be more possibilities for co-present experiences.

Budgets for travel time may also be strained due to more spatially compact social networking where access to or provision of public transportation is restricted. An individual's travel intention by young people plays an essential role in the traveling destination choosing process. Tourists' attitudes and preferences toward a specific tourist destination are often used to determine travel intentions [11]. Travel intention is habitually based on the tourists' attitude and preference toward a particular tourist destination. Rational and practical conditions are seen as critical measures of tourists' behavior, particularly their attitude and preference [12]. To put it another way, functional and psychological factors frequently affect one's desire to travel, which leads to actual travel behavior. For that reason, tours should promote marketing more to attract more young visitors.

The following parts of this study included the literature review, research method, result and discussion, and the conclusion.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. LITERATURE REVIEW

A new generation is making its way into the limelight, and the rest of the world is taking note. These people are Generation Z, and they were born between 1996 and 2015 [13]. According to the Digital Tourism Think Tank, by 2020, Gen Z will account for 40% of all customers. Gen Z is reaching adulthood, and they are excited to fulfill their wanderlust. Furthermore, even though Gen Zers may not have a strong influence in a particular field, they significantly impact their family vacation. Stillman of GenZGuru discovered that children schedule more than 70% of multigenerational trips. According to a 2017 Contiki survey, 98 percent of 732 Gen Z respondents agreed that travel experiences are essential in life [14]. Overall, this multicultural generation is not only socially aware but also internationally aware. They consider themselves people of the planet, unconstrained by geographical limits, and, happily for travel advisors, they want to see it all for themselves.

The Technology Acceptance Model (*hereafter* TAM) is the most widely used in analyzing people's behavior intentions [15]. TAM is an extension of reasoned action theory with perceived behavior control to improve behavioral intention to predict an individual's actual behavior. TAM expressed why a person uses the system for their daily activities through the relationship between the perceived usefulness, perceived ease of use, attitude, behavioral intention, and user behavior [16]. The blooming of Web 2.0 has changed consumer behavior. The user-generated contents, which were evaluated directly on the sellers' websites or social network pages after experiencing the products or services by customers, become reliable sources for other customers who intend to use the service or product. Therefore, if customers believe the user-generated contents, they would easily buy or use the goods or services. Trust was recommended as the positive attitude of a person towards others or somethings.

On the other hand, the prior studies pointed out that user-generated content will benefit the other users as economic benefits as functional, monetary benefits [18]; and social benefits as relational, social, and emotional benefits [17]. From the critical evaluation above and the TAM model adoption in the tourism industry, this research proposed the research model with three components, including the perceived usefulness with an economic and social perspective, the trust of the user-generated content as the consumer attitude, and the inbound travel intention of Z travelers. The conceptual model was presented in Figure 1.

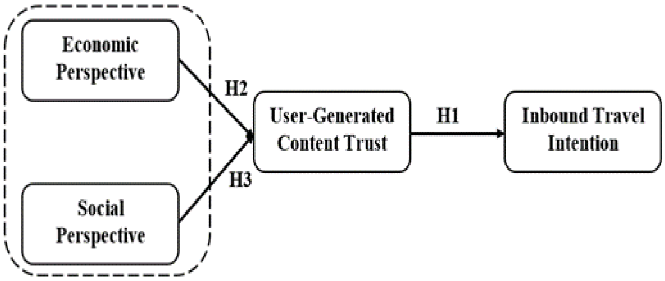


Fig. 1. The conceptual model

A. Inbound travel intention

If there is an opportunity to act, the purpose will manifest itself in actions [18]. Consequently, behavioral intention refers to the probability of engaging in a specific type of action in a given situation. As this perspective is extended to tourism, the traveler's desire or probability to visit a destination is referred to as travel purpose [19]. The greater the intention of one to travel, the higher the possibility they will travel. It is also seen as the possibility, commitment, and motivation to travel to a specific tour destination [20]. Recent studies within the field of tourism that adopt TPB's viewpoint are as follows [21]. Travel intention is habitually based on the tourists' attitude and preference toward a particular tour destination [22]. Rational and practical factors influence tourist behavior, especially attitude and preference. To put it another way, functional and psychological factors frequently affect one's desire to travel, which leads to actual travel behavior. The psychological variables were linked to the tourists' emotions, marked by extreme feelings episodes [23], whereas the functional variables were found to be related to the destination's particular qualities and atmosphere [23].

B. User-generated-content trust

The word "travel 4.0" is a modern idea that emphasizes the growing role of social media in travel and new customers in the hospitality and tourism industries [24, 25]. The growth of user-generated content is affecting travel consumer decisions. Gretzel, et al. [26] stated that almost nearly half of travel buyers said they used user-generated content in their trip planning, and nearly a third said they found the helpful knowledge. Since they can easily access travel and tourism via social media and rate tourism products and services on online online platforms, these consumers are better educated.[26]. Most readers believe that travel reviews are more likely than travel service providers to provide up-to-date, pleasant, and accurate information [25]. Peer reviews are perceived as superior by frequent travelers, who are more likely to be favorably affected. According to the above, more than half of users check online feedback while planning a vacation. The value of feedback for the accommodation product is extreme, whereas the significance for other travel products is much lower [27]. Reviews are particularly important for the accommodation product, with relevance for other travel products much smaller. Hence, the hypothesis H1 was proposed:

H1: User-generated content trust has a positive impact on the inbound travel intention Of Z Travelers

C. Economic perspective

In the last two years, the Covid-19 pandemic has made a significant impact on the global economy [28]. he Covid-19 pandemic has had a huge effect on the global economy over the past two years. The person does not benifit from the travel time (whether it be pleasure or conducting work-related tasks to reduce work time infringement on personal time). As a result, we can use the concept of travel time as a gift to refer to network involvement. In this case, the economic imperative of reducing travel time will not always be true (or be so fully applicable). People can take longer routes or agree longer travel times in some cases so that they have enough time to listen to music, rest, or think about a job problem [29]. he Covid-19 pandemic has had a huge effect on the global economy over the past two years. The person does not benifit from the travel time (whether it be pleasure or conducting work-related tasks to reduce work time infringement on personal time). As a result, we can use the concept of travel time as a gift to refer to network involvement. In this case, the economic imperative of reducing travel time will not always be true (or be so fully applicable). People can take longer routes or agree longer travel times in some cases so that they have enough time to listen to music, rest, or think about a job problem [30]. The roots of backpacker travel must be understood in the context of tourism's more enormous past. Examining historical practices such as the Grand Tours of 17th and 18th century Europe, tramping and the youth hostel movement, the idea of non-institutionalized tourism, and the meanings of terms like "drifter" tourism, youth tourism, and alternative tourism are all helpful. The travel behavior of the wealthy, well-educated youth of the late Victorian century, who set out on adventure trips to explore the secret, strange, and exotic life of faraway countries and unknown people, has impacted new backpacking. These eager explorers also readily welcomed extreme difficulties and even embraced their hosts' way of life [31]. Therefore, this study proposed the hypothesis H2 as:

H2: Economic perspective has a positive impact on user-generated content trust Of Z Travelers

D. Social perspective

Social space is physical surroundings or location, denoted social space, which can be considered at many levels. From a travel perspective, one is interested in the main life venues that individuals connect through travel. The implications of telecommuting for travel have received considerable attention [32]. As Fischer [33] hypothesized, "limitation on the potential social relations available to individuals leads to fewer communal social relations". In the age of information being transmitted in less than a second, any sentence or action becomes a trend. Especially for young people, generation Z, who like free life, exploring, and traveling is an overall experience. The most used tools are the Internet and social networks. Social networking networks are now online applications that allow individuals to socialize by sharing digital content [34]. The majority of major social media firms quickly established advertisement channels after their initial launches. According to academics and industry experts, social networking has become an important advertisement venue for advertisers [7]. Social networking networks are now online applications that allow individuals to socialize by sharing digital content. The majority of major social media firms quickly established advertisement channels after their initial launches. Social networking has become a vital advertisement venue for advertisers, according to academics and industry experts. Consequently, it is necessary to understand better the relationship between the Z generation's social contact and travel intention.

H3: Social perspective has a positive impact on user-generated content trust Of Z Travelers

III. RESEARCH METHOD

This research was conducted based on a mixed research method that combines qualitative research methods and quantitative research to increase its credibility and significance [35]. In the qualitative research method, a focus group discussion was conducted with the participation of 8 experts, including university lecturers, tourism company managers, who are the expert in research fields. Qualitative research results adjusted the scales' content for adoption in a tourism survey; simultaneously, it confirmed the research constructs were appropriate for evaluating perceived usefulness, attitude, and inbound tourism intention. All 12 items measuring six research constructs were accepted. Questionnaires for quantitative research were also established through qualitative research. Next, the study uses quantitative research methods to test the proposed model and research hypotheses. Table 1 showed the measurement scales in this study.

The study used a non-probabilistic purposive sampling method. Participants were travelers belong to Generation Z, who like to travel inbound. The total number of the collected respondent survey was 578 through the online questionnaire. After the initial cleaning phase, 12 imperfect copies were eliminated, so the amount of survey data officially used for quantitative analysis was 566. About the gender, 315 participants were male (55.65%), there were 251 female respondents (44.35%). Of 566 participants, 214 people belong from 15 to 17 years old, accounting for 37.81%; the others are from 18 to 23 years old, accounting for 62.19%.

The data is processed by Smart PLS 3.7.8 software to analyze the Partial Least Squares Structure Equation Model (PLS-SEM). The procedures to be performed evaluated the measurement model (reliability, validity) and the PLS-SEM (R^2,f^2, Q^2, VIF, path coefficient). Thereby, the study would discuss research results and propose managerial implications to enhance Generation Z's inbound travel intention.

TABLE I. MEASUREMENT SCALE

Item	***Source***
Economic perspective (*hereafter* ECP)	
I could have got the functional values (have an enjoyable trip) from the online user-generated contents (ECP1)	Chu, et al. [36]
I could have got the monetary benefits (save the money because of booking a cheaper hotel) from the online user-generated contents (ECP2)	
I could have got the sales promotions (have a discount for the room) from the online user-generated contents (ECP3)	
Social perspective (*hereafter* SOP)	
I could have got the relational benefits (have a new partner for the trips) from the online user-generated contents (SOP1)	Chu, et al. [36]
I could have got the social values (have an agreement from the others) from the online user-generated contents (SOP2)	
I could have got the emotional values (have enjoyment from reading the news) from the online user-generated contents (SOP3)	
User-generated Content Trust (*hereafter* UGCT)	
I am reliable to the online user-generated contents as traveling (UGCT1)	Chari, et al. [37]
I am affected (enjoyable/likable) in the online user-generated contents as traveling (UGCT2)	
I have the willingness to rely on the online user-generated contents as traveling (UGCT3)	
Inbound Travel Intention (*hereafter* IBTI)	
I will save time and money within 12 months for inbound traveling (IBTI1)	Park, et al. [38]
I will have at least an inbound travel with friends/family within 12 months (IBTI2)	
An inbound journey is my first choice for traveling in the future (IBTI3)	

IV. RESULT

Based on the Partial Least Squares (PLS) algorithm results, the evaluation criteria' values are calculated. Research results in Table 2 showed that the scales have reliable because Cronbach's Alpha coefficients (CA) were more significant than 0.7. Simultaneously, the outer loading coefficients were more significant than 0.708, Composite Reliability (CR) was higher than 0.7, and the Average Variance Extracted (AVE) was more significant than 0.5. Therefore, the proposed research constructs' measurement scale achieves convergent validity.

TABLE II. THE RELIABILITY AND CONVERGENT VALIDITY

Constructs	***CA***	***Outer loading***	***CR***	***AVE***
ECP	0.856	[0.746 - 0.897]	0.903	0.701
SOP	0.882	[0.949 - 0.959]	0.919	0.739
UGCT	0.836	[0.787 - 0.868]	0.913	0.701
IBTI	0.878	[0.791 - 0.871]	0.912	0.674

Besides, the study tested the discriminant validity between the research constructs, which was evaluated via Heterotrait – Monotrait (HTMT). The HTMT threshold is 0.85. The results of Table 3 showed that the enormous HTMT value was 0.735. Consequently, the constructs in the study gain discriminant validity.

TABLE III. DISCRIMINANT VALIDITY

	ECP	***SOP***	***UGCT***
SOP	0.453		
UGCT	0.612	0.276	
IBTI	0.735	0.245	0.513

Table 4 shows the Variance inflation factor (VIF) of all the independent variables less than 2 (the largest VIF is 1.27). Therefore, the research model does not have the phenomenon of multicollinearity occurred. Besides, the coefficient of R^2 of the IBTI was 0.61, which means that UGCT explained 61% of the change in IBTI. Additionally, R^2 of UGCT was 0.553; hence, 55.3% of the change in UGCT depends on the ECP and SOP. Moreover, the values of the effect coefficient f^2 of the independent variables for the dependent variable are all greater than 0.02, of which, $f^2_{SOP->UGCT} = 0.15$, so SOP had a moderate impact on the UGCT; $f^2_{ECP->UGCT} = 0.627$, so ECP had an enormous impact on the UGCT; $f^2_{UGCT->IBTI} = 1.564$, so UGCT had an enormous impact on the IBTI. The prediction coefficient Q^2 of the theoretical model is also valid at 0.474 and 0.486, which were greater than 0.

TABLE IV. THE RESULT OF THE PLS-SEM EVALUATION

	VIF		***f^2***		***R^2***	***Q^2***
	IBTI	UGCT	IBTI	UGCT		
ECP		1.27		0.627		
IBTI					0.61	0.474
SOP		1.27		0.15		
UGCT	1.00		1.564		0.553	0.486

The PLS-SEM model's analysis results in Table 5 also showed that all the path coefficients between the independent and dependent variables are significant, with 99% confidence. The results showed that all three exogenous variables have original weights greater than 0. According to the H1 hypothesis, UGCT had a positive impact on IBTI ($\beta = 0.781$; p-value < 0.001); therefore, hypothesis H1 is supported. Meanwhile, H2 is also supported when the results show a positive ECP effect on UGCT ($\beta = 0.594$; p-value < 0.001). Similar to hypothesis H3, showing positive effects of SOP on UGCT ($\beta = 0.253$; $p < 0.001$); therefore, H3 is supported.

TABLE V. THE PATH COEFFICIENT IN PLS- SEM

	Beta	***Standard Deviation***	***P Values***	***H***	***Result***
UGCT -> IBTI	0.781	0.031	0.000	H1	Accepted
ECP -> UGCT	0.594	0.056	0.000	H2	Accepted
SOP -> UGCT	0.253	0.055	0.000	H3	Accepted

V. DISCUSSION

Research results have shown a positive relationship between the factors in the research model. The research results have inherited the previous studies and have applied to the tourism industry in Vietnam.

Firstly, user-generated content trust has a positive impact on inbound travel intention. Kartajaya, et al. [39] pointed out that it usually goes through the information search phase before a customer decides. Therefore, the content shared by others has an important role and affects customer travel intention. It is difficult for young travelers, who do not have much money or time, to have a good trip if they have no authentic information source. These sharing contents can make a positive impact on customer decisions for choosing destinations. The exciting travel destinations or special foods may be the most considerable concern of Z travelers; hence, the reliable user-generated content in social media becomes the right motivation for them to decide to travel.

Secondly, the result pointed out that the economic perspective positive impact on the user-generated content trust. The economic perspective of user-generated content in travel includes functional values, monetary benefits, and sales promotions. Many Z travelers have based on the recommended websites or social network pages to evaluate or choose a beautiful destination; consequently, they have exciting or interesting trips with their friends or family. The emergence of websites offering travel services has added a channel of communication between visitors. The information exchange of room's or service's price helps travelers save money when choosing accommodations with fair prices. During the fierce competition between businesses, the dissemination of information about sale promotions programs will be a valuable weapon for businesses to create a competitive advantage. Information spreading via social networking is one solution that makes electronic word of mouth happen faster. Customers, who regularly access and follow travel-related content, will receive promotions created by service providers.

In addition to the effect of economic values on trust, social values also impact Generation Z travelers' trust when referring to user-generated content pages before and while traveling. The social perspective includes the relational benefits, social value, emotional values. The sharing of travel information on social media can receive the thousand comment or shared by people who love traveling. It can connect the strangers and satisfy the belonging needs as the Maslow's Needs Hierarchy. Young travelers can even initiate direct conversations with others after reading information about a destination written by people who have been there.

Moreover, making accurate assessments of place information also creates respect and admiration from other people. The travel community's high value to the dedicated blog owner is also an essential recognition of social value. Besides, for Generation Z who are of school age and just graduated from university, the amount of time spent traveling is not much; therefore, they can temporarily indulge themselves through reading posts written by people with real travel experience. These are posts full of pictures of tourist destinations and accurate information about services during travel, which will create confidence for travel enthusiasts. Although not comparable to authentic travel, however, this approach also creates excitement for young travelers and can be seen as a motivation for them to motivate them to take their trips soon.

VI. Conclusion

The Covid-19 pandemic has limited, even greatly affected the tourism industry's development; however, human tourism's demand has not decreased. A bright outlook awaits the tourism industry in the post-Covid-19 period. Businesses need to understand the visitors' behavior in choosing a destination or factors that influence the tourist's inbound tourism intentions to take these advantages. Hence, this study aimed to determine the relationship between the economic perspective, social perspective, user-generated content trust, and inbound travel intention of Generation Z travelers. The research results have some theoretical contributions, and some managerial implications were proposed for the tourism business managers to increase the business performance.

About the theory, there are many studies based on the TAM model in the tourism industry; however, this study is based on the TAM model and extended with the economic and social perspective to forecast the traveler intention. This research result pointed that the economic and social perspectives are the perceived usefulness in the user-generated content of tourism, and they have a positive impact on the user-generated content trust. Moreover, the measurement scales have high reliability and validity; they can be used for further studies.

Based on the research results, this study proposed some managerial implications for the tourism company's managers to enhance the inbound travel intention via user-generated content. With the vigorous development of the industrial 4.0 era, almost everyone uses social networks. Due to the pandemic COVID-19, people are restricted from going out, especially overseas travel. Therefore, inbound travel has a chance to develop. Social networking is a place to give opportunities to reach people who intend to travel based on finding tourist destinations, finding hotels, homestays. Identify the Generation Z people's target customers to help businesses identify the ideal customers, matching what businesses can offer. Stimulate travelers' excitement with beautiful landscape images, videos of trips, delicious food, scenic spots, or historic sites. Create groups on the social network to categorize travelers with similar interests to a particular tourist destination so that travelers can share experiences, photos, and videos of their trips. Create a fan page on the social network that provides information about tourist destinations, travel knowledge, advertises promotions of tours, and promotions of hotels, restaurants. Improve the tourism industry's service quality, make commitments to tourists, and reduce potential tourists' risks. Promote local brands, highlight unique local features to attract tourists. Invest in beautiful and quality images and videos to introduce exciting and unique tourist destinations to customers.

There are some limitations in this study, and they will be opportunities for further research. However, the research based on the TAM model focused on the relationship between usefulness, attitude, and behavioral intention, even though the relationship between economic, social, and inbound travel has not been exploited. The group cohort difference, i.e., Generation X, Y, X, will affect the relationship between the research model's constructs; however, this study did not reach this issue. Further studies could add more factors in the conceptual model as the perceived ease of use or external factors to fulfill the theory. Furthermore, the comparison between the group cohort should clearly understand the traveler behavior in all age groups.

References

[1] B. Pan, T. MacLaurin, and J. C. Crotts, "Travel blogs and the implications for destination marketing," *Journal of travel research,* vol. 46, no. 1, pp. 35-45, 2007.

[2] X. Xu and S. Pratt, "Social media influencers as endorsers to promote travel destinations: an application of self-congruence theory to the Chinese Generation Y," *Journal of travel & tourism marketing,* vol. 35, no. 7, pp. 958-972, 2018.

[3] I. P. Tussyadiah and D. R. Fesenmaier, "Mediating tourist experiences: Access to places via shared videos," *Annals of tourism research,* vol. 36, no. 1, pp. 24-40, 2009.

[4] G. W.-H. Tan, V.-H. Lee, J.-J. Hew, K.-B. Ooi, and L.-W. Wong, "The interactive mobile social media advertising: an imminent approach to advertise tourism products and services?," *Telematics and Informatics,* vol. 35, no. 8, pp. 2270-2288, 2018.

[5] B. T. Khoa and H. M. Nguyen, "Electronic Loyalty In Social Commerce: Scale Development and Validation," *Gadjah Mada International Journal of Business,* vol. 22, no. 3, pp. 275-299, 2020.doi:10.22146/gamaijb.50683

[6] H. A. Voorveld, G. Van Noort, D. G. Muntinga, and F. Bronner, "Engagement with social media and social media advertising: The differentiating role of platform type," *Journal of advertising,* vol. 47, no. 1, pp. 38-54, 2018.

[7] A. Saxena and U. Khanna, "Advertising on social network sites: A structural equation modelling approach," *Vision,* vol. 17, no. 1, pp. 17-25, 2013.

[8] B. T. Khoa, T. D. Nguyen, and V. T.-T. Nguyen, "Factors affecting Customer Relationship and the Repurchase Intention of Designed Fashion Products," *Journal of Distribution Science,* vol. 18, no. 2, pp. 198-204, 2020.

[9] A. Garg, "Travel risks vs tourist decision making: A tourist perspective," *International Journal of Hospitality & Tourism Systems,* vol. 8, no. 1, pp. 1-9, 2015.

[10] N. A. A. Jalil, A. Fikry, and A. Zainuddin, "The Impact of Store Atmospherics, Perceived Value, and Customer Satisfaction on Behavioural Intention," *Procedia Economics and Finance,* vol. 37, pp. 538-544, 2016.

[11] T. Lam and C. H. Hsu, "Predicting behavioral intention of choosing a travel destination," *Tourism management,* vol. 27, no. 4, pp. 589-599, 2006.

[12] R. Wu and J.-H. Lee, "The use intention of mobile travel apps by Korea-visiting Chinese tourists," *The Journal of Distribution Science,* vol. 15, no. 5, pp. 53-64, 2017.

[13] B. T. Khoa, "The role of Mobile Skillfulness and User Innovation toward Electronic Wallet Acceptance in the Digital Transformation Era," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, Bandung - Padang, Indonesia, 2020, pp. 30-37: IEEE, 2020.

[14] M. Dimock, "Defining generations: Where Millennials end and Generation Z begins," *Pew Research Center,* vol. 17, pp. 1-7, 2019.

[15] F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," Doctor of Philosophy, Sloan School of Management, Massachusetts Institute of Technology, USA, 1986.

[16] B. T. Khoa, L. T. Huynh, and M. H. Nguyen, "The Relationship between Perceived Value and Peer Engagement in Sharing Economy: A Case Study of Ridesharing Services," *Journal of System and Management Sciences,* vol. 10, no. 4, pp. 149-172, 2020.doi:10.33168/JSMS.2020.0210

[17] A. Senders, R. Govers, and B. Neuts, "Social media affecting tour operators' customer loyalty," *Journal of Travel & Tourism Marketing,* vol. 30, no. 1-2, pp. 41-57, 2013.

[18] D. Albarracin, B. T. Johnson, M. Fishbein, and P. A. Muellerleile, "Theories of reasoned action and planned behavior as models of condom use: a meta-analysis," *Psychological bulletin,* vol. 127, no. 1, p. 142, 2001.

[19] Y.-C. Chen, R.-A. Shang, and M.-J. Li, "The effects of perceived relevance of travel blogs' content on the behavioral intention to visit a tourist destination," *Computers in Human Behavior,* vol. 30, pp. 787-799, 2014.

[20] T. Ahn, Y. Ekinci, and G. Li, "Self-congruence, functional congruence, and destination choice," *Journal of Business Research,* vol. 66, no. 6, pp. 719-723, 2013.

[21] M. T. Al Ziadat, "Applications of planned behavior theory (TPB) in Jordanian tourism," *International Journal of Marketing Studies,* vol. 7, no. 3, p. 95, 2015.

[22] H.-B. Chen, S.-S. Yeh, and T.-C. Huan, "Nostalgic emotion, experiential value, brand image, and consumption intentions of customers of nostalgic-themed restaurants," *Journal of Business Research,* vol. 67, no. 3, pp. 354-360, 2014.

[23] U. R. Orth, Y. Limon, and G. Rose, "Store-evoked affect, personalities, and consumer emotional attachments to brands," *Journal of Business Research,* vol. 63, no. 11, pp. 1202-1208, 2010.

[24] M. Mariani, "Big data and analytics in tourism and hospitality: a perspective article," *Tourism Review,* 2019.

[25] B. T. Khoa and T. Khanh, "The Impact of Electronic Word-Of-Mouth on Admission Intention to Private University," *Test Engineering and Management,* vol. 83, no. (May -June 2020), pp. 14956-14970, 2020.

[26] U. Gretzel, K. H. Yoo, and M. Purifoy, "Online travel review study: Role and impact of online travel reviews," 2007.

[27] B. T. Khoa, "Electronic Loyalty in the Relationship between Consumer Habits, Groupon Website Reputation, and Online Trust: A Case of the Groupon Transaction," *Journal of Theoretical and Applied Information Technology,* vol. 98, no. 24, pp. 3947-3960, 2020.

[28] B. T. Khoa, "The Perceived Enjoyment of the Online Courses in Digital Transformation Age: The Uses - Gratification Theory Approach," presented at the 2020 Sixth International Conference on e-Learning (econf), Sakheer, Bahrain, 6-7 Dec. 2020, 2020. doi:10.1109/econf51404.2020.9385490

[29] L. S. Redmond and P. L. Mokhtarian, "The positive utility of the commute: modeling ideal commute time and relative desired commute amount," *Transportation,* vol. 28, no. 2, pp. 179-205, 2001.

[30] R. J. Hjorthol, L. Levin, and A. Sirén, "Mobility in different generations of older persons: The development of daily travel in different cohorts in Denmark, Norway and Sweden," *Journal of Transport Geography,* vol. 18, no. 5, pp. 624-633, 2010.

[31] E. Cohen, "Nomads from Affluence: Notes on the Phenomenon of Drifter-Tourism1," *International Journal of Comparative Sociology,* vol. 14, no. 1-2, pp. 89-103, 1973.

[32] P. L. Mokhtarian, "A synthetic approach to estimating the impacts of telecommuting on travel," *Urban studies,* vol. 35, no. 2, pp. 215-241, 1998.

[33] C. F. Fischer, "Hartree--Fock method for atoms. A numerical approach," 1977.

[34] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business horizons,* vol. 53, no. 1, pp. 59-68, 2010.

[35] Z. Dörnyei, *Research methods in applied linguistics : quantitative, qualitative, and mixed methodologies* (Oxford applied linguistics). Oxford ; New York, N.Y.: Oxford University Press, 2018.

[36] S.-C. Chu, T. Deng, and H. Cheng, "The role of social media advertising in hospitality, tourism and travel: a literature review and research agenda," *International Journal of Contemporary Hospitality Management,* 2020.

[37] S. Chari, G. Christodoulides, C. Presi, J. Wenhold, and J. P. Casaletto, "Consumer trust in user ‐ generated brand recommendations on Facebook," *Psychology & Marketing,* vol. 33, no. 12, pp. 1071-1081, 2016.

[38] S. H. Park, C.-M. Hsieh, and C.-K. Lee, "Examining Chinese college students' intention to travel to Japan using the extended theory of planned behavior: Testing destination image and the mediating role of travel constraints," *Journal of Travel & Tourism Marketing,* vol. 34, no. 1, pp. 113-131, 2017.

[39] H. Kartajaya, P. Kotler, and D. H. Hooi, *Marketing 4.0: Moving From Traditional To Digital* (World Scientific Book). Hoboken, NJ: John Wiley & Sons, 2019, pp. 99-123.

Segment-Routing Analysis: Proof-of-Concept Emulation in IPv4 and IPv6 Service Provider Infrastructures

Gustavo D. Salazar-Chacón
Facultad de Ingeniería
Pontificia Universidad Católica del Ecuador – PUCE
Quito, Ecuador
gsalazar787@puce.edu.ec

Andy R. Reinoso García
Facultad de Ingeniería
Pontificia Universidad Católica del Ecuador - PUCE
Quito, Ecuador
arreinoso@puce.edu.ec

***Abstract*—With the complexity and overload of IP connectivity and the variety of protocols that each technology and applications has been introducing to keep the human beings connected to the modern Internet, especially in times of COVID-19 pandemic, it is a priority for Service Providers to seek a simplification of their infrastructures to make them more scalable, efficient, simplified, and profitable. In this Proof-of-Concept (PoC) paper, Segment Routing (SR) is presented as an alternative to interconnect business networks in IPv4 and IPv6 scenarios where Headquarter-Branches connectivity is required. This paper begins by defining SR and its fundamentals. To test theory in practice, three scenarios are emulated and compared: MPLS-LDP, SR for IPv4 infrastructures, and finally, SR for IPv6 networks. The aim of the present paper is to provide a more understandable vision of Segment Routing and its operation in ISP networks.**

Keywords—Segment-Routing, MPLS-L3VPN, LDP, SRv6, IS-IS

I. Introduction

Due to the appearance of new communication protocols and technologies such as 5G, IoT [1], telepresence, Artificial Intelligence, Big Data requirements, microservices, cybersecurity concerns and Software-Defined Networking, the processing carried out by All-IP dual-stack WAN Infrastructures is becoming overloaded [2], thus not allowing to scale and being resilient in the near future. Therefore, a new ISP architecture easier to manage, operate and with automation and programmability in mind is required in the Digital Transformation era.

The use of MPLS in the ISP side has covered current market needs so far, but, with the exponential growth of information handling and processing, this business and technological model may be threatened by not being able to satisfy new information transmission requirements, which translates into non-profitable Service Providers (SP) [3].

It is, at this point, essential to describe the characteristics of Segment Routing protocol as it is conceived to be developed in an IPv6 environment with IPv4 support, and due to the simplicity of its operation, is in fact a protocol to consider, especially for ISPs that need highly qualified personnel in managing MPLS architectures.

The emulation of a Service Provider network with Segment Routing is important, because this kind of research will allow to analyze and evaluate SR as a new MPLS proposal in order to determine the behavior of this sort of network, establishing its advantages and constrains of the so called “MPLS evolution”.

Once the comparison between MPLS-LDP, SR for IPv4 and SRv6 has been made, it will be shown the improvements that SR introduces, and of course, its feasibility of implementation in the real world, thereby establishing a reference to SP operators that guarantees to carry out a gradual migration with savings in operational and economic costs (CAPEX and OPEX).

The development of this work deals with the following topics:

First, the current situation and state of the art in WAN technologies is presented, describing the evolution in terms of usage of the preferred SP infrastructures, their challenges in production environments and their evolutions with Segment Routing and SD-WAN.

Later, an analysis of Segment Routing protocol is presented as a proposal towards the simplification of a network that is going to be easier to manage and operate, presenting its advantages and disadvantages compared to the current MPLS-LDP paradigm.

Finally, the emulation of a network in an SP environment is performed to interconnect two remote sites with three different scenarios: MPLS L3VPN, SR-MPLS and SRv6. Testbeds and comparisons of these scenarios have been made to determine the feasibility of SR implementation with its different data planes.

The work finishes with the most relevant conclusions and recommendations, which are extracted from the development of this investigation.

II. State of the Art in ISP Networks

The evolution of data transport technologies has its beginnings in the development of Computing in the 60's, along with the foundations of Information Theory that have had great contributions by Claude Shannon, Harry Nyquist and Ralph Hartley at the early 20th Century [4].

From there, a series of continuous evolution in data transport technologies happened, being MPLS the turning point [5].

Multiprotocol Label Switching (MPLS) is one of the most implemented technologies used by Service Providers to send packets, mainly because of its high-performance in forwarding data from one router to another based on a set of labels instead of doing it through IP network addresses.

Label distribution is carried out thanks to LDP (Label Distribution Protocol), where LSRs (Label Switching Routers) share information in order to reach other LSRs. With

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

RSVP (Resource Reservation Protocol), resources can be reserved in the entire route of any LSP (Label Switching Path) for a specific traffic flow, this is known as MPLS-TE (Traffic Engineering).

On the other hand, unlike traditional MPLS and the complications required to operate it, Segment Routing allows to implement IP routing in a flexible and scalable manner, so the origin router chooses a route and inserts it as an ordered list of segments, not depending on LDP or RSVP-TE signaling, hence the fact that it uses segments to send in the origin.

SR can operate on an MPLS fabric or IPv6 data plane, together with its own services such as L3VPN.

The implementation of SR in current providers depends on having an adequate platform in hardware and software to allow the evolution of the network to operate with this type of packet forwarding.

Segment Routing reinforces the idea that an explicit route is an ordered set of instructions put into the packet, with routers executing these instructions as they are sent. Each instruction is called a segment, and has its own number called Segment ID (SID), encoded as a stack of labels in MPLS, with each label representing a particular segment, so the MPLS label values can carry individual segment IDs inside the Segment Routing Domain, as seen in Fig. 1

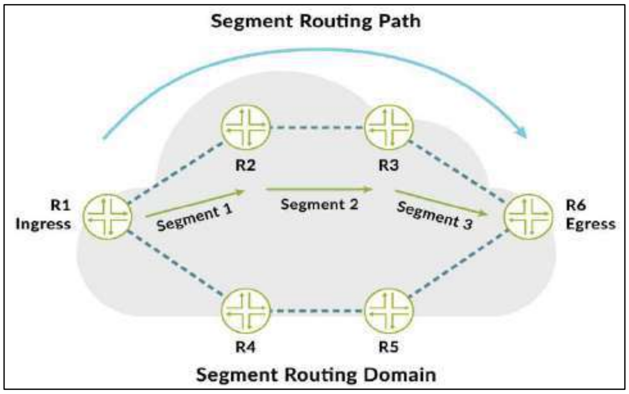


Fig. 1. Segment Routing Domain [6]

In the next section, Segment Routing operation is explained in greater detail since it is the main topic of this work.

SD-WAN bases on incorporating SDN paradigm and NFV (Network Function Virtualization) technology to a WAN network [7]; in other words, the process of creating WAN tunnels on fast and reliable manner from an API (Application Programming Interface), so the WAN network has a standardized, flexible, and scalable architecture [13].

Characteristics of SD-WAN environments are:

- Policy management: Configure and manage parameters such as QoS [8], security, access, among other policies like PBR (Policy-Based Routing [20]).
- Network by application: Where virtual networks are defined regardless of the physical part, and therefore policies can be applied for each virtual network.
- Dynamic assignment of services: Adding services in traffic flows for a given time, such as redirecting traffic to an element in the network or to a data center.
- Dual uplink: Allows to use the active and backup link according to the services or traffic flows.
- Hybrid Cloud Approach: Allows extending IT capabilities to a data center in a dynamic, simple and transparent way.
- Auto request/self-management: Through a website or Front-End API, the client can execute or request policies, changes, insights or reports.
- Monitoring and reports: As indicated, through the customer portal, information can be obtained in real time about the configuration, network topology, traffic patterns, troubleshooting, and whatever the customer deems necessary (great insights).

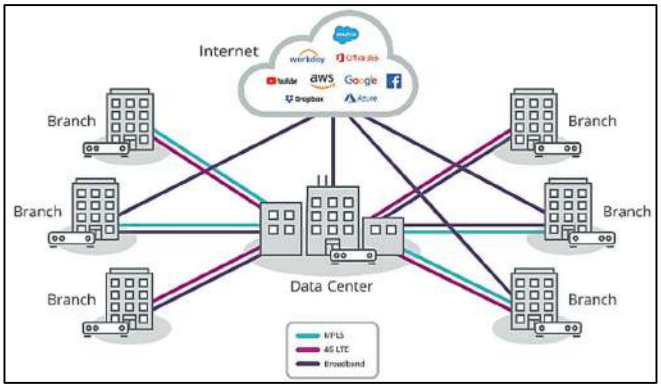


Fig. 2. SD-WAN Architecture [9]

III. Segment Routing Fundamentals

The SR's initial idea was proposed by Cisco Systems engineer Clarence Filsfils in November 2012 and IETF who formed the SPRING (Source Packet Routing in Networking) Working Group in October 2013. RFC 8402, known as Segment Routing Architecture, was developed in July 2018.

Being Segment Routing the latest technology in changing the way packets are handled in network infrastructures like Internet cores, DCs or between DCs, this protocol is a simpler way to forward packets, control networks, schedule routing, packet processing, and to implement traffic engineering policies, using the concept of source routing and bases its implementation in forwarding data plane on top of two technologies: MPLS and IPv6 as its Data Planes.

The MPLS data plane allows carry multiple labels, acting as Segment Identifiers (SIDs), IPv6 also does the same through the Segment Routing Header (SRH), carrying multiple segments as well. The MPLS data plane places the next segment represented by the tag closest to the MAC header and removes this tag after using it to determine the next hop, so the current tag and the next one will always be in the same relative place to each other.

On the other hand, SRv6 carries the current segment as a destination IPv6 address, with the rest of the segments in the SRH, the last segment being the closest to the MAC header, and the next segment in the stack is the furthest from the MAC header; since these locations vary, a pointer is used that indicates the location of the next segment, in such a way that no segment is removed from the stack, which means a great difference in speed in relation to data plane processing. This approach is represented in Fig. 3

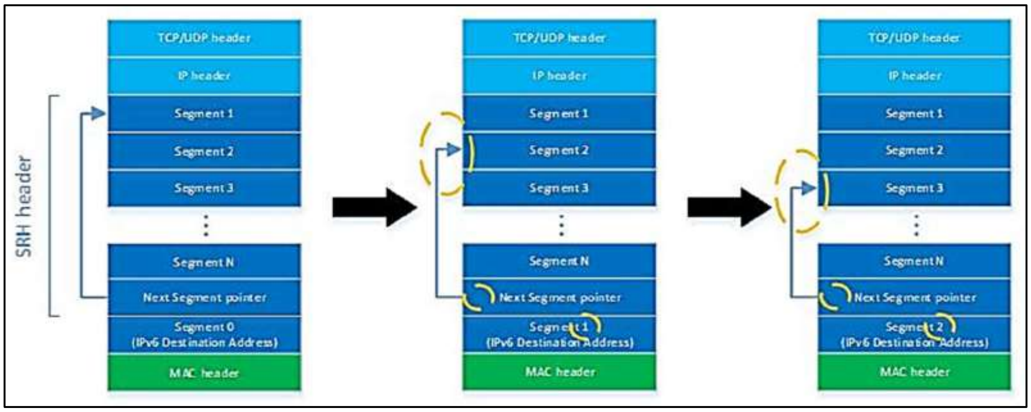


Fig. 3. SR Stack – SR Header [10]

In Segment Routing there are two main classes of segments:

- Global Segment: It is an SID-value valid throughout the entire SR domain, in such a way that each node knows this value and assigns the same action for the associated instruction in its LFIB (Label Forwarding Information Base). The range used is <16000-23999>. It is known as Segment Routing Global Block (SRGB) and this range is specific to each manufacturer.
- Local Segment: it is an SID-value that has local meaning, and only the source router can execute the associated instruction. These values are not in the SRGB range, but there are in the locally configured label range.

In a SR Topology, the LFIB (Label Forwarding Information Base) table is distributed by the Link-States IGPs like IS-IS and OSPF.

One of the differences between SR and MPLS-LDP is that the forwarding table (nodes + adjacencies) remains constant despite the number of routes in a Full-Mesh ISP topology.

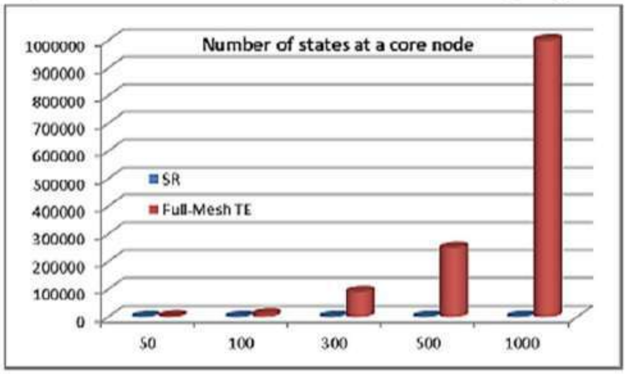


Fig. 4. SR vs MPLS-TE routing states [11]

IV. Proof-of-Concept Emulation of L3VPN in MPLS, Segment Routing amd SRv6 Networks

In this section, an emulation of a network architecture to communicate two sites by a Service Provider is presented. Three scenarios are shown to compare the feasibility and characteristics of Segment Routing IPv6 vs. MPLS-LDP.

A. MPLS L3VPN

Nowadays, it is the main model and framework deployed by SPs to establish communication between two or more client sites, but in production environments, there are different business clients managed by a single node thanks to Virtual Routing Forwarding (VRF) technology. This model consists of a client network formed by Customer Edge (CE) devices that are connected to a Service Provider through BGP routing protocol, and the Provider network formed by Provider Edge (PE) and Provider (P) devices using Level 2 Intermediate System-to-Intermediate System (IS-IS) protocol as IGP to exchange global routes, and Multiprotocol Label Switching (MPLS) to forward label-based packets. The Service Provider Cloud can forward Virtual Private Networks (VPN) traffic thanks to Multiprotocol-BGP (MP-BGP). This architecture permits PE takes part of the client routing with an optimum routing, and PE would process different routes for each client, so, it can address overlapping IPs [14].

The P devices do label switching and they do not know VPN routing at all, as well as the CE devices do not know the P device existence, so the provider network is transparent to the client.

To carry client routes from one PE to another across the MPLS domain, it is necessary to enable BGPv4 protocol through MP-BGP sessions, in which a 64-bits prefix called Route Distinguisher (RD), transforms a 32-bits address into a unique 96-bits address known as VPNv4/VPNv6 address-family. The RD are configured inside a VRF on all PEs. Route Targets (RT) permit that a VRF can participate in more than one VPN; RTs are announced as MP-BGP extended communities with the use of **import RT/export RT** commands.

To identify a remote client, the egress PE generates the 4-bytes VPN label through MP-BGP to the ingress PE which is used as an intern label at the MPLS process. The VPN label indicates to the last MPLS equipment the specific destiny of a packet for a particular VPN. This process can be observed in Fig. 5.

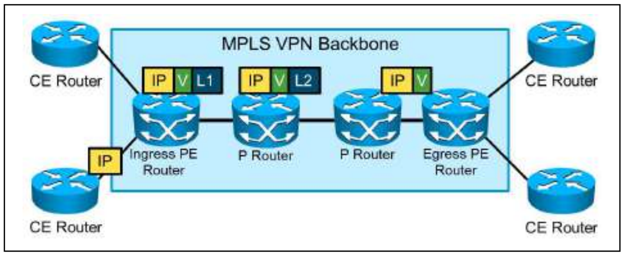


Fig. 5. VPN label operation

1) L3VPN over MPLS Architecture

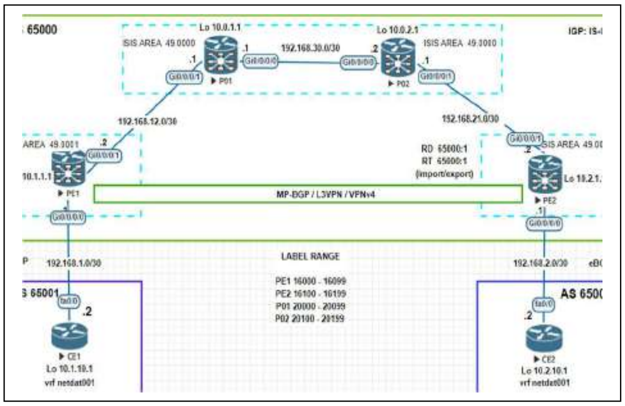


Fig. 6. MPLS L3VPN Architecture

In Fig. 6, there are two sites to be connected and one ISP Cloud acting as Transit network with the following Autonomous System (AS) numbers: 65000, 65001 and 65002, respectively. Sites are connected to the SP via eBGP. At the SP cloud there are two PEs and two Ps devices where L3VPN over MPLS is deployed. Every device has its own Loopback interface.

2) Configuring basic MPLS L3VPN

This emulation uses Cisco ASR9000 with 6.0.1 and 6.1.3 IOSXR releases and Cisco 3725 devices with 12.4(15) IOS release.

a) MPLS configuration: After configuring IS-IS, MPLS must be enable, in addition with an MPLS label range per each device (Fig. 7).

```
mpls ldp
 router-id 10.2.1.1
 interface GigabitEthernet0/0/0/1
mpls label range table 0 16100 16199
```

Fig. 7. PE1's MPLS configuration

b) eBGP configuration: This configuration lets connect a site with the Service Provider Cloud.

```
router bgp 65000
  vrf netdat001
  rd 65000:1
  address-family ipv4 unicast
  !
  neighbor 192.168.2.2
   remote-as 65002
   update-source GigabitEthernet0/0/0/0
   address-family ipv4 unicast
    route-policy PERMITE_TODO in
    route-policy PERMITE_TODO out
```

Fig. 8. PE1's eBGP configuration

c) MP-BGP configuration: Configuration to carry and exchange routes between PE devices. VPNv4 address family is used, as seen in Fig. 9

```
router bgp 65000
 address-family vpnv4 unicast
 !
 neighbor 10.1.1.1
  remote-as 65000
  update-source Loopback0
  address-family vpnv4 unicast
   next-hop-self
```

Fig. 9. PE1's MP-BGP configuration

d) VRF configuration: This step lets differentiate between more than one VRF that could exist in a PE device.

```
vrf netdat001
 address-family ipv4 unicast
  import route-target
   65000:1
  export route-target
   65000:1
interface GigabitEthernet0/0/0/0
  vrf netdat001
 ipv4 address 192.168.2.1 255.255.255.252
```

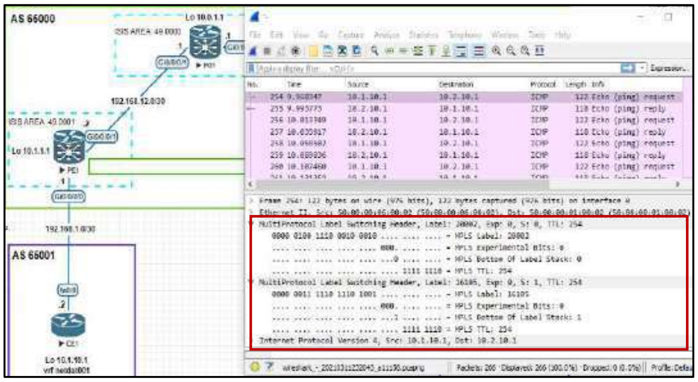
Fig. 10. PE1's VRF configuration

3) Verification of MPLS L3VPN: With the **show bgp vpnv4 unicast vrf <vrf_name>** command, prefixes that are members of the VPN are shown in Fig. 11.

```
RP/0/0/CPU0:PE1#sho bgp vpnv4 unicast vrf netdat001
Sat Feb 27 17:54:11.399 UTC
BGP router identifier 10.1.1.1, local AS number 65000

Status codes: s suppressed, d damped, h history, *
valid, > best
              i - internal, r RIB-failure, S stale, N
Nexthop-discard
Origin codes: i - IGP, e - EGP, ? - incomplete
   Network            Next Hop   Metric LocPrf Weight
Path
Route Distinguisher: 65000:1 (default for vrf netdat001)
*> 10.1.10.1/32       192.168.1.2    0      0 65001 i
*>i10.2.10.1/32       10.2.1.1       0 100  0 65002 i
```

Fig. 11. PE1's L3VPN verification

Finally, a ping can be executed between branch offices to confirm that a connectivity works (Fig.12).



```
No.   Time         Source      Destination   Protocol  Length  Info
  254 9.968347     10.1.10.1   10.2.10.1     ICMP      122 Echo (ping) request
  255 9.995773     10.2.10.1   10.1.10.1     ICMP      118 Echo (ping) reply
  256 10.013740    10.1.10.1   10.2.10.1     ICMP      122 Echo (ping) request
  257 10.030917    10.2.10.1   10.1.10.1     ICMP      118 Echo (ping) reply
  258 10.056987    10.1.10.1   10.2.10.1     ICMP      122 Echo (ping) request
  259 10.089036    10.2.10.1   10.1.10.1     ICMP      118 Echo (ping) reply
  260 10.107460    10.1.10.1   10.2.10.1     ICMP      122 Echo (ping) request

> Frame 254: 122 bytes on wire (976 bits), 122 bytes captured (976 bits) on interface 0
> Ethernet II, Src: 50:00:00:06:00:02 (50:00:00:06:00:02), Dst: 50:00:00:01:00:02 (50:00:00:01:00:02)
v MultiProtocol Label Switching Header, Label: 20002, Exp: 0, S: 0, TTL: 254
    0000 0100 1110 0010 0010 .... .... .... = MPLS Label: 20002
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...0 .... .... = MPLS Bottom Of Label Stack: 0
    .... .... .... .... .... .... 1111 1110 = MPLS TTL: 254
v MultiProtocol Label Switching Header, Label: 16105, Exp: 0, S: 1, TTL: 254
    0000 0011 1110 1110 1001 .... .... .... = MPLS Label: 16105
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...1 .... .... = MPLS Bottom Of Label Stack: 1
    .... .... .... .... .... .... 1111 1110 = MPLS TTL: 254
> Internet Protocol Version 4, Src: 10.1.10.1, Dst: 10.2.10.1

wireshark_-_20210311232043_a11150.pcapng    Packets: 266 · Displayed: 266 (100.0%) · Dropped: 0 (0.0%)    Profile: Default
```

Fig. 12. MPLS-LDP WireShark Traffic Capture

B. Segment Routing MPLS

It is called SR-MPLS because it uses MPLS data plane fabric as an "underlay" infrastructure, so its operation is the same as MPLS-LDP. It is important to known that a segment equals to a label and a list of segments equals to a label stack with the advantage that LFIB table remains constant. SR-MPLS does not use Label Distribution Protocol (LDP). By default, if a device has been configured with MPLS and SR, it is going to prefer MPLS, so SR must be selected in the configuration.

Nowadays, there are many SR use cases, and real-world implementations in large ISP, such as Bell Canada, in which Segment Routing meets all its next-generation network requirements and is adopted as part of its phased Network 3.0 transformation project according to [18] [21].

Fig. 15. L

1) SR-MPLS Architecture

The scenario presented in Fig. 13 is Segment Routing over MPLS Cloud. Inside ISP Cloud, IS-IS is the cloud routing protocol; IS-IS needs an extension to keep the SID database, as well as a VRF to differentiate traffic for a specific client, and MP-BGP tunnel to establish end-to-end communication between PE devices.

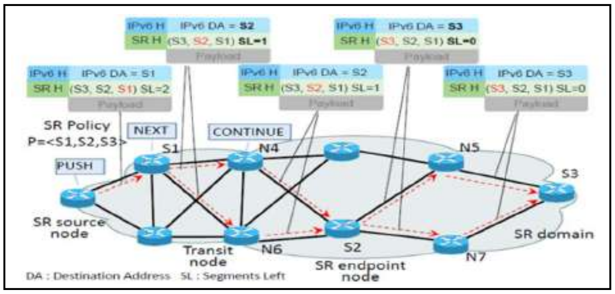


Fig. 13. L3VPN over SR-MPLS Architecture

2) Configuring SR MPLS

Because SR is built over MPLS, the operation of IS-IS, as well as the exchange of VPN through MP-BGP, are the same as in MPLS-LDP.

a) SR MPLS configuration: It is set under **router isis** instance on all SP devices, and a prefix-sid which starts at 16000, must be configured.

```
router isis 1
 net 49.0001.0100.0100.1001.00
 address-family ipv4 unicast
  metric-style wide
  segment-routing mpls
 !
 interface Loopback0
  address-family ipv4 unicast
   prefix-sid index 0
```

Fig. 14. PE1's SR MPLS configuration

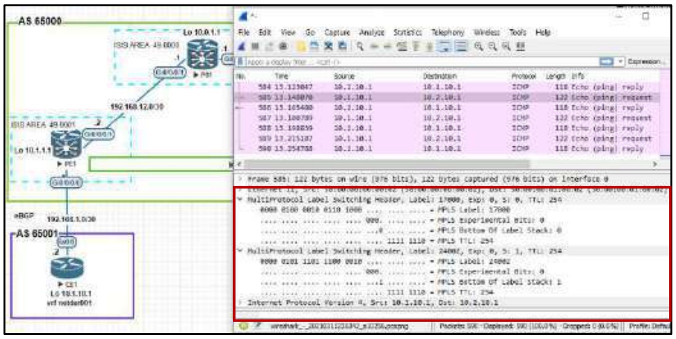
3) Verification of SR MPLS

With **show mpls interface detail** and **show cef** commands, can see the no use of LDP [15].

```
RP/0/0/CPU0:PE1#sh mpls int giga 0/0/0/1 detail
Sun Feb 28 22:52:17.405 UTC
Interface GigabitEthernet0/0/0/1:
        LDP labelling not enabled
        LSP labelling not enabled
        MPLS ISIS enabled
        MPLS enabled
```

Fig. 15. No use of LDP in SR Topology

With a ping, is possible to check segment assignment to SR MPLS and VPN label.



```
No.  Time        Source      Destination  Protocol  Length Info
584 13.123047   10.2.10.1   10.1.10.1    ICMP      118 Echo (ping) reply
585 13.148070   10.1.10.1   10.2.10.1    ICMP      122 Echo (ping) request
586 13.165400   10.2.10.1   10.1.10.1    ICMP      118 Echo (ping) reply
587 13.180789   10.1.10.1   10.2.10.1    ICMP      122 Echo (ping) request
588 13.198859   10.2.10.1   10.1.10.1    ICMP      118 Echo (ping) reply
589 13.215187   10.1.10.1   10.2.10.1    ICMP      122 Echo (ping) request
590 13.254788   10.2.10.1   10.1.10.1    ICMP      118 Echo (ping) reply

> Frame 585: 122 bytes on wire (976 bits), 122 bytes captured (976 bits) on interface 0
> Ethernet II, Src: [illegible]
v MultiProtocol Label Switching Header, Label: 17000, Exp: 0, S: 0, TTL: 254
    0000 0100 0010 0110 1000 .... .... .... = MPLS Label: 17000
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...0 .... .... = MPLS Bottom Of Label Stack: 0
    .... .... .... .... .... .... 1111 1110 = MPLS TTL: 254
v MultiProtocol Label Switching Header, Label: 24002, Exp: 0, S: 1, TTL: 254
    0000 0101 1101 1100 0010 .... .... .... = MPLS Label: 24002
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...1 .... .... = MPLS Bottom Of Label Stack: 1
    .... .... .... .... .... .... 1111 1110 = MPLS TTL: 254
> Internet Protocol Version 4, Src: 10.1.10.1, Dst: 10.2.10.1
```

Fig. 16. SR-MPLS for IPv4 WireShark Traffic Capture

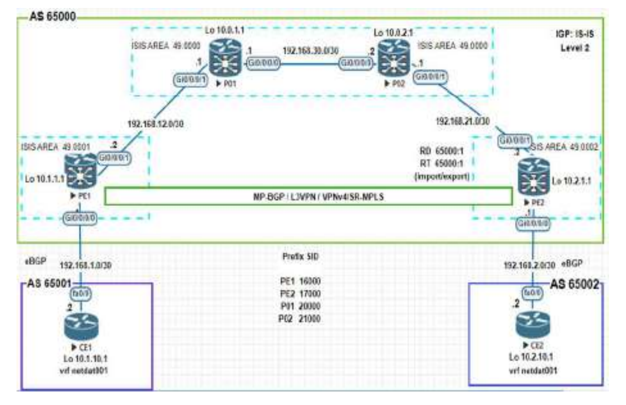


Fig. 17. SR-MPLS Topology

C. Segment Routing SRv6

An IPv6 address is used as SID, so the source router codifies the path to the destination as an ordered segment list or IPv6 addresses, which are placed in a new header called SRH. Along the path of SRv6 that a packet could take, there are nodes with different functionality, according to [16]:

- Source node and End Node: It generates IPv6 packets and place it in SRH. The End Node is the final node.
- Transit node: Appears on SRv6 path, but do not examines the SRH.

1) SRv6 Architecture

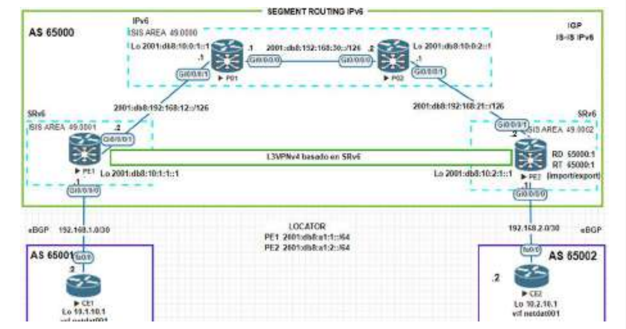


Fig. 18. L3VPN over SRv6 Architecture

To implement this scenario, at least. 6.6.1 IOS XR release is needed [16], and for this case, 6.0.1 and 6.1.3 releases have been used.

2) Configuring SRv6

a) SRv6 basic locator: It is global enabled, and a locator with its prefix must be configured, this locator is announced by IS-IS protocol.

```
RP/0/0/CPU0:PE1(config)#segment-routing srv6
RP/0/0/CPU0:PE1(config-srv6)#locators
RP/0/0/CPU0:PE1(config-srv6-locators)#locator myLoc1
RP/0/0/CPU0:PE1(config-srv6-locators)#prefix
2001:db8:a1::/64
```

Fig. 19. SRv6 locator configuration

b) SRv6 IS-IS: This protocol needs an extension to support SRv6 SID, allowing to learn local and remote locator prefixes to install it in the RIB, FIB. It is important to say that only one address-family of IS-IS can support one segment routing way, SR MPLS or SRv6[16].

c) SRv6 L3VPNv4: This configuration allows L3VPNv4 over SRv6 (Fig 20 and Fig. 21).

```
RP/0/0/CPU0:PE1(config)#router isis 1
RP/0/0/CPU0:PE1(config-isis)#address-family ipv6
unicast
RP/0/0/CPU0:PE1(config-isis-af)# segment-routing srv6
RP/0/0/CPU0:PE1(config-isis-srv6)#locator mvLoc1
```

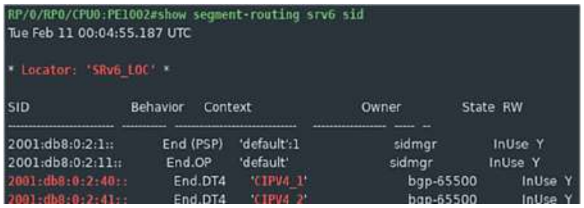
Fig. 20. ISIS Configuration in SRv6

```
RP/0/0/CPU0:PE1(config-bgp-af)#vrf netdat001
RP/0/0/CPU0:PE1(config-bgp-vrf)# rd 65000:1
RP/0/0/CPU0:PE1(config-bgp-vrf)#address-family ipv4
unicast
RP/0/0/CPU0:PE1(config-bgp-vrf-af)#segment-routing srv6
RP/0/0/CPU0:PE1(config-bgp-vrf-af-srv6)#alloc mode per-
vrf
```

Fig. 21. VRF configuration with per label allocation

3) Verification of SRv6

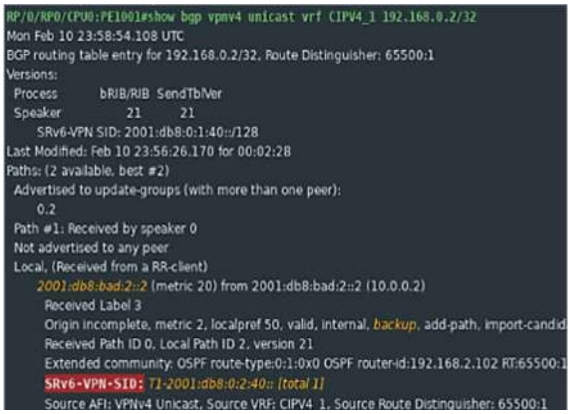
Once SRv6 has been enabled, the **show segment-routing srv6 sid** command verified SRv6 operation as seen in Fig. 22.



```
RP/0/RP0/CPU0:PE1002#show segment-routing srv6 sid
Tue Feb 11 00:04:55.187 UTC

* Locator: 'SRv6_LOC' *

SID                  Behavior  Context                Owner        State RW
-------------------- --------- ---------------------- ------------ ----- --
2001:db8:0:2:1::     End (PSP) 'default':1            sidmgr       InUse Y
2001:db8:0:2:11::    End.OP    'default'              sidmgr       InUse Y
2001:db8:0:2:40::    End.DT4   'CIPV4_1'              bgp-65500    InUse Y
2001:db8:0:2:41::    End.DT4   'CIPV4_2'              bgp-65500    InUse Y
```

Fig. 22. SRv6 SID verification [19]

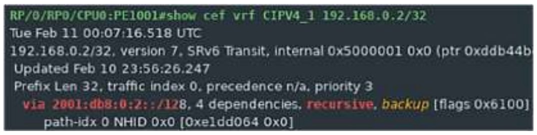
It is possible to know the prefixes learned inside a VRF in the PE device, and the SRv6-VPN-SID assigned with the **show bgp vpnv4 unicast vrf vrf-name prefix** command.



```
RP/0/RP0/CPU0:PE1001#show bgp vpnv4 unicast vrf CIPV4_1 192.168.0.2/32
Mon Feb 10 23:58:54.108 UTC
BGP routing table entry for 192.168.0.2/32, Route Distinguisher: 65500:1
Versions:
  Process        bRIB/RIB  SendTblVer
  Speaker              21        21
      SRv6-VPN SID: 2001:db8:0:1:40::/128
Last Modified: Feb 10 23:56:26.170 for 00:02:28
Paths: (2 available, best #2)
  Advertised to update-groups (with more than one peer):
      0.2
  Path #1: Received by speaker 0
  Not advertised to any peer
  Local, (Received from a RR-client)
      2001:db8:bad:2::2 (metric 20) from 2001:db8:bad:2::2 (10.0.0.2)
        Received Label 3
        Origin incomplete, metric 2, localpref 50, valid, internal, backup, add-path, import-candid
        Received Path ID 0, Local Path ID 2, version 21
        Extended community: OSPF route-type:0:1:0x0 OSPF router-id:192.168.2.102 RT:65500:1
        SRv6-VPN-SID: T1-2001:db8:0:2:40:: [total 1]
        Source AFI: VPNv4 Unicast, Source VRF: CIPV4_1, Source Route Distinguisher: 65500:1
```

Fig. 23. Prefixes learned inside VRF [19]

Also, to verify that traffic is transported over SRv6 data plane, **show cef vrf-name prefix** command is used.



```
RP/0/RP0/CPU0:PE1001#show cef vrf CIPV4_1 192.168.0.2/32
Tue Feb 11 00:07:16.518 UTC
192.168.0.2/32, version 7, SRv6 Transit, internal 0x5000001 0x0 (ptr 0xddb44b
 Updated Feb 10 23:56:26.247
 Prefix Len 32, traffic index 0, precedence n/a, priority 3
   via 2001:db8:0:2::/128, 4 dependencies, recursive, backup [flags 0x6100]
      path-idx 0 NHID 0x0 [0xe1dd064 0x0]
```

Fig. 24. SRv6 data plane for a VPNv4 prefix [19]

Fig. 25 shows the SRv6 emulated topology, and the addresses assigned in PE devices to support of VPNv4 over SRv6 are shown.

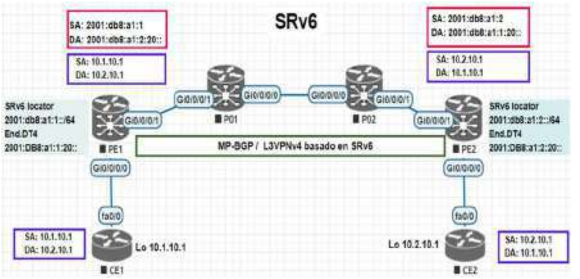


Fig. 25. SRv6 end-to-end connectivity

Finally, Fig. 26 shows the SR adoption, with companies and ISPs like Bell, SoftBank, Telefónica, among others, taking the leadership in implementing SR worldwide.

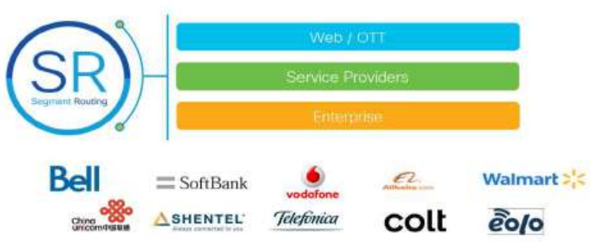


Fig. 26. SR Adoption [21]

V. PoC Analysis

The described scenarios show the operation of MPLS-LDP and SR architecture. The next table compares the architectures to determine the advantages of SRv6 over the other mentioned technologies.

TABLE I. MPLS vs SR MPLS vs SRv6 – Comparison

MPLS	SR MPLS	SRv6
TRAFFIC LABELING		
By Labels	By SIDs	By IPv6 address
Label Stack	SR Policy	Segment List in SRH
Topmost Label	Active Segment	IPv6 address in Dest. Address
OPERATIONS		
Push	Push	Equals SR MPLS adding IPv6 to Segment List en SRH.
Swap	Continue	Forward to IPv6 Dest. Address.
Pop	Next	Decreases Segment Left and active segment copied in IPv6 Dest. Address.
TECHNOLOGY SUPPORT		
Support by all modern devices	Requires software update, it uses MPLS data plane	Hardware changes are needed and software update to support SRv6 data plane
In operation with all technical support	SR knowledge to operate over MPLS data plane	Requires SRv6 new concept study. It is full IPv6 deployed in the backbone
MAIN CHARACTERISTICS		
Number of states increases with the number of routes	LFIB table remains constant, regardless of the routes in a Full-Mesh topology; Therefore, it considerably reduces the number of states.	
Uses LDP protocol for label distribution and route establishment, as well as	It does not use LDP, since a segment is an instruction, it uses the concept of the SID prefix, taking	It is based on the SRv6 SID, where the route is encoded in an ordered list of segments or IPv6 addresses that are placed on the SRH. It does not use RSVP or tunneling, its algorithm is

RSVP for TE optimization	advantage of the MPLS data plane	native according to source routing
Dual Stack Support	Dual Stack Support	IPv6 Support
Uses the LDP NSR (Non-Stop Routing) feature, which provides a rapid recovery mechanism in control plane from failures	50msec in prefix protection against link failures, nodes with the use of TI-LFA (Topology Independence-Loop Free Alternate). It is also supported in SR / LDP scenarios. It is calculated by the routers during the IGP process. Covers 100% of any topology	
It supports QoS in packet classification	It satisfies QoS requirements by ensuring applied bandwidth in the AB attribute of the node or adjacency segments in the SR TE route.	

VI. Conclusions

After a deep research and analysis, it was concluded that SRv6 bases its operation on the concept of source-routing, therefore, it simplifies routing and processing within the service provider's backbone by providing a simple, scalable and effective solution through network programming from the source that defines the route to the destination through IPv6 addresses in the SRH header.

With the emulation of the network architecture in the different L3VPN scenarios that connect two remote sites through a SP backbone, it has been shown that the MPLS operation is carried out through the LDP protocol with the label assignment when passing through each node; while SR-MPLS from scenario 2, takes advantage of the MPLS data plane, eliminating the need of LDP, introducing SID prefixes and writing an ordered list of instructions or SR Policy placed in the packet header; and SRv6 from scenario 3, establishes a pure IPv6 backbone on an IPv6 data plane that is based on source-routing through the registration of IPv6 addresses that are formed based on the SRv6 SIDs that are placed in the SRH header, with which the routing within the service provider becomes simple, effective and fast.

SRv6 enhances and improves latency, jitter and QoS characteristics by ensuring bandwidth that is applied to the node's attributes or adjacency segments in an SR-TE Segment Routing without using RSVP protocol, also enhances Fast Re-Route (FRR) for failovers, supports Secure VPNs, and network automation with Network Function Virtualization and Software Defined Networking.

SR-MPLS (IPv4 support), requires a software update on SR nodes, while SRv6 requires a change in hardware and software which represents an investment for ISPs, as well as training in human talent, but, the benefit of these changes will generate great interest in the operators and IT customers by supporting new network applications that demand a large amount of information processing with SDN in mind.

Acknowledgment

The authors thank Pontificia Universidad Católica del Ecuador PUCE, and its Engineering Faculty for supporting this investigation.

References

[1] G. D. Salazar Ch, C. Venegas and L. Marrone, "MQTT-Based Prototype Rover with Vision-As-A-Service (VAAS)in an IoT Dual-Stack Scenario," *2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG)*, Quito, Ecuador, 2019, pp. 344-349, doi: 10.1109/ICEDEG.2019.8734341.

[2] G. D. Salazar Ch., C. Venegas, M. Baca, I. Rodríguez and L. Marrone, "Open Middleware proposal for IoT focused on Industry 4.0," *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA)*, Barranquilla, Colombia, 2018, pp. 1-6, doi: 10.1109/CCRA.2018.8588117.

[3] E. F. Naranjo and G. D. Salazar Ch, "Underlay and overlay networks: The approach to solve addressing and segmentation problems in the new networking era: VXLAN encapsulation with Cisco and open source networks," *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, Salinas, Ecuador, 2017, pp. 1-6, doi: 10.1109/ETCM.2017.8247505.

[4] G. D. Salazar Ch., C. Hervas, E. Estevez and L. Marrone, "High-Level IoT Governance Model Proposal for Digitized Ecosystems," *2019 International Conference on Information Systems and Software Technologies (ICI2ST)*, Quito, Ecuador, 2019, pp. 79-84, doi: 10.1109/ICI2ST.2019.00018

[5] G. D. Salazar Ch., E. F. Naranjo and L. Marrone, "SDN-Ready WAN networks: Segment Routing in MPLS-Based Environments," *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2018, pp. 173-178, doi: 10.1109/UEMCON.2018.8796613

[6] Juniper Networks, "What is Segment Routing" [online]. Available: https://www.juniper.net/us/en/products-services/what-is/segment-routing/, [Accessed: 28-feb-2021].

[7] J. E. Vaca P. and G. D. Salazar-Chacón., "VXLAN-IPSec Dual-Overlay as a Security Technique in Virtualized Datacenter Environments," *2020 IEEE ANDESCON*, Quito, Ecuador, 2020, pp. 1-6, doi: 10.1109/ANDESCON50619.2020.9272160.

[8] G. Salazar (2016). Fundamentos de QoS -Calidad de Servicio en Capa 2 y Capa 3. Cisco

[9] Silver Peak, "What is SDWAN?" [online]. Available: https://www.silver-peak.com/sd-wan/sd-wan-explained, [Accessed: 28-feb-2021].

[10] B. Gafni. "On Segment(ed) Routing" [online]. Available: https://blog.mellanox.com/2018/10/segment-routing-using-mpls-ipv6-srv6/, [Accessed: 28-feb-2021}.

[11] D. Jaksic. "Segment Routing in Service Provider Networks" [online]. Available: https://www.cisco.com/c/dam/m/hr_hr/training-events/2018/cisco-connect/pdf/Segment_Routing_in_Service_Provider_Network_-_Dejan_Jaksic.pdf, [Accessed: 28-feb-2021].

[12] Cisco Systems, "Implementing Cisco Service Provider Next-Generation Core Network Services", San José, CA, 2014

[13] G. D. Salazar-Chacón and L. Marrone, "OpenSDN Southbound Traffic Characterization: Proof-of-Concept Virtualized SDN-Infrastructure," *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, 2020, pp. 0282-0287, doi: 10.1109/IEMCON51383.2020.9284938.

[14] Cisco Systems, "Implementing Cisco Service Provider Next-Generation Core Network Services", San José, CA, 2014.

[15] G. Salazar, E. Naranjo and L. Marrone., "SDN-Ready WAN networks: Segment Routing in MPLS-Based Environments," 9th IEEE Annual Ubiquitous Computing Electronics & Mobile Communication Conference(UEMCON), pp. 173-178, November 2018.

[16] Cisco Systems , "Segment Routing Configuration Guide for Cisco ASR 9000 Series Routers, IOS XR Release 6.6.x", San José, CA, 2019.

[17] P. Camarillo, "SRv6 Network Programming", Madrid, Madrid, 2019.

[18] P. L. Ventre, S. Salsano, M. Poverini, A. Cianfrani, A. Abdessalam, C. Filsfils, P. Camarillo y F. Clad, "Segment Routing: a Comprehensive Survey of Research Activities, Standardization Efforts and Implementation Results," June 2020.

[19] A. Gonzalez, "SRv6 a new Hope. Case 1: SRv6-Based IPv4 L3VPN," [online]. Available: https://www.linkedin.com/pulse/srv6-new-hope-case-1-srv6-based-ipv4-l3vpn-asier-gonzalez-diaz/, [Accessed: 28-feb-2021].

[20] G. Salazar Chacón, and G. Chafla Altamirano. "Empleo de Path-Control Tools en una red empresarial moderna mediante Políticas de Enrutamiento". 3C Tecnología, vol. 4, no. 1, pp. 1-18, 2015.

[21] A. Donzelli, "Introduction to Segment Routing". Cisco Live 2019. BRKRST-2124

Process monitoring vehicle for SRAM critical path using ring oscillator in 7 nm Finfet

Mohammad Anees
Memory Design Group
Xilinx Pvt. Ltd
Hyderabad, India
manees@xilinx.com

Kumar Rahul
Memory Design Group
Xilinx Pvt. Ltd
Hyderabad, India
kumarr@xilinx.com

Sourabh Aditya Swarnkar
Memory Design Group
Xilinx Pvt. Ltd
Hyderabad, India
sourabh@xilinx.com

Santosh Yachareni
Memory Design Group
Xilinx Pvt. Ltd
Hyderabad, India
santoshy@xilinx.com

***Abstract*— In the nano-scale fabrication process, one of the major concerns is the impact of the process variation on the functionality of the design. At lower technology nodes, the impact from the second-order elements which were not significant is playing an important role in defining the designs. SRAM being at the forefront of every process node, it is important to assess the impact of the process in the early stages to make the required design changes. Self-time path in SRAM is crucial in defining the functionality, power, and performance. In the early stages of SRAM development, the proposed ring oscillator (RO) design helps in estimating the post-fabrication impact. This paper discusses the design and implementation of the SRAM critical path-based ring oscillator in the 7nm FinFet process and comparing the behavior with a more conventional ring oscillator. The impact from the process parameter variation is captured in the form of frequency variations. It is observed that the process tracking capability of conventional ring oscillator differs by ~17% compared with SRAM critical path behavior.**

Keywords*—** *Bitcell, Critical-Path, FinFet,* ***PMV *(Process Monitoring Vehicle),* ***RO*** *(Ring Oscillator),* ***SRAM*** *(Static Random-Access Memory)****,*** ***STP*** *(Self-Time Path),* ***SAE*** *(Sense Amplifier Enable)*

I. Introduction

Deep submicron scaling of process nodes is done to add more functional blocks on-chip by incorporating more transistors. At sub 28nm, the effects of process variations are significant. These effects are having more influence on the performance and functionality of the design in submicron nodes. Process variations are caused by the deviation of process parameters from their desired values. This is caused due to the limited controllability of a fabrication process. As device dimensions continue to scale towards ultra-deep submicron range typically less than 100 nm, manufacturing equipment’s become less reliable in controlling the design parameters. Intrinsic parameter fluctuations play an increasingly important role in submicron devices. In technology nodes 100nm and less, margins shrink due to reduction in supply voltage [1]. SRAMs are scaled aggressively to fit more bits per unit area. Scaling the transistor dimensions comes at the cost of complex process and elevated variability in the transistors [2] and hence reduces the overall yield. In FinFets, the gate of the device is elevated, hence the transistor behavior becomes more complex. This results especially in bitcells requiring new techniques to incorporate the variation of the bitcells in the ring oscillator [3]. Process monitor vehicle (PMV) block which includes ring oscillator helps to estimate the process variations in fabrication. SRAMs are more prone to variation since the bitcells dimensions are scaled more compared to logic gates. In the 7nm process, the SRAM (6T) is ~15% smaller than the smallest inverter (INV). Reduction in SRAM supply voltage reaching towards the threshold voltage makes it more vulnerable to variation [4]. The number of path delay faults observed during the first silicon debug are significantly [5] higher at high-frequency operation. Process variations of 10%-30% across wafers and 5%-20% across different die significantly affect the behavior of devices and interconnects [6].

Critical paths define the operating frequency of the block. Hence it is required to optimize the critical path and study impact from process variation at the early stages of silicon debug [7]. To capture the impact from fabrication at an early stage of SRAM development we have designed the self-suffice ring oscillator with SRAM critical path as part of the oscillator. This helps us in characterizing the SRAMs critical path which contains the STP (Self-Time Path). STP is an important part of the SRAM design. The STP design is critical and impacts the functionality, power, and speed of the memory. Speed characterization is used in many places in chip design, this will help both yield and business aspects of chip design. In 7nm Xilinx FPGA we can have several hundreds of megabytes of SRAMs. SRAMs on the mid-sized FPGA are the largest contributors to the critical path. Section II briefly discusses the conventional ring oscillator. Section III discusses the proposed design. Section IV discusses results.

II. Conventional (Ring Oscillator)RO design

Commonly used ring oscillator design comprises of inverters, nand/nor logic arranged in an odd number of stages.

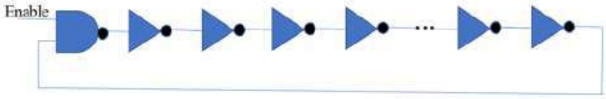


Fig. 1 Conventional design of RO.

Fig. 1 design is used for comparing the frequency response of the self-time path-based ring oscillator. The inverter-based ring oscillator is designed with the 7nm FinFet stdcell library. Ring oscillator designs are testable at the early stages of the process where the first level of metals can be used for initial process development as well as

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

process monitoring in manufacturing stages. Different configurations of ring oscillator structures are embedded inside the design itself and can be used to diagnose design and fabrication process impact on the design such as the proposed in the paper [8].

Some of the ring oscillator-based test structures are observed here [9] [10] [11] [12]. The frequency of the ring oscillator is used as a metric to quantify the process. There are different configurations of inverters that are chosen to improve their sensitivity to the process. It includes usage diode-connected configuration [13]. SRAM's critical path contains bitcell based self-timed path and many custom logic circuits which don't contain characterized standard cells. To model the SRAM's critical path behavior proposed design is embedded inside the SRAM, this will be a more accurate estimate of fabrication impact.

III. Proposed Design

The proposed design doesn't contain any chain of inverters or other gates with odd stages. Toggling here is achieved by a combination of self-time path and ring logic design. There are many methods used to characterize the process vs speed. To study the impact of the process bitcells are used while constructing the ring oscillator [14][15][16]. The proposed design includes tracking bitcells and the custom logic in the critical path.

SRAMs critical path contains tracking bitcell, which are modified bitcell used for the tracking of bitcell operation. These are an essential part of SRAMs and are very critical in compiler memories. The STP includes many custom logics along with tracking bitcells. Fig. 2 shows the self-timed path containing tracking bitcells in rows and columns. The self-time path is designed to track the worst bitcell in the bitcell array. The worst bitcell to access for the operation will be furthest from the control block. Tracking is required for word line in row and bit line in a column. The self-time path is responsible for triggering the sense signal to enable the sense amplifier in IO for the read operation. If the STP is trigger sense signal too early, then the result of the sense amplifier can resolve in an unknown state. Since sense amplifier requires certain differential voltage on bitlines to read the correct state of the bitcell. If the STP is fired too late then the bitcell will discharge too much charge from the bit line causing more power to be burned. When the word line is turned ON the entire row is activated hence any more delay than required will result in burning access power. This also results in higher cycle time since the process takes more time to complete the read/write cycle. Hence impacting the frequency of operation.

Fig. 3 shows SRAM and ring oscillator stage. SRAM output, sense amplifier enables (SAE) is taken out from the SRAM and given as input to the ring oscillator stage. By incorporating this into the oscillator structure we can accurately model SRAMs critical path behavior. The ring oscillator is responsible to generate the clock pulses depending on the input SAE signal. SAE is a pulse generated from the STP, it is processed in the ring oscillator to generate the set and reset signal for the SR latch. This generates the pulses proportional to the incoming SAE signal. The output of RO is feedback to SRAM as the clock. Fig. 4 shows the standard 2-input active low enabled SR latch which is used for the clock generation. Internally, care is taken to avoid any unwanted states which can result in functional failure. Fig. 5 shows the implemented ring oscillator and SRAM.

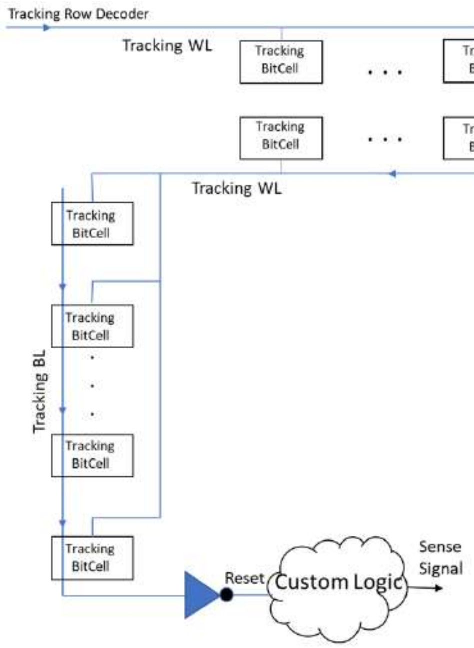


Fig. 2 Tracking bitcells used in Self-Timed Path

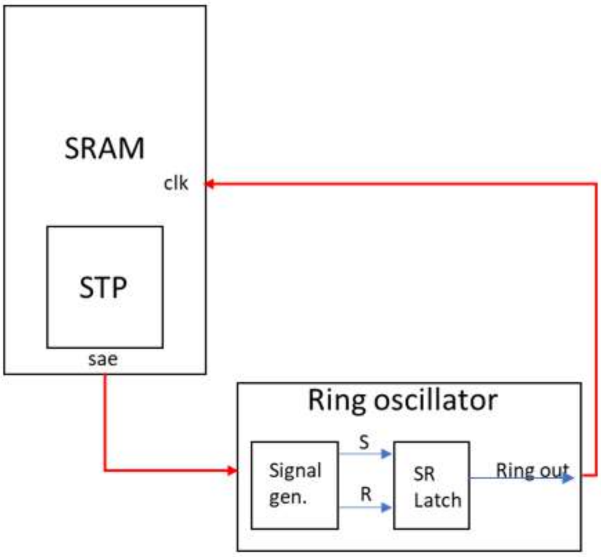


Fig. 3 SRAM and ring oscillator connected to enable toggling

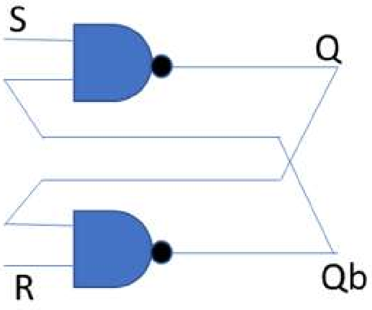


Fig. 4 Two-input NAND active low input SR latch

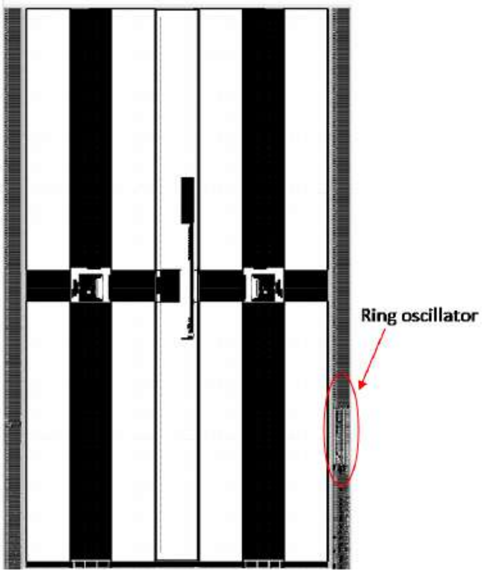


Fig. 5 Ring oscillator and SRAM layout

IV. Results

Fig. 6 and 7 show the sample frequency response of SRAM at the TT process. We can see at low voltage and high voltage there is variation in the behavior of the ring oscillator. This is due to temperature inversion caused at low voltage. SRAM RO is having a high frequency at high temperatures for low voltage and high frequency is achieved at low temperature for high voltages. This is highlighted in a red circle at two places in Fig. 6.

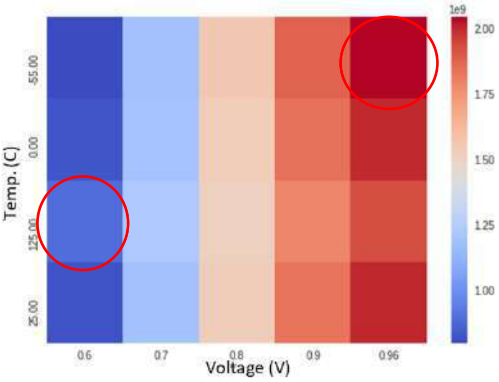


Fig. 6 SRAM PMV Frequency response at TT

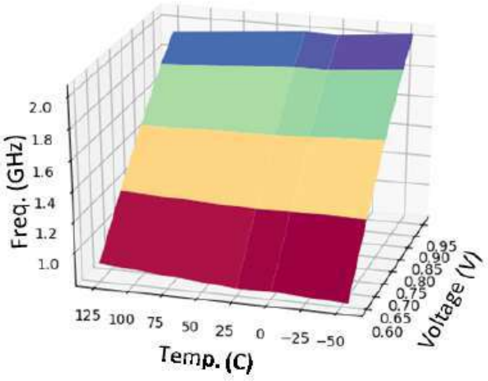


Fig. 7 SRAM PMV Temp. inversion effect at TT

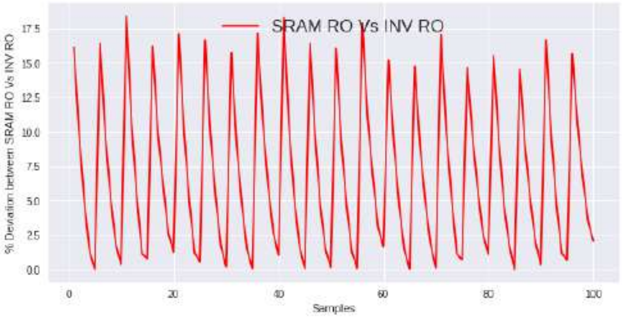


Fig. 8 100 sample points from SRAM and inverter RO with normalized frequency for comparison

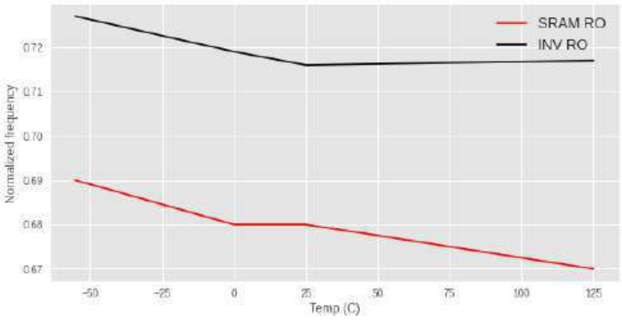


Fig. 9 Normalized SRAM and inverter RO Frequency response with temperature.

Fig. 8 contains normalized 100 samples of frequency at different processes, voltage, and temperature. 0% indicates the SRAM and inverter RO are inline and any deviation from their behavior is captured. There is a roughly ~17% variation between SRAM RO and inverter RO across process, voltage, and temperature. This gap needs to be reduced If the other ring oscillator-based design needs to be used to track the SRAM. Fig. 9 shows one of the 100 cases where the behavior of SRAM and inverter RO with temperature is not aligning with each other. This is mainly due to the presence of bitcells used in the self-time path. Bitcells behave differently compared to conventional logics since their dimensions are smaller and more prone to variations. In a compiler, it is preferred to have more variable delay and smaller fixed delay due to logic, since the STP should be able to track the bitcell for various memory configurations.

TABLE 1. σ/μ FOR DIFFERENT PROCESSES ACROSS VOLTAGE AND TEMPERATURE.

Process	*Sigma(σ)/Mean(μ)*		*% Deviation in the variation*
	SRAM RO	*INVERTER RO*	
TT	0.292	0.244	16.29
SS	0.322	0.276	14.25
SF	0.293	0.245	16.29
FS	0.289	0.244	15.31
FF	0.257	0.212	17.57

Table 1 shows the σ/μ variation of the SRAM and inverter ring oscillator. Process corners considered for the analysis are (Typical-NMOS, Typical-PMOS) TT, (Slow-NMOS, Slow-PMOS) SS, (Slow-NMOS, Fast-PMOS) SF, (Fast-NMOS, Slow-PMOS) FS, (Fast-NMOS, Fast-PMOS) FF. Variation in the σ/μ between the SRAM and inverter RO shows that these are not aligned with each other and are apart by ~17%. Embedding ring oscillator inside SRAM will yield more realistic process tracking since the ring oscillator takes the SRAM critical path into considerations.

V. Conclusion

In this paper, we have designed a self-time path-based ring oscillator incorporating the SRAM critical path. This helps in estimating the process impact on the SRAM more accurately. As nodes are shrinking beyond 7nm impact of second-order effects is becoming more prominent and it is critical to evaluate their impact early stages in the design. SRAM ring oscillator helps in tuning the design for functionality, power, and performance. Increasing the number of bits per unit area and having reduced power without compromising the reliability, performance is very critical in SRAM design. Similar more process-aware designs need to be implemented to track the process impact on design. Future work will involve updating the design to capture the effects of bias temperature instability and hot carrier injection in the ring oscillator.

References

[1] A. Asenov, S. Kaya and A. R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," in IEEE Transactions on Electron Devices, vol. 50, no. 5, pp. 1254-1260, May 2003, doi: 10.1109/TED.2003.813457.

[2] M. Anees, K. Rahul, S. Yachareni and S. A. Swarnkar, "Study on impact of process on Bitcell design in FinFets," 2020 IEEE 14th Dallas Circuits and Systems Conference (DCAS), Dallas, TX, USA, 2020, pp. 1-5, doi: 10.1109/DCAS51144.2020.9330648.

[3] M. Anees, K. Rahul, S. A. Swarnkar and S. Yachareni, "Behaviour Shockley and Sakurai Models in 7nm FinFet," 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, 2020, pp. 1-4, doi: 10.1109/IEMTRONICS51293.2020.9216418.

[4] M. Anees, K. Rahul and S. Yachareni, "Study of bitcell qualification using current and voltage based metric's in 7nm FinFet Technology," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2020, pp. 101-105, doi: 10.1109/ICSTCEE49637.2020.9276815.

[5] H. Balachandran, K. Butler N. Simpson, "Facilitating Rapid First Silicon Debug," in Proc. ITC'02, pp. 628-637, Oct 2006.

[6] M. Nourani and A. Radhakrishnan, "Testing On-Die Process Variation in Nanometer VLSI," in IEEE Design & Test of Computers, vol. 23, no. 6, pp. 438-451, June 2006, doi: 10.1109/MDT.2006.157.

[7] X. Wang, M. Tehranipoor and R. Datta, "Path-RO: A novel on-chip critical path delay measurement under process variations," 2008 IEEE/ACM International Conference on Computer-Aided Design, San Jose, CA, 2008, pp. 640-646, doi: 10.1109/ICCAD.2008.4681644.

[8] M. Bhushan, A. Gattiker, M. B. Ketchen and K. K. Das, "Ring oscillators for CMOS process tuning and variability control," in IEEE Transactions on Semiconductor Manufacturing, vol. 19, no. 1, pp. 10-18, Feb. 2006, doi: 10.1109/TSM.2005.863244.

[9] M. Nourani and A. Radhakrishnan, "Testing On-Die Process Variation in Nanometer VLSI," in IEEE Design & Test of Computers, vol. 23, no. 6, pp. 438-451, June 2006, doi: 10.1109/MDT.2006.157.

[10] Z. Abuhamdeh, B. Hannagan, J. Remmers and A. L. Crouch, "A Production IR-Drop Screen on a Chip," in IEEE Design & Test of Computers, vol. 24, no. 3, pp. 216-224, May-June 2007, doi: 10.1109/MDT.2007.59.

[11] H. Dao, "Process evaluation, validation, and monitoring with ring oscillator scribelane modules," 2016 27th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), Saratoga Springs, NY, USA, 2016, pp. 218-219, doi: 10.1109/ASMC.2016.7491129.

[12] M. Nourani and A. Radhakrishnan, "Testing On-Die Process Variation in Nanometer VLSI," in IEEE Design & Test of Computers, vol. 23, no. 6, pp. 438-451, June 2006, doi: 10.1109/MDT.2006.157.

[13] Y. An, D. Jung, K. Ryu, H. S. Yim and S. Jung, "All-Digital ON-Chip Process Sensor Using Ratioed Inverter-Based Ring Oscillator," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, no. 11, pp. 3232-3242, Nov. 2016, doi: 10.1109/TVLSI.2016.2550603.

[14] Ming-Chien Tsai et al., "Embedded SRAM ring oscillator for in-situ measurement of NBTI and PBTI degradation in CMOS 6T SRAM array," Proceedings of Technical Program of 2012 VLSI Design, Automation and Test, Hsinchu, 2012, pp. 1-4, doi: 10.1109/VLSI-DAT.2012.6212587.

[15] Q. Tang and C. H. Kim, "Characterizing the Impact of RTN on Logic and SRAM Operation Using a Dual Ring Oscillator Array Circuit," in IEEE Journal of Solid-State Circuits, vol. 52, no. 6, pp. 1655-1663, June 2017, doi: 10.1109/JSSC.2017.2681809.

[16] M. Igarashi, Y. Uchida, Y. Takazawa, M. Yabuuchi, Y. Tsukamoto and K. Shibutani, "Study of Local BTI Variation and its Impact on Logic Circuit and SRAM in 7 nm Fin-FET Process," 2019 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2019, pp. 1-6, doi: 10.1109/IRPS.2019.8720508.

Improving Connectivity at Ships and Planes Through Land-based BTS and Point-to-Point Antennas

Shahriar Khan
Department of EEE
Independent University, Bangladesh
Dhaka, Bangladesh
skhan@iub.edu.bd

Muhit Kabir Sarneabat
Chemical Engineering Division
Institute of Electrical Engineers,
Dhaka, Bangladesh
smkabir87@gmail.com

Asma Khatun
Department of Electrical Engineering
Independent University, Bangladesh
Dhaka, Bangladesh
asma636dhaka@gmail.com

***Abstract*—Online connectivity speeds are increasing rapidly on land, but are still very slow on modern ships and planes. Ships and planes depend mostly on satellite communication, which is divided up between their many passengers. This study attempts to find and overcome bottlenecks to the connectivity in ships and planes. Land-based BTS can connect to ships and planes, which can connect to other ships and planes, creating an ad hoc network. Planes can act as repeaters, overcoming the curvature of the Earth. Point-to-point microwave communication (rather than cellular concept) can link ships with ground BTS and other ships, tens of km away. Tracking antennas on ground BTS and ships, can follow planes in the air. Antenna size and movement on planes must be kept minimal, as they increase drag and fuel costs. Real-time location information can be used to point tracking antennas. There can be either moving antennas, or a large number of narrow-range antennas switching with the movement of the ship or plane. In the far oceans, submarine cable connected, solar-powered BTS can be installed for connecting to nearby ships and planes. With multiple channels available between distant ships and planes, packets of information can choose the best available path for transmission (like the internet). Point-to-point communication may be difficult between ships, and the cellular concept can be used instead. Today's technical capabilities suggest we are on the verge of a breakthrough in better connectivity of ships and planes hundreds of km from the shore. New competing technologies will emerge, of which the best will survive. Policies and rules must be established, so as to encourage competition. It is hoped that this paper will help in the research and development of the new technologies for better connectivity.**

***Keywords*—connectivity, ship, plane, airplane, dish, antenna, directional, ad hoc network, BTS, point-to-point, microwave, tracking, antenna, microwave.**

I. Introduction

Connectivity speeds for smart phones and computers have been increasing exponentially over the last few decades, in agreement with Moore's law. Communication at up to 2 Mbps is possible with the Mars Rover hundreds of millions of miles away (20 minutes for round-trip of light). Surprisingly, in this day and age, connectivity at modern ships and planes is still very slow. Most passengers may have no connectivity or slow connectivity in the 18 hours on intercontinental flights, or for days or weeks on cruise ships. The isolation of the crew of cargo ships and oil tankers, in the seas for weeks and months, is worsened by slow connectivity with the rest of the world.

This paper attempts to identify and overcome the bottlenecks to connectivity on ships and planes. Ships and planes largely depend on dividing among passengers the already low bandwidth of satellites [1].

A. Low connectivity on Planes and Ships

Large commercial planes, such as the A380 and Boeing 777 often provide no connectivity to economy class passengers. Limited connectivity may be available upon purchase. The Wi Fi offered (free or paid) on the plane may at best allow voice communication and low-resolution picture transfer [2],[3]. Transmission of video is almost impossible. Passengers are advised to shrink pictures before sending, and to turn off background apps that refresh and synchronize. Apps such as *WhatsApp*, *Youtube* and *Netflix* are usually blocked.

Connection of planes with geostationary satellites means there is a time delay for the signals to travel 36,000 km to the satellite and back. Also, satellites do not reach planes traveling close to the poles, which happens often for inter-continental travel.

Only very recently are planes beginning to connect to ground based BTS with antenna located on the base of the plane. The plane antennas may move from BTS to BTS as they fly over the ground [4], [5], [6].

B. Low connectivity on Ships

The total number ships in the seas may be small compared to the total number of planes in the air. However, passengers and crew may spend weeks or months in isolation in the sea, which is only worsened with low connectivity.

In cruise ships, economy class passengers spend days or weeks getting slow and intermittent connectivity only with extra payment. Passengers discover that a vacation on a cruise ship also means a vacation from the internet. Even on modern liners, disconnection of internet starts from the passenger's entry into the ship, and ends upon disembarkation from the ship.

: 978-1-6654-4067-7/21/$31.00 ©2021 IEEE

C. *Ad-hoc networks*

Ad hoc networks, or temporary or changing networks, are an established technology today [7]. Ad hoc networks can be made from smart phones, or from autonomous vehicles talking to each other to prevent accidents [8],[9]. Ad-hoc networks have been proposed for interconnecting planes (Aeronautical Ad hoc Network), at least partly for the sake of safety [4] [5] [6].

This paper further develops these ideas and goes a step further in proposing an ad-hoc network from a combination of ships, planes and ground-based BTS.

Good connectivity between ships and planes means, that high demand on any ship or plane at any time, can be met by distributing the demand over a wide number of ships and planes. This means slow satellite connections to a number of ships and planes in a network can meet the temporary high connectivity demand of a single ship or plane.

D. *Potential for Improvement*

Planes flying over the ground should be made to have better connectivity with ground Base Transceiver Stations (BTS). Ships close to land should have better connectivity with shoreline BTS. Also, ships have the potential to become mobile BTS for planes and other ships.

Land-based BTS can connect to ships and planes, which can connect to other ships and planes, creating a large ad hoc network. The large number of planes on common flight paths can connect to each other, helping create such a network. Often ships travel in common shipping routes, meaning they too can help form an ad hoc network. Ships and planes can connect with each other, to form a large ad hoc network.

Point-to-point microwave antennas can link and track ships with ground BTS and other ships. Similar antennas on ground BTS and ships can follow planes real-time in the air. Real-time publicly available location information can be used to point tracking antennas. In the far oceans, BTS can be set up on islands, connected with submarine cables, and powered with solar panels

In 2008, the in-flight broadband company *Gogo* (then known as Aircell) launched its first onboard Wi-Fi service on a Virgin America plane. The 3 Mbps connection was enough for a few laptops (streaming video was prohibited). But now, this bandwidth is inadequate with every passenger having at least one device to connect to many apps and websites, *Gogo* had a monopoly on US in-flight Wi-Fi, with a network covering the whole country.

Today's technical capabilities suggest we on the verge of a Wild West of new technologies. At least for the next few years, there will be some competing technologies for enhancing connectivity of ships and planes. A major revolution in the connectivity of ships and planes may be in the near future, and it is hoped that this paper will help in the design and implementation.

II. Cellular Communication and Base Transceiver Stations

We start by comparing cellular and point-to-point communication, which will help in determining what would be appropriate for ships and planes. The cellular concept may be useful for creating a network for planes, which may be find it difficult to use point-to-point communication.

The cellular phone revolution of today was made possible by the cellular concept and the reuse of frequencies. Usually, cells are created with (a) omni-directional antennas, (b) or 120° antennas or (c) or 60° antennas

Base Transceiver Stations (BTS) may use omni-directional antennas, if the need is to cover a few cell phones over a wide area (figure below).

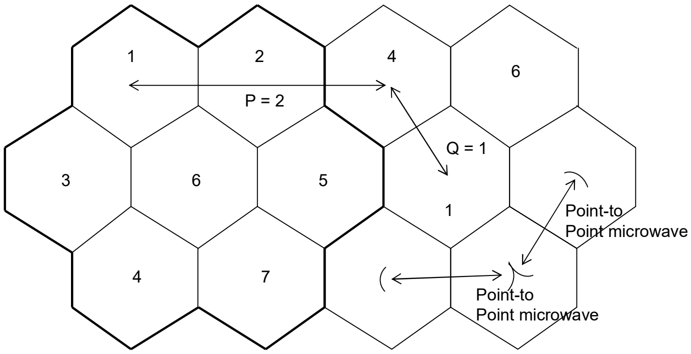


Fig. 1. Using omnidirectional antennas allows hexagonal cells and the reuse of frequencies. Number of frequencies, $N = P_2 + PQ + Q^2$. Point-to-point communication may be done between cells with microwave links or fiberoptic cable.

In the above figure, from theory, the number of frequencies required is a function of $P = 2$, and $Q = 1$, which are adjacent distances between cells.

$$N = P^2 + PQ + Q^2$$

$$= 2^2 + 2 \text{ x } 1 + 1^2 \quad = 7$$

Point to point communication between cells may be done with fiberoptic lines or point-to-point microwave links (figure above).

For higher concentrations of cell phones, three 120° antennas or six 60° antennas may be used to cover all round 360°, in effect, sectoring the cells (figure below).

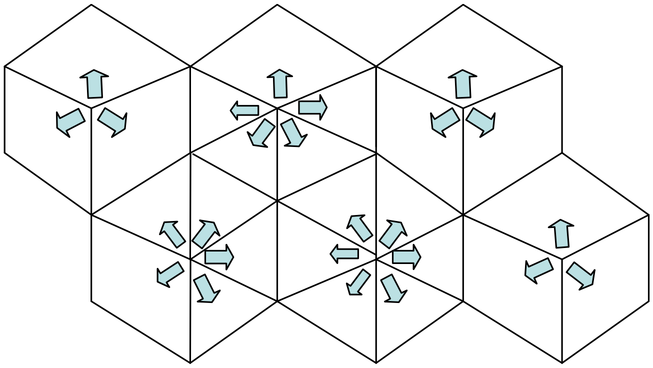


Fig. 2. The cells may be sectored with 120° antennas and 60° antennas.

BTS are often connected point-to-point with microwave towers, up to 50 - 80 km apart; a distance limited by the curvature of the Earth. Such point-to-point microwave has possibilities for connecting ships with other ships and with planes.

Owing to the tens-of-km distances between planes and between ships, the point-to-point concept may be applicable. Narrow-direction BTS antennas may have good potential for communication with planes and shoreline ships up to tens of km away (figure below).

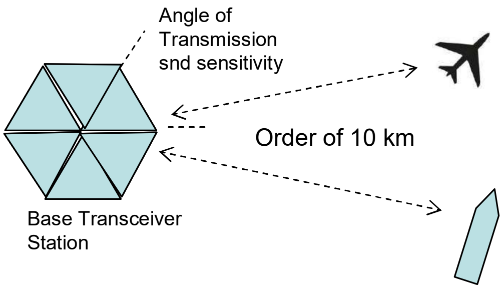


Fig. 1. Top view of a BTS shows six 60° directional antennas, covering all of 360°. Narrow-direction BTS antenna have good potential for faraway planes and shoreline ships. Advantage is planes and ships have much available power for transmission and reception (unlike cell phones)

Planes and ships have much power available for transmission and reception (unlike cell phones). Also, ships can use directional antennas (unlike cell phones).

TABLE 1. COMPARISON OF CELL PHONES, PLANES, SHIPS AND BTS ON LAND

	Power for Transmission & Reception	**Antenna**	**Tracking and Moving**	**Point to Point Commu-nication**	**Cellular communication**
Cell phone	Only battery power	Omni-directional	Unsuitable	Not suitable	Very suitable
Plane	High power available	Small size makes directional difficult	Moving and tracking difficult	Difficult, due to small antenna	Difficult because of tens of km distance
Ship	High power available	Omni-directional and /or unidirectional	Possible at top of ship	Very suitable, but need tracking antenna	Difficult because of tens of km distance
BTS on Land	High power available.	Omni direction, 60°, and uni-directional	Movement very possible	May need tracking antenna for ships and planes	Suitable for cell phones.

III. EXISTING TECHNOLOGY AND LITERATURE REVIEW

The latest aircraft (e.g. Boeing 787) already uses about 26 antenna placed around the aircraft. An antenna on top (jetwave antenna) may connect to satellites.

A. Literature Review

Considering that aircraft and ship connectivities provide great potential for growth, there are few scientific papers on the subject.

Connecting ships with each other and with shoreline BTS has been proposed [10],[11],[12].

Integrated connectivity of ground BTS, ships and planes has not been encountered in the literature.

IV. IMPROVING CONNECTIVITY OF PLANES

According to communication theory and practice, there are a number of possibilities for connecting planes.

A primary consideration is that antenna placed on planes tend to bulge out from the aircraft and produce drag, and consume extra fuel. New antennas may be difficult to install, as they may require working on the aluminum aircraft. A small sized antenna, rather than large, would be preferred. A stationary rather than moving antenna would be preferred. Large-sized moving antenna would be especially hard to install. These suggest the use of the cellular concept between planes.

The distances of planes from each other may be in the range of many tens of kilometers, meaning both microwave and RF would be possible. But the high speed and the associated Doppler shift may been a source of problems.

A. *Aircraft to Ground Connection*

There is much potential at this time, for connecting planes with ground BTS. (figure below).

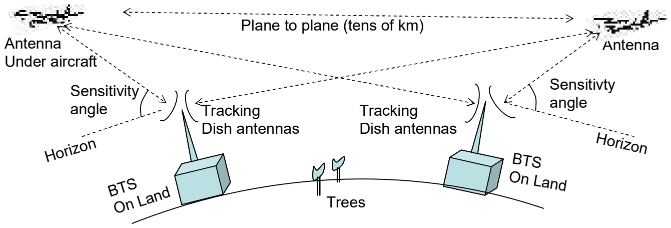


Fig 2 . Planes can connect with ground-based BTS with narrow-angle directional antennas. Moving antennas at the BTS can track and move with real-time location information. BTS antenna may point up for ships.

Planes will mostly be above the horizon of land BTS and ships, requiring their antenna to point upwards.

Connection with BTS is only possible when planes are over land. When far out at sea, planes, BTSs may not be available, but planes can still connect to ships, which can act like ground based BTS. A BTS can also be installed on isolated islands in the oceans.

B. *Moving versus Stationary Antennas*

Moving antennas have inherent problems, such as moving parts, possibility of breakage, and greater maintenance requirements [13]. In communications, moving antennas have traditionally been avoided in favor of stationary antennas.

In satellite communication, the preference is for stationary antennas connecting to geostationary satellites 36,000 km from the Earth, rather than moving antennas connecting to Low Earth Orbit (LEO) or Medium Earth Orbit (MEO) only a few hundreds or thousands of km away.

A moving antenna can be substituted by multiple stationary narrow-angle directional antennas, switching electronically from one to another (with no movement) as the plane (or ship) moves from the path of one antenna to another (figure below).

In the BTS below, there are ten 36° antennas, for all around 360° coverage. The antennas can switch from one to another, tracking the movement of the plane.

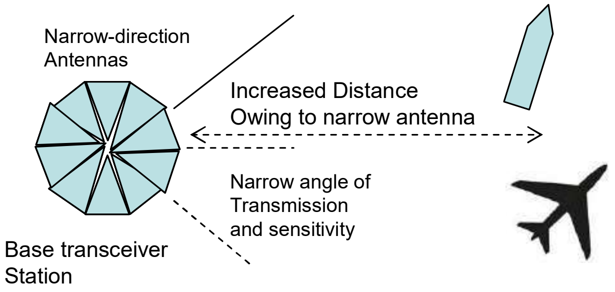


Fig. 3. Using narrow-angle directional antennas (36°) switching with movement of plane (or ship), will allow stronger transmission and greater sensitivity to signals, without movement of the antennas.

C. *Tracking with GPS and height Information*

The tracking of planes with antennas would be possible with location information from GPS. The location of the plane is also available from real time trackers [14] (figure below).

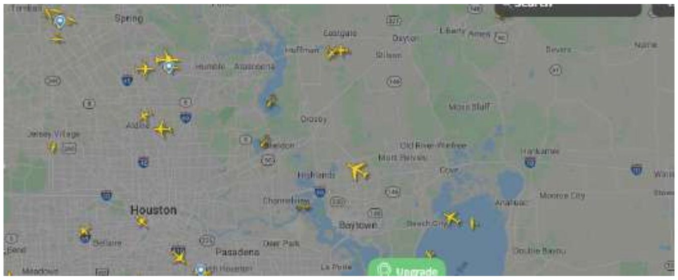


Fig 4. Planes in the Houston - Galveston area (FlightRadar24.com, April 4, 2021, 2:00 am, Dhaka time)

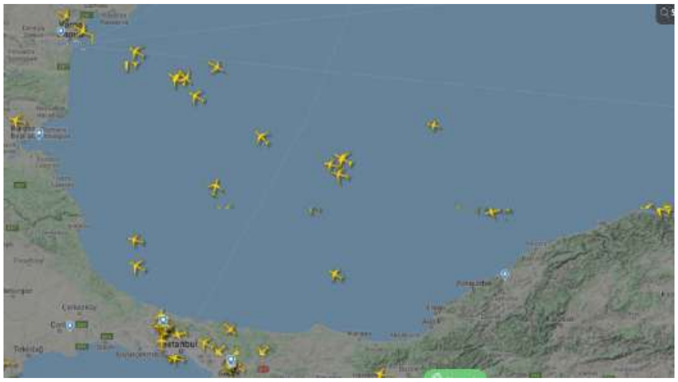


Fig 5. Plnnes over the Mediterranean, with Bulgaria on the left (flightradar24.com, on April 5, 5:00 am, Eastern time, US). Movement along common flight paths is visible.

In the above real-time location of planes near Bulgaria and the Mediterranean, planes are seen on common flight paths, indicating distances close enough for connectivity and forming an ad hoc network.

The above pictures are during the Covid pandemic, and the density of flights is much lower than during other years.

Planes can act as repeaters, overcoming the curvature of the Earth, to reach distant ships and planes (figure below).

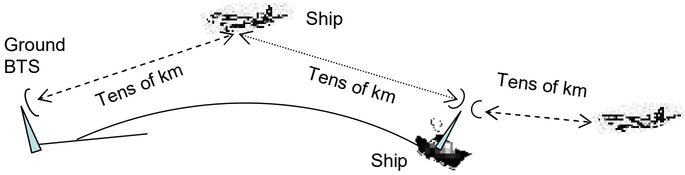


Fig. 6. A plane can act as a repeater for microwave communications, overcoming the line-of-sight requirement for microwave towers.

Using the plane as a repeater, connectivity may be possible for ships and planes about a hundred km from the shore.

V. Connectivity for Ships

The numbers of ships in the seas are far less than the number of planes in the air at any time.

When planes travel over land, they can connect to ground based BTS. But ships cannot be on the land, meaning they can connect to ground BTS only when they are close to the shore. A ship and a shoreline BTS may be connected with a point-to-point microwave dish antenna of range up to many tens of km. A ship may even act like a BTS at sea.

Ships are spacious and weighty enough to have large moving antennas. However, too high an antenna will thwart the ship's original design height specification to pass under bridges.

The roll, pitch and yaw of the ship must be compensated for:

(a) by having a wider-angle dish antenna to compensate for the movement of the ship.

(b) by having compensatory movement of the antenna along the three axes, to compensate for the movement of the ship (figure below).

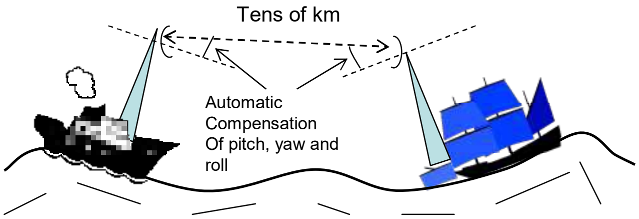


Fig. 7. Point-to-point antenna of ships may need compensation for pitch, roll, and yaw.

A problem arises far into seas and oceans, when there are no ground-based BTS to connect to. Connection to ground is not possible when passing over large bodies of water.

Many ships travel on common routes. So it is possible that ships close to the shore may interconnect with distant ships on the same route to become a network of ship-based BTSs.

The positions and velocities of ships can be seen real time in some tracking websites (figure below for Straits of Gibraltar) [15].

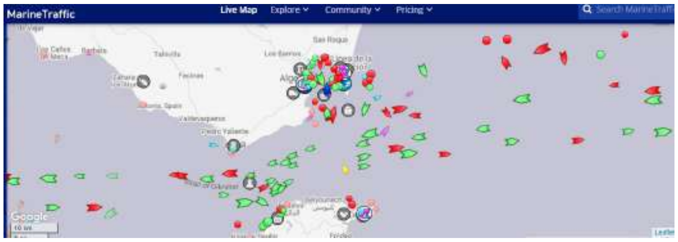


Fig 8. Ships in the Straits of Gibraltar (MarineTraffic.com, April 4, 2021, 2:00 am, Dhaka time)

VI. Integrating Ships, Planes and Ground BTS

Having seen the interconnection of planes with each other, and ships with each other, we now look at a more integrated system involving ground BTS, ships and planes.

The high speed connectivity from land BTS can be transferred to nearby ships and planes, which can then transmit to other ships and planes.

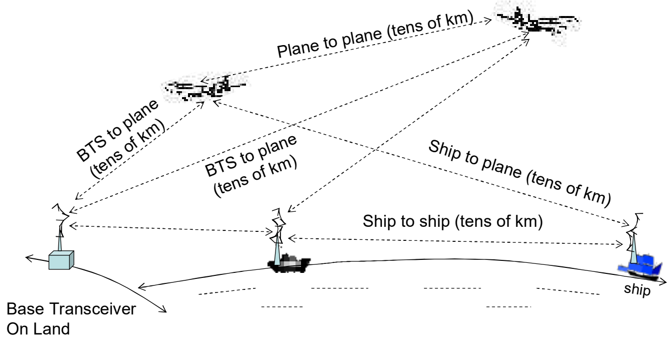

Plane to plane (tens of km)
BTS to plane
(tens of km)
BTS to plane
(tens of km)
Ship to plane (tens of km)
Ship to ship (tens of km)
Base Transceiver
On Land
ship

Fig. 9. Integrating connectivity of ground BTS, ships and planes with each other forming a large ad hoc network. The best path can be chosen for transmission, from a number of paths.

When there are a number of paths with ships and planes available for connectivity, the best path can be chosen, similar to the concept of the internet.

A. *Summary of Connections*

Summarizing the possible connections between land BTS, ships and planes;

(A) Ground BTS to aircraft connectivity. Dish antenna on ground can be large (and moving) and must point a few degrees above the horizon. Plane antenna will be small.

(B) Aircraft to aircraft connectivity. Can be narrow-direction or omni-directional. Antennas are limited in size and movability, as antennas protrude from aircraft, contributing to drag and fuel loss.

(C) Aircraft to Ship connectivity. Ship antennas must point above by a few degrees, and be compensated for yaw, pitch and roll.

(D) Ground BTS to Ship connectivity. Microwave connection can allow ship to act as new BTS. Only available when ship is a few tens of kilometers from the shoreline.

(D) Ship to Ship connectivity. Antenna must compensate for roll, pitch and yaw.

These are summarized in the table below.

TABLE 1. Tupes of connections for Ships and Planes

	Over land	**Close to land (a few tens of km)**	**Medium distance from Land (100 km)**	**Very far from land (100s of km)**
Planes	Can connect to Ground BTS with small antennas	Connection to land BTS.. Can act as repeater for BTS.	Secondary connection to ships and planes	Connect to ships. Connect to island BTS, if possible
Ship	(not applicable)	To BTS, Point to point Microwave	Secondary connection to ships, connected to land BTS	Connect to island BTS, if possible.

TABLE 2. COMPARISON OF ANTENNAS

	BTS	**Planes**	**Ships**
Antenna Features	Can be elaborate	Must be compact, as antennas increase drag	Moderately compact
Installing New Antenna	Relatively easy	Difficult	Moderately easy
Directionality angle	Can be very narrow	Wide angle	Not-so-narrow
Moving Antenna	Moving possible	Difficult	Moving possible

B. *In the Open sea, with BTS on an Island*

Connectivity becomes difficult when the ships or planes are far from land (hundreds of km) in the open seas. In this case, an available island with can be connected with a solar-powered BTS and a submarine cable. BTS connections on uninhabited island may be problematic, as salty dust may accumulate on solar panels.

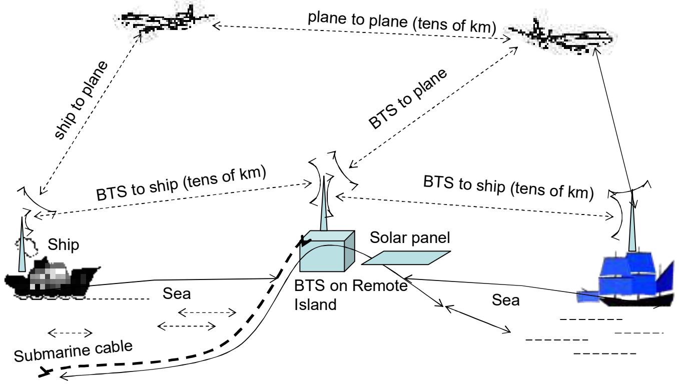


Fig. 10 . In the open sea, far from the mainland, a fiberoptic-connected solar-powered BTS can be set up on an island, for, connecting to ships and planes

With point-to-point distances between microwave at 50 - 80 km, it may be possible to extend the ad hoc network a hundreds of km into the sea (figure below).

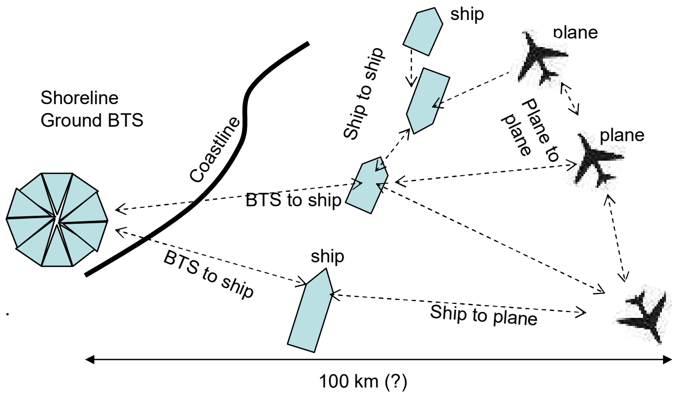


Fig. 11. Top view of a network of BTS, ships and planes, that can go perhaps a hundred km or more into the sea, depending on the connectivity of microwave.

C. *Challenges and Limitations*

As with any new technology, there are numerous challenges and limitations that must be overcome..

1. The large distances between ships and between planes may require specialized technology.

2. The proposed technology of this paper depends on the continuously tracking (or moving) narrow-angle dish antenna on ships and on ground-based BTS. Such tracking and movement may be difficult for antenna on aircraft.

3. Both ships and planes have enough power to have powerful transmitters and receivers.

4. Tracking of antennas requires moving parts and continuous movement, which could be a problem. But movement can be avoided by having numerous narrow-angle dish antennas.

5 As there may be considerable roll on ships, their antennas may have to have wider-angle dish antennas. Or there can be compensating motion of the antenna to keep it horizontal and pointing in the right direction. Rewriting, antennas must:

(a) have wider directionality to compensate for the rolling and yawing of ships.

(b) or move and compensate so as to keep connection with BTS or ships or planes.

6. The antennas on planes generate wind resistance, and may be required to be smaller and non-moving.

7. Short distances between ships and planes may allow nonmoving antennas with wide directionality. Long distances between ships and planes may require moving and tracking antennas with very narrow directionality.

8. The main problem with moving antennas is that that they must keep moving all the time, and that salty dust will lead to rapid corrosion and will require frequent maintenance.

VII. COMMERCIAL IMPLEMENTATION

With all the general strategies presented above, how are they to be implemented in practice over the next few years? Commercial airlines and passenger ships may be given an incentive not only to buy connectivity for themselves, but to act as part of a larger connectivity backbone. It is in the interest of the ships and planes to make the ad hoc networks as large as possible, because any ship or plane may be at the outer edges of this ad hoc network.

Opportunities and legislation should be provided so that a maximum number of companies can compete with each other to provide connectivity services. Rules and standards should encourage competition. Allowing multiple players would avoid monopoly and syndication, which are easy traps to fall into, considering the global backdrop.

CONCLUSION

Ships and planes today mostly rely on the very limited bandwidth of satellites, which gives slow connectivity to their many passengers.

Land BTS are connected to high-speed microwave or fiberoptics, meaning planes and ships close to the shore can connect to land BTS. Planes in proximity, especially in common flight paths can form ad-hoc networks, Ships can form ad hoc networks, especially with shoreline BTS. Ships in common routes can help in forming ad hoc networks. With multiple channels available between distant ships and planes, the best path can be chosen for transmission; which is also the principle of the internet.

Microwave communication with point-to-point antenna can be mostly used, in preference to cellular communication, Antenna on planes must be compact to reduce drag, meaning specialized compact antenna must be used. In comparison, dish antenna on BTS and on ships can be large and moving (for tracking).

Numerous narrow-angle non-moving dish antennas can switch automatically to track a moving ship or plane. Or else, an antenna can keep moving to track a ship or plane. Ships and planes have the extra power for all this communication, which a cell phone does not have.

Isolated islands far into the sea can have a BTS installed with the help of submarine cable and solar panels.

Today's technical capabilities suggest we on the verge of a Wild West of new technologies for connectivity on ships and planes. A major revolution in the connectivity may be just on the horizon, allowing connectivity hundreds of km from the shore. It is hoped that this paper will help in the design and implementation of these new technologies. A number of competing technologies are likely to emerge, followed by survival of the fittest. New policies and standards should encourage competition and discourage monopoly and syndication.

ACKNOWLEDGEMENTS

Thanks are due to undergraduate student Toufiqul Islam Bilash, for his enthusiasm and cooperation.

REFERENCES

[1] Shahriar Khan, Telecommunication Engineering, ISBN: 978-984-33-6164-6, , S. Khan, Dhaka, Bangladesh, June 2016,

[2] "How does WiFi work at 35,000 feet, and why don't all airlines offer it? The Telegraph, Jan 30, 2017.

[3] Elizabeth Manneh, How Does Inflight Wifi Work, Anyway? Reader's Digest, Updated: Mar. 18, 2021

[4] Sadman Shahriar, Md. Erfanul Haque Sajib, Md. Toufiqul Islam Bilash, Shahriar Khan, Ad Hoc Network for Ships and Planes (Student poster), SS 12 Innovation Challenge and Maker Fair 2018 (IEEE BDS Pilot), Independent University, Bangladesh, July 20, 2018.

[5] H. Li, B. Yang, C. Chen, "Connectivity of Aeronautical Ad hoc Networks," 2010 IEEE Globecom Workshops, 6-10 Dec. 2010, Miami, Florida, USA.

[6] E. Sakhaee, A. Jamalipour, Nei Kato, "Aeronautical ad hoc networks," IEEE Wireless Communications and Networking Conference, WCNC 2006., Las Vegas, NV, USA, 2006, pp. 246-251.

[7] J. Wu, I. Stojmenovic, "Ad hoc networks." Computer, 2004 Aug 2, 37 (2), pp 29-31.

[8] V. Naumov, T. R. Gross. "Connectivity-aware routing (CAR) in vehicular ad-hoc networks." IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications 2007 May 6, pp. 1919-1927.

[9] J. J. Blum, A. Eskandarian, L. J. Hoffman, "Challenges of intervehicle ad hoc networks," IEEE Transactions on Intelligent Transportation Systems, vol. 5, no. 4, Dec. 2004, pp. 347-351.

[10] S. Pathmasuntharam, J. Jurianto, P. Kong, Y. Ge, M. Zhou and R. Miura, "High Speed Maritime Ship-to-Ship/Shore Mesh Networks," 2007 7th International Conference on ITS Telecommunications, Sophia Antipolis, 2007, pp. 1-6.

[11] C. Yun, Y. K. Lim, "ASO-TDMA: ad-hoc self-organizing TDMA protocol for shipborne ad-hoc networks," EURASIP Journal on Wireless Communications and Networking. Dec;2012 (1), pp. 1-3.

[12] M. T Zhou,. H. Harada, "Cognitive maritime wireless mesh/ad hoc networks." Journal of Network and Computer Applications, 35(2), 2012, pp. 518-526.

[13] Ardupilot.org, Ardupilot, Antenna Tracking, Accessed April 5, 2021.

[14] flightradar24.com, Live Flight Tracker - Real-Time Flight Tracker Map, Accessed April 2021.

[15] MarineTraffic.com, Global Ships Tracking Intelligence, Accessed April 2021

Detecting Secret Messages in Images Using Neural Networks

Nour Mohamed
Research Institute of Science & Eng.
University of Sharjah
Sharjah, United Arab Emirates
u18105848@sharjah.ac.ae

Tamer Rabie
Dept. of Computer Eng.
University of Sharjah
Sharjah, United Arab Emirates
trabie@sharjah.ac.ae

Ibrahim Kamel
Dept. of Computer Eng.
University of Sharjah
Sharjah, United Arab Emirates
kamel@sharjah.ac.ae

Khawla Alnajjar
Dept. of Electrical Eng.
University of Sharjah
Sharjah, United Arab Emirates
kalnajjar@sharjah.ac.ae

***Abstract*—Image-based applications are widely spread nowadays. Steganography is a type of data hiding methods that manage covering the presence of a confidential communication between two ends. That is obtained by concealing the secret medium into a cover medium to deliver a stego medium that ought to be unnoticeable to a third party. The contrary of steganography, image steganalysis, is about identifying the presence of stego images. Image steganography schemes are becoming more and more secure every day. Cyber criminals can utilize these schemes to conduct a secret and malicious communication. Therefore, image steganalysis is of great importance to interfere such communications. In this paper, a transform domain based steganalysis scheme is proposed that utilizes the architecture of AlexNet, which is an object classification DL scheme. Some modifications are performed on the existing AlexNet architecture to enhance the detection performance of the model. Experiments showed that FB-GAR steganography scheme was successfully detected with an accuracy of 74.72%. Also, the relationship between the capacity and the quality of stego images was studied in this paper for both FB-GAR and J-UNIWARD. Furthermore, the relationship between the correlation of the cover images and the capacity and quality of stego images was discussed.**

***Index Terms*—Steganography, Steganalysis, DCT, CNN, Accuracy.**

I. Introduction

A type of data hiding methods is steganography that manages to cover the presence of a confidential communication between two ends. In steganography, a stego medium that ought to be unnoticeable to a third party is produced by hiding the secret medium in a cover medium. Establishing this secret communication by steganography can be done by utilizing text, audio, or pictures as the cover medium [1]. The two main categories of steganography are spatial domain steganography and transform domain steganography. First, spatial steganography embeds the secret into the pixels of the cover image. Secondly, transform domain steganography embeds the secret into the frequency coefficients of the cover image. Throughout the past few years, there was a noticeable increase in the complexity of steganography schemes and therefore resulted in profoundly imperceptible stego images [2].

On the other hand, recognizing the presence of stego images and subsequently uncovering the mysterious communication between two entities is called image steganalysis, which is the contrary of steganography [3]. Spatial and transform steganalysis are the types of image steganalysis. With the expanding utilization of steganography by cyber criminals, image steganalysis is also advancing to prevent their malicious actions. A type of Neural Networks (NN) called Convolutional Neural Networks (CNNs) is being used recently in image steganalysis. Nevertheless, research in spatial domain steganalysis has commanded more notice than transform domain steganalysis [3].

Recognizing the utilization of steganography schemes that use Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), or Discrete Wavelet Transform (DWT) in the hiding procedure is the main goal of Transform domain steganalysis. DCT steganalysis has a well-known sub-classification called JPEG steganalysis. In addition, the quantized DCT coefficients of a cover image is where the embedding of secret data in JPEG image steganography happens. Besides, spatial domain steganalysis distinguishes embedding changes made to the pixels directly.

The odds of image steganography being utilized increment considerably more than other media because images are broadly utilized in the current communication. Many different cover image formats like BMP, GIF, JPEG and so forth can be used in image steganography. JPEG format is used by a lot of image capturing tools. A well-known JPEG image Steganographic scheme that is publicly available is J-UNIWARD [4].

The aim of this paper is to utilize the architecture of an existing CNN that was designed for a certain purpose and use it for image steganalysis and evaluate its performance. There-

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

fore, this paper presents a transform domain based steganalysis scheme that utilizes the architecture of Alexnet [5], which is an object classification DL scheme that classifies objects into 1000 classes. Some modifications are performed on the Alexnet architecture to enhance the detection performance of the model. The model aims to detect two steganography schemes, J-UNIWARD [4] and FB-GAR [6] namely.

The rest of this paper is arranged as follows: the background information is discussed in Section II, while Section III presents the steganalysis scheme and discusses the experimental results and setup. Finally, concluding remarks are found in Section IV.

II. Background

Iin image and signal processing, a broadly used transformation is the Discrete Cosine Transform (DCT), because of its “energy compaction” property. Most of the image data is located in the top-left area of the domain, that is the low-frequency coefficients region [7]. An extension of the 1D-DCT is widely used in image processing, the 2-dimensional DCT (2D-DCT), since images are two-dimensional signals. 2D-DCT is useful in steganalysis since it seperates the low frequency from the high frequency content in an image.

Artificial Intelligence (AI) has shown significant advancement in the course of recent many years. It has significantly decreased the gap between the tasks that ought to be performed by people and machines. Deep Learning (DL) is a subcategory of AI algorithms [8]. One class of DL algorithms is CNN. The input of a CNN is an image, and then it specifies adaptable weights to specific characteristics of the image to differentiate between different input images. Orthodox classification algorithms utilize hand-engineered filters. On the other hand, CNNs are capable of learning these features. Furthermore, CNNs need less pre-processing in comparison with other classification algorithms. The CNN used throughout this paper is AlexNet.

XuNet, a 20-layer CNN for JPEG steganalysis was proposed in [9]. The pre-processing of the input images was done using a fixed high-pass DCT filter. The shortcut connection in [10] was utilized to attain a deep structure, while instead of pooling layers, (3x3) convolutions with a stride of two were used. Chen et al. in [11] presented a CNN for JPEG steganalysis that is phase-aware. They divided the feature maps into 64 parallel channels, and used a few high-pass filters including KV filter, a point filter, and two Gabor filters. BOSSbase [12] and BOWS2 [13] databases were used in the experiments that showed improved detection performance.

Two designs were presented by Yang et al [14]; the first design is 32-layer CNN architecture whereas the second design is an ensemble architecture (CNN-SCA-GFR). The 32-layer CNN module reuses features by concatenating all features from the preceding layers to improve the flow of gradient and data as well as reduce the amount of parameters used. The CNN-SCA-GFR model is a combination of their proposed deep CNN and the traditional SCA-GFR technique [15]. BOSSbase [12], BOWS2 [13], and ImageNet [16] databases were used and J-UNIWARD [4] and UERD [17] steganography techniques. The ensemble design obtained better performance than the 32-layer CNN because of the directional features from the SCA-GFR technique with Gabor filtering also from the CNN model, the non-directional feature maps.

III. Experiments

Alexnet [5] is an eight layer deep convolutional neural network that was designed for object classification. The network was trained using more than one million image from the ImageNet database. A pretrained version can be found online that can classify input images of size of (227x227) into 1000 classes. These classes include keyboards, pencils, and a lot of animals. Details on the architecture can be found in [5].

In this paper, the AlexNet layer layout was used as the basis model to detect the use of steganography. The layers contain convolutional layers, max pooling layers, ReLU layers, and cross channel normalization layers. The layers were trained on the BOSSbase and BOWS2 datasets as displayed in section III-B. The first layer was modified to an image input size of (256x256) and only one channel as opposed to the original size which was (227x227) and 3 channels. Another modification was done in the fully connected layer since it obtained 1000 classes, therefore it was changed to classify into two classes (stego and cover). In addition, DCT was used to pre-process the data before training and testing the network. The reason behind using DCT is because it distinguishes between the high frequency content of the image and the low frequency content, which helps the netwrok in classifying the images and detecting the use of steganography. The definition of the 2D-DCT for an input image A and output image B is [18] [19]:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} cos \frac{\pi(2m+1)p}{2M} cos \frac{\pi(2n+1)q}{2N}$$

$$where\ \ 0 \le p \le M-1\ \ and\ \ 0 \le q \le N-1$$

$$where\ \ \alpha_p = \begin{cases} \frac{1}{\sqrt{(M)}} & p = 0 \\ \sqrt{(\frac{2}{M})} & 1 \le p \le M-1 \end{cases}$$

$$and\ \ \alpha_q = \begin{cases} \frac{1}{\sqrt{(N)}} & q = 0 \\ \sqrt{(\frac{2}{N})} & 1 \le q \le N-1 \end{cases}$$

M and N are the row and column size of A, respectively. (1)

All experiments were done using MATLAB. The training parameters used were: stochastic gradient descent as the solver with 0.9 momentum, initial learn rate of 0.001, mini batch size of 32, and 15 training epochs. The training data is 70% of the total data.

A. Evaluation Metrics

One of the parameters that is controlled and changed throughout this work is the capacity. The capacity of a steganographic technique is how much hidden data is in the stego image. Naturally, the higher the capacity, the easier it

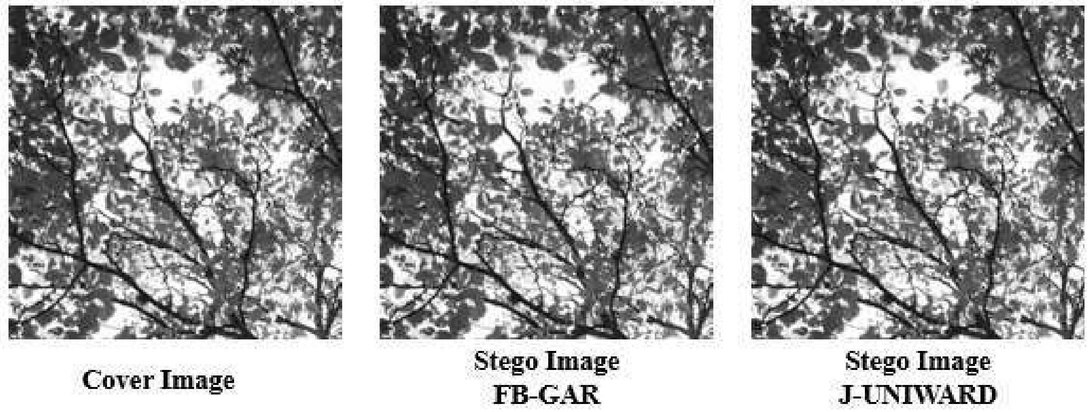


Fig. 1: An example of the stego images generated by FB-GAR and J-UNIWARD steganography schemes for a capacity of 1 bpp.

becomes to detect anomalies in the stego image. The capacity is measured in bits per pixels (bpp) as well as bits per non zero AC coefficients (bpnzAC). The formula below is used to calculate capacity if the embedding process is done in grayscale while the capacity is multiplied by three if the hiding process is done in RGB colorspace.

$$Capacity\,(Bpp) = \frac{(Secret\,width \times Secret\,height \times 8)}{(Cover\,width \times Cover\,height)} \tag{2}$$

The Peak-Signal-to-Noise-Ratio (PSNR) uses the original cover image as reference and compares it with the stego image. PSNR measures the quality of reconstruction in an image. It represents the ratio of the highest power of a signal over the power of noise. The formula for PSNR is:

$$PSNR\,(dB) = 10Log_{10}\left(\frac{255^2}{MSE}\right) \tag{3}$$

Where MSE is the mean square error of comparison between the stego and cover image.

Structural SIMilarity (SSIM) is another metric for comparison between the stego and cover images. It demonstrates whether the structure of the stego image has changed or not in comparison to the structure of the cover image.

The main performance parameter used to evaluate the efficiency of the steganalysis scheme is the detection accuracy, which is defined by equation 1.More details on the paramters can be found in [20]:

$$Detection\ Accuracy\,(\%) = \frac{(TP + TN)}{(P + N)} \tag{4}$$

The classification error probability is another metric that has been used within the steganalysis society and it is calculated as:

$$P_E = min_{P_{FA}} \frac{1}{2}(P_{FA} + P_{MD}) \tag{5}$$

Where P_{FA} represents the false-alarm (false positive) probability and P_{MD} represents missed-detection (false negative) probability.

B. Database

The databases used for the training process were modified from the BOSSBase [12] and BOWS2 [13] databases. Each of these databases consists of 10,000 images, each image was divided into four blocks using MATLAB resulting in a total of 80,000 images of size 256 x 256. The division of the images was based on two reasons: to increase the database size and to avoid losing the details in the images by resizing them to a very small size.

The created database was used to generate cover/stego pairs using two steganography techniques: J-UNIWARD [4] and FB-GAR [6]. The FB-GAR scheme was modified to work on gray-scale images as opposed to the original RGB. It was also modified to limit the capacity to ensure fair comparison between the two steganography schemes given that it was originally proposed to maximize the capacity. The fixed block size used in the FB-GAR scheme is 128 x 128. The capacity used to generate the datasets ranged from 0.1 bpp to 1.0 bpp in 0.1 intervals. Figure 1 displays an example of stego images generated by both steganography schemes.

C. Results

Experiments were done to examine the quality of stego images using both steganography techniques. Figure 2 displays the three cover images used in the experiments. These images portray a wide range of frequencies, cover 1 is highly correlated, cover 2 is medium, while cover 3 is uncorrelated.

We can conclude from Figure 3 that for highly correlated images (a) J-UNIWARD outperforms FB-GAR in both comparisons, while for uncorrelated images (c) it still outperforms FB-GAR in terms of PSNR but the gap in PSNR is less, and FB-GAR outperforms J-UNIWARD in terms of SSIM when the capacity exceeds 0.7 bpp. An important observation is that

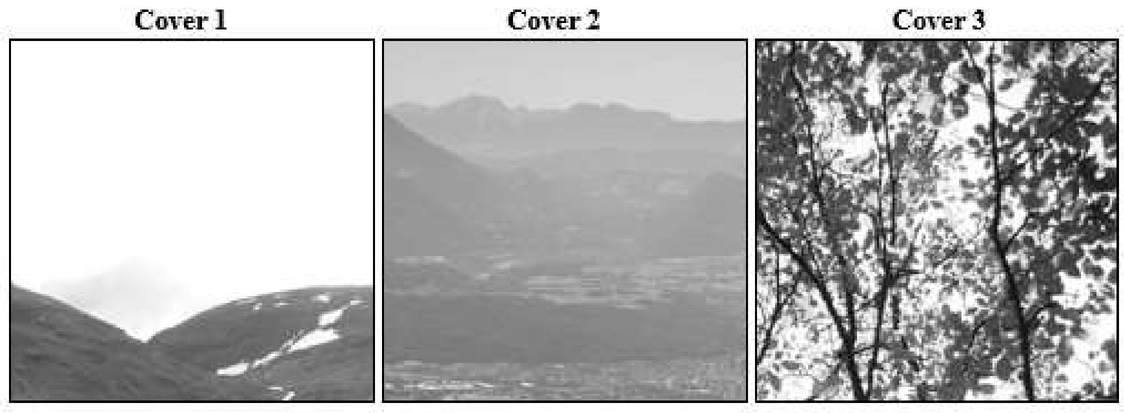


Fig. 2: The different cover images used to compare the quality of the stego images generated by J-UNIWARD and FB-GAR.

both schemes perform better with correlated cover images, since they both utilize DCT in the hiding process.

After training the network, 30% of the data was used to test the performance of the network. Figure 4 summarizes the performance in terms of detection accuracy and error vs capacity for FB-GAR. It is evident from the graphs that the higher the capacity, the higher the detection accuracy, and the lower the detection error Pe. The AlexNet layout has succeeded to detect the FB-GAR steganography scheme with accuracy of 74.72% for 1 bpp. On the other hand, J-UNIWARD was not detected throughout the experiments since it is more secure than FB-GAR. In addition, some promising experiments are being conducted to detect J-UNIWARD after modifying the AlexNet architecture.

IV. Future Work and Conclusions

Image steganography schemes are becoming more and more secure every day. Cyber criminals can utilize these schemes to conduct a secret and malicious communication. Therefore, image steganalysis is of great importance to interfere such communications. The relationship between the capacity and the quality of stego images (PSNR and SSIM) was studied in this paper for both FB-GAR and J-UNIWARD. Also, the relationship between correlation of the cover images and the capacity and quality of stego images were examined. AlexNet architecture was used to detect the use of the two mentioned steganography schemes, and FB-GAR was successfully detected with an accuracy of 74.72% for 1 bpp. For future work, some promising experiments are being conducted to detect J-UNIWARD after modifying the AlexNet architecture.

References

[1] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*. Morgan kaufmann, 2007.

[2] M. Chaumont, “Deep learning in steganography and steganalysis,” in *Digital Media Steganography*. Elsevier, 2020, pp. 321–349.

[4] V. Holub, J. Fridrich, and T. Denemark, “Universal distortion function for steganography in an arbitrary domain,” *EURASIP Journal on Information Security*, vol. 2014, no. 1, p. 1, 2014.

[3] T.-S. Reinel, R.-P. Raúl, and I. Gustavo, “Deep learning applied to steganalysis of digital images: A systematic review,” *IEEE Access*, 2019.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[6] T. Rabie and I. Kamel, “High-capacity steganography: a global-adaptive-region discrete cosine transform approach,” *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 6473–6493, 2017.

[7] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[8] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] G. Xu, “Deep convolutional neural network to detect j-uniward,” in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 2017, pp. 67–73.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] M. Chen, V. Sedighi, M. Boroumand, and J. Fridrich, “Jpeg-phase-aware convolutional neural network for steganalysis of jpeg images,” in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 2017, pp. 75–84.

[12] P. Bas, T. Filler, and T. Pevnỳ, “” break our steganographic system”: the ins and outs of organizing boss,” in *International workshop on information hiding*. Springer, 2011, pp. 59–70.

[13] BOWS2. Web Page. (2007). [Online]. Available: http://bows2.ec-lille.fr/

[14] J. Yang, X. Kang, E. K. Wong, and Y.-Q. Shi, “Jpeg steganalysis with combined dense connected cnns and sca-gfr,” *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 8481–8495, 2019.

[15] T. D. Denemark, M. Boroumand, and J. Fridrich, “Steganalysis features for content-adaptive jpeg steganography,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1736–1746, 2016.

[16] ImageNet. Web Page. (2012). [Online]. Available: http://www.image-net.org/

[17] L. Guo, J. Ni, W. Su, C. Tang, and Y.-Q. Shi, “Using statistical image model for jpeg steganography: uniform embedding revisited,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2669–2680, 2015.

[18] A. K. Jain, *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[19] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still image data compression standard*. Springer Science & Business Media, 1992.

[20] N. Mohamed, T. Rabie, and I. Kamel, “A review of color image steganalysis in the transform domain,” in *2020 14th International Conference on Innovations in Information Technology (IIT)*. IEEE, 2020, pp. 45–50.

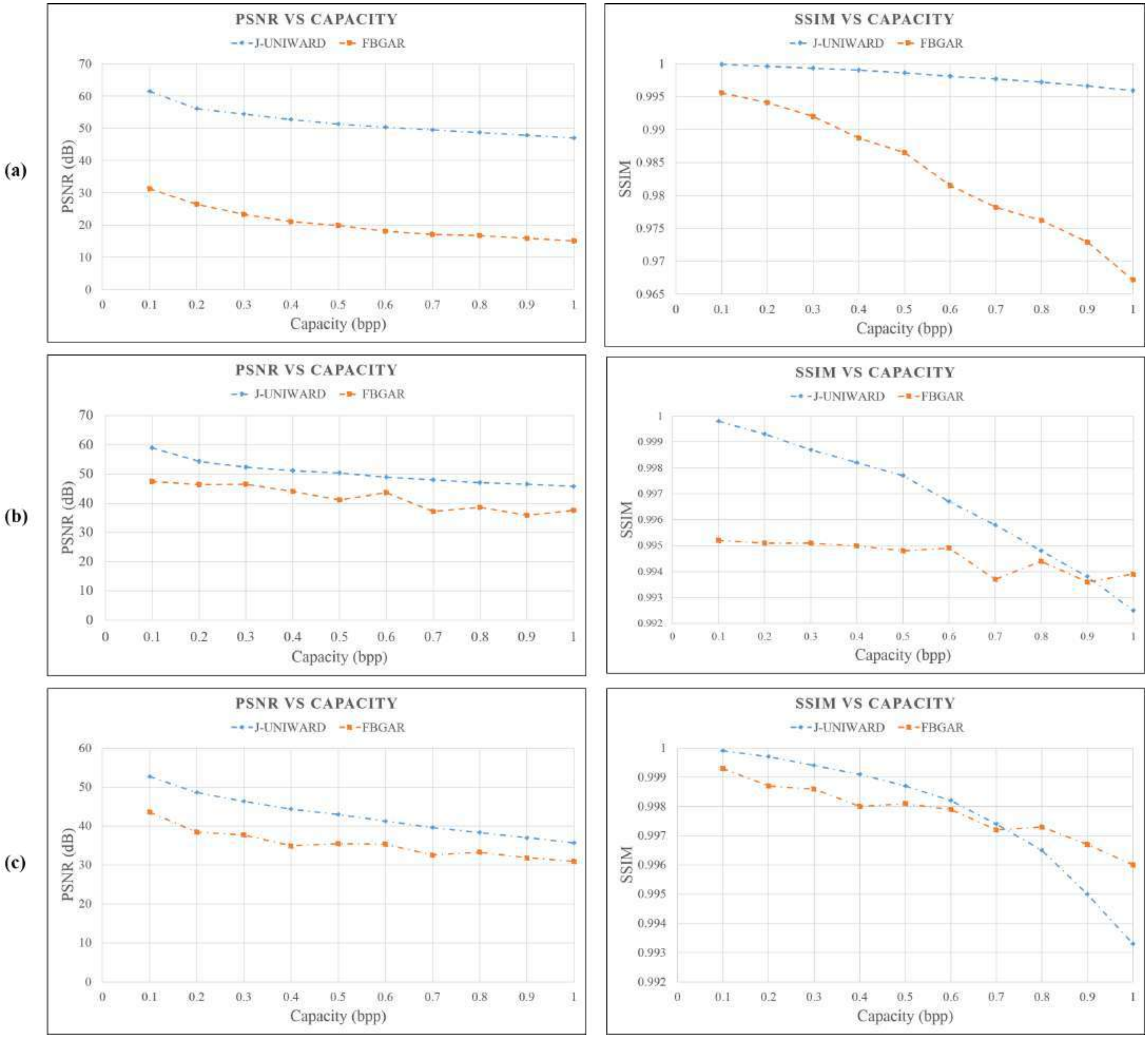
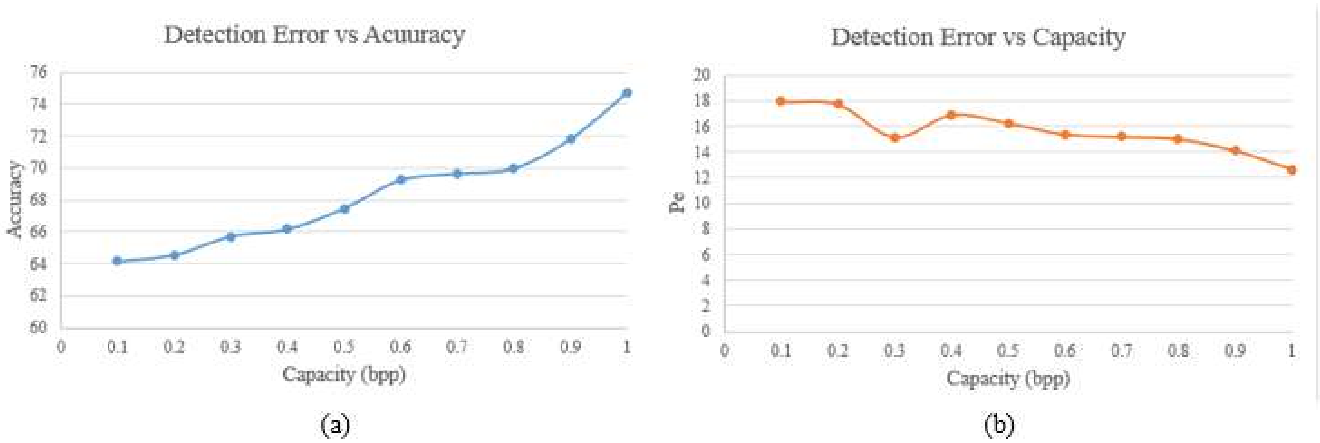


Fig. 3: (a) Comparison between J-UNIWARD and FB-GAR in terms of PSNR vs capacity and SSIM vs capacity for cover 1. (b) Comparison between J-UNIWARD and FB-GAR in terms of PSNR vs capacity and SSIM vs capacity for cover 2. (c) Comparison between J-UNIWARD and FB-GAR in terms of PSNR vs capacity and SSIM vs capacity for cover 3.



(a)

(b)

Fig. 4: (a) The network performance in terms of detection accurcay. (b) The network performance in terms of detection error.

Habitability of Exoplanets using Deep Learning

Rutuja Jagtap, Unzela Inamdar,
Shivani Dere, Maziya Fatima
Computer Science and Engineering
MIT School of Engineering, MIT ADT University
Pune, India

Prof. Nikhilkumar B Shardoor
Assistant Professor, Dept. of CSE,
MIT School of Engineering, MIT ADT University
Pune, India

Abstract - In observational astronomy, researchers have reached a point where large datasets from sky surveys are regularly published. Sending probes to land on asteroids, collect material, and ship it back to Earth is one strategy for space exploration. Astronomical data is rapidly increasing in size and complexity as space and ground-based telescopes are developed and deployed. Machine learning has gained popularity among astronomers in recent years, and it is now used to solve a variety of tasks, including classification, regression, clustering, outlier detection, time series analysis, association law, and so on. The study on previous work of papers on exoplanets and their habitability show machine learning methods including Support Vector Classification, Random forest, K- Nearest Neighbor, and so on were used in the majority of the papers. Until now, only a few deep learning methods have been studied. As a result, this research work overcomes different challenges faced by astronomers dealing with large data and seeking relevant knowledge for each objective using Deep Learning techniques. Furthermore, deep learning techniques are capable of managing complex data with ease. To begin, this paper proposes ASTRONET, a deep learning architecture, to look for exoplanets that are habitable based on their planet gravity, eccentricity, mass, radius, and other characteristics. This research paper explains the entire process of finding exoplanets and categorizing them based on their habitability. Finally, to establish a knowledge base of parameters that affect the habitability of exoplanets.

Keywords – Exoplanets, Deep Learning, Astronet, Classification, Artificial Intelligence, TESS (Transiting Exoplanet Survey Satellite), Astronomy, Machine Learning, Habitability, Convolution Neural Network (CNN)

I. Introduction

Most of us wonder if there is life outside of our planet? We reside in a universe that is colossal beyond our imagination. The problem of detecting anomalies in massive, high-volume astronomical datasets is discussed in this paper, and a solution based on machine learning algorithms is studied and the most impactful and intense deep learning technique is proposed.

The NASA Astrobiology Roadmap has established an understanding of the origin and distribution of habitable planets and moons in the Galaxy as a major research theme (Des Marais et al., 2008). We require more detailed and cohesive models to determine the conditions for the habitability of planets to understand how sustainable the atmosphere and living conditions are or could be, to generate detectable biosignatures, and determine the similarities to the conditions that support life on Earth.

The habitable zone for exoplanets was first presented and modelled in detail by [14], who explained how physics and chemistry play a vital role in the exoplanet study. The immense diversity of exoplanets, as well as the estimated variation in their atmospheres in terms of mass and composition, has sparked a deep desire to rethink planetary habitability.

So, in this paper, the model gives an overview of existing machine learning techniques for assessing exoplanet habitability and suggests the ASTRONET deep learning approach to achieve our objectives.

II. Literature Survey

This paper discusses the effectiveness of various Machine Learning methods in classifying exoplanets into classes of thermal habitability and characterizing them based on possible habitability. The challenges solved are supremacy in the non-habitable planetary sample dataset using under-sampling methods (of dominant class samples) and over-sampling methods [1].

They proposed to build a Machine Learning model to automate Kepler cumulative object of interest data classification and then deploy it. KNN, Random Forest, and SVM models were used, validated, and checked for precision, accuracy, and recall to create a better model for classification. And for model Deployment- Flask API, Azure Cloud were used. It

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

has a Comprehensive ML pipeline: Engineer data, train, and test models. A robust model is required to handle errors [2].

The purpose of this project was to use the data as training data and use planetary and stellar characteristics to construct a machine learning model that can anticipate habitable planets. Data source: From the explanation of NASA Exoplanet Archive Attribute of Kepler data, 14 stellar and planetary features were discovered. Few of them are "Planetary Radius", "Isolation Flux", "Equilibrium Temperature", "Orbital Period", "Distance from parent Star", "Stellar Temperature" SVM with rbf kernel used. But a robust model needed to cope with huge and complex data [3].

In this paper, they used various supervised learning algorithms to predict the habitability of recently observed exoplanets. Used various models like CART, SVM, FNN, Random Forest, Logistic Regression, and Naïve Bayes. A Regression tree was created to anticipate the value of ESI for a planet. It was not suitable for unlabelled data and image data. Further optimization of the model is possible [4].

Applied Deep Learning algorithms to identify Kepler candidate transit cases. Used Kepler DR24, TESS. CNN models used are Astronet: Baseline Model, Exonet: Revised Model, and Exonet-XS: Decreased Model Size. Stellar parameters include metallicity, radius, mass, stellar effective temperature, surface gravity, and density. The model uses two views of phase-folded light curves as inputs, which is followed by completely connected layers that output a value between 0 and 1 that predicts whether transit is a planet or not. Astronet was prone to overfitting [5].

The model can differentiate between false positives and genuine exoplanets, with good accuracy. The training set was derived from the NASA Exoplanet Archive's Autovetter Planet Candidate List. It requires a high training period and high computational power. But it requires a high training period and high computational power [6].

The probability of an observation being an exoplanet is estimated using a number of datasets and classification models. For predicting exoplanets' existence, the proposal aims at using three classifiers. By using different classification algorithms like SVM, ANN, and Naive Bayes Classifiers can obtain different results. It is not suitable for unlabelled data [7].

It concentrates on the implementation of various algorithms of Machine Learning on NASA's Kepler data for the prediction of exoplanet habitability disposition. The proposed model will be able to operate on data generated by different ground and space observatories and classify exoplanet candidates as habitable or non-habitable. It involves the execution of supervised Machine Learning algorithms which include K Nearest Neighbor, Logistic Regression Naive Bayes, Decision Tree, and Random Forest. The binary classification of objects as "FALSE POSITIVE" or "CONFIRMED" exoplanets is taken into account by the model. The decision tree algorithm is a suitable model for predicting habitability [8].

The traditional Machine Learning models like SVM, Random Forest, Decision Tree, etc. used to predict the habitability of exoplanets had certain drawbacks like time-consuming and variation in outcomes.

So, the proposed Deep Learning model i.e "Astronet" is a modern technique backed with high computation capabilities that will provide error-free and unbiased outcomes.

III. Problem Statement

"Exoplanet Habitability" A planet outside the Solar System is known as an exoplanet or extrasolar planet. Habitability refers to an area or region where life can sustain. Every day, a large amount of astronomical data is produced. NASA is a well-known space agency that has similar data. Satellites, radio telescopes, orbital telescopes, space stations, and other instruments are used to collect this information. The instruments assist in the search for new exoplanet information. Additionally, they gather data on the physical states of the exoplanet. The characteristics of exoplanets will be studied using deep learning and machine learning. Further, the model will be able to discover what the key features of an exoplanet are that make it ideal for sustaining life.

IV. Objective

To determine the habitability of exoplanets based on different properties of exoplanets using deep learning algorithms and techniques.

V. Proposed system

There is numerous research work done in machine learning techniques for the classification of habitability of exoplanets. In this paper, the research proposes a deep learning model for the detection and classification of planets based on their habitable behavior. ASTRONET deep learning model is used to predict anomalies in astronomical bodies by classifying whether the studied newly existing planet is habitable or not. The design of the overall module including process flowchart, algorithm, detailed architecture of ASTRONET is explained.

Prediction of habitability of exoplanets using Deep Learning algorithm: Habitability here refers to the quality of having similar properties of being adequate enough to live in. Habitability can be determined using the exoplanet properties like mass, radius, eccentricity, orbital inclination, gravity, metallicity. Exoplanets are planets that do not orbit around our sun but do orbit around a star. So, we put forward the idea of predicting the habitability of exoplanets using the previous huge amount of data present with the help of deep learning techniques.

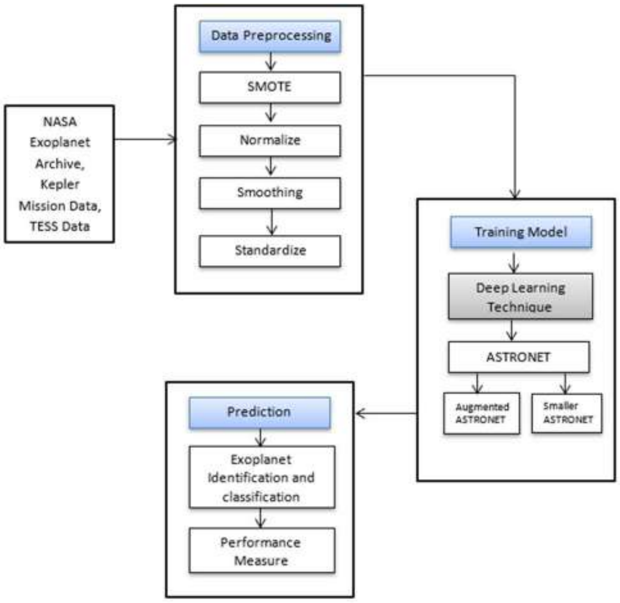


Fig 1. Flow Chart explains the total flow of Exoplanet Classification

ASTRONET:

The proposed ASTRONET architecture to identify and detect potentially habitable exoplanets that support a variety of planet and star characteristics. NASA's Exoplanet Program, NASA's Kepler mission results, and data from the Transiting Exoplanet Survey Satellite (TESS)[9].

These data must be compelled to be efficient, which means they must be reviewed for false-positive signals, such as those caused by stellar eclipses and instrumental noise, which outnumber true planet transit signals.

Let's start with the information gathered by Kepler's telescope, which was used to detect the presence of a planet. A light-weight curve is a graph that displays the brightness of a star (as determined by Kepler's photometer) over time. When a planet moves in front of a star, it blocks some of the suns, causing the measured brightness to drop and then rise again.

Now let's understand the model flowchart (Fig 1) in detail.

A. DATA GATHERING

We start with gathering information from various available data sources. We fetch the data from NASA Archive, Kepler Mission Data, and TESS (Transiting Exoplanet Survey Satellite) data.

B. DATA PREPROCESSING

It is essential to perform data preprocessing before feeding our model with data. This will result in high-quality data or valuable knowledge, which will have a direct effect on our model's ability to learn.

1.SMOTE

SMOTE stands for Synthetic Minority Oversampling. It is a method for eliminating data imbalances. It is introduced to minimize dependency on majority class values.

2.Normalization

The input dataset includes several features with varying ranges and normalization assists in getting them all to a similar scale. The values in the range [0,1] are rescaled.

3. Smoothing

Filters are used to smooth things out. Filters provide robustness in the face of noisy data. This paper aims to increase the accuracy of the data without distorting the signal tendency by averaging out neighboring data points.

4. Standardization

Standardization helps in generalizing the data. By generalizing the numerical conditions of inconsistent data, standardizing the dataset makes the training process more well behaved.

C. TRAINING MODEL

The ASTRONET is proposed which is similar to Deep Convolutional Network [10] to find exoplanets. The diagram below illustrates the Astronet Architecture briefly.

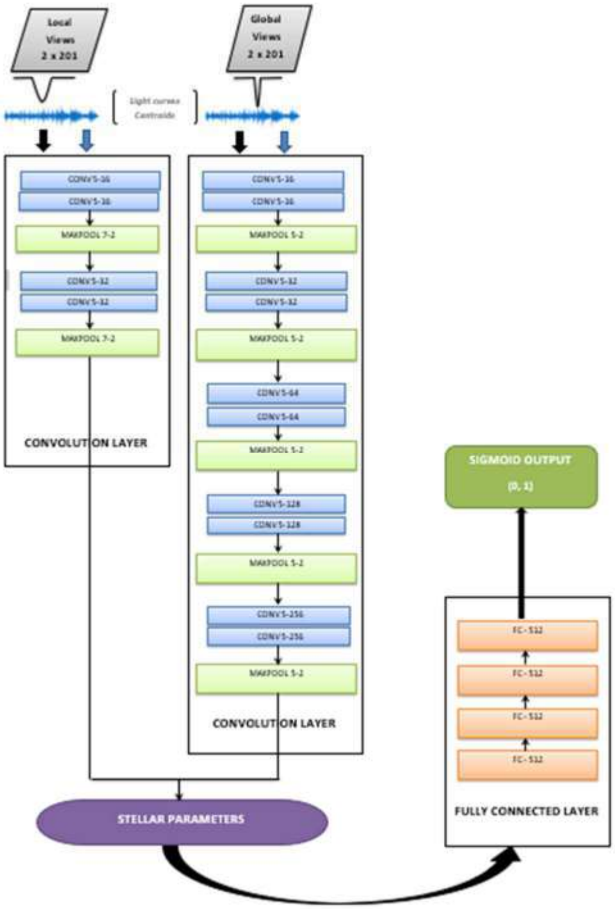


Fig 2. Astronet Model Architecture

The basic architecture is similar to the Convolutional Neural Network.

Step 1: Light Curves

The global and local views of each phase-folded TCE of Light curves [13] are used as the initial input. Each TCE represents a possible exoplanet transit with a particular period, epoch, and length.

Step 2: Convolution Layers

This section contains filters that assist in feature extraction. Every hidden layer of ASTRONET architecture uses ReLU. ReLU is a linear rectifier activation function that produces a linear graph by eliminating all negative values.

Step 3: Max Pooling

By proving an abstract form of representation, Max

Pooling reduces overfitting. It also decreases the number of parameters to learn and provides basic translation invariance to the internal representation, lowering the computational cost.

Step 4: Stellar Parameter

The performance of Max Pooling is combined and fed into stellar parameters. Stellar parameters are important concepts in science. Efficient temperature (Teff), surface gravity (log g), mass (m), density (d), metallicity, and other parameters are among them. These parameters form a set of standardized parameters for data validation.

Step 5: Fully Connected Layer

Stellar Parameter output is given as input to fully connected layers. It compiles the data extracted by previous layers to give the final output.

Step 6: Sigmoidal Output

In CNNs, the sigmoid function is the most common activation function. Finally, the astronet architecture finishes with a sigmoidal feature that generates a number of outputs (0,1). It either classifies the input as true positive exoplanet transit or as true negative exoplanet transit. The graph of the sigmoid function is in the form of an 'S,' and it has a finite limit.

$S(x) = 1 + \frac{1}{e^{-x}}$, where S(x) is the mathematical representation of a sigmoidal function.

Two Astronet variants can be designed and created during the training process.

1. *Augmented Astronet*
2. *Smaller Astronet*

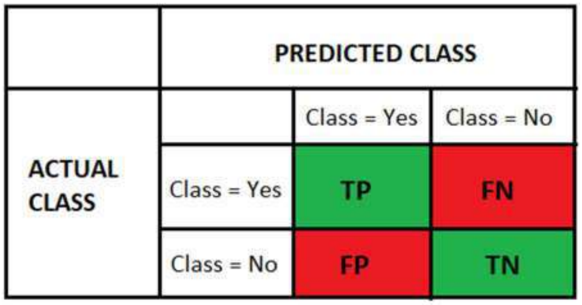
These variants have fewer convolution layers and max-pooling layers. As a result, the model size is reduced, requiring less training time and computing resources.

D. PREDICTIONS

Hyperparameters: Hyperparameters are used to refine the output metrics' values. The required learning rate, epochs, and batch size are all taken into account here. The value may differ from model to model based on the amount of noise present in the dataset. In some cases, drop-out is also added to reduce model overfitting.

Once the model is built it is very important to evaluate how good the model performs. We will evaluate our model using the below matrix and then finally predict in which category the studied exoplanet belongs to i.e habitable or not habitable.

Performance Matrix or Confusion Matrix: This matrix is generated to evaluate the model performance of classification models. It shows the following values-



		PREDICTED CLASS	
		Class = Yes	Class = No
ACTUAL CLASS	Class = Yes	TP	FN
	Class = No	FP	TN

Fig 3. Confusion Matrix

True Positive (TP) - These are correctly predicted positive values that means prediction of actual class is yes and predicted class is also yes. E.g if the actual class says this exoplanet is habitable and the predicted class tells you the same thing.

True Negative (TN) - These are the correctly predicted negative values which means that the value of the actual class is no and value of predicted class is also no. E.g., if the actual class says this exoplanet is not habitable and the predicted class tells you the same thing.

False Positives (FP) – When actual class is no and predicted class is yes. E.g., if the actual class says this exoplanet is not habitable and the predicted class tells you it is habitable.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g., if the actual class says this exoplanet is habitable and the predicted class tells you it is not habitable.

1. *Accuracy*—the number of accurate classifications made in a given period of time. It is the ratio of correctly predicted observations to the total observations.

Accuracy = TP+TN/TP+FP+FN+TN

2. *Recall or sensitivity* — Ratio of true planets identified correctly. Recall is the ratio of correctly predicted positive observations to all observations in actual class.

Recall = TP/TP+FN

3. *Exoplanet precision or positive predictive value*—the percentage of input TCEs identified as exoplanets that are actually planets. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/TP+FP

4. F1 score - F1 Score is the weighted average of Precision and Recall. It measures the test accuracy. It scores maximum value when it tends to reach 1 and minimum value when it tends to reach 0.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Optimizer: To render an algorithm's cost function as minimal as possible. The ASTRONET model recommends using Adam Optimizer (Adaptive Moment Estimation). Adam is a combination of two optimization techniques i.e RMSprop and Stochastic Gradient Descent with momentum. Adam optimizer is chosen as it has less memory requirement, more computational efficiency, easy to implement and well suited for problems with large datasets.

Mathematical approach: Adam algorithm first computes the gradient g_t w.r.t parameters Θ, then computes and stores first and second order moments of gradient, m_t and v_t respectively, as

$$mt = \beta 1 * mt + (1 - \beta 1) * gt$$

$$vt = \beta 2 * vt - 1 + (1 - \beta 2) * gt_2$$

Where β1 and β2 are hyper-parameters that ϵ [0,1], β1 = 0.9 and β2=0.999 are ideally considered. At t=1 m_0 and v_0 are zero. Since initially the values are biased towards zero, we counter it by updating it mt' and vt' as

$$mt' = mt/(1 - \beta t1)$$

$$vt' = vt/(1 - \beta t2)$$

Finally, the parameters updated are computed as

$$\theta t = \theta t - 1 - \alpha * mt'/(\sqrt{vt'} + \epsilon)$$

where is a small stability constant, with standard value of $\epsilon = 10^{-8.}$

Classification: Finally, the model would classify whether the studied planet is a habitable exoplanet or not.

Conclusion

Our ASTRONET model would be an automatic approach to examine catalogs of interplanetary objects. By learning to recognize abnormalities in these bodies in any set, you will be able to recognize anomalies in any set. This model will save researchers a lot of time by supplying them with useful knowledge in just a few seconds. Our mission is to dig deep into planetary bodies in order to discover factors that will enable life to sustain on those planets, as well as to collect a vast amount of data for research purposes.

The approach of deep learning has been accomplished to obtain the major objectives of Exoplanet classification and detection. Firstly, a complex deep learning model is proposed that will give accurate predictions. Secondly, Temperature, humidity, eccentricity, radius, weight, metallic materials, transit signals, and a number of other characteristics can all be used to determine whether or not an exoplanet is habitable and lastly, various performance measures are stated to evaluate the model predictions. To achieve these goals, a comprehensive ASTRONET will be developed and implemented.

References

[1] Suryoday Basak, Surbhi Agrawal, Snehanshu Saha, Abhijit Jeremiel Theophilus, Kakoli Bora, Gouri Deshpande, Jayant Murthy, “Habitability Classification of Exoplanets: A Machine Learning Insight “,2018

[2] Brychan Manry, George Sturrock, Sohail Rafiqi, “Machine Learning Pipeline for Exoplanet Classification”, 2019

[3] Rajeev Mishra, “Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning”, 2017

[4] Karan Hora, “Classifying Exoplanets as Potentially Habitable Using Machine Learning”, 2018

[5] Megan Ansdell, Yani Ioannou, Hugh P. Osborn, Michele Sasdelli, (2018 NASA Frontier Development Lab Exoplanet Team), “Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning”, 2018

[6] Christopher J. Shallue1, Andrew Vanderburg, “Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90”, 2018

[7] Piyush Gawade, Akshay Mayekar, Ashish Bhosale, Dr. Sanjay Jadhav, “Finding New Earths Using Machine Learning & Committee Machine”, 2020

[8] Shivam Pratap Singh, Devendra Kumar Misra, “Exoplanet Hunting in Deep Space with Machine Learning”, 2020

[9] Ricker, Winn, Vanderspek, et al. (2014), "Transiting Exoplanet Survey Satellite," Journal of Astronomical Telescopes, Instruments, and Systems, 014003.

[10] Shallue & Vanderburg (2017), “Identifying Exoplanets with Deep Learning,” The Astronomical Journal, 155, 94.

[11] Saha, S., Agrawal, S., Manikandan, R., Bora, K., Routh, S., Narasimhamurthy, A.: ASTROMLSKIT: A New Statistical Machine Learning Toolkit: A Platform for Data Analytics in Astronomy. arXiv:1504.07865 [cs.CE] (April 2015)

[12] Classifying Exoplanets as Potentially … 211 10. Smith, M.R., Martinez, T.: Improving classification accuracy by identifying and removing instances that should be misclassified. In: The 2011 International Joint Conference on Neural Networks (IJCNN), pp. 2690–2697. (2011)

[13] Wolszczan, A., “Searches for planets around neutron stars”, Celest. Mech. Dyn. Astr., vol. 68, p. 13, 1997.

[14] J. F. Kasting, D. P. Whitmire, R. T. Reynolds, Icarus 101, 108 (1993).

[15] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 321–357

[16] Peng, N., Zhang, Y., Zhao, Y., 2013. A SVM-kNN method for quasar-star classification. Science China Physics, Mechanics and Astronomy 56, 1227–1234.

[17] Borucki, William J., et al.: Kepler planet-detection mission: introduction and first results. Science. 327(5968), 977–980 (2010)

[18] Bloom, J., Richards, J.: Data mining and machine-learning in time-domain discovery & classification. Adv. Mach. Learn. Data Min. Astron. (2011)

[19] Schulze-Makuch, D., Méndez, A., Fairén, A.G., von Paris, P., Turse, C., Boyer, G., Davila, A.F., António, M.R.D.S., Irwin, L.N., Catling, D.: A Two-Tiered Approach to Assess the Habitability of Exoplanet (2011)

[20] Petigura, E.A., Howard, A.W., Marcy, G.W.: Prevalence of Earth-size planets orbiting Sun-like stars. In: Proceedings of the National Academy of Sciences of the United States of America. arXiv: (October 31, 2013)

Current-Mode Self-Sensing Temperature Sensor Using DC-CCII for Optoelectronic Devices

Shahrzad Ghasemi, Soliman A. Mahmoud
Electrical Engineering Department
University of Sharjah
Sharjah, United Arab Emirates
{u18105750, solimanm}@sharjah.ac.ae

***Abstract*— In this paper, a digitally controlled second-generation current conveyor (DC-CCII) and a transimpedance amplifier (TIA) will be employed to implement CMOS temperature sensor interface. The light emitting diode (LED) is used in optogenetic devices for optical stimulation. One of the major concerns of these devices is overheating that can cause cell damage. This work will utilize the LED used for optical stimulation as its own temperature sensor, in order to overcome the overheating concerns in optogenetics. LED reverses current represent temperature-sensitive parameter (TSP) and it can be used to sense the junction temperature. Therefore, the junction temperature can be used to measure the surface temperature. To ensure stable LED biasing and measure the LED reverse current, the CCII circuit will be utilized. In addition, the CMOS temperature sensor interface is operating under a ±0.75 V voltage supply. The standby power consumption of the circuit ranges between 240 µW at a gain for the DC-CCII equals to 0 dB and 700 µW at a gain for the DC-CCII equals to 16.9 dB. The proposed sensor interface has been implemented using LTspice with 0.25-µm CMOS technology.**

Keywords— DC-CCII, LED, optogenetics, sensor interface; TIA, TSP.

I. Introduction

Healthcare and biomedical devices have been improved as a treatment for different neurological disorders. Epilepsy is one of the neurological disorders that can be controlled by implantable optogenetic devices [1]–[3]. The optical stimulation in an optogenetic device can be achieved by human brain light exposure to light sources such as LED. Neural implantable devices are used as a treatment for a wide range of neurological disorders such as deafness, blindness, and motion disorders. For neurological motion disorders such as Parkinson disease, depression, and epilepsy deep brain stimulation (DBS) can be used for recording and stimulation [4]. Therefore, the device safety itself is equally important. The safety of those devices is represented by accidental electric discharge or overheating of the device.

Considering implantable devices overheating, the surface of an optogenetic device must remain below the threshold temperature. The optogenetic device can be used with keeping the device overheating and brain temperature within an acceptable range. In µLED-based optogenetic device, the main challenge is the dissipated heat that can damage human brain tissue. The optogenetic stimulation required acceptable power and acceptable spatial with keeping the brain tissue temperature rise below 0.5 °C. Therefore, a 2.0 °C temperature rise in the probe during stimulation leads to a 0.5 °C rise in the brain tissue [5], [6]. Consequently, to design optogenetic probes for seizures, controlling the temperature rise must be considered. In order to monitor the temperature permanently, the temperature sensor is required in implantable devices.

Several temperature sensors have been implemented using a resistance temperature detectors (RTDs) or by adding CMOS temperature sensor realizations [7]–[10]. These temperature sensors require additional surface space on implantable devices. The LED can be used as its own temperature sensor by employing reverse current (I_R) of the LED as a TSP [11], [12]. The method aimed to use the LED when it is not illuminated. The sensor interface has been developed using CCII for biasing the LED to ensure stable biasing and current sensing. The circuit is realized for an operational frequency less than 130 kHz. Therefore, the reverse current can be used to measure the junction temperature and I_R can be known as TSP [13]. The relation between temperature and reverse current can be driven as

$$I_R \alpha\, e^{\frac{1}{T}}. \qquad (1)$$

Consequently, the additional temperature sensor is not required for the device sensing.

The CCII circuit has been employed in different types of application, Such as wideband waveform generator, filters, oscillators and other instrumentation systems such as different types of sensor interfaces [14]–[18]. In this work, a digitally controllable temperature sensor interface has been implemented using a digitally controllable second-generation current conveyor (DC-CCII) and a transimpedance amplifier (TIA). The reverse current of the LED is nonlinearly and strongly sensitive to the biasing voltage. Therefore, the CCII is used to bias the LED with constant voltage and the TIA is utilized to convert the CCII output current to voltage and amplify the voltage to be suitable for ADC input dynamic range. The sensor interface is designed to be suitable with different optogenetic implantable devices using the current summing network (CSN) at terminal Z of the CCII.

The rest of the paper is organized as follows. The sensor interface circuit and CMOS realization of the self-sensing temperature sensor is firstly covered in Section II. The simulation and results of the proposed circuit using 0.25-µm CMOS technology is then explained in Section III. In Section IV, the paper conclusion has been proposed.

II. Proposed CMOS Self-sensing Temperature Sensor

As mentioned in the introduction, the sensor interface could be realized using DC-CCII cascaded with a TIA. The proposed temperature sensor interface CMOS realization is shown in Fig. 1.The DC-CCII receives the biasing voltage at terminal Y, then it buffers the voltage to terminal X to bias the LED. The current of terminal Z follows the input current received at terminal X which is digitally controlled for gain greater than one. The TIA is used to convert and amplify the Z terminal output current to match the dynamic range of the analog-digital convertor (ADC) block.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

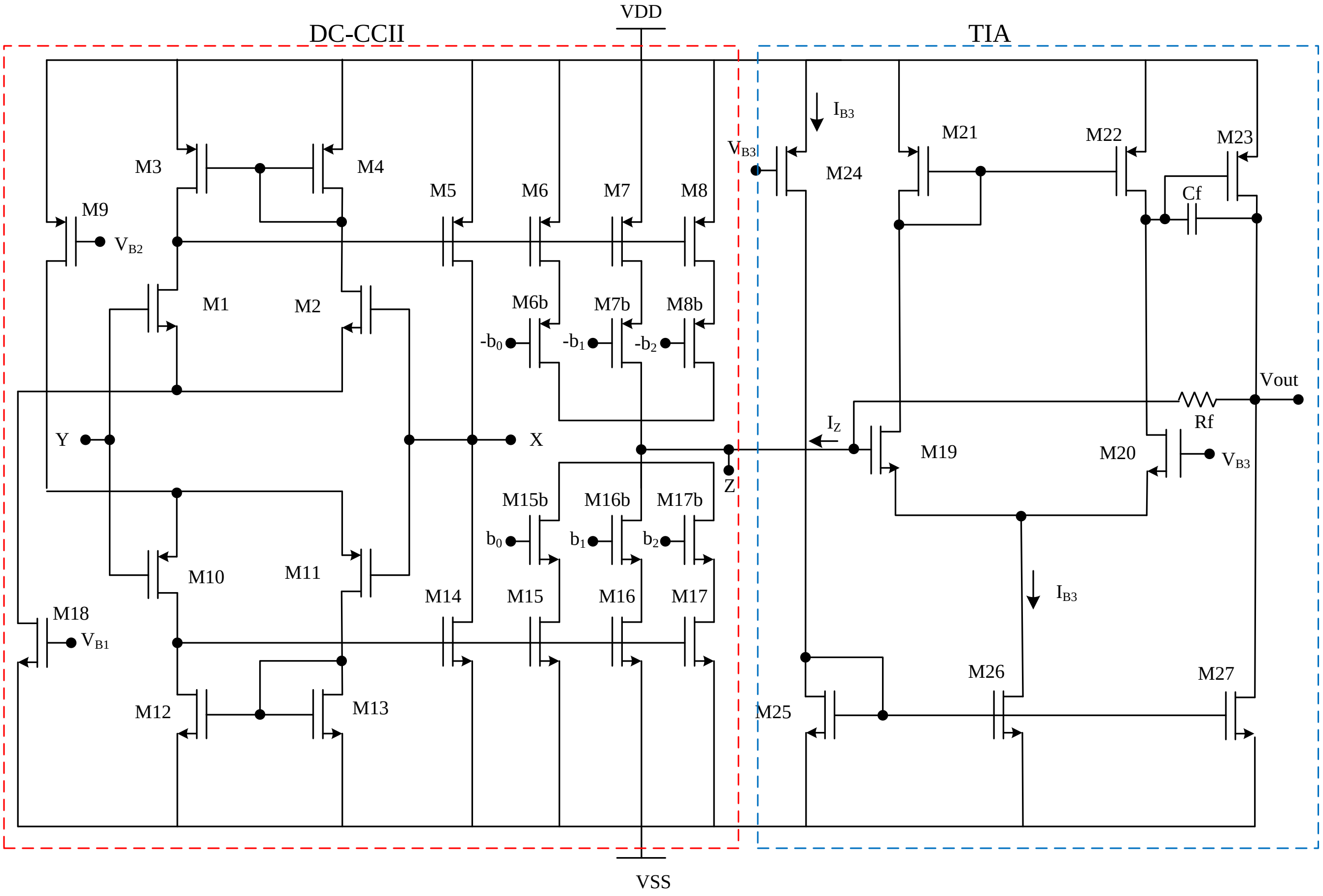


Fig. 1. Proposed CMOS sensor interface.

Based on the CCII introduced by [19], the CCII circuit's input stage has been realized using two complementary differential pairs. It consists of NMOS matched differential pair (M1, M2) and PMOS matched differential pair (M10, M11). These matched differential pair are used to provide voltage follower between terminal X and terminal Y. The biasing currents for the differential pairs are provided by transistors (M9, M18).

Two current mirrors consist of (M3, M4) and (M12, M13) provides constant tail current biasing in order to ensure the rail-to-rail operation. High Y terminal impedance has been provided and the current is equals zero. The matched transistors' pair (M5, M14) and digitally controllable transistors (M6, M15), (M7, M16), and (M8, M17) are used to convey the X terminal current to the Z terminal. The push-pull output stage at the X terminal has been formed using transistors (M5, M14) and it reduce the power consumption.

The DC-CCII realization provides gain greater than and equal to one. The following matric equation describes the DC-CCII with current gain α:

$$\begin{bmatrix} I_Y \\ V_X \\ I_Z \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \alpha & 0 \end{bmatrix} \begin{bmatrix} V_Y \\ I_X \\ V_Z \end{bmatrix}. \qquad (2)$$

By replacing the Z terminal transistors in the CCII with transistor arrays using CSN, the current transfer gain α can be controlled in the DC-CCII circuit. The current gain is expressed as:

$$\frac{I_Z}{I_X} = \sum_{i=0}^{n-1} b_i 2^i, \qquad (3)$$

where b_i indicates the digital control bit and n is equal to 3. Therefore, the proposed circuit is digitally controlled using 3-bits control words.

The output current of the CCII is the input of the TIA. The TIA implemented using two-stage CMOS transconductance operational amplifier (OTA) [20]. The TIA circuit was implemented with differential pair (M19, M20). Transistor M21 mirrors the current to M22 to get single output. Transistors M24, M25, and M26 are used as a current source of the differential pair transistors. The third stage is consisting of transistors (M23, M27) represent the output stage of the TIA. The output stage of the TIA is class A where the current of M27 is constant. The overall transimpedance $\frac{V_{out}}{I_Z}$. All transistors are operating in saturation region.

III. Simulation Results

The proposed CMOS DC-CCII circuit and sensor interface circuit performance was tested and verified by performing LTspice simulations under ±0.75 V voltage supply using 0.25-μm CMOS technology parameters and transistor aspect ratios given in TABLE I.

TABLE I. SENSOR INTERFACE'S TRANSISTORS ASPECT RATIOS

Transistor	Width (µm)	Length (µm)
M1, M2, M10, M11	15	1
M3, M4, M12, M13	0.5	0.25
M9	1	0.5
M18	10	0.5
M5, M6, M14, M15	25	0.25
M7, M16	50	0.25
M8, M17	100	0.25
M19, M20	14	0.6
M21, M22	8.5	0.6
M23	37.5	0.6
M24	9	1.2
M25, M26	7	0.6
M27	33	0.6

The differential pairs (M1, M2) and (M10, M11) are biased through transistors (M9, M18) by setting M9 and M18 biasing voltage -219 mV and 105 mV, respectively. Fig. 2 shows the output voltage at terminal X versus the input voltage at terminal Y where terminal X is terminated with 10 kΩ. Fig. 3 shows the X terminal offset voltage variation versus X terminal input current when Y terminal voltage is equal to zero. The DC-CCII circuit offset voltage is below 12.5 mV and the X terminal input resistance (R_X) is below 18.5 Ω. The bandwidth of the voltage transfer gain is 8.07 MHz 3-dB. The circuit provides total harmonic distortion (THD) for low frequencies around 0.091%.

By including all combinations of 3-bits control word excepting zero, the DC-CCII realization with gains greater than and equal to one is examined. Fig. 4 represents the DC transfer characteristics for an input range of ±10 µA. For the gain equal to one, the X terminal and Z terminal currents are equal. Then, the current of the Z terminal increases with gain α. The magnitude frequency response of the DC-CCII with gain α is shown in Fig. 5.

Considering temperature dependency, the circuit has been tested for different temperatures. In the optical stimulation phase the temperature increases, then in the second phase which is the sensing phase the temperature decreases as shown in Fig. 6. The Z terminal current of the DC-CCII represents the output current of the circuit and the input current of the TIA circuit. The output current for different gains is shown in Fig. 7. The TIA convert and amplify the CCII output current to match the dynamic range of the ADC block. The output voltage is amplified by gain of 200×10^3 V/A. The output voltage of the sensor interface by applying current with temperature variation is shown in Fig. 8. The sensor interface has a 3-dB bandwidth around 291.6 kHz. The input referred noise spectral density of the sensor interface is less than 39 $\mathrm{pA}/\sqrt{Hz}$. The standby power consumption of the circuit ranges between 240 µW and 700 µW. For DC-CCII gain α equal to one, the power consumption is 240 µW. While for gain α equal to 7, the power consumption is 700 µW. The proposed sensor interface simulated results summary is given in TABLE II.

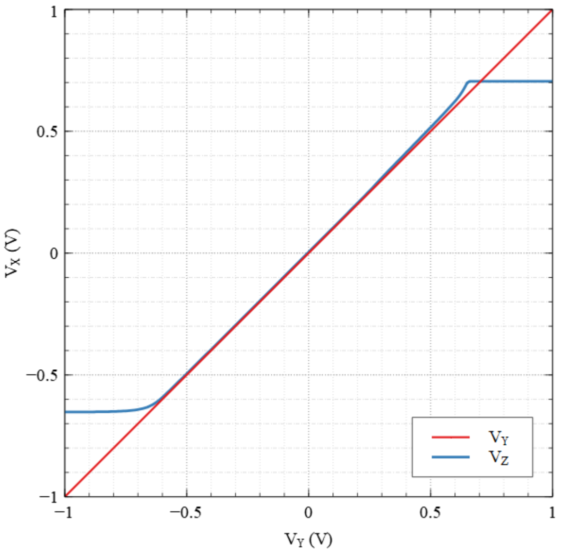


Fig. 2. Terminal X output voltage versus terminal Y input voltage.

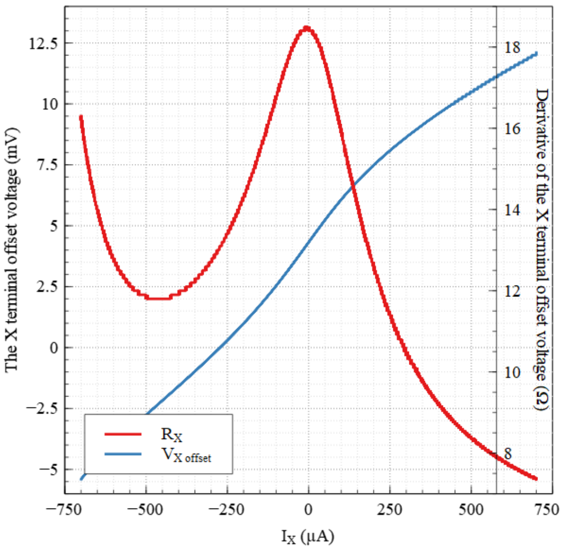


Fig. 3. Terminal X offset voltage and resistance.

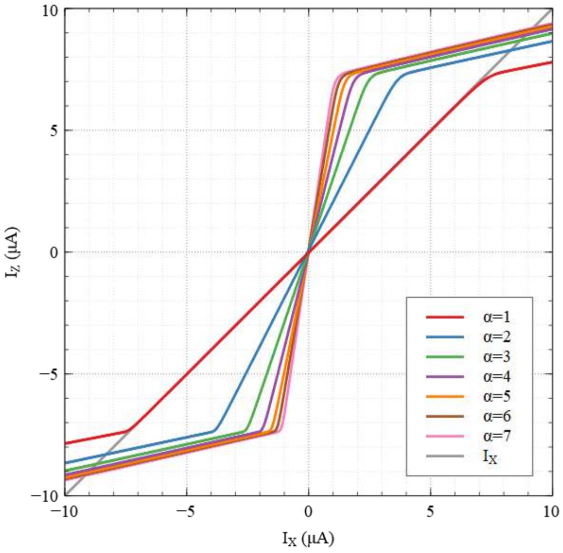


Fig. 4. DC transfer characteristics of the Output current with gain α for α=1-7.

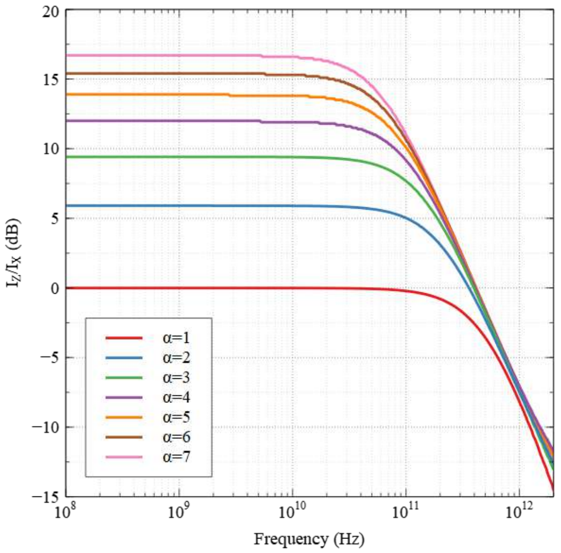


Fig. 5. DC-CCII magnitude response with gain α for α=1-7.

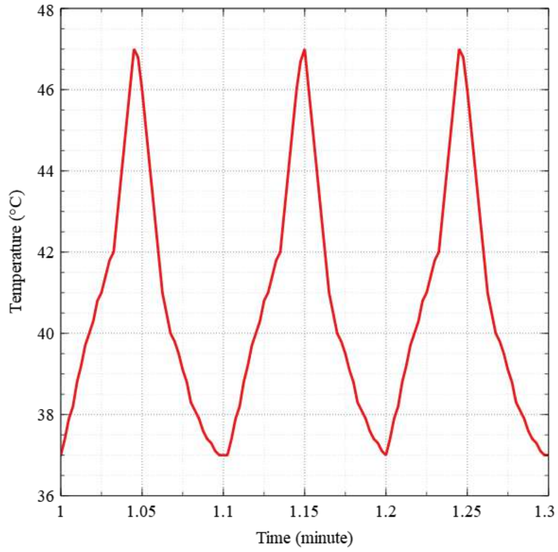


Fig. 6. Junction temperature versus time.

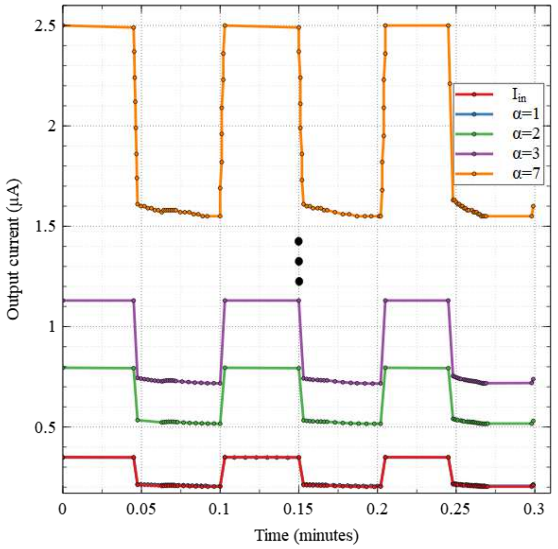


Fig. 7. The DC-CCII output current with temperature variation versus the input current.

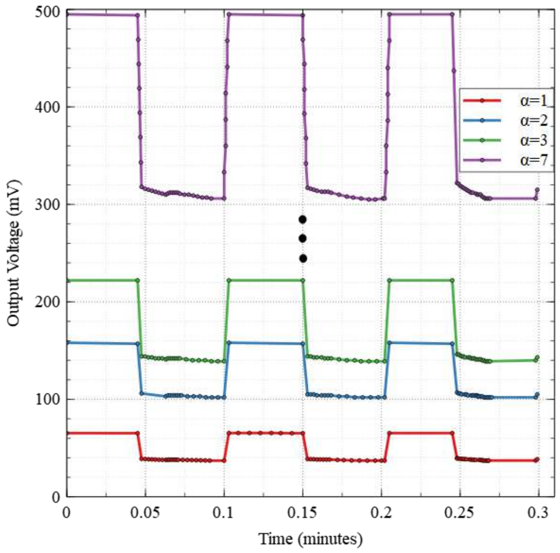


Fig. 8. The time response of sensor interface output voltage.

TABLE II. SENSOR INTERFACE SPECIFICATIONS

Parameters	Proposed Sensor Interface
CMOS technology (µm)	0.25
Power supply (V)	±0.75
Transimpedance gain (V/A)	1.4×10^6
3-dB BW (kHz)	291.6
Power consumption (µw)	240-700
Current driving capability (µA)	±1.23
Input referred noise (below 3 MHz)	39 pA/$\sqrt{Hz}$

IV. CONCLUSION

A CMOS temperature sensor interface has been implemented in 0.25-µm CMOS technology under supply voltage ±0.75 V in order to measure the implanted LED's temperature in optogenetic devices. The reverse current of the LED has been used as a TSP to measure the temperature variation. Therefore, the LED used for optical stimulation has been employed as a self-sensing element. This method is used to reduce the implantable optogenetic device's area by using the LED in stimulation and sensing phases. The CMOS sensor interface has been designed using a digitally controllable CCII to bias the LED and convey the reverse current received from the LED after controlling the current by 3-bits digital word. The circuit used to convey the digitally controlled current to a high gain TIA which is used to convert and amplify the signal. The output of the proposed DC-CCII has a 10 nA offset and the total standby power consumption is less than 700 µW. The achieved circuit characterization results show the CMOS sensor interface is suitable for implantable optogenetic devices which requires low voltage and accurate current conveyor. Future work may include the reduction of the offset current and modification of the digital control circuit. This can be obtained by increasing number of bits and reducing the power consumption.

REFERENCES

[1] S. A. Mahmoud, A. Bamakhramah, and S. A. Al-Tunaiji, “Low-noise low-pass filter for ECG portable detection systems with digitally programmable range,” Circuits, Syst. Signal Process., vol. 32, no. 5, pp. 2029–2045, 2013.

[2] T. B. Nazzal, S. A. Mahmoud, and M. O. Shaker, “A 200-nw

7.6-enob 10-ks/s sar adc in 90-nm cmos for portable biomedical applications," Microelectronics J., vol. 56, pp. 81–96, 2016.

[3] S. I. Khan and S. A. Mahmoud, "Highly linear CMOS subthreshold four-quadrant multiplier for Teager Energy Operator based Sleep Spindle detectors," Microelectronics J., vol. 94, p. 104653, 2019.

[4] P. Degenaar et al., "Optobionic vision-a new genetically enhanced light on retinal prosthesis," J. Neural Eng., vol. 6, no. 3, p. 35007, 2009.

[5] T. M. Seese, H. Harasaki, G. M. Saidel, and C. R. Davies, "Characterization of tissue morphology, angiogenesis, and temperature in the adaptive response of muscle tissue to chronic heating.," Lab. Invest., vol. 78, no. 12, pp. 1553–1562, 1998.

[6] C. Childs, "Human brain temperature: regulation, measurement and relationship with cerebral trauma: part 1," Br. J. Neurosurg., vol. 22, no. 4, pp. 486–496, 2008.

[7] S. Goncalves et al., "LED Optrode with Integrated Temperature Sensing for Optogenetics," Micromachines, vol. 9, no. 9, p. 473, 2018.

[8] J. Wang, H. Xie, T. Chung, L. L. H. Chan, and S. W. Pang, "Neural probes with integrated temperature sensors for monitoring retina and brain implantation and stimulation," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 9, pp. 1663–1673, 2017.

[9] P. C. Crepaldi, T. C. Pimenta, and R. L. Moreno, "A CMOS low-voltage low-power temperature sensor," Microelectronics J., vol. 41, no. 9, pp. 594–600, 2010.

[10] C. Deng, Y. Sheng, S. Wang, W. Hu, S. Diao, and D. Qian, "A CMOS smart temperature sensor with single-point calibration method for clinical use," IEEE Trans. Circuits Syst. II Express Briefs, vol. 63, no. 2, pp. 136–140, 2015.

[11] F. Dehkhoda, A. Soltan, N. Ponon, A. Jackson, A. O'Neill, and P. Degenaar, "Self-sensing of temperature rises on light emitting diode based optrodes," J. Neural Eng., vol. 15, no. 2, p. 26012, 2018.

[12] F. Dehkhoda, A. Soltan, N. Ponon, A. O'Neill, A. Jackson, and P. Degenaar, "A current-mode system to self-measure temperature on implantable optoelectronics," Biomed. Eng. Online, vol. 18, no. 1, pp. 1–15, 2019.

[13] B. Wu et al., "Junction-temperature determination in InGaN light-emitting diodes using reverse current method," IEEE Trans. Electron Devices, vol. 60, no. 1, pp. 241–245, 2012.

[14] A. H. Madian, S. A. Mahmoud, and A. M. Soliman, "New 1.5-V CMOS second generation current conveyor based on wide range transconductor," Analog Integr. Circuits Signal Process., vol. 49, no. 3, pp. 267–279, 2006.

[15] S. A. Mahmoud and E. A. Soliman, "Digitally programmable second generation current conveyor-based FPAA," Int. J. Circuit Theory Appl., vol. 41, no. 10, pp. 1074–1084, 2013.

[16] S. A. Mahmoud, "Fully differential CMOS CCII based on differential difference transconductor," Analog Integr. Circuits Signal Process., vol. 50, no. 3, pp. 195–203, 2007.

[17] S. A. Mahmoud, "New Fully-Differential CMOS Second-Generation Current Conveyer," Etri J., vol. 28, no. 4, pp. 495–501, 2006.

[18] S. A. Mahmoud, H. O. Elwan, and A. M. Soliman, "Low voltage rail to rail CMOS current feedback operational amplifier and its applications for analog VLSI," Analog Integr. Circuits Signal Process., vol. 25, no. 1, pp. 47–57, 2000.

[19] T. M. Hassan and S. A. Mahmoud, "Fully programmable universal filter with independent gain-ω0-Q control based on new digitally programmable CMOS CCII," J. Circuits, Syst. Comput., vol. 18, no. 05, pp. 875–897, 2009.

[20] G. Palmisano, G. Palumbo, and S. Pennisi, "Design procedure for two-stage CMOS transconductance operational amplifiers: A tutorial," Analog Integr. Circuits Signal Process., vol. 27, no. 3, pp. 179–189, 2001.

Encoder-Decoder Model for Automatic Video Captioning Using Yolo Algorithm

Hanan Nasser Alkalouti
Faculty of Computing and Information Technology
King Abdul-Aziz University
Jeddah, Saudi Arabia
halkalouti@stu.kau.sa.edu

Dr. Mayada Ahmed AL_Masre
Faculty of Computing and Information Technology
King Abdul-Aziz University
Jeddah, Saudi Arabia
Malmasre@kau.edu.sa

Abstract— **Humans can use informed visual perception to generate sentences by bridging the gap between the recognition of visual features (images) and linguistic expression (words) describing these images. Videos are an example of visual perception; humans can describe the content of the video in meaningful sentences based on understanding their contents as a caption for the video. However, automating the video caption process is a challenging task as it confronts the model with two problems are: object detection and generating a sentence.**

This research aims to develop a model that automates video captioning based on Encoder-Decoder using a deep learning algorithm following these two steps. Firstly, using the KATNA model to select the most significant frames from the video and remove redundant ones. Secondly, combining the two deep learning algorithms YOLO and LSTM. The You Only Look Once (YOLO) algorithm recognizes objects in the video frames and the Long Short-Term Memory (LSTM) algorithm generates the video caption.

The proposed model describes the video's content in a meaningful sentence and it shows good accuracy and efficiency, it applies YOLO on the MSVD dataset unlike other video captions using other deep learning techniques.

Keywords—(Deep Learning, Natural Language Processing (NLP), Video captioning, You Only Look Once (YOLO)).

I. INTRODUCTION

The information over the internet is growing exponentially hour by hour, and visual content understanding becomes an interesting search area in computer vision, and video captioning is one of its applications. There are a huge number of videos uploaded around the world from different areas at one second through the Internet; these videos need to be arranged and classify based on their captions to facilitate reaching them. It is a challenging task to automate the video caption process; tasks related to video captioning are still considered a challenging research topic. The way video stream is structured and how they are dependent on temporal sequencing, multiple frames, varied objects, and actions and generating accurate sentence complicate the captioning process. Deep learning has currently shown up its efficiency in different areas and it has currently transformed computer vision studies and applications, video caption is one of its applications. Research experimenting with video caption using deep learning techniques develop various language models using LSTM (Long short-term memory), RNN (recurrent neural network), CNN (convolutional neural network), GRU (Gated Recurrent Unit), and TPGN (Tensor Product Generation Network).Recently, deep learning showed up its capability to deal with visual and text contents. Based on that, we propose a deep learning model constructs of encoder-decoder architecture; we compare the performance and the accuracy of using YOLO on the MSVD dataset and other deep learning techniques on the MSVD dataset. Our main objective is to develop an Encoder-Decoder video caption that utilizes YOLO as an encoder, unlike other models that use other deep learning techniques. , and compare final results with previous models using other deep learning techniques.

II. LITERATURE REVIEWS

Video captioning is a process for describing video content using one or more sentences [1]; It translates visual contents to natural language explanation. It starts by recognizing objects and relates them in sentences [2]. Figure1 shows the general process of video captions using Encoder-Decoder. An overview is given in what follows:

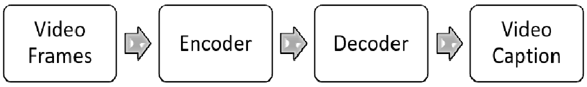


Figure 1. General Structure of Encoder-Decoder Video Caption.

Researchers informed by DL techniques; currently, adapt RNN models, such as LSTM[3] and GRU[4], to act as a decoder of the video clip and learned to generate natural language sentences instead of using a specified template.

In [5], the researchers proposed a RESNET-50 CNN-LSTM Encoder-Decoder for video captioning and an LSTM Encoder-Decoder for sentence generation. There are a different number of video frames; they take samples of each 10 frames to reach an average of 40 processed video frames. In the beginning, the researchers changed the layers' structure of the (CNN) by using the residual function F(x) to enhance performance, and they used it to produce feature vectors of each video frame. A stacked LSTM used for encoding the visual features, and another for decoding the feature output of the CNN to natural language. They used LSTM on both sides of the model as Encoder and decoder; unlike our model, we use it as Decoder and YOLO as Encoder to enhance performance. Besides, it processes a huge number of frames, which consumes time.

In [3], the researcher proposes to enhance the accuracy of video captioning by including Temporal Deformable Convolutional in both Encoder-Decoder. The Temporal Deformable Convolutional (TDConvED) supports combining information of features for a long time by adding fully convolutional to each Encoder-Decoder. In the Encoder, the CNN extracts a feature of video frames to be

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

fed into (TDConvED), model uniformly samples 25 frames of each video. This results in video intervals within context. Mean pooling is used to represent these contexts and send them to the decoder. In the Decoder, stacked shifted convolutional blocks are used to produce a word for each representation. It uses a temporal attention mechanism to help the decoder focus on selected frames based on their weights to produce video captions. It is feed-forward, which means the result from the current layer does not depend on results from previous layers; it affects the accuracy of the final result (caption) in the opposite of using RNN techniques such as LSTM in our model which is back-forward.

In[6], researchers propose a Multimodal Memory Model (M3) for video captions, they proposed a shared memory for both visual (frames) and textual (sentences) and they guide visual attention on described elements to solve visual-textual alignments. The researchers experimented with two datasets: MSVD and MSR-VTT, they uniformly sample 98 video frames to 28, and 149 video frames to 40. Findings demonstrated that their method, when evaluated using BLEU and METEOR, did outperform most of the previously reported methods. It takes a huge number of frames which consume time and there are no specific criteria to choose process frames.

In [2], the Model generates captions based on spatial-temporal attention (STAT); which focuses on some important frames and regions within the video, not whole video frames. Firstly, the encoder network extracts global feature (frame level) using 2d CNN, motion features (frame level) using 3D CNN, and local features (object level (actions)) using faster RCNN from each video frame. Then, the spatial attention mechanism detects the most relevant objects in video frames based on increasing attention weights (sum of global, local, and motion features). After that, the temporal mechanism tracks trajectories of objects detected by spatial and frames; they select frames and regions and send them to the decoder. Finally, LSTM generates sentences with high probabilities of words using beam search. It uses MSVD and MSRVTT-10 datasets, and models evaluated by BLEU, METEOR, and CIDEr.They used RNN techniques on both sides of a model as Encoder and decoder; unlike our model, we use it as a Decoder and YOLO as Encoder to enhance performance.

In [7], they were the first to pick informative frames to be processed by Encoder-Decoder based on Pick Net (two-layer feed-forward neural network) mechanism. Firstly, it transforms each color frame to grayscale and resizes to a small size to produce a "glance" version of frames. Then, subtract the current glance from the previous one (the first frame took to compare with it), results in a flat to fixed-size vector to produce binomial distribution to decide to drop it or keep it. Secondly, kept frames are access to CNN encoder to extract features from them. After that, they use a gated Recurrent Unit (GRU) decoder to generate sentences; effective and performance of generating caption is affected by several selected frames. The model uses Microsoft Video Description (MSVD) and the MSR Video-to-Text (MSRVTT) datasets and researchers evaluated using BLEU, ROUGE, METEOR, and CIDEr. It reduces processing time; it takes between 6 to 8 frames, but other factors should be considered in selecting frames for accuracy.

Many types of research applied deep learning to the MSVD dataset; but they did not apply with YOLO, they use other deep learning techniques as mentioned. Therefore, the researchers in this paper compare their results with the result of the proposed model that YOLO applies to MSVD.

III. Methodology

The proposed model has four main elements: dataset, keyframe extraction, object detection, and sentence generation process. These elements join a sequential process to generate a video caption (Figure 2).

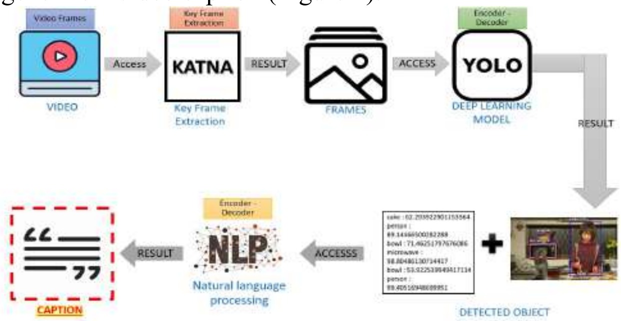


Figure 2: Architecture of Proposed Model.

For the dataset, we choose MSVD; it contains 1970 open muted videos collected from YouTube. The average duration of each video is 5 -25 seconds and it is about one action. There are 41 captions for each video on average. Researchers mostly use it for video captioning, and it is divided into training, testing, and validation sets[9]. Video caption based MSVD is done through the next three processes:

A. Key Frames Extraction (using KATNA):

keyframe extraction is a technical process for capturing meaningful frames from videos. Videos contain heavy contents and there are repeated frames; instead of processing all of them and consuming time, it considers only frames show changes in the video; it is helpful and solves heavy processing problem [10] Model uses KATNA as keyframe extraction; it is open-source code written by python, and it does video extracting frames; it provides summary frames' of video content based on five elements are: LUV color space, degree of brightness, cluster of K-Means, Entropy or contrast filter and blur detection of extracted frames. It has been tested with different types of videos format [11]. As shown in Figure 3, it captures only 3 frames from a video, its duration is 6 seconds. The number of captured frames depends on the video duration and changes that appear within frames.

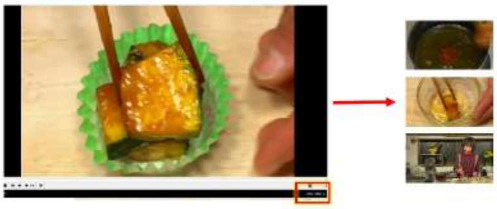
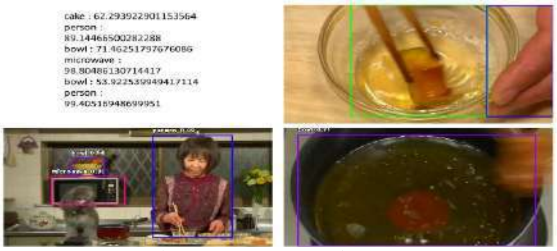


Figure 3: KATNA Result.

B. *Object Detection (using YOLO*

Detecting objects becomes an interesting subject in many areas; fast and accurate detecting is an important factor of any technique. The model needs it to construct sentences based on detected objects.

YOLO is one of the accurate and fast real-time detecting objects technique, it is based on predicting many objects, classifying the type of them and it shows accuracy percent. It does not slide the whole image and it is less error background detection than other deep learning techniques [12]. Figure 3 shows detected objects from YOLO in the model, it supports working with multiple images in one run and it takes 6-12 seconds to process an image based on CPU (it would run faster on GPU) [13].

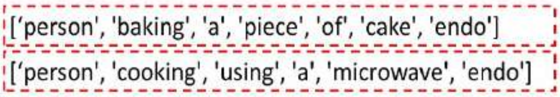


```
cake : 62.293922901153564
person :
89.14466500282288
bowl : 71.46251797676086
microwave :
98.80486130714417
bowl : 53.922539949417114
person :
99.40518948699951
```

Figure 4: YOLO Result.

C. *Sentence Generation based NLP (using LSTM):*

It is a process to construct sentences based on some words. LSTM is a type of RNN, it is back- forward model; which means its current result from the current layer depends on results from previous layers[14]. LSTM tries to relate between result words in a text file; it searches in trained sentences by scanning 3 words at a time in sentences, to find detected objects words; because LSTM is backward – forward technique, every result depends on the previous one. Then, it produces sentences showing the relationship between detected objects words to choose one of them as a video caption. As shown in Figure 5:



```
['person', 'baking', 'a', 'piece', 'of', 'cake', 'endo']
['person', 'cooking', 'using', 'a', 'microwave', 'endo']
```

Figure 5: LSTM Result.

We develop a model based on tensor flow and Keras, we use the Sequential model for NLP; which means, it is plain layers and takes\produces one input\output tensor [14]. It is an unsupervised model; it has been trained on created sentences by us. It contains – sentences, it has different types of objects. The sequential model has three types of layers are embedding, LSTM, and dense; we create six Sequential models contains the same types of layers. Each model search for a given word in created sentences. Embedding is the first layer, it converts integers to fixed-size vectors[15]. LSTM is an RNN type, it is a forward-backward technique. Dense connects LSTM layers, it implements activation function, and it is softmax activation[16]. Figure 6 shows the architecture of NLP:

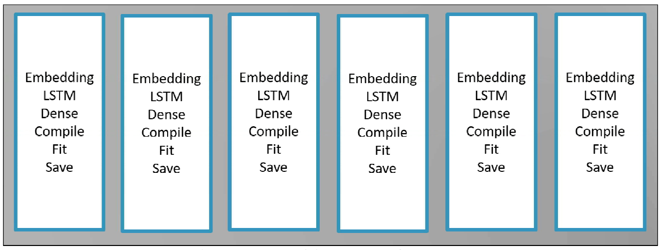


Figure 6: Architecture of LSTM.

IV. Discussion

Automatic video caption requires two important steps are: detecting objects and classifying their types (Encoder), plus step of generating sentence (Video Caption) based on detected objects as shown in Figure 1. To perform these steps, we propose model constructs of Key Frame Extraction (KATNA), object detection (YOLO), and generating sentences (LSTM).

At first, KATNA up to five frames based on video duration. It compares each consecutive two frames by KATNA elements as mentioned in the previous section, then it detects changes in frames within video and captures (frames (images)). After that, YOLO detects objects in captured frames by a surrounding box around objects with classification type plus percent accuracy; all this information is saved in a text file. Finally, LSTM reads detected objects words from the text file, then it generates an English sentence consists of subject and verb as a video caption.

V. Evaluation

Two evaluation methods implemented to measure the performance and accuracy of the model:

A. *the Metric for Evaluation of Translation with Explicit Ordering (METEOR):*

It has been used in many video captioning and description evaluation projects, especially, with short clips. It has 5 versions that calculate each hypothesis's alignment to its reference pair [17]. The proposed model has been evaluated by METEOR version 1.5, and the evaluation result of the proposed model is 0.35, it is a sufficient result. The following table shows a summary of some researchers who applied deep learning techniques to the MSVD dataset.

Table 1: Comparison with other Models.

Paper	**Percentage in METEOR on MSVD dataset**
Less Is More: Picking Informative Frames for Video Captioning [7].	0.33
M3: Multimodal Memory Modelling for Video Captioning [6].	0.2658
STAT: Spatial-Temporal Attention Mechanism for Video Captioning [2].	0.33
Temporal Deformable Convolutional Encoder-Decoder Networks for Video Captioning [3].	0.308
Proposed model	0.35

B. Human Evaluation

These criteria (Object Detection Quality, Sentence is Readable, Informativeness, Meaning preservation) have been formed in four questions and it has been sent to 20 persons for 3 different videos. Then, the average of the user's answers has been plotted.

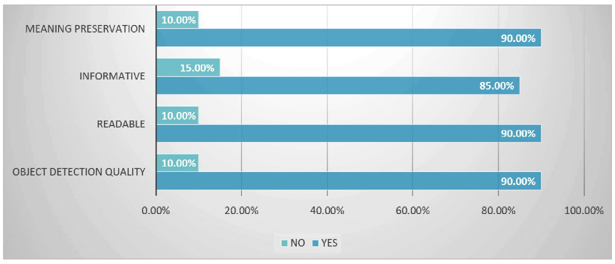


Figure 7: Human Evaluation Result.

VI. Conclusion and Future Work

In summary, video caption translates visual contents to text words sequentially (frame by frame and word by word), it is based on understanding video frames and transforms them to sentence (caption). Deep learning techniques have been greatly utilized in the field of video captioning research, which motivates researchers to develop a variety of video captioning framework which can automatically generate sentence (caption).

We develop a model using deep learning, which combines KATNA, YOLO, and LSTM. The model uses KATNA as a keyframe extraction, it chooses frames that show changes in video and remove redundant ones. The model applies YOLO on the MSVD dataset for object detection, and it generates English sentences using LSTM.

The model shows good accuracy and performance, it applies YOLO on the MSVD dataset, unlike the previous models that used other deep learning techniques. In the future, video captions will be measured by METEOR metric and human evaluation and compared with previous models using other deep learning techniques.

In the future, we think to train our model on more datasets; to be evaluated with different data. Also, we think to implement video caption-based Arabic language using (CAMeL Tools: An Open Source Python Toolkit for Arabic NLP). Finally, we think to evaluate the proposed model with different evaluation metrics such as BLEU, ROUGEL, and Diversity.

References

[1] . Li, B. Zhao, and X. Lu, ‘MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning’, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, Aug. 2017, pp. 2208–2214, doi: 10.24963/ijcai.2017/307.

[2] C. Yan *et al.*, ‘STAT: Spatial-Temporal Attention Mechanism for Video Captioning’, *IEEE Trans. Multimed.*, vol. 22, no. 1, pp. 229–241, Jan. 2020, doi: 10.1109/TMM.2019.2924576.

[3] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, ‘Temporal Deformable Convolutional Encoder-Decoder Networks for Video Captioning’, *ArXiv190501077 Cs*, May 2019, Accessed: Aug. 28, 2020. [Online]. Available: http://arxiv.org/abs/1905.01077.

[4] Chenyang Zhang and Yingli Tian, ‘Automatic video description generation via LSTM with joint two-stream encoding’, in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec. 2016, pp. 2924–2929, doi: 10.1109/ICPR.2016.7900081.

[5] R. A. Rivera-Soto and J. Ordonez, ‘Sequence to Sequence Models for Generating Video Captions’, p. 7.

[6] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, ‘M3: Multimodal Memory Modelling for Video Captioning’, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 7512–7520, doi: 10.1109/CVPR.2018.00784.

[7] Y. Chen, S. Wang, W. Zhang, and Q. Huang, ‘Less Is More: Picking Informative Frames for Video Captioning’, *ArXiv180301457 Cs*, Mar. 2018, Accessed: Aug. 28, 2020. [Online]. Available: http://arxiv.org/abs/1803.01457.

[8] ‘Rivera-Soto and Ordonez - Sequence to Sequence Models for Generating Video C.pdf’. Accessed: Feb. 01, 2021. [Online]. Available: http://cs231n.stanford.edu/reports/2017/pdfs/31.pdf.

[9] J. Xu, T. Mei, T. Yao, and Y. Rui, ‘MSR-VTT: A Large Video Description Dataset for Bridging Video and Language’, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5288–5296, doi: 10.1109/CVPR.2016.571.

[10] M. K. Asha Paul, J. Kavitha, and P. A. Jansi Rani, ‘Key-Frame Extraction Techniques: A Review’, *Recent Pat. Comput. Sci.*, vol. 11, no. 1, pp. 3–16, Feb. 2018, doi: 10.2174/2213275911666180719111118.

[11] Alok, ‘Video Key Frame Extraction With katna’, *Medium*, Oct. 22, 2019. https://medium.com/@Aloksaan/video-key-frame-extraction-with-katna-11971ac45c76 (accessed Aug. 28, 2020).

[12] J. Redmon and A. Farhadi, ‘YOLO9000: Better, Faster, Stronger’, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[13] L. Zhao and S. Li, ‘Object Detection Algorithm Based on Improved YOLOv3’, *Electronics*, vol. 9, no. 3, Art. no. 3, Mar. 2020, doi: 10.3390/electronics9030537.

[14] K. Team, ‘Keras documentation: Embedding layer’. https://keras.io/api/layers/core_layers/embedding/ (accessed Mar. 09, 2021).

[15] ‘tf.keras.layers.Embedding | TensorFlow Core v2.4.1’, *TensorFlow*. https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding (accessed Mar. 09, 2021).

[16] ‘tf.keras.layers.Dense | TensorFlow Core v2.4.1’, *TensorFlow*. https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dense (accessed Mar. 09, 2021).

[17] “https://www.cs.cmu.edu/~alavie/METEOR/index.html#Publications.” .

A Testbed for a three dimensional Pico–Sphere Satellite–Simulator(T3Dpilare)

Muhammad Faisal, Florian Wolz, Felix Dengel, Sergio Montenegro
Institute of Aerospace Information Technology,
Julius-Maximilians-University Würzburg,
Am Hubland D-97074 Würzburg, Germany.
muhammad.faisal@uni-wuerzburg.de

***Abstract*—This paper presents the mechanical design, assembly and construction of T3Dpilare. The principle target for this research was to design and built a three–axis floating–satellite simulator with six degree of freedom which is placed in a sphere for free movement. This three axis float–sat will be used to test the algorithms for three dimensional attitude control of satellites on earth. The simulator is named as 'T3Dpilare'. T3Dpilare is quite unique in its design, principle and mode of operation. It is a compact satellite simulator for the students and researchers. It offers a simple platform for on–ground testing of satellites and their subsystems in a quite frictionless environment to simulate the effect of zero gravity in the space.**

***Index Terms*—Aerospace, Avionics, Satellite Simulator, Attitude Determination and Control.**

I. Introduction

T3Dpilare is an air–bearing sphere, which is able to rotate freely in all three directions, internally rotation should be created by the virtue of the torque produced by the reaction wheels, same mechanism is also used in real satellites. The challenging situation for these kind of satellite simulator is the gravitational force of earth. If the centre of gravity of the sphere does not match with the physical geometric centre of the sphere then there would be a force arm which results in an unwanted torque, it is therefore required to shift the centre of the gravity of the sphere in the direction of its geometric centre. In order to full-fill this requirement there should be three moving masses in each direction (X,Y,Z), The position of the moving masses are controlled by the stepper motors. To determine the rotations a nine degree of freedom(DoF) IMU (Inertial Measurement Unit) should be used, it has a three dimensional rotation speed as well as a three dimensional acceleration sensor. The accelerometer shows us the direction of the gravitational vector and the position of the sphere, therefore it is necessary to determine the centre of of gravity of the sphere by using 6DoF IMU, where the the centre of gravity of the sphere is controlled by means of stepper motors and movable masses. The project had following requirements.

- The complete structure of T3Dpilare must reside in a hollow glass sphere of 30cm diameter.
- Structure should include three DC and three stepper motors.
- Each dc–motor should be able to rotate the float–sat in its particular axis so that the rotation in all 3–axis should be possible.
- The structure should be designed in such a manner that the stepper motor should be able to balance the centre of gravity by moving the attached masses .

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

In section two the state–of–the–art of the related projects are described. In section three the planning and the Computer Aided Design (CAD) of the simulator is presented which is followed by section four where the assembly and the integration of the different parts of the simulator are presented with explanation about the on–board electrical and electronics modules. In the last section the overall conclusion of the design and development of the T3dpilare is discussed with the recommendations for the future version of the simulator.

II. State–of–the–Art

In the field of Aerospace technology several projects are conducted to simulate the small satellites. In this section a brief description is provided about the similar efforts people have done for space research and development:

A. Floating Satellite (Float–Sat)

The Chair of Aerospace Information Technology at the University of Würzburg (Germany) has developed a Floating Satellite (Float–Sat) system. This system is used as a part of the exercises offered in the aerospace laboratory. It explains the approach for building small satellites to the students to learn and get familiar with basic satellite subsystems. It also provides a way to develop and test different control algorithms and strategies for different kind of space missions in an almost frictionless environment. The FloatSat system consists mainly of a mechanical structure that contains the basic satellite subsystems with one reaction wheel mounted at the centre of the horizontal plane of the structure. This reaction wheel is used to control the orientation of the satellite in one dimension. The detail of the FloatSat project can be found at [1].

B. Three Degree of Freedom (3DoF) test–bench for CubeSats

The University space centre of Montpellier and Nime, France has created a test–bench for CubeSats with three degree of freedom. In this test–bench a CubeSat is held in a frame with the hollow air–bearing, there is an external frame which is attached to a support and have an arm that is joined with an inner frame that contains hollow air–bearing pucks. When the sliding air–caps are attached to the bearing pucks the test-bench is able to rotate the CubeSat in all three axes [2].

C. Simsat: A Ground Based Platform For Demonstrating Satellite Attitude Dynamics And Control

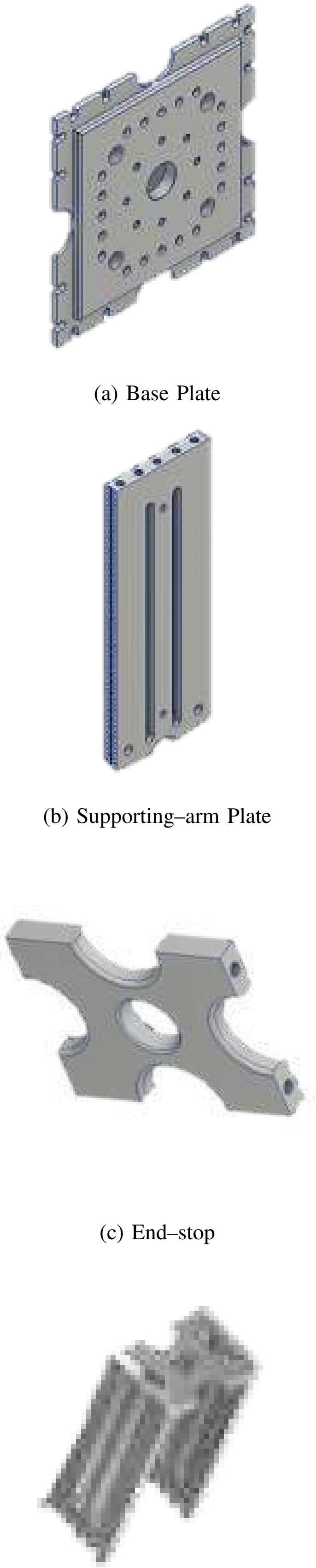
A simulation satellite is developed in the joint venture of United States Air Force Academy and Air Force Institute of Technology. Simsat was based on a spherical rotor which was placed on a air pedestal.The satellite's components were attached around the surface of the spherical rotor in such a manner that whole spacecraft was balanced [3]. This idea was presented almost twenty years ago. It provides an excellent opportunity to the students to practically learn the operations of a satellite in real time in a laboratory. Although the objective of Simsat and T3Dpilare were same but the T3Dpilare is also meant to be very simple, compact and small in size.

The goals of all the simulators described above were quite same but their structure and Attitude Determination and Control System (ADCS) are different. In the paper [4] the author has described many more simulators people have invented for the simulations of an aerospace vehicle.

III. Planning

The T3Dpilare consists mainly of a mechanical structure that contains the basic satellite subsystems with three reaction wheels attached with a dc–motor mounted in alignment with the X, Y and Z axis. These reaction wheels are used to control the orientation of the satellite in three dimensions. In–line with each dc–motor there exist a stepper motor with a moving mass to adjust the centre of Mass (CoM) of the structure to balance it perfectly.

In order to facilitate a simple, compact and symmetrical design , computer aided design (CAD) tool Inventor$^{\text{TM}}$ was used to develop the structure of the T3Dpilare. The mechanical parts were designed to create the housing for three DC motors and three stepper motors, DC motors housing was created to put the motor with the reaction wheel. The housing for the stepper motor was created to put the stepper motor with the moving mass. This moving mass approach was adopted to adjust the centre of gravity during the operation of the satellite. The design was conducted so that wiring remains simple. In the following two figures some of the CAD designs of the components of T3Dpilare are presented. The housing for the batteries was designed in such a way that they can be easily charge and replace. Each battery is one cell battery. There are four housing designs to accommodate three batteries in each housing. In total there are twelve batteries in T3Dpilare. In the following figure the fully assembled design is presented. The reaction wheels were facing towards the cube and the centre housing was reserved for the stack of the controlling boards which includes the micrcontroller, dc motor driver, stepper motor controller, battery charging circuit and also an IO–expansion board.



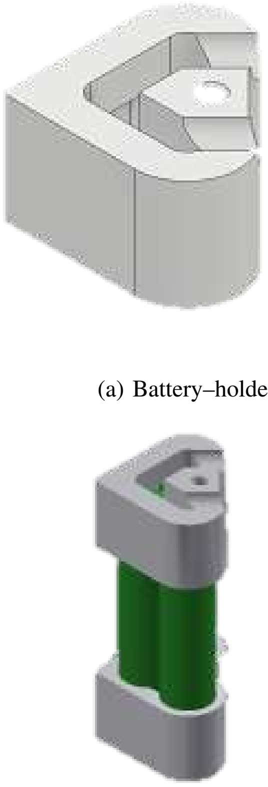
(a) Base Plate

(b) Supporting–arm Plate

(c) End–stop

(d) Housing for dc and stepper motors

Fig. 1. Different Parts of the structure of the Simulator.



(a) Battery–holder

(b) Batteries in the housing.

Fig. 2. Mechanical Structure for the batteries.

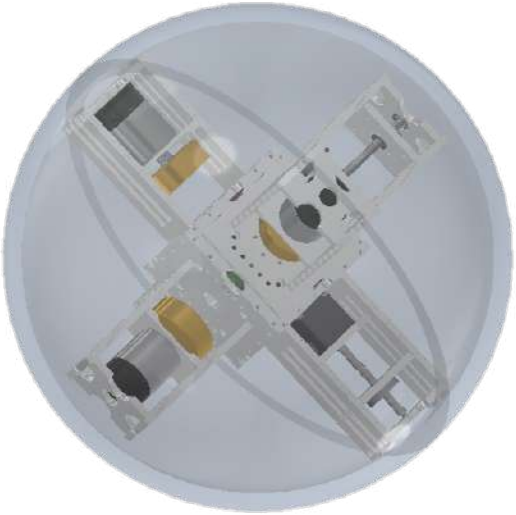


Fig. 3. Fully assembled CAD design of T3Dpilare

Diameter	3.5 cm
Thickness	0.8 cm
Weight	64.8 gm

Table I: Properties of reaction wheel.

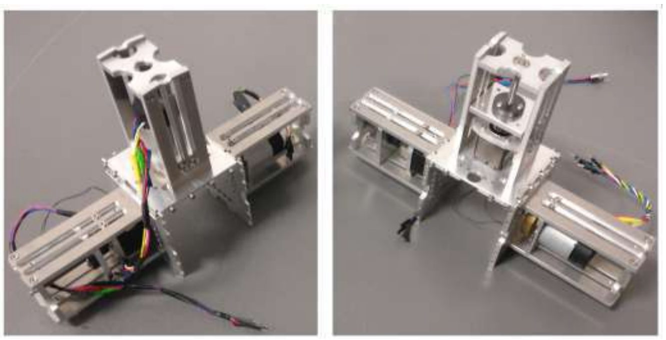


Fig. 4. Partially assembled structure of T3Dpilare

IV. Execution

After designing, all the parts of the structure were manufactured with Aluminium with the CNC machine. Aluminium was chosen because of the stability requirement. The reaction wheel that is meant to produce the torque to create angular momentum in each direction is made of brass. The density of this type of brass is 8620 $\frac{Kg}{m^3}$. This reaction wheel has following characteristics. As the design was conducted in a symmetrical fashion the assembly and the integration was also followed in the same way as shown below: The centre housing of the T3Dpilare was assigned for the microcontroller, driver boards for motors, IMU , IO–expansion and also the battery. The wiring of the components is conducted in a very careful way to avoid any loose wire that potentially creates the noise, endangers the electronics and also disturbs the operation of the T3Dpilare.

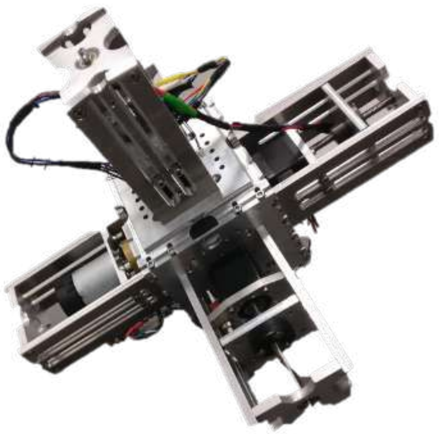


Fig. 5. Fully assembled structure of T3Dpilare

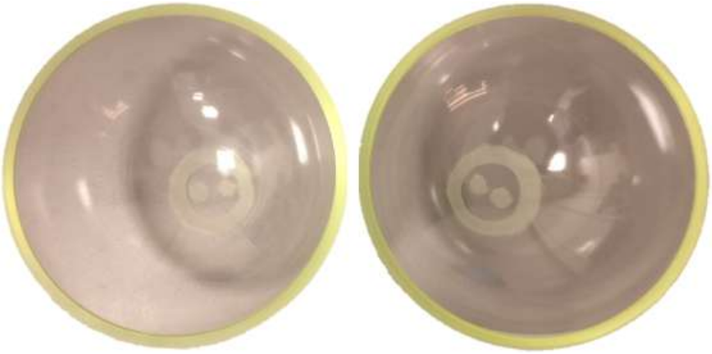


Fig. 6. Floating sphere

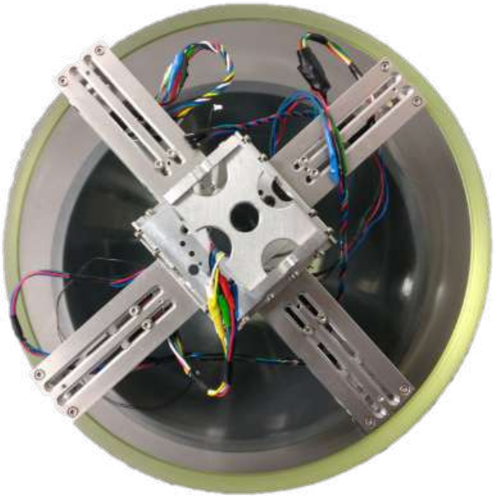


Fig. 7. Aerial view of T3Dpilare in the sphere on a floating platform

A. Floating Sphere

The T3Dpilare's structure presented in the above section is placed in a hollow sphere of the glass. The glass sphere is chosen over the plastic because the metal structure was heavy and it was not possible to properly operate with the plastic sphere. During the testing some non–uniformity was found on the surface of the glass then it was scratched to make it uniform.

Inner Diameter	30 cm
Thickness	12 cm

Table II: Properties of sphere

B. Vacuum Generation

In order to perform the proper operation of the T3Dpilare the vacuum is created in the sphere with a Deck Purge Box from the company Global Ocean Design^{TM} [5]. This box in combination with a purge port removes all the air and moisture from the sphere and creates vacuum inside to resemble the space environment around the concealed structure of T3Dpilare.

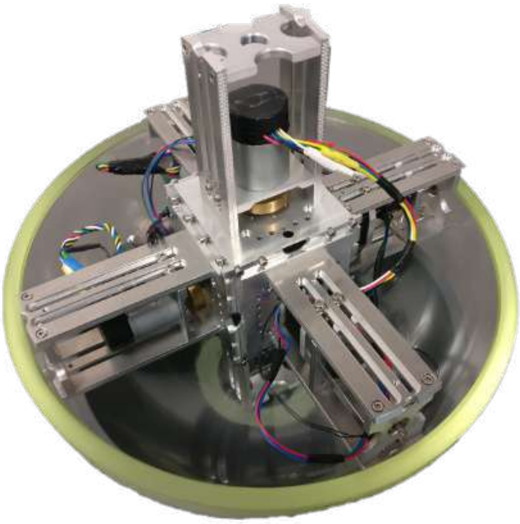


Fig. 8. Side view of T3Dpilare in the sphere on a floating platform

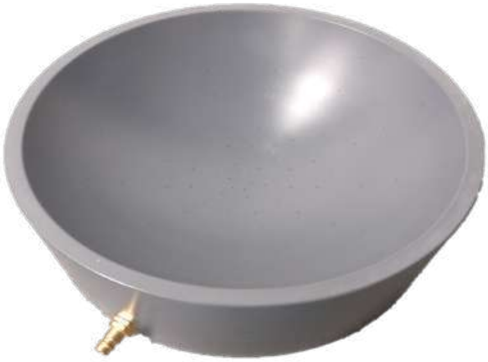


Fig. 9. Spherical Air Bearing Unit

C. Floating Platform

T3Dpilare's structure is placed into an Acrylic glass hemi–sphere shell that it is floating inside a Spherical Air Bearing Unit (SABU). The air bearing unit requires pressurized air input with a flow rate that may vary, depending on the mass of the floating unit. The pressurized air is fed in between the bearing surfaces through very small holes that are distributed on the unit's surface and discharge in the atmosphere through the edges of the bearing. This unit is made from Polyvinyl Chloride (PVC) material and it was designed to have a uniform distribution of the air flow and the pressure throughout the lubricating film between the unit and the hemisphere shell as shown in the figure 9. At the time of assembly it was observed that all the specifications of the mechanical designs were meeting the operating conditions, as shown in figure 10

D. Electrical Specification

There are four most important electrical components in the project, The dc motor interfacing, stepper motor interfacing,

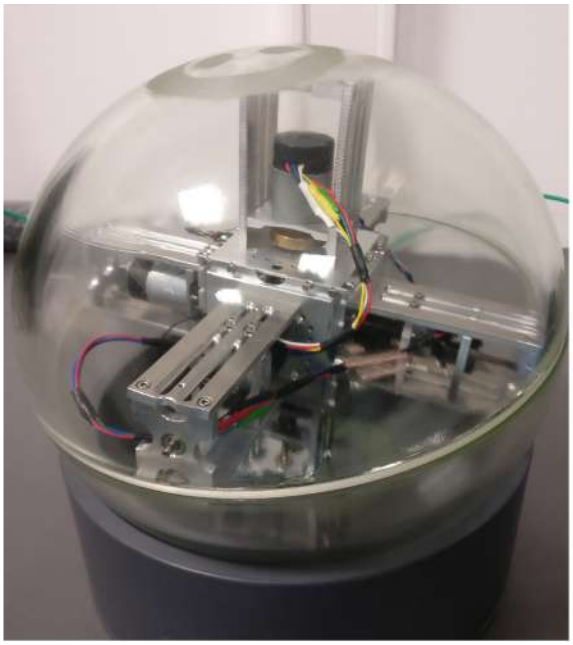


Fig. 10. T3Dpilare in operation

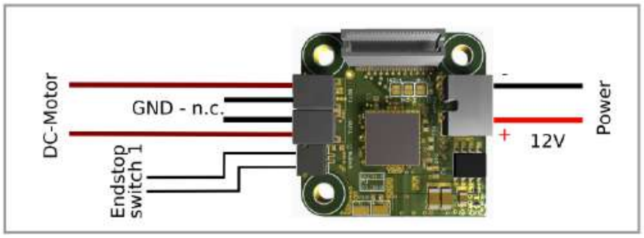


Fig. 11. DC motor driving board

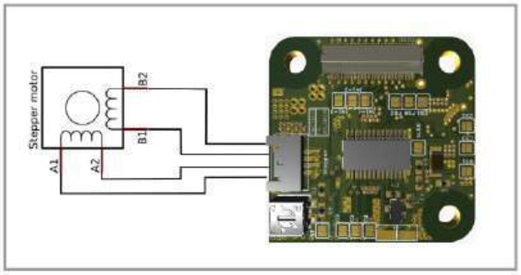


Fig. 12. Stepper motor Driving board

IMU and the power supply to the satellite. Each dc motor provides the rotation in a particular direction. These rotations are used to control the different modes of operation of the satellite. The IMU used in this project is LSM9DS1 from STMicroelectronics^{TM}, it has a 3-axis accelerometer, 3-axis gyroscope, and 3-axis magnetometer in a single chip therefore it is able to provide nine degree of freedom in motion sensing [6].

1) On–Board Computer:: The on–board computer of this T3Dpilare is based on stm32f407 micrcontroller from STMicroelectronics^{TM}, it has a Cortex–M4 core [7]. Department of Aerospace Information Technology in the University of Wuerzburg has developed a very compact and customized state–of–the–art development kit based on this microcontroller for these kind of nano and CubeSat projects. This kit is developed in the form of stack. There is a unique possibility to add several extension boards on this kit. The first module contains the main board then there is an extension board which contains the driver for stepper motor and dc motor, then comes the extension board the IMU with the WIFI module and at last there is a IO–expansion board. The distance between two boards in the stack is 5mm. The thickness of each board is 1.6 mm.

2) Battery:: Battery was selected on the basis of the power requirement as well as on the basis of the geometry of the T3Dpilare. There are twelve Lithium Ion 18650 Cell (2600mAh) Batteries.

3) DC–Motor Interface: : There are three dc–motors present in the structure therefore it was required to design a dc motor driver. The dc motor driver was based on a MC33926 full H–bridge, which can be operated in the range of 5 Volt to 28 Volt and it can provide almost 3 Ampere current(with 5A peak to peak value) [8]. Power to the dc–motor is controlled via the pulse width modulation (PWM) generated by the internal timer circuit of the microcontroller, the PWM output is passed to the H–bridge board which controls the motor operation.

4) Stepper Motor Interface:: In order to balance the structure in all three directions there is a stepper motor attached in–line with each dc motor. A moving mass is attached with each stepper motor. The control software of the T3Dpilare should be able to rotate each stepper motor such that the attached mass is moved to a specific position so that structure does not vibrate and remains stable. The stepper motor carrier board is based on L6470 dSPIN motor driver from STMicroelectronics^{TM}. It can be operated in the range of 8 to 45 V and it can provide upto 7.0 ampere of current. The driver has an SPI interface [9].

5) Circuit Diagram:: The most important part of the electrical circuit of T3Dpilare is the battery charging circuit. The battery setup of the project was providing 14.8 Volt but each board in the stack is required to be operated on 12V. Therefore a step–down voltage regulator is used to reduce the voltage from 14.8 volt to 12 volt with 15 ampere of current. A power switch is connected between the voltage regulator and the charging circuit which allows the battery to be charged in isolation i.e without being connected with the rest of the circuit. The driver boards for the dc and stepper motors, board with debugging port, IO–expansion board as well as the main board with microcontrollers are designed discretely in such a manner that they can be connected together on top of each other to form a stack. This kind of setup has improved the harness efficiency by eliminating the use of many individual wires and this stacking of boards was also necessary to place all the boards in the centre housing of the T3Dpilare.

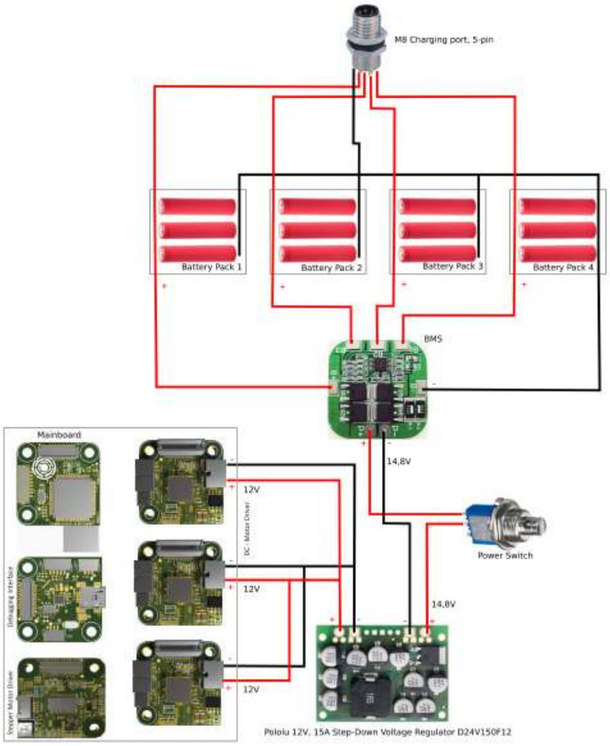


Fig. 13. Electrical Circuit diagram of T3Dpilare

V. Conclusion and Future Work

This research has produced a ground–based platform to understand the sub–systems and components of a space vehicle. It provides means to experiment, check and validate different control algorithm for the ADCS of satellites. The whole target of this project was to develop the structure of a pico–satellite with six degree of freedom . Additionally the electronics for the motors, batteries and sensors was also developed. The T3Dpilare is an state of the art satellite simulator which is designed and built to support research and education. The pitch, roll and yaw of the simulator can be controlled for the different modes of operations of the satellites. The structure is enclosed in a hollow–sphere where the vacuum is created with the vacuum generator and the structure is being floated on a hemispherical air–bearing unit which is providing a frictionless platform for the simulator to be operated in an environment which is similar to the space where the gravity is zero. The future work for this project is related to the development of the control algorithms for the autonomous operation of ADCS and also the testing of different machine learning and evolutionary algorithms for the ADCS.

References

[1] Prof. Dr. Sergio Montenegro & M.Sc. Eng. Atheel Redah. Floatsat –floating satellite. http://www8.informatik.uni-wuerzburg.de/wissenschaftforschung/floatsat/.

[2] Irina Gavrilovich, Sébastien Krut, Marc Gouttefarde, François Pierrot & Laurent Dusseau. Test bench for nanosatellite attitude determination and control system ground tests. In *4S: Small Satellites Systems and Services Symposium*, Porto Petro, Spain, May 2014.

[3] Steven Tragesser & Gregory Agnes. Simsat: A ground based platform for demonstrating satellite attitude dynamics and control. Annual Conference, Montreal, Canada, June 2002.

[4] Schwartz Jana, Peck Mason & Hall Christopher. Historical review of air-bearing spacecraft simulators. Journal of Guidance, Control, and Dynamics, 2003.

[5] Global Ocean Design. Deck purge box. https://www.globaloceandesign.com/deck-purge-box-dpb-107.html.

[6] STMicroelectronics. Lsm9ds1: inemo inertial module:3d accelerometer, 3d gyroscope, 3d magnetometer. https://www.st.com/resource/en/datasheet/lsm9ds1.pdf.

[7] STMicroelectronics. Stm32f405xx,stm32f407xx: Microcontrollers. https://www.st.com/resource/en/datasheet/stm32f407vg.pdf.

[8] Freescale Semiconductor. 5.0 a throttle control h–bridge, document number: Mc33926, rev. 10.0, 8/201. https://www.pololu.com/file/0J233/MC33926.pd.

[9] STMicroelectronics. L6470: A fully integrated microstepping motor driver with motion engine and spi. https://www.st.com/en/motor-drivers/l6470.html.

Topology Optimization of KUKA KR16 Industrial Robot Using Equivalent Static Load Method

Lakshmi Srinivas G
Ph.D. Scholar, Department of Mechanical Engineering
Birla Institute of Technology and Science
Hyderabad, Telangana, India
P20170411@hyderabad.bits-pilani.ac.in

Arshad Javed
Assistant Professor, Department of Mechanical Engineering
Birla Institute of Technology and Science
Hyderabad, Telangana, India
arshad@hyderabad.bits-pilani.ac.in

***Abstract*— The application of industrial manipulators or robots is increasing annually due to its productivity, superiority, and accuracy. The manipulator-links (arms) play an important role in supporting the structure of robots and bares the various loadings such as; inertial, bending, torsional, etc. A topology optimization method is the best way to distribute the useful material in the design space; thereby, needless densities can be removed. This approach reduces the mass and makes the industrial manipulator components as energy efficient. At the same time, it also maximized the stiffness of the robotic components and helped to improve the dynamic performance and stability. In this work, a topology optimization technique is implemented on upper arm and forearm of the KUKA KR16 industrial manipulator by considering volume fraction within a range of 0.3 to 0.9. The complete robot model is designed using SOLIDWORKS 2018 software and imported into Altair Inspire software for multibody dynamic analysis and topology optimization. To make the study realistic, the complete model is operated in dynamic environments within the working range and obtained boundary conditions such as; joint forces and torques. These boundary conditions are used as load cases for topology optimization study. The objective function is chosen as maximization of stiffness or minimization of compliance. To get the better manufacturability of the final topology shape controls of the geometry applied such as; symmetry, split draw and extrusion surface, etc. The obtained topology is analyzed using post-simulation methods and obtained performance values of displacement and Von-Mises stress. Within stated range, the volume reduction of 30-50% gives better performance values compared to the before optimization.**

Keywords*— *Dynamic loading; Industrial manipulator; KUKA KR16; Robotic-arm; Stiffness; Topology optimization

I. Introduction

The usage of industrial serial robots is increasing annually in various fields such as; automobile manufacturing, medical, engineering, aerospace, product inspection, etc., due to their productivity and accuracy [1]. The industrial manipulator links play an important role in supporting structure of the robot [2]. Topology optimization is the proven method to distributes useful material in the design space; thereby, needless densities are removed, which doesn't carry a significant load in the design space [3]. Using this approach, practitioners can produce energy-efficient mechanical components, and it also improves dynamic performance and structural stability [4].

From previous research, the usage of topology optimization approach in the arena of industrial manipulators was started in 2006 [5]. Lohmeier et al. optimized the leg of humanoid topologically with the help of *Optistruc* commercial software without considering dynamic forces [6], [7]. Most researchers optimized industrial robotic arms considering static loading conditions by arranging the manipulator in maximum stretching or worst position [8]–[13]. However, when the industrial manipulator is functioning in oprating range, static loading conditions were not sufficient. To make the study realistic, consideration of dynamic loading study is a must. Rare efforts are made in design optimization of industrial robots considering dynamic loads. Albers et al. optimized one degree of freedom (DOF) manipulator-link topologically using TOSCA commercial software [14], [15]. Multi-objective topology was performed on serial manipulator considering dynamic analysis with HyperWorks or SOLIDWORKS software [16], [17]. Srinivas and Javed conducted topology optimization of 1 and 3-DOF industrial manipulator considering dynamic loading conditions [18], [19]. In this method, the dynamic loading conditions were evaluated with respect to the angle by uniform angular position. Recently, the same authors proposed a novel approach to minimize the computational time of topologies generation using non-uniform angular position based on deflection threshold values [20]. Equivalent static load method (ESLM) is proven technique used to solve the topology optimization in a dynamic environment [21], [22].

This work focused on topology optimization of the KUKA KR16 industrial manipulator considering equivalent static load method. The entire model is designed in SOLIDWORKS software and imported into Altair Inspire software for multibody simulation. Force and torque values at the upper arm and forearm of the manipulator are analyzed. This data is used as input loads for topology optimization. The objective function is chosen as maximization of stiffness. Topologies are generated for various volume fractions ranging from 0.3 to 0.9 with a step value of 0.2. The performance values are captured for various optimization links such as displacement and Von-Mises stress. The obtained results are compared with the manipulator without topology optimization.

The rest of the manuscript is organized as; the methodology of topology optimization and ESLM are presented in section 2. Dynamic analysis of the manipulator-links is provided in section 3. Simulation results are detailed in section 4.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. Topology Optimization Approach

Topology optimization is a well-established method in structural optimization. It is an effective and superior approach when compared to other two optimizations like shape and size [23]. It distribute the material in the design region considering applied loads, volume fraction, and density variables. In this method, the design region is discretized and every element is allocated a density value. The densities allocation is carried out using the global stiffness '$\mathbf{K_{SIMP(z)}}$' matrix using popular method, i.e., solid isotropic material with penalization (SIMP) approach as follows:

$$\mathbf{K}_{\mathbf{SIMP}(z)} = \sum_{e=1}^{n} [z_{\min} + (1 - z_{\min}) z_e^P]\mathbf{K_e} \qquad (1)$$

where '$\mathbf{K_e}$' is stiffness matrix of the element. 'z_e', and '$z_{\min}$' are the density and minimum density of the element, respectively. '*P*' is the penalization variant often selected as 3. '*n*' is a total elements in the design region. To distribute material in design space, the objective function of optimization loop is considered as maximization of stiffness or minimization of compliance (*C*). Applied loads, constraints and boundary conditions are selected as constraints as follows:

$$\left.\begin{array}{rl} \min\limits_{z} : & C(z) = \mathbf{U^T K U} \\ s.t \;\; : & V(z) = V \times f \\ & \mathbf{P} = \mathbf{KU} \\ & 0 < z_{\min} \le z_e \le 1 \end{array}\right\} \qquad (2)$$

where '*f'* is the volume fraction of design region and '**U**' is displacement of each element. '**P**' is the applied load vector. '*V(z)'* and '*V*' are targeted and stating volume of the design space.

A. Equivalent Static load Method

The ESLs are a set of static forces that produce the similar displacement effect as static study at time period for non-linear dynamic investigation [24], [25]. In ESL process, multiple load cases are calculated with respect to the time is as follows:

$$\mathbf{M}(z)\ddot{D}_N(t) + \mathbf{K_N}(z, D_N(t))D_N(t) = P(t) \qquad (3)$$

where 'M(z)' is mass matrix for design variable, '*P(t)*' is the force at time period value. '$\mathbf{K_N}$' is stiffness matrix for non-linear investigation. The displacements from non-linear study (D_N) are multiplied by stiffness matrix of linear study ($\mathbf{K_L}$) to capture the ESL at time step, as follows:

$$P_{ESL}(t_a) = \mathbf{K_L} D_N(t_a) \qquad (4)$$

where 'F_{ESL}' is the load or force at the given instance (t_a). The several loading conditions are successfully handled by structural optimization without much computational time. The optimization is conducted using ESL will update the design variables based on non-linear analysis.

B. Modeling of KUKA KR16 Industrial Manipulator

The multi-purpose industrial robot, KUKA KR16, was used for different applications such as; welding, cutting, material handling, and painting, etc. The industrial robot consist of six DOF and a pose repeatability of ± 0.04 mm. The various arms and joints of manipulator are detailed in Fig. 1. The KUKA KR16 robot can reach a maximum distance of 1612 mm .The rated payload and approximate weight are 16 kg and 245 kg, respectively.

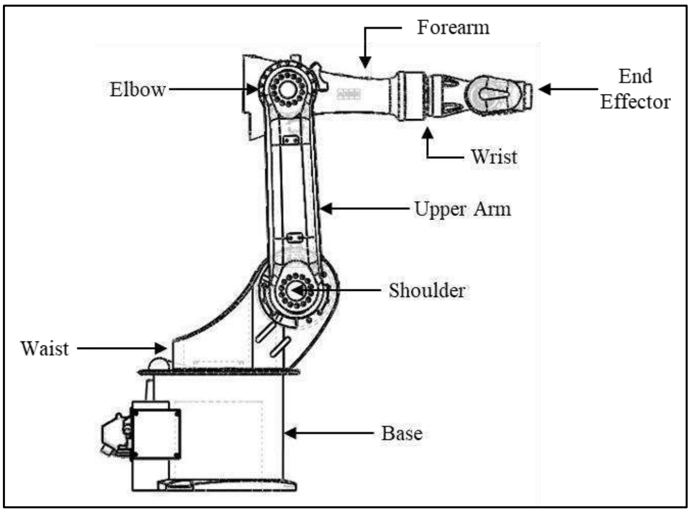


Fig. 1. The Wireframe diagram of KUKA KR16

The KUKA KR16 industrial manipulator has rotary joints and extension for end-effector at wrist. A KR C4 controller can operate the manipulator. The operating range for six DOF or axis of industrial manipulator and its maximum speeds are presented in Table 1.

TABLE I. Operating Range of Different Axis for KUKA KR16

Axis No.	Operating range (degrees)	Maximum speed
Axis 1	+185 to -185	200°/s
Axis 2	+65 to -185	175°/s
Axis 3	+175 to -138	190°/s
Axis 4	+350 to -350	430°/s
Axis 5	+130 to -130	430°/s
Axis 6	+350 to -350	630°/s

C. Dynamic analysis of Robot

The entire model of the KUKA KR16 manipulator is modeled using SOLIDWORKS 2018 software. The design is imported into ALTAIR Inspire multibody simulation software for dynamic analysis. The manipulator is considered at worst condition maximum reach point, as shown in Fig. 2. From Figure, user can also find the joint axes of the manipulator [26]. The joints are defined in the motion analysis, such as; base as fixed joint and axes are rotary joints. At center of mass, the acceleration due to gravity is applied, i.e., 9.8 m/s^2. Two servo motors are attached in the simulation at shoulder and elbow for the motion of upper arm and forearm of the manipulator.

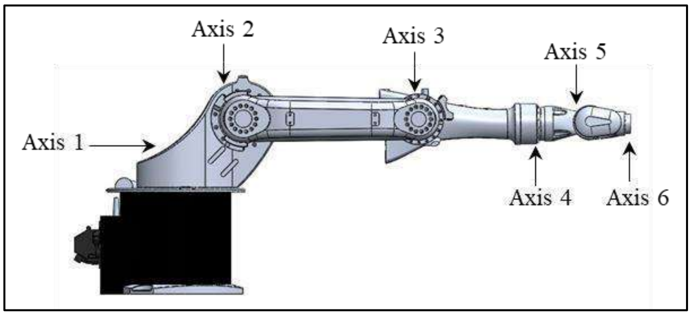
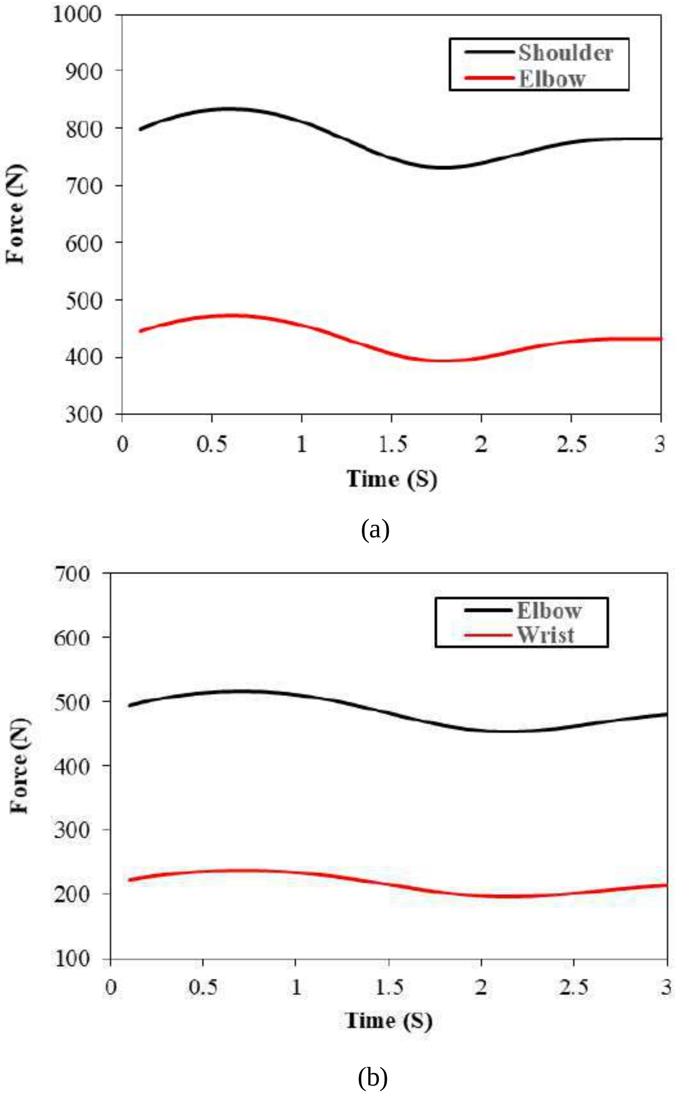


Fig. 2. Maximum stretching or worst position of a manipulator.

The ramp step velocity profile is chosen with a magnitude of 10 rpm and 8.3 rpm for upper arm and forearm, respectively. When the robotic arm is operated within working range 180^0 and 150^0, the force and torque values are captured with respect to time. The joint force values for upper arm and forearm are shown in Fig. 3.

The joint force values of shoulder and elbow for upper arm are shown in Fig. 3 (a). The joint force values of elbow and wrist for forearm are shown in Fig. 3 (b). The minimum magnitude of the force is obtained at 90^0 angular positions. This data is captured based on ESL method using Inspire simulation software. Maximum force of upper arm is recorded as 832.6 N and 464.2 N for shoulder and Elbow, respectively. The forearm maximum force is recorded as 512.6 N and 234.2 N for Elbow and wrist, respectively.

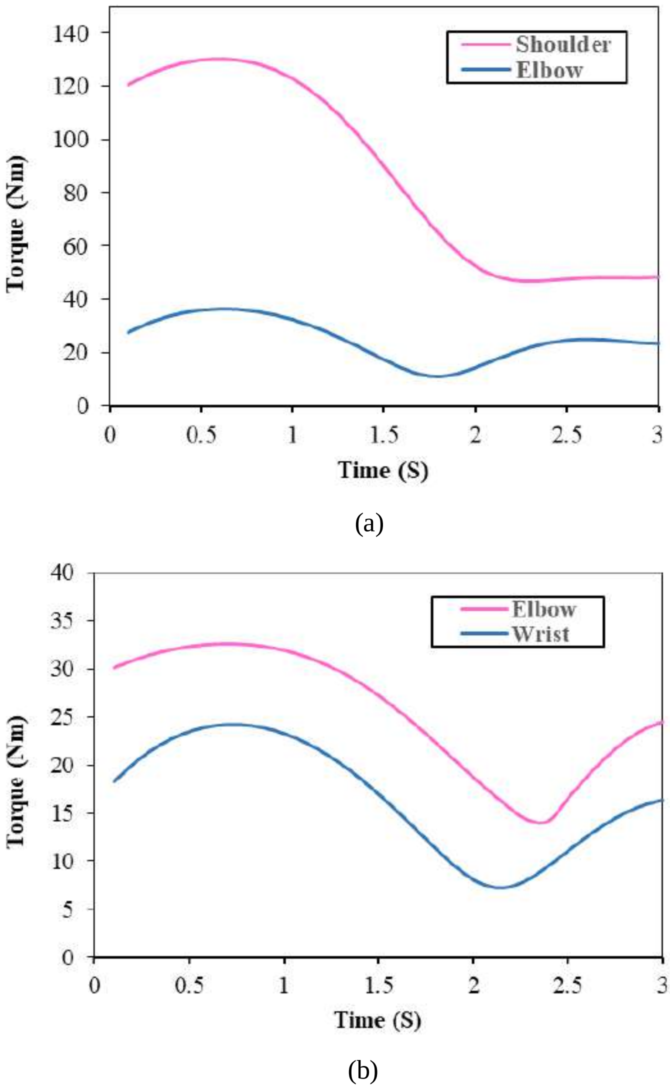
Similarly, the joint torque values are also captured for upper arm and forearm, as shown in Fig. 4. For upper arm, maximum torque is recorded as 124.5 Nm and 36.8 Nm at shoulder and Elbow, as shown in Fig. 4 (a). For forearm maximum torque is recorded as 33.6 Nm and 24.2 Nm at Elbow and wrist, as shown in Fig. 4 (b). The obtained force and torque values are used as input load cases for topology optimization. The process of obtaining topology optimization, shape controls, design space, and objective function of robotic arms are detailed in the next section.



(a)

(b)

Fig. 3. Force w.r.t. time for (a) Upper arm and (b) Forearm.



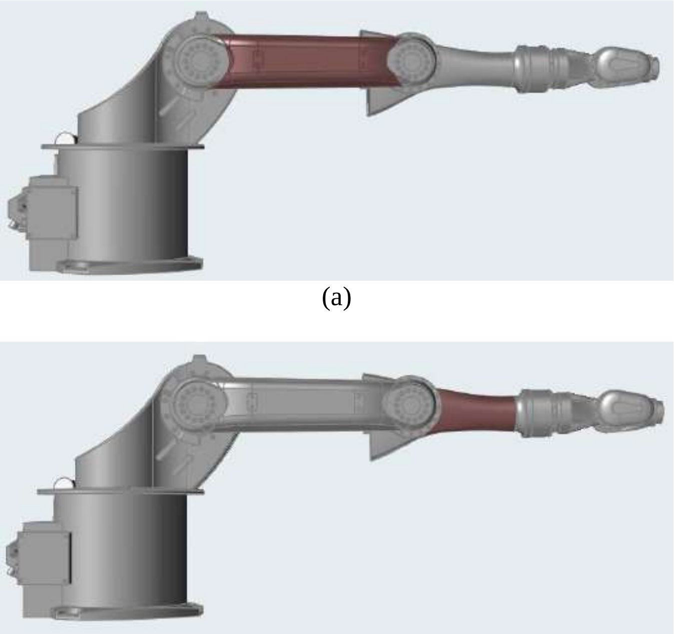
(a)

(b)

Fig. 4. Torque w.r.t. time for (a) Upper arm and (b) Forearm.

III. Topology Optimization of Robotic Arms

The implementation of dynamic loading conditions is very important in topology optimization method when the industrial manipulator is in motion. The obtained results of force and torque in motion analysis using ESL method are used as load cases for topology optimization method. The design space of the robotic components which are subjected to the topology optimization process is shown in Fig. 5.


(a)

(b)

Fig. 5. Design space of (a) Upper arm and (b) Forearm.

The robotic arms are divided into two parts such as; design space and non-design space. Only the design space is subjected to the topology optimization process remaining parts have not participated in the optimization loop. The joints are excluded from the topology optimization and come under non-design space. This feasibility is provided for assembly and placing different measuring instruments at that location. The design space for the upper arm and forearm is highlighted in the Figure. The topology optimization method distributes the useful material based on the objective function from the design space, i.e., maximization of stiffness.

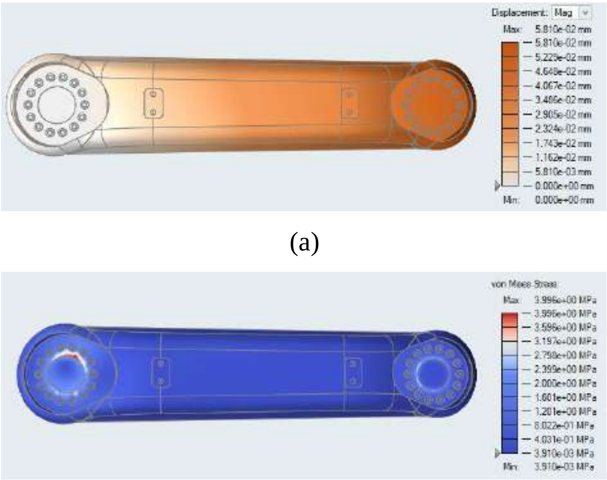
Initially, the design density elements are considered very small to avoid the singularity problem of Stress. The non-linear dynamic study is conducted to achieve linear stiffness and non-linear displacement. The boundary conditions obtained in the motion analysis are used for topology optimization and used as linear static loads w.r.t. the time. Based on the objective function, the design space is updated by useful material and eliminates needless densities. At the required volume fraction, the final manipulator-links archive the better dynamic performance. If the attained results are converged, the loop terminates, or else the design variables update and continue from the start. The results of the robotic arms are detailed in the next section.

IV. Results and Discussions

The obtained results of optimized topologies for upper arm and forearm using Altair Inspire software are provided here. The volume fraction is considered from 0.3 to 0.9 with a step value of 0.2. The displacement and Von-Mises stress values of the industrial manipulator are recorded.

A. *Before Topology optimization*

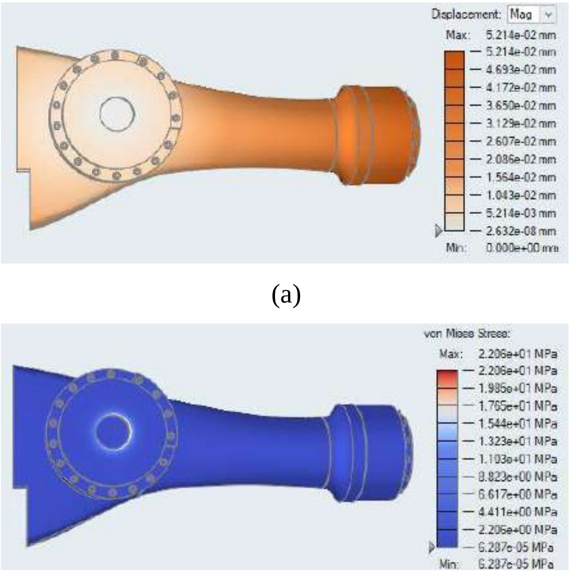
The robotic arms are dynamically analyzed and recorded the performance values before the application of topology optimization. The performance values of upper arm are shown in Fig. 6. The maximum deflection and Stress are recorded as 0.0581 mm and 3.99 MPa. Similarly, the performance values of forearm before topology optimization is shown in Fig. 7. The maximum deflection and Stress are recorded as 0.0521 mm and 22.06 MPa.



(a)

(b)

Fig. 6. Deflection and Stress of Upper arm before optimization



(a)

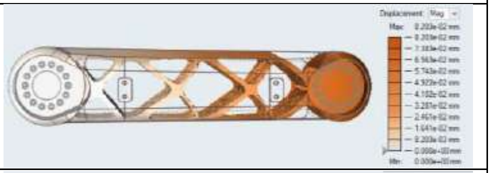
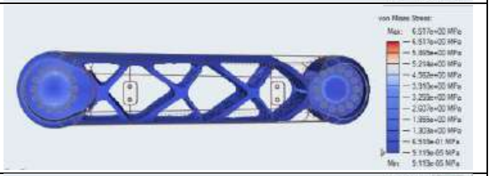
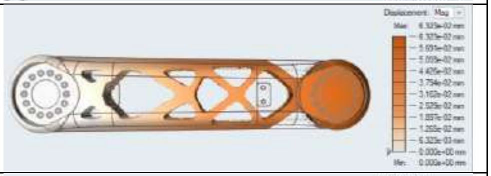
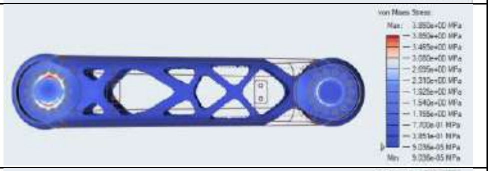
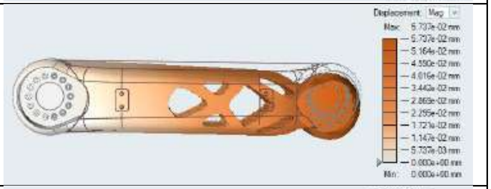
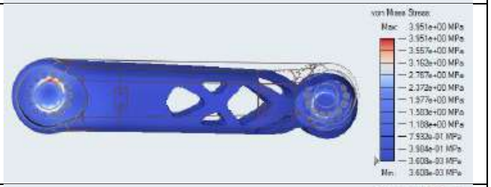
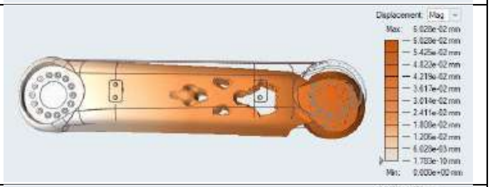
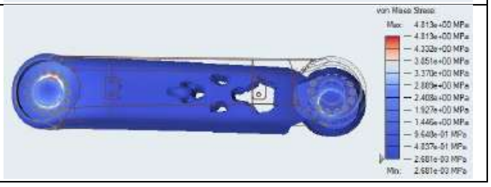
(b)

Fig. 7. Deflection and Stress of Forearm before optimization

B. *After Topology optimization*

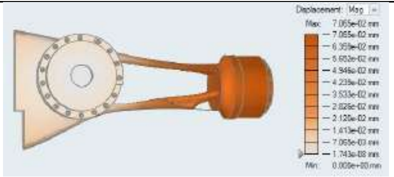
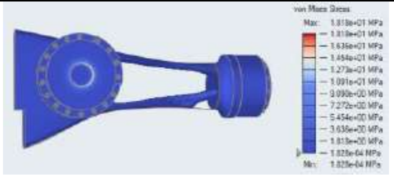
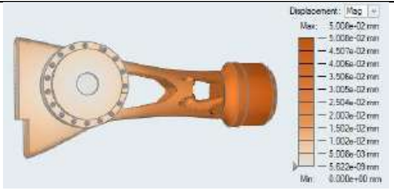
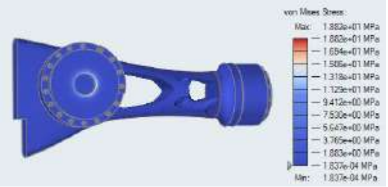
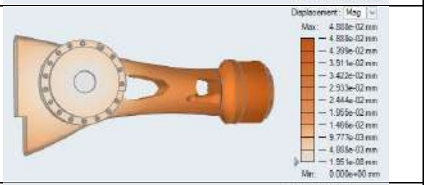
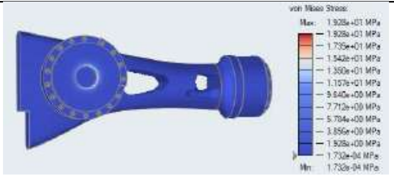
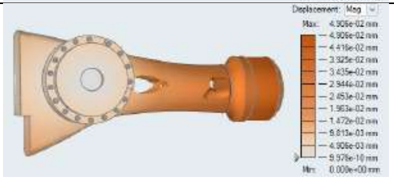
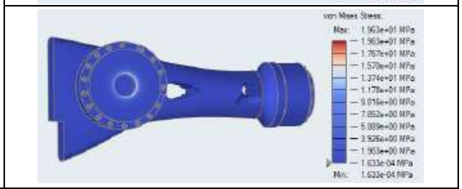
As stated earlier, based on the ESL method, boundary conditions are updated at the joints for topology optimization process. The objective function was chosen as maximization of stiffness for volume fraction ranging from 0.3 to 0.9. The shape controls are applied to achieve better manufacturing conditions such as; symmetry and split draw. The obtained topologies of upper arm for various volume fractions along performance values are shown in Table 2.

TABLE II. TOPOLOGY OPTIMIZATION OF THE UPPER ARM

VF	Performance values (Displacement and Stress)
0.3	 
0.5	 
0.7	 
0.9	 

From Table 2, at volume fraction 0.7, the performance values of displacement and Stress are 0.0573 mm and 3.95 MPa. It is provided slightly 1.3% better results compared to the solid manipulator-link, i.e., before topology, at the same time, volume reduced by 30%. Similarly, the obtained topologies of forearm for various volume fractions along performance values are shown in Table 3.

TABLE III. TOPOLOGY OPTIMIZATION OF THE FOREARM

VF	Performance values (Displacement and Stress)
0.3	 
0.5	 
0.7	 
0.9	 

From Table 3, at volume fraction 0.7, the performance values of displacement and Stress are 0.0488 mm and 19.28 MPa. It is attained better performance with 6.2% for displacement and 14.4% for Stress compared to the solid manipulator-link. The topology obtained at volume fraction 0.5 also attained better performance. However, the performance values of displacement obtained at volume fraction 0.3 are slightly higher. Practitioners can select the required volume fraction based on the performance. From the results, user can reduce the upper arm volume by 30% and forearm volume by 50% with better performance of industrial manipulator.

V. Conclusions

This paper mainly focused on multibody topology optimization of the KUKA KR16 industrial manipulator to improve its stability and dynamic performance. The complete model was designed in SOLIDWORKS and imported to Altair Inspire software for motion and topology optimization study. The industrial manipulator initially considered at maximum stretching position, and joints were defined as fixed or rotary. The upper arm and forearm were operated in a working range of 180^0 and 150^0 with a rated speed of 10 and 8.3 rpm. The force and torque were captured at joints and used as load cases for topology optimization study. Maximization of the stiffness was considered as an objective function. The volume fraction is selected from 0.3 to 0.9 with a step value of 0.2. For better manufacturability, shape controls were incorporated, such as; symmetry and split draw. The upper arm attained the better results of 1.6% compared to the solid manipulator-link at 0.7 volume fraction. Similarly, the forearm achieved better performance results up to the reduction of the 50% volume fraction. The practitioners can reduce the volume of upper arm and forearm by 70% and 50% with better dynamic performance. Users can select the required volume fraction based on the necessary performance values. This method is also applicable for other mechanical or structural members subjected to dynamic loading conditions.

Acknowledgment

This work was funded by the "Department of Science and Technology– Science and Engineering research board, India" (Grant Number: ECR/2017/000799).

References

[1] K. Zhou and P. Yao, “Overview of recent advances of process analysis and quality control in resistance spot welding,” *Mech. Syst. Signal Process.*, vol. 124, pp. 170–198, 2019.

[2] M. Liang, B. Wang, and T. Yan, “Dynamic optimization of robot arm based on flexible multibody model †,” vol. 31, no. 8, pp. 3747–3754, 2017, doi: 10.1007/s12206-017-0717-9.

[3] G. Carabin, “A Review on Energy-Saving Optimization Methods for Robotic and Automatic Systems,” 2017, doi: 10.3390/robotics6040039.

[4] M. P. Bendsoe, *Optimization of structural topology, shape, and material*. 1995.

[5] A. Albers, S. Brudniok, J. Ottnad, C. Sauter, and K. Sedchaicham, “Upper Body of a new Humanoid Robot - the Design of ARMAR III,” pp. 308–313, 2006.

[6] S. Lohmeier, T. Buschmann, M. Schwienbacher, H. Ulbrich, and F. Pfeiffer, “Leg Design for a Humanoid Walking Robot,” pp. 536–541, 2006.

[7] S. Lohmeier, T. Buschmann, and H. Ulbrich, “Humanoid robot LOLA,” *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 775–780, 2009.

[8] H. Hagenah, W. Böhm, T. Breitsprecher, M. Merklein, and S. Wartzack, “Modelling, construction and manufacture of a lightweight robot arm,” *Procedia CIRP*, vol. 12, pp. 211–216, 2013, doi: 10.1016/j.procir.2013.09.037.

[9] B. Denkena, B. Bergmann, and T. Lepper, “Design and optimization of a machining robot,” *Procedia Manuf.*, vol. 14, pp. 89–96, 2017, doi: 10.1016/j.promfg.2017.11.010.

[10] H. Zhang, Y. Huang, Z. Mo, and X. Zhang, “Mechanism design and analysis for a lightweight manipulator based on topology optimization methods,” *Lect. Notes Electr. Eng.*, vol. 408, pp. 467–477, 2017, doi: 10.1007/978-981-10-2875-5_39.

[11] M. Bugday and M. Karali, “Design optimization of industrial robot arm to minimize redundant weight,” *Eng. Sci. Technol. an Int. J.*, vol. 22, no. 1, 2019, doi: 10.1016/j.jestch.2018.11.009.

[12] A. J. G. Lakshmi Srinivas, “Numerical Simulation and Experimental Study on Lightweight Mechanical Member,” in *Advanced Engineering Optimization Through Intelligent Techniques*, Advances in Intelligent Systems and Computing, Springer Singapore, 2020, pp. 631–641.

[13] G. L. Srinivas and A. Javed, “Numerical evaluation of topologically optimized ribs for mechanical components,” *Mater. Today Proc.*, no., 2020, doi: 10.1016/j.matpr.2019.12.292.

[14] A. Albers, J. Ottnad, H. W. P. Haeussler, and A. T. Process, “Methods for Lightweight Design of Mechanical Components in Humanoid Robots,” vol. 49, no. 721, pp. 609–615, 2007.

[15] A. Albers and J. Ottnad, “System Based Topology Optimization as Development Tools for Lightweight Components in Humanoid Robots,” pp. 674–680, 2008.

[16] X. Chu, H. Xu, and G. Shao, “Multi-objective Topology Optimization for Industrial Robot,” no. August, pp. 1919–1923, 2016.

[17] G. L. Srinivas and A. Javed, “Multibody dynamic optimization for upper arm of industrial manipulator,” *AIP Conf. Proc.*, vol. 2281, no. October, 2020, doi: 10.1063/5.0027965.

[18] G. L. Srinivas and A. Javed, “Topology optimization of industrial manipulator-link considering dynamic loading,” *Mater. Today Proc.*, vol. 18, pp. 3717–3725, 2019, doi: 10.1016/j.matpr.2019.07.306.

[19] G. Lakshmi Srinivas and A. Javed, “Topology optimization of rigid-links for industrial manipulator considering dynamic loading conditions,” *Mech. Mach. Theory*, vol. 153, p. 103979, 2020, doi: 10.1016/j.mechmachtheory.2020.103979.

[20] J. A. Srinivas GL, “A novel method to synthesize a single topology for dynamically loaded members,” *J. Mech. Sci. Technol.*, vol. 35, no. 4, pp. 1–9, 2021, doi: 10.1007/s12206-021-03-y.

[21] S. B. Jeong, S. Yoon, S. Xu, and G. J. Park, “Non-linear dynamic response structural optimization of an automobile frontal structure using equivalent static loads,” *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.*, vol. 224, no. 4, pp. 489–501, 2010, doi: 10.1243/09544070JAUTO1262.

[22] E. Tromme, V. Sonneville, J. K. Guest, and O. Brüls, “System-wise equivalent static loads for the design of flexible mechanisms,” *Comput. Methods Appl. Mech. Eng.*, vol. 329, pp. 312–331, 2018, doi: 10.1016/j.cma.2017.10.003.

[23] O. S. M.P. Besndsoe, *Topology optimization theory, methods and applications*. Springer Singapore, 2003.

[24] H. Dong, J. P. Leiva, P. Adduri, T. Miki, and T. Fukuoka, “Large Scale Structural Optimization using GENESIS ® , ANSYS ® and the Equivalent Static Load Method.”

[25] B. S. Kang and G. J. Park, “Optimization of flexible multibody systems Using the Equivalent Static Load Method,” *CISM Int. Cent. Mech. Sci. Courses Lect.*, vol. 511, no. 4, pp. 375–426, 2009, doi: 10.1007/978-3-211-99461-0_18.

[26] M. Dahari and J. D. Tan, “Forward and inverse kinematics model for robotic welding process using KR-16KS KUKA robot,” *2011 4th Int. Conf. Model. Simul. Appl. Optim. ICMSAO 2011*, no. September 2014, 2011, doi: 10.1109/ICMSAO.2011.5775598.

[27] N. G. Chalhoub and A. G. Ulsoy, “Control of a flexible robot arm: Experimental and theoretical results,” *J. Dyn. Syst. Meas. Control. Trans. ASME*, vol. 109, no. 4, pp. 299–309, 1987.

Limitations and Challenges of Fog and Edge-Based Computing

Dheeraj Basavaraj
Department of Electrical and Computer Engineering
California State University, Fresno
CA, USA
bdheeraj@mail.fresnostate.edu

Shahab Tayeb
Department of Electrical and Computer Engineering
California State University, Fresno
CA, USA
tayeb@csufresno.edu

***Abstract*—Fog computing has set standards in networking and overcome complexities in cloud computing. However, despite its many benefits, the inherent distributed model of fog computing architecture introduces a new attack surface cyber-attacks. It is important to detect and prevent such attacks on the network to make the network reliable. In this paper, an overview of the various challenges and solutions proposed by many researchers is provided. This paper particularly focuses on solutions in securing networks by proposing several methods while developing an efficient intrusion detection system. Finally, observations are made based on the results of various proposed solutions and the existing research gap is also presented.**

I. INTRODUCTION

The improvement of computational and communication technologies such as cloud, edge, the fog has resulted in a smart system in many applications such as smart city, smart health, smart vehicles, etc. Deployment of these smart systems and providing IoT services are not possible by ruling out characteristics like reliability, security, and sustainability [7]. In the case of any smart system, there will be massive data collected from various sensor nodes that are requested by the user. The processing of a massive amount of data also limitation such as latency, response, and data security is the biggest challenge for cloud-only based systems.

In terms of network bandwidth scalability and latency edge computing performs better than cloud computing which prompted to design of edge-based applications over cloud-based applications in smart systems.

Due to the wide distribution of IoT networks are prone to cyber-attacks. It is key to secure the systems from this kind of attack by implementing the detection mechanisms that will detect abnormal behavior and protect sensitive data and manages the resource of the system.

In this study, we review Fog/Edge computing, its limitations/challenges and there advantage over cloud-based systems in communication and computational technologies.

Additionally, we report several methods for data confidentiality, access and authentication. However, even with this numerous mechanism, IoT networks are prone to multiple attacks in the network.

978-1-6654-4067-7/21$31.00 ©2021 IEEE

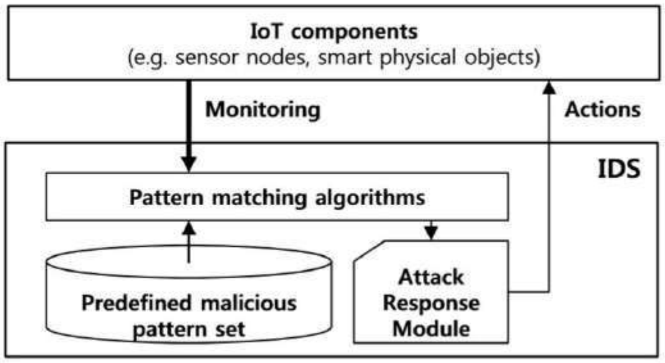

IoT components
(e.g. sensor nodes, smart physical objects)
Monitoring
Actions
IDS
Pattern matching algorithms
Predefined malicious pattern set
Attack Response Module

Fig. 1. Architecture of Intrusion detection system

The rest of this paper is organized as follows: Section 2 provides an overview of the literature on security, latency, energy and resource utilization, and speed. Section 3 discusses the methodology and the datasets. Section 4 summarizes the findings of this survey and section 5 gives a concise conclusion.

II. LITERATURE REVIEW

Signature-based IDS detect cyber-attacks using predefined patterns which are classified as malicious that are present in the database of the system as shown in Figure 1. The method uses multiple pattern matching algorithms to detect malicious activity in the network. Proposed methods need to find all the patterns in set P. The pattern consists of alphanumeric characters that monitors flow of packets across the network.

A. Security

A summary of the highlights of this section is provided in Table 1.

Oh et al. [8] proposed a pattern for malicious detection to secure IoT networks. They significantly decreased memory usage in the process of pattern-matching by proposing a new method called auxiliary shifting and early decision scheme. Results found to be very efficient in early decision scheme of the malicious pattern. However, they were not successful in detecting and classifying other attacks like False data inject, DOS etc. Also, a malicious user may try

TABLE I
TABLE1 : SECURITY COMPARATIVE REVIEW

Citation	Methods	Metric	Result
[10]	Fat learning network (FLN), Particle swam optimization	Accuracy with nerons and classified classes	Accuracy increases as hidden layer increases
[17]	Generic algorithm, Support Vector Machine	Classification	Detection Rate 96.38, False alarm rate 0.0
[18]	Generic algorithm(GA),Whale optimization algorithm(WOA)	Sensibility, accuracy, precision	Accuracy 96 with 400 samples
[22]	Intrusion label inside decode layer	Recall,F-1 Score,False positive	Accuracy 99 percentage

patterns each time which challenges node to detect such kind of attack. Although this limitation is improved in Ali et al. [10] by proposing a fast learning network(FLN) with particle swarm optimization (PSO) which are applied in the classification parameter. Even though the results were conceiving, the system complexity applied to sensor node is high because of low computational and resource capabilities.

The above limitations were improved by combining a hybrid genetic algorithm (GA) and an SVM with PSO for feature subset selection proposed by Moukhafi et al. [17] for the Intrusion detection system. This system reported in 100 percentage accuracy in differentiating DoS attack from another attack. However, this method fails to discriminate normal behaviour from other types of attacks with a low accuracy rate. The above limitations were targeted in Vajayanand et al. [18] to improve accuracy in classification for SVM-based classifier by proposing a hybrid feature delegation module based on GA and mutual information (MI). This proved in the experimental results showing SVM-based classifier performs better than an artificial neural network(ANN). Accuracy results of 96 percentage are reported when the classifier was trained with 400 samples and this proves that using these hybrid models high accuracy is obtained. Unfortunately considering battery usage and computational cost this scheme is not suitable.

For optimum resource allocation, least-square SVM was introduced for intrusion detection systems in Kabir et al. [19]. This scheme combines datasets of training and testing. Later it determines the amount of training and testing sets. Lastly, the samples are directly selected for the classifier. Even though results obtained are satisfactory due to the missing features in the limits of the dataset training models. Soon after on the basic of automata or finite state machine, one more IDS was introduced in M. Roesch et al [26]. The characterization of the network was done based on the automata transition that was used to detect intrusion attacks. In this method three kinds of attacks were experimented i.e, jamming attack, false attack and reply attack. Few attacks like false injection and DoS were not addressed.

Another method to detect intrusion was proposed in li et al. [20] which involves two steps for efficient detection of intrusion. In the primary step, numerous binary classifiers were used to classify the sample. In the second step, using k-nearest neighbours (k-NN) algorithm the samples are classified if output obtained by the previous step is vague. Results show high accuracy at high computation and high resource utilization to deploy numerous classifier.

Further, based on feature selection, the optimized parameter of SVM and weight Tao et al. [21] presented the genetic algorithm (GA). This algorithm first selects the features set and then optimizes the parameters of the SVM. After many instances, the trained classifier is used to classify and detect anomalies in a communication network. Later M. Lopez et al. [22] based on conditional variational autoencoder (CVAE) initiates intrusion detection system. Here labels of each sample are inputted for the decoder block of VAE. Although results were efficient due to the overhead of adding the extra input the decoder block, the system is complex. To overcome these limitations another efficient IDS based on long short-term memory recurrent neural network was proposed in [23]. This technique clearly shows they are efficient and overcome other machine learning techniques which includes k-NN and SVM.

Helmer et al. [24] introduced lightweight IDS. In this technique, mobile supervisors would collect data by moving around the network and update to a mediator to get the current status of the network. Even though, this unique type of architecture is successful in limiting the major problems of centralized network it has many challenges in terms of delay while transmission of data and high response time from the mediator. Meanwhile, another lightweight IDS was implemented by Li et al. [25]. In the first phase, the SVM-based classifier algorithm selects the feature.

All the above methods perform efficiently on detecting instructions in networks but they have high resource utilizations. It is important to focus on the design that requires minimal energy, lower computational cost and uses little memory in network nodes.

Bushra Zaheer Abbasi et al. [5] surveys about the security in present IoT technology and analyse the challenges and also suggest a few solutions to mitigate the limitations. Although it proposes numerous solutions there is a drawback in energy and resource allocations. After analysing the solutions and challenges in IoT systems authors suggest Hybrid version to optimize the security systems.

T.D Dang et al. [13] approached in two ways to resolve the issues in data security. Region-Based Trust-Aware (RBTA) and Fog-based Privacy-aware Role Based Access control

(FPRBAC) are the two approaches that tracks and trace location and many also detect attack vectors like Intrusion attack, Man in the middle attack and attacks with polices and Role-based attacks. Metric used in this method is Execution time and these approaches pose Feasibility and Efficiency.

Jun Wu et al. [3] proposed FCC i.e, Fog computing content-Aware filtering in the social network applications. This technique utilizes ICN i.e, an information-centric social network that uses names instead Ip-address as addressing mechanism. This method performs an efficient security service. This is implemented in NDN i.e, Network data networking simulator. Although, it has significant results inaccuracy rate, latency, speed and control overhead the system is too complex to implement. Another technique was proposed by MD. Wazid et al. [4] for securing social networks. In this method key management and user authentication methods were implemented to protect the CIA cycle i.e, confidentiality, integrity and availability. Author targets on access control to protect the sensitive data from unauthorised users. Although this method is different from the above ICN method in terms of performance it overcomes the limitation of ICN. This method can be implemented in NS2 simulator.

Mingming Cui et al. [6] proposed an authentication scheme for safety road conditions that have real-time content transmission. Here the network can fully exchange and share information. This scheme can be used for authentication, tracking the vehicle, and for user anonymity. This scheme is implemented using Java programming language using JPBC Library in intel core processor. This scheme is evaluated using four hash functions using the user key generation. Results obtained in this scheme are in real-time with low computation. Another scheme was proposed in the area of IOV, M. Bousselham et al. [15] which points out the failure in the security of the cryptographic algorithm. For this purpose, it employs two methods Decoy technology DT and User behaviour profiling (UBP). Here UBP profiles the user data based on actions performed by vehicles in IoV and DT does on-demand generation of decoy document. Results show that UBP and DT have a greater number of false-positive and provides an unprecedented level of security in vehicular networks in VCC.

B. Latency

Nathan Wheeler [1] made a comparative survey of fog enabled and fog-less system architecture and explains problems related to latency, computational speed and energy consumption in these two architectures. The paper shows how raspberry pi can be used as an edge node. Here AWS is used as cloud and performance is evaluated. Request time and response time during service request are the metrics used to evaluate the performance. Results show there is a massive response time difference between the two architectures. Thus, the author concludes fog computing architecture is reliable and efficient also mentions response time can in the future be reduced by using SDN in the context.

C.T. Do et al. [12] targeted problems related to resource allocation in video streaming. A proximal algorithm was proposed to minimize latency in fog applications this method involves dynamic scaling at runtime that result in achieving high bandwidth and minimal resource utilization. Although this algorithm is efficient it does not address limitations of internet computational incapabilities. This limitation is focused by M. Abedi et al. [14]. Authors try to resolve the computational capabilities by using ANN which resulted in efficient results in term of delta and computational capabilities.

Moreover, Sukhpal Singh Gill et al. [9] addresses problem-related to Network bandwidth, latency, Response time and energy consumption in smart home systems. Here Resource management Technique for smart home (ROUTER) using IFogsim and iCloud tools to evaluate the performance of the system. Although results reported there is reduced network bandwidth, energy consumption and increase response time and latency. But in terms of scalability design is not suitable for huge network and are not reliable.

C. Energy Utilization and Resource Allocation

Mouro Tortonesi et al. [2] made a functionality survey on prioritizing resource allocation for data analysis to reduce computational time and resource usage in IoT applications. Paper demonstrates the usage of fog computing in the smart city. The author proposed two approaches for improving these above limitations first by implementing an information-centric server and then by prioritizing vol factors for resource allocation. This is done by OpenCV open-source library with evaluation metric vol factor and CRIOS. Significant improvement in CPU time and efficiency of the system is obtained. This approach is easy and simple although securing the system can be considered for future study [27].

Another method is presented in Adila Mebrek et al. [16]. This method uses an Evolutionary Algorithm for optimisation of memory usage and lowers energy consumption. This algorithm is choose based on the objects which are among BIP and genetic algorithm (GA) algorithms. Algorithm accesses energy, processes data and update to the cloud. Results show latency increases with the number of nodes increases.

D. Speed

C. Alippi et al. [11] proposed a re-programmable framework for Wireless sensor Networks that operate in the real-time scenario. This framework gives the advantage of programming remotely and improved computational speed.REEL framework

TABLE II
TABLE 2: LATENCY COMPARATIVE REVIEW

Citation	Methods	Metric	Result
[1]	Fog and fogless architecture	Req. and resp. time	Resp. time is high
[12]	Altering direction method	Resource usage	BW is twice cap
[14]	ANN	Process time	Reduced internet traffic
[9]	Resource management technique	Bandwidth	N/w BW reduction

is implemented in Contiki and execution time, Memory utilization was the evaluation metrics. Although results show significant low in execution time and increased efficiency in real-time computation but ROM occupation is large which results in large resource allocation.

III. METHODOLOGY

Here we will be discussing the brief overview of methods used by researchers while designing Intrusion detection system, Table 2 and Table 3 shows the comparative view of research papers in security and latency metric.

Below are the few methods proposed in the research papers.

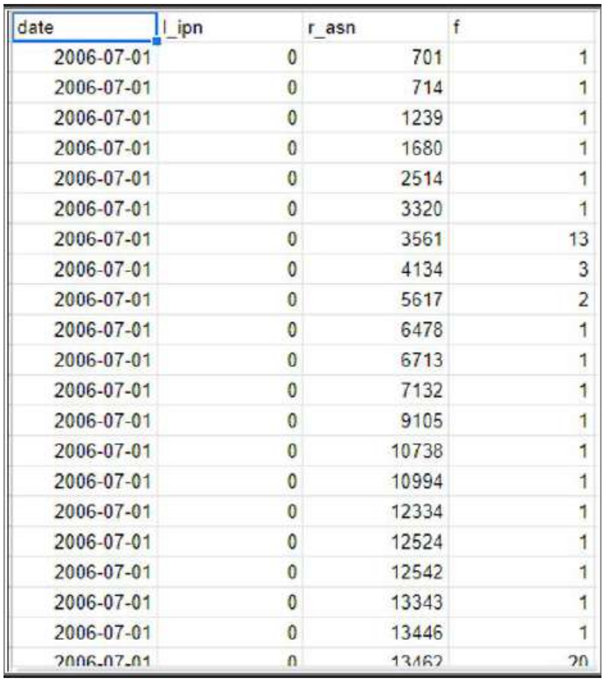
- **Auxiliary shift**: There are two stages
  a. Pre-processing: Construct three tables i.e, Shift table, hash table, and prefix table.
  b. Pattern matching stage: It will match the pattern with the flow of the packet with a predefined malicious pattern in the database.
- **Early decision**: Here the malicious patterns are sorted in ascending alphabetical order to reduce the execution time while matching the pattern.
- **Genetic Algorithm**: This algorithm selects the fittest particle based on the fitness value evaluated based on the RBF kernel function in SVM to get the next offspring using methods like crossover and mutation.
- **Support vector machine**: It is supervised by ML which classifies two groups of problems.It uses the RBF kernel function to categories the data.
- **Particle swarm optimization**: It searches for optimal solutions in the given solutions through agents like particles.
- **Artificial Neural Network**: It consists of unit or node called artificial neurons that are used to train to classify the attack types.
- **k-NN algorithm**: Algorithm that is used to classify attack types by training the unit.

A. Data-sets

- **KDD99 and NSL-KDD**:
  a. Analysed based on number of output classes, training and testing models. b. Models can be applied are
  i. Clustering
  ii. Classification
  iii. Feature selection and reduction algorithm
  c. KDD99 is extracted from DARPA
  d. NSL-KDD obtained from KD99 by eliminating dataset.
  e. Output classes are DoS, Probe, R2L ,U2R and Normal
  f. Dataset has 24 attack and 14 attack types training and testing respectively.
  Table 3 . shows few feature selected in KDD99 algorithm.

TABLE III
TABLE 3: FEATURES IN KDD DATASET

No.	Feature name
1	duration
2	protocol type
3	service
4	src byt
5	destbyt
6	Flag
7	root shell count
8	is user login
9	count



date	l_ipn	r_asn	f
2006-07-01	0	701	1
2006-07-01	0	714	1
2006-07-01	0	1239	1
2006-07-01	0	1680	1
2006-07-01	0	2514	1
2006-07-01	0	3320	1
2006-07-01	0	3561	13
2006-07-01	0	4134	3
2006-07-01	0	5617	2
2006-07-01	0	6478	1
2006-07-01	0	6713	1
2006-07-01	0	7132	1
2006-07-01	0	9105	1
2006-07-01	0	10738	1
2006-07-01	0	10994	1
2006-07-01	0	12334	1
2006-07-01	0	12524	1
2006-07-01	0	12542	1
2006-07-01	0	13343	1
2006-07-01	0	13446	1
2006-07-01	0	13462	20

Fig. 2. Network traffic dataset

- **Network traffic**:
  a. This data set is used as input to one of the models.
  b. It has approximately 500k real-time network traffic data
  c. It has approximately 21k rows and over 10 workstations local IPS
  d. This data-set is monitored over three months duration. Each row consists of data, local Ip, remote asn, and count of connections in a day.
  The snap of the network dataset is shown in Figure 2.

IV. FINDINGS

Following are a summary of observations made while reviewing the paper regarding the solution w.r.t security of the

TABLE IV
TABLE 4: COMPARATIVE ANALYSIS

Paper cited	Dataset	Accuracy	Detection Rate
[1]	KDD99	96.01 percentage	96.38 percentage
[21]	KDD99 CUP	99.56 percentage	90.66 percentage
[20]	NSL-KDD	97.8 percentage	97.72 percentage

IoT.

- The detection rate of attacks is normally high with hybrid models compared to other models.
- Few models with pattern matching algorithms possess low in classifying attacks.
- Majority of the efficient proposed solution uses SVM for evaluation of fitness function.
- Performance evaluation of several proposed solutions is done by using KDD99 datasets.

V. CONCLUSION

Reviewed problems faced in IoT in various metrics such as latency, security, speed, energy, and resource allocation to understand the challenges in IoT also compared the solutions for specific problem categories. Furthermore, observations are made based on the results of efficient proposed solutions. Finally, efficient proposed solutions are studied to fill the gap and improve the performance of the system for future work.

Based on the results obtained from the proposed solutions, various enhancements in implementations can be made to produce efficient performance of the system. Firstly, by combing efficient algorithm in terms of efficient memory and execution time usage that override the limitation of memory usage and latency in few proposed solutions. Secondly, by using updated data set for efficient performance of Intrusion detection system application in real time.

REFERENCES

[1] Riya, Gupta, N. & Dhurandher, S.K. Efficient caching method in fog computing for internet of everything. Peer-to-Peer Netw. Appl. (2020).

[2] Mauro Tortonesi, Marco Govoni, Alessandro Morelli, Giulio Riberto, Cesare Stefanelli, Niranjan Suri, Taming the IoT data deluge: An innovative information-centric service model for fog computing applications, Future Generation Computer Systems

[3] J. Wu, M. Dong, K. Ota, J. Li and Z. Guan, "FCSS: Fog-Computing-based Content-Aware Filtering for Security Services in Information-Centric Social Networks," in IEEE Transactions on Emerging Topics in Computing, vol. 7, no. 4, pp. 553-564, 1 Oct.-Dec. 2019, doi: 10.1109/TETC.2017.2747158

[4] Wazid, M., Das, A. K., Kumar, N., & Vasilakos, A. V. (2019). Design of secure key management and user authentication scheme for fog computing services. Future Generation Computer Systems, 91, 475-492.

[5] B. Z. Abbasi and M. A. Shah, "Fog computing: Security issues, solutions and robust practices," 2017 23rd International Conference on Automation and Computing (ICAC), Huddersfield, 2017, pp. 1-6, doi: 10.23919/IConAC.2017.8082079

[6] M. Cui, D. Han and J. Wang, "An Efficient and Safe Road Condition Monitoring Authentication Scheme Based on Fog Computing," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 9076-9084, Oct. 2019, doi: 10.1109/JIOT.2019.2927497.

[7] S. Tayeb, S. Latifi and Y. Kim, "A survey on IoT communication and computation frameworks: An industrial perspective," 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2017, pp. 1-6. doi: 10.1109/CCWC.2017.7868354

[8] D. Oh, D. Kim, and W. W. Ro, "A malicious pattern detection engine for embedded security systems in the Internet of Things," Sensors, vol. 14, no. 12, pp. 24188–24211, 2014

[9] Gill, S. S., Garraghan, P., & Buyya, R. (2019). ROUTER: Fog enabled cloud based intelligent resource management approach for smart home IoT devices. Journal of Systems and Software, 154, 125-138.

[10] M. H. Ali, B. A. D. Al Mohammed, A. Ismail, and M. F. Zolkipli, "A new intrusion detection system based on fast learning network and particle swarm optimization," IEEE Access, vol. 6, pp. 20255–20261, 2018.

[11] C. Alippi, R. Camplani, M. Roveri and L. Vaccaro, "REEL: A real-time, computationally-efficient, reprogrammable framework for Wireless Sensor Networks," SENSORS, 2011 IEEE, Limerick, 2011, pp. 1193-1196, doi: 10.1109/ICSENS.2011.6126919.

[12] C. T. Do, N. H. Tran, Chuan Pham, M. G. R. Alam, Jae Hyeok Son and C. S. Hong, "A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing," 2015 International Conference on Information Networking (ICOIN), Cambodia, 2015, pp. 324-329, doi: 10.1109/ICOIN.2015.7057905.

[13] T. D. Dang and D. Hoang, "A data protection model for fog computing," 2017 Second International Conference on Fog and Mobile Edge Computing (FMEC), Valencia, 2017, pp. 32-38, doi: 10.1109/FMEC.2017.7946404.

[14] M. Abedi and M. Pourkiani, "Resource Allocation in Combined Fog-Cloud Scenarios by Using Artificial Intelligence," 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), Paris, France, 2020, pp. 218-222, doi: 10.1109/FMEC49853.2020.9144693.

[15] M. Bousselham, N. Benamar and A. Addaim, "A new Security Mechanism for Vehicular Cloud Computing Using Fog Computing System," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-4, doi: 10.1109/WITS.2019.8723723

[16] A. Mebrek, L. Merghem-Boulahia and M. Esseghir, "Efficient green solution for a balanced energy consumption and delay in the IoT-Fog-Cloud computing," 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, 2017, pp. 1-4, doi: 10.1109/NCA.2017.8171359.

[17] M. Moukhafi, K. El Yassini, and S. Bri, "A novel hybrid GA and SVM with PSO feature selection for intrusion detection system," Int. J. Adv. Sci. Res. Eng., vol. 4, pp. 129–134, May 2018

[18] R. Vijayanand, D. Devaraj, and B. Kannapiran, "A novel intrusion detection system for wireless mesh network with hybrid feature selection technique based on GA and MI," J. Intell. Fuzzy Syst., vol. 34, no. 3, pp. 1243–1250, 2018.

[19] E. Kabir, J. Hu, H. Wang, and G. Zhuo, "A novel statistical technique for intrusion detection systems," Future Gener. Comput. Syst., vol. 79, pp. 303–318, Feb. 2018.

[20] L. Li, Y. Yu, S. Bai, Y. Hou, and X. Chen, "An effective two-step intrusion detection approach based on binary classification and K-NN," IEEE Access, vol. 6, pp. 12060–12073, 2018

[21] P. Tao, Z. Sun, and Z. Sun, "An improved intrusion detection algorithm based on GA and SVM," IEEE Access, vol. 6, pp. 13624–13631, 2018

[22] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in IoT," Sensors, vol. 17, no. 9, p. 1967, 2017.

[23] F. Jiang et al., "Deep learning based multi-channel intelligent attack detection for data security," IEEE Trans. Sustain. Comput., to be published.

[24] G. Helmer, J. S. K. Wong, V. G. Honavar, L. Miller, and Y. Wang, "Lightweight agents for intrusion detection," J. Syst. Softw., vol. 67, no. 2, pp. 109–122, 2003.

[25] Y. Li, J.-L. Wang, Z.-H. Tian, T.-B. Lu, and C. Young, "Building lightweight intrusion detection system using wrapper-based feature selection mechanisms," Comput. Secur., vol. 28, no. 6, pp. 466–475, 2009.

[26] M. Roesch et al., "Snort—Lightweight Intrusion Detection for Networks," in Proc. Lisa, vol. 99, 1999, pp. 229–238

[27] Tayeb, S., Mirnabibaboli, M., & Latifi, S. (2018). Cluster head energy optimization in wireless sensor networks. Software Networking, 2018(1), 137-162.

Defense in Depth approach on AES Cryptographic Decryption core to Enhance Reliability

Gayatri Yendamury
Robert Bosch Engineering and Business Solutions Private Limited
Coimbatore, India
Gayatri.Yendamury@in.bosch.com

N Mohankumar
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India
n_mohankumar@cb.amrita.edu

***Abstract*— Security is need of the hour in today's world since the cyber physical systems are prone to malicious attacks. Advanced Encryption Standard is a cryptographic algorithm which is utilized extensively but is sensitive to dangerous attacks due to advances in technology. This paper administers Defense in Depth approach at system level on AES cryptographic core using an effective logic locking technique. AES lock block is inserted judiciously in subprocess of AES decryption algorithm. This approach achieves output corruption of 70% when incorrect password is provided at input due to which the probability of guessing the information and reverse engineering the architecture is reduced to greater extent. AES Lock block is highly efficient to secure the AES cryptographic decryption core. This work successfully shields AES cryptographic decryption core using Defense in Depth approach.**

Keywords*— *Defense in Depth, Hardware Security, Design for security, Logic locking, AES Decryption Core, AES Lock Block.

I. INTRODUCTION

Security is a major concern these days. Amidst the era of fabless companies, for the cost saved in manufacturing a significant effort has to be put in to curb variety of security threats like IP Piracy, Reverse Engineering and Hardware Trojan and a steep increase in the number of untrustworthy electronics which are encountered due to deployment of IC fabrication in the global design flow. Hardware security is an essential tool being worked upon recently to restrain the huge number of losses caused to the semiconductor industry due to the above-mentioned problems to mitigate malicious threats from attacker's end, Defense in Depth approach stands tall as it implements mechanisms in layered fashion. Design for Trust techniques (DfTr) [1] are also utilized which are broadly classified into active and passive. IC metering, Fingerprinting, Watermarking, Logic locking, IC camouflaging are examples of DfTr techniques. Out of them, logic locking is one the various active techniques which has been garnering more attention from the research world recently because of versatility, lesser limitations than others and its ability to protect against an attack in anywhere in the supply chain.

Logic locking method involves insertion of additional hardware and numerous extra key-inputs, in addition to the original inputs, into the design which are driven by an on-chip tamper proof memory. The added locking hardware makes sure that the design details cannot be recovered with the help of reverse engineering and ensures that the output is incorrect if the correct key is not provided at the input. There are various sequential and combinational logic locking techniques. The former ones' result in incorrect output since it is corrupted due to the presence of incorrect keys. Various combinational logic locking techniques implement XOR/XNOR key gates, AND/OR gates, multiplexers, or combinations of these gates [2]. Kundi et al. AES encryption core is implemented on Spartan-3 FPGA. The design includes MUX in the intermediate states to fed back data into encryption core [3]. The implementation provided fast encryption core. Santoosh et al. elaborates that AES encryption and decryption algorithm was carried out by implementing more than one round parallelly [4]. It increased throughput and offered high security.

In this work, Defense in Depth approach is administered to prevent malicious attacks like key sensitization, brute force attack and reverse engineering and secure critical data using lock locking technique on AES decryption core. Section II provides brief on the state of art technology. Section III elaborates on the AES decryption core, AES lock block core and Performance metrics. Section IV presents the results justifying all our claims. Section V presents the conclusion of this work.

II. OVERVIEW

Logic locking techniques are utilized to protect against malicious threats. Stripped functional behaviour logic locking technique is proposed by Yasin et al. which mentions that certain amount of functional behaviour is latent in form of secret key [5]. A stripped-functionality logic locking (SFLL) technique which strips the functionality of the design and hides it in the form of a secret key. The stripped functional behaviour is recovered only through on-chip restore process. There are a number

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

of ways to attack the design. SAT attack which is a recent and a fatal one which has the ability to decipher the right key of almost all logic locking techniques. Xie et al. explains one of the effective methods to counteract this is proposed in [6]. Delay locking determines functionality and time profiles. A key with incorrect timing will end up in violation of time and malfunction of the circuit. A delay key gate which is tunable is also proposed which can overshadow both functionality and timing profile of IC design. Sweeney et al. proposes Latch-based logic locking technique [7]. This method alters data flow and logic in the circuit by inserting latches. Karmakar et al. describes Logic Encryption strategy [8]. This strategy inserts key gate in such a way that quality of security improves and enhances key-interdependence due to incorporation of key-dependency module. Torrance et al. discusses Reverse engineering techniques [9]. These techniques extract design information which is confidential. Yasin et al. introduces Strong logic locking method that has been implemented in such a way that the key gates are inserted where it is non mutable. Drawback of this method is that it is laborious to find interfering key locations and does not guarantee output corruption. The study also states that existing SAT-based attacks can be averted possibly using one-way random functions. The technique of weighted logic locking [11], [12], [13] was proposed recently in which multiple key inputs are given to every control gate, usually at locations of highest fault impact, instead of the conventional technique where a single key locks the entire circuit. To gain security, Rekha et al. has implemented logic locking concept to protect the clock line of I2C protocol by securing data [14]. When loaded onto an on-chip memory, the secret keys restore the original functionality of the design and creates mismatch between the reverse- engineered netlist and original design. Baby et al. proposes a technique known as LUT based dynamic obfuscation. It ensures that the functionality is latent from untrustworthy stages of design flow [15]. Benchmark circuits were analysed related to number of cycles and hamming distance. Power consumption and area overhead is also decreased.

III. PROPOSED METHODOLOGY

The motive behind this work is to secure the critical data and prevent malicious attacks by administering the defense in depth approach at system level. This approach surges redundancies which reduces momentum of attacker resulting in enhanced reliability. This work is divided into two phases where first phase elaborates the defense in depth approach and second phase evaluates its performance using specific parameters.

A. Defense of Depth

In this work, Defense in Depth approach is applied in two stages. In the first stage, logic locking technique is used to authenticate and grant access control to implement AES algorithm and in the second stage, security is propounded by performing AES cryptographic decryption algorithm. The attacker has to obtain access in order to access the crypto primitive algorithm (AES) before key guessing resulting in high protection of the data.At gate level, in each round, the control signals from logic lock block are connected to different logic cells. This increases random behaviour and disables the attacker from differentiating the correct and incorrect output control signals. This approach adds multiple layers of security and strengths it against malicious attacks. Figure 1 depicts the Defense in Depth approach adopted in this project.

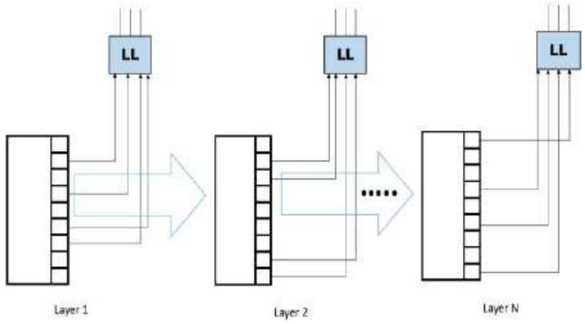


Fig. 1. Defense in Depth Approach

AES Decryption. AES algorithm is widely used symmetric cryptographic algorithm as it is more secure and has low cost for implementation. AES decryption algorithm is iterative in nature. The transformation occurs in every round of the decryption algorithm. Each round consists four subprocess. The last round is an exception as it excludes the Inverse Mix Column subprocess. AES Decryption algorithm has 4 main components. Plain text, Secret key, Cipher text and Decryption algorithm. AES Decryption algorithm consists of four subprocesses. Inverse Shift rows, Inverse Substitution Bytes, Add round key, Inverse Mix column. AES core is vulnerable to attacks. Absence of strong secure system has led to such malicious attacks and insertion of trojans. Numerous techniques are implemented to detect the presence of Trojans. But the call is for a safe and smart technique to safeguard the IC by adding extra layers of security using logic locking technique.

There are three types of AES decryption algorithms depending on key size. AES supports key size of 128, 192 and 256 bits. The number of rounds increases with respect to key size. In AES algorithm, the input text data size always remains 128 bits. The organization of AES decryption algorithms are illustrated in Table I.

TABLE I. ORGANIZATION OF AES

Type	Key Size	Data Size	Rounds
AES 128	128	128	10
AES 192	192	128	12
AES 256	256	128	14

Logic Locking. Logic Locking is an emerging technology in the sphere of Design for security. Purpose behind using logic locking technique in this work is to protect against vulnerable attacks. In this technique, the attacker is unaware of the presence of logic lock block and hence obtains corrupted text while performing attacks. This technique reduces the probability of the model being prone to attacks. Lock module is inserted to ensure that the correct output is obtained only in presence of correct password. The output is corrupted when incorrect password is provided at input.

Logic lock block is inserted in each subprocess of AES Decryption algorithm. It is inserted judiciously at system level in such a manner where the probability of differentiating it from the original circuit is minimal. The 16-bit lock key is flashed in tamper proof memory (D flash). The lock block consists of non-linear weighted hexadecimal component and a comparator. Comparator compares the weighted hexadecimal with the input password.

The naming convention of the block is in systematic form. In the naming convention, key size is mentioned first followed by B which indicates Block and then followed by numerical value which in turn indicates the subprocess as illustrated in the Table II. For example, 192-B3 indicates that the logic lock block is inserted in the AddRound Key subprocess of AES-192 algorithm. The schematic of AES algorithm with lock block using Defense in Depth is depicted in Figure 2.

TABLE II. NAMING CONVENTION

Value after B	Corresponding Sub-Process
1	Inverse Shift Row
2	Inverse Substitution Byte
3	AddRound Key
4	Inverse Mix Column

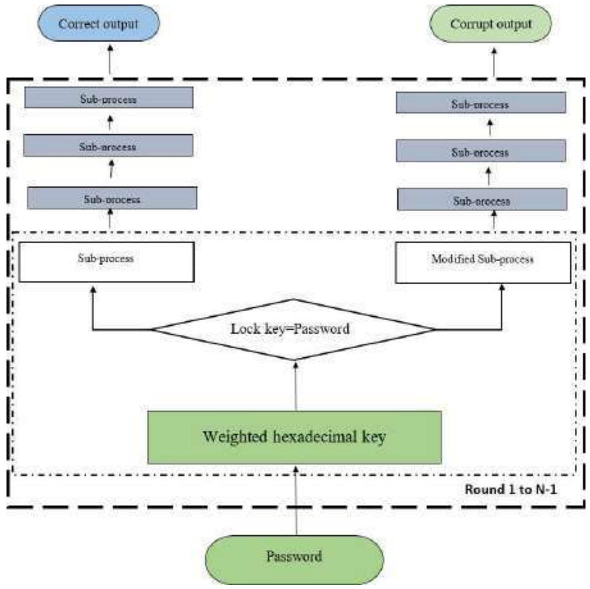


Fig. 2. AES lock block in Defense in Depth approach

1) *Insertion of Lock Block in Inverse Shift row:* The lock block accepts the password and maps the bits to weighted hexadecimal bits and compares it with the lock key in tamper proof memory. The block unlocks and executes the remaining subprocess without alteration if the password provided at input is correct. If the password provided is incorrect, the inverse shift rows sub process is altered by varying the offsets to shift the rows cyclically.

2) *Insertion of Lock Block in Inverse Substitution Byte:* The lock block accepts the password and maps the bits to weighted hexadecimal bits and compares it with the lock key in tamper proof memory. The block unlocks and executes the remaining subprocess without alteration if the password provided at input is correct. If the password provided is incorrect, the inverse substitution sub process is altered by replacing the mapping elements in the substitution box.

3) *Insertion of Lock Block in Add Round Key:* The lock block accepts the password and maps the bits to weighted hexadecimal bits and compares it with the lock key in tamper proof memory. The block unlocks and executes the remaining subprocess without alteration if the password provided at input is correct. If the password provided is incorrect, the AddRound key subprocess is altered by including additional XOR function to perform XOR operation with a random constant.

4) *Insertion of Lock Block in Inverse Mix Column*: The lock block accepts the password at input and maps the bits to weighted hexadecimal bits and compares it with the lock key in tamper proof memory. The block unlocks and executes the remaining sub process without alteration if the lock key provided at input is correct. If the password provided is incorrect, the inverse mix column sub process is altered by performing XOR operation on the resultant value of multiplicative inverse.

B. *Performance Metrics*

Performance metric such as distance metrics, power, time and Information gain are evaluated to estimate the performance of the Defense in Depth approach by examining Lock block. Distance metrics such as Hamming distance, Levenshtein distance and Jaro distance are utilized to estimate the performance of the algorithm and evaluate the extent to which the output is corrupted when incorrect password is provided at the input.

Hamming Distance. Hamming distance is the difference between the number of bits in output state matrix and input state matrix. Substitution of bits are carried out for evaluating the hamming distance.

Levenshtein Distance. Levenshtein distance is the difference between the number of bits in the input state matrix and output state matrix. Insertion, deletion and substitution operations are carried out to transform input state matrix to output state matrix for evaluating Levenshtein distance.

Jaro Distance. Jaro distance is difference between the number of bits in input state matrix and output state matrix. Transposition is carried out to transform from input state matrix to output state matrix for evaluating Jaro distance.

Power and Time. Power and time consumed is measured by performing side channel power analysis. Side channel power analysis is performed on Chip Whisperer Lite board using the measurement setup.

Information Gain. Information gain is a measure of the information obtained. Entropy is lack of order. Information gain is the difference of entropy before transformation and after transformation. The formula for Information gain is given by formula, Information Gain = [entropy(parent)] – [average entropy(children)]. Entropy is computed using the formula, Entropy = $- \sum P_i \log_2 P_{i.}$

IV. RESULTS AND ANALYSIS

This work is implemented on AES symmetric cryptographic decryption core with key size of 128,192 and 256 bits. Table III depicts the correct output and corrupted output values of AES crypto primitive for varying inputs when Lock block is inserted.

TABLE III. SAMPLE CASES

Case I	Correct Password	0x0C, 0x0B
	Incorrect Password	0x0D,0x0E
	Weighted hexadecimal	0x0A,0x0F
	Key	0xb267516182ea2d6aba7f518890fcf4c3
	Input	0x935248d120cd90bfcd11587991a6b023
	Output	0xdc4e3f68ddb1c6e1ddc8236e5f562aa
	Corrupted output_128AR	0xadc5c7640060ef51968bee4664ecc8b0
	Corrupted output_128MC	0x6dbe5aa938416feeb1b78e51d2784e68
	Corrupted output_128SB	0x23c7c820a97bc328756ff28a07745ac4
	Corrupted output_128SR	0xc01e90698bf037c85536405784b07fa
Case II	Correct Password	0x0A, 0x0B
	Incorrect Password	0x0C,0x0D
	Weighted hexadecimal	0x0D,0x0E
	Key	0x5520617265207468652062657374206966e20776f26c6420a
	Input	0xd8f3a72fc3cdf74dfaf6c3e6b97b2fa6
	Output	0x55a33826068337bbf99165c6530d134c
	Corrupted output_192AR	0xc66d9bd03393c575350902a77fa995a7
	Corrupted output_192MC	0xe934ba0a54cd0cd0be58be710bcc1184
	Corrupted output_192SB	0x84f6ca442db2d935378695ded3b20888
	Corrupted output_192SR	0x5f15645c80789c83467c9e4d55c9f425
Case III	Correct Password	0x0C, 0x0D
	Incorrect Password	0x0A,0x0B
	Weighted hexadecimal	0x09,0x0A
	Key	0x6c6f6769636c6f636b696e6769736173656372657 4636f6e63657074206f75747
	Input	0x26f39bbca19c0fb7c72e7e3063927313
	Output	0x29f7741db1f4a824f2d6ab7d18051543
	Corrupted output_256AR	0x13e758ed901536abf0673ecf1843230e

	Corrupted output_256MC	0x229a61ac9c057f356dac209279259e1a
	Corrupted output_256SB	0x077550bfae6a7741d960b22381375f03
	Corrupted output_256SR	0xf0f74252fc4d997d8ce3fcf1e34eb3a0

In figures 3,4 and 5, x-axis indicates the key size and subprocess where the lock block is judiciously inserted and the y-axis indicates the dissimilarity measure of input text and output text. The uncorrupted output is obtained by unlocking the lock block if correct password is provided at input whereas corrupted output is obtained when the lock block remains locked if incorrect password is provided at input.

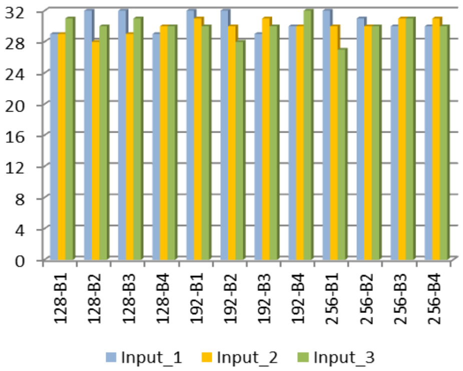


Fig. 3. Analysis of Hamming distance

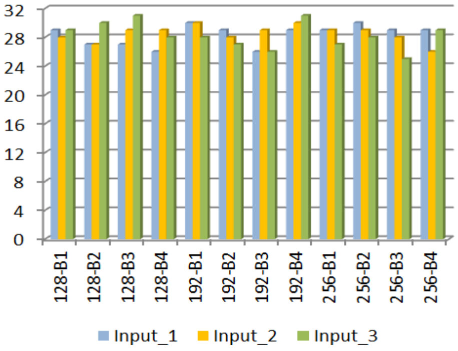


Fig. 4. Analysis of Levenshtein distance

Figure 3 depicts the analysis of Hamming distance in AES 128,192 and 256 when AES Lock block is judiciously inserted in the each of the four subprocess of AES Decryption algorithm for three different set of input cipher text and input secret key. From the figure, it can be inferred that the output text is entirely corrupted with Hamming distance greater than 80%. Figure 4 depicts the analysis of Levenshtein distance in AES 128,192 and 256 when AES Lock block is judiciously inserted in each of the four sub process of AES Decryption algorithm for three different set of input cipher text and input secret key. From the above figure, it can be inferred that the output text is entirely corrupted with Levenshtein distance greater than 80%.

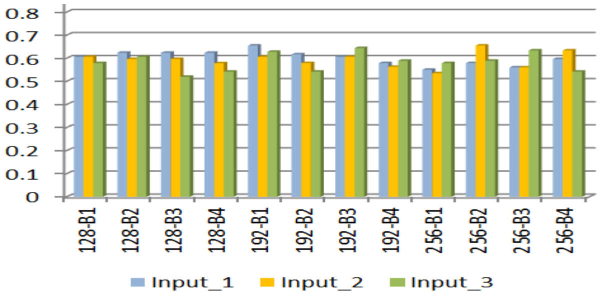


Fig. 5. Analysis of Jaro distance

Figure 5 depicts the analysis of Jaro distance in AES 128,192 and 256 when AES Lock block is judiciously inserted in each of the four subprocess of AES Decryption algorithm for three different set of input cipher text and input secret key. From the Figure, it can be inferred that the output text is entirely corrupted with Jaro distance greater than 50%.

It can be inferred that the hamming distance, Levenshtein distance and Jaro distance is above 50%. Logic locking technique is resilient to brute force attack as this technique is over shadowing the vulnerabilities due to added security control access layer. Further this technique is sturdy enough to resist key sensitization attack since the conversion of input password to non-linear weighted hexadecimal tightens the security by making it difficult for the attacker to identify correlation. between input password and weighted hexadecimal. Also, Reverse engineering is inhibited by mystifying the attacker through layered structure of defense in depth approach.

Further, power consumption and execution time are measured using ChipWhisperer Lite board. The power consumed and execution time while running conventional AES Decryption core and AES Decryption core with AES lock block is depicted in Figure 6 and Figure 7 respectively.

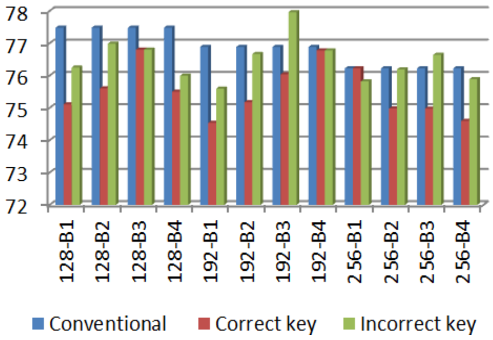


Fig. 6. Analysis of Power Measurements

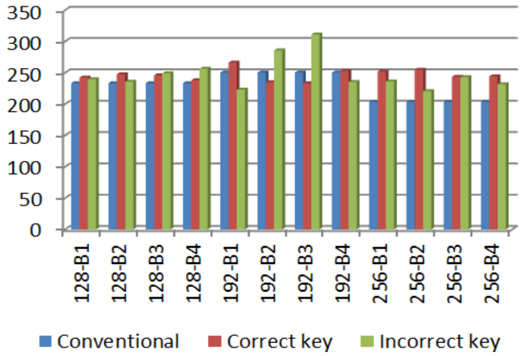


Fig. 7. Analysis of Execution time

Figure 6 depicts the analysis of Power when AES lock block is judiciously inserted in each of the four subprocess of AES 128, 192 and 256 when correct password and incorrect password are provided at input. The x-axis indicates the key length and subprocess where it is inserted and the y-axis indicates the power consumption. From the plot, it can be clearly inferred that the difference between power consumed by conventional AES core and AES core with AES lock block is less than 4 % (negligible differences of µW). Figure 7 depicts the analysis of execution time when AES lock block is judiciously inserted in each of the four subprocess of AES 128,192 and 256 when correct password and incorrect password are provided at input. The x-axis indicates the key length and subprocess where it is inserted and the y-axis indicates the execution time of the circuit.

The difference between execution time of conventional AES core is less than AES core with AES lock block core by 6% for key sizes 128 and 192, and 20% for key size of 256 (negligible differences in ms). Hence, AES core with AES lock block does not drastically increase the power consumption and execution time. So, it is a difficult task for attackers to obtain the password using side channel attacks.

Information gain has been evaluated considering the power consumption attribute. Information gain for AES 128,192 and 256 has been evaluated which has depicted in Table III. It can be inferred from the table that the information gain obtained is less. It is extremely difficult to rely on the information obtained from power consumed to differentiate the correct output and incorrect output when Lock block is inserted.

TABLE IV. INFORMATION GAIN

	Entropy	**Average Entropy**	**Information Gain**
AES 128	0.992774	0.853473	0.139301
AES 192	0.896038	0.775885	0.120153
AES 256	0.746234	0.651909	0.094326

V. CONCLUSION

In this work, the AES lock block has been successfully designed and judiciously inserted into AES cryptographic decryption core in each of the four subprocess of AES decryption algorithm. Power consumed and execution time was measured in XMEGA microcontroller with difference of less than 6% and 20% respectively. Information gain evaluated is also minimal to perform attacks. This technique achieved 80% Hamming distance, 80% Levenshtein distance and 50% Jaro distance. This technique is resilient to Key sensitization attack, reverse engineering and brute force attack. Hence, it is validated that the proposed method is an efficient defense in depth approach which uses logic locking technique. This approach is an effective security measure to avoid unauthorized access on AES Decryption core.

REFERENCES

[1] Yasin, Muhammad & Mazumdar, Bodhisatwa & Rajendran, Jeyavijayan & Sinanoglu, Ozgur. (2019). Hardware Security and Trust: Logic Locking as a Design-for-Trust Solution: Design and Implementation. 10.1007/978-3-319-93100-5_20.

[2] M. Yasin and O. Sinanoglu, "Evolution of logic locking," 2017 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), Abu Dhabi, 2017, pp. 1-6, doi: 10.1109/VLSI- SoC.2017.8203496.

[3] Kundi, Dur-e-Shahwar & Zaka, Saleha & Qurat-Ul-Ain, & Aziz, Arshad. (2009). A compact AES encryption core on Xilinx FPGA. 1 - 4. 10.1109/IC4.2009.4909251.

[4] S. K. R, S. R, M. A. M, P. K. M. S and R. M, "Design of High Speed AES System for Efficient Data Encryption and Decryption System using FPGA," 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization

Techniques (ICEECCOT), Msyuru, India, 2018, pp. 1279-1282, doi: 10.1109/ICEECCOT43722.2018.9001535.

[5] Muhammad Yasin, Abhrajit Sengupta, Mohammed Thari Nabeel, Mohammed Ashraf, Jeyavijayan (JV) Rajendran, and Ozgur Sinanoglu. 2017. Provably-Secure Logic Locking: From Theory to Practice. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17). Association for Computing Machinery, New York, NY, USA, 1601–1618.

[6] Y. Xie and A. Srivastava, "Delay locking: Security enhancement of logic locking against IC counterfeiting and overproduction," 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, TX, 2017, pp. 1-6, doi:10.1145/3061639.3062226.

[7] J. Sweeney, V. Mohammed Zackriya, S. Pagliarini and L. Pileggi, "Latch-Based Logic Locking," 2020 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), San Jose, CA, USA, 2020, pp. 132-141, doi:10.1109/HOST45689.2020.9300256.

[8] R. Karmakar, N. Prasad, S. Chattopadhyay, R. Kapur and I. Sengupta, "A New Logic Encryption Strategy Ensuring Key Interdependency," 2017 30th International Conference on VLSI Design and 2017 16th International Conference on Embedded Systems (VLSID), Hyderabad,2017, pp. 429-434, doi: 10.1109/VLSID.2017.29.

[9] R. Torrance and D. James, "The state-of-the-art in semiconductor reverse engineering," 2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC), New York, NY, 2011, pp. 333-338.

[10] M. Yasin, J. J. Rajendran, O. Sinanoglu and R. Karri, "On Improving the Security of Logic Locking," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 35, no. 9, pp. 1411-1424, Sept. 2016, doi: 10.1109/TCAD.2015.2511144.

[11] J. Rajendran, Y. Pino, O. Sinanoglu and R. Karri, "Logic encryption: A fault analysis perspective," 2012 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, 2012, pp. 953-958, doi: 10.1109/DATE.2012.6176634.

[12] N. Karousos, K. Pexaras, I. G. Karybali and E. Kalligeros, "Weighted logic locking: A new approach for IC piracy protection," 2017 IEEE 23rd International Symposium on On-Line Testing and Robust System Design (IOLTS), Thessaloniki, 2017, pp. 221-226, doi: 10.1109/IOLTS.2017.8046226.

[13] S. Krishnan, M. K. N. and N. D. M., "Weighted Logic Locking to Increase Hamming Distance against Key Sensitization Attack," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 29-33, doi: 10.1109/ICECA.2019.8821880.

[14] Rekha, S. & Reshma, B. & Dilipkumar, N. & Crocier, A. & N., Mohankumar. (2020). Logically Locked I2C Protocol for Improved Security.10.1007/978-981-15-2612-1_67.

[15] Baby J., Mohankumar N., Nirmala Devi M. (2020). Reconfigurable LUT-Based Dynamic Obfuscation for Hardware Security. In: Sengodan T., Murugappan M., Misra S. (eds) Advances in Electrical and Computer Technologies. Lecture Notes in Electrical Engineering, vol 672 Springer, Singapore. https://doi.org/10.1007/978-981-15-5558-9_81.

INVARIANT CONTINUATION OF DISCRETE MULTI-VALUED FUNCTIONS AND THEIR IMPLEMENTATION

Anvar Kabulov
Department of "Applied mathematics
and intellectual technologies
National University of Uzbekistan
named after Mirzo Ulugbek
Tashkent, Uzbekistan
anvarkabulov@mail.ru

Ibrokhimali Normatov
Department of "Applied mathematics
and intellectual technologies
National University *of Uzbekistan*
named after Mirzo Ulugbek
Tashkent, Uzbekistan
ibragim_normatov@mail.ru

Erkin Urunbaev
Department of "Mathematics"
Samarkand State University
Samarkand, Uzbekistan
urin54@rambler.ru

Firdavs Muhammadiev
Department of "Applied mathematics
and intellectual technologies
National University *of Uzbekistan*
named after Mirzo Ulugbek
Tashkent, Uzbekistan
furik7live@mail.ru

***Abstract*–The article discusses partially defined discrete functions and investigates the problem of logical continuation of such functions in the class of canonical normal forms. A theorem is proved on the invariant continuation of a multivalued function in the class of canonical forms, which does not depend on the adopted coding. An algorithm for constructing a set of invariant points independent of the adopted encoding and software for such an invariant continuation are constructed.**

Key words: encoding, logic, conjunction, disjunction, sets, quasi-Boolean invariant, synthesis, equivalent.

INTRODUCTION

Extension of a discrete function by the invariance principle. Let $S^n = S_1 \times S_2 \times ... \times S_n$ the Descartes product where $S_i,\ i = \overline{1, n}$ is an arbitrary collection of real numbers. Consider a formula (metric) $\rho(\alpha, \beta) = \sum_{i=1}^{n} |\alpha_i - \beta_i|$. A ball $S_R^{\tilde{\alpha}}$ with center $\tilde{\alpha}$ and radius r at S^n is the set of all $\tilde{\beta} \in S^n$ such that $\rho(\tilde{\alpha}, \tilde{\beta}) \le r$ [1-3].

Let $F(x_1, x_2, ..., x_n)$ is a discrete function, taking k values $\{0, 1, ..., k-1\}$ be defined on the set $M \subseteq S^n : F(\tilde{x}) = j,$ if $\tilde{x} \in M_j,\ (j = \overline{0, k-1}),$ $M = \bigcup_{i=0}^{m} M_i$ and $M_i \bigcap M_j = \varnothing$. Function $F'(\tilde{x})$ is called a continuation $F(\tilde{x})$, if $F'(\tilde{x}) = j$, for $\tilde{x} \in M_j,$ $M_j \subseteq M'_j,$ $M_i \bigcap M_j = \varnothing,\ \left(j = \overline{0, k-1}\right)$ for $i \ne j.$

It is easy to see that you can find different function extensions $F(\tilde{x})$, that are not equal to each other. Let us define the set of points where the continuation of the function $F(\tilde{x})$ does not depend on the adopted encoding [4-15].

Let us define an elementary function (e.f.)

$$\mathfrak{A}_{\tilde{\alpha},r}^{r}(\tilde{x}) = \begin{cases} \gamma, & \text{if } \ \tilde{x} \in S_r^{\tilde{\alpha}}, \\ 0, & \text{otherwise.} \end{cases}$$

The ball $S_r^{\tilde{\alpha}}\left(\mathfrak{A}_{\tilde{\alpha},r}^{\gamma}\right)$ is called admissible for $F(\tilde{x})$ and

$$M_j, \text{ if } S_n^{\tilde{\alpha}} \bigcap M_j \ne \varnothing, \quad S_n^{\tilde{\alpha}} \subseteq S^n \setminus \bigcup_{j=1}^{\gamma-1} M_j.$$

Further we will investigate the representation of the function $F(\tilde{x})$ in the form of the maximum of the e.f., the domain of which is chosen to be balls.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Let M_i - a ball $S_{r_i}^{\tilde{\alpha}_i}$ with radius r_i and center at $\tilde{\alpha}_i$, $i=\overline{0, n}$.

Consider a point $\tilde{\alpha}$ to preserve the encoding of the ball $S_r^{\tilde{\alpha}}$ and a point $\tilde{\delta}_j$ on a segment $(\tilde{\alpha}, \tilde{\alpha}_j)$, $j=\overline{1, n}$, $j \neq i$ such that $\rho(\tilde{\alpha}, \tilde{\delta}_j)=\rho(\tilde{\alpha}, \tilde{\alpha}_j)-r_j$. It is easy to see that, if $\rho(\tilde{\alpha}, \tilde{\delta}_j) \geq \rho(\tilde{\alpha}, \tilde{\alpha}_i)-r_i$ for any $j \neq i$, then $\tilde{\alpha}$ will be a point that preserves the code of the set M_i.

Let a set of discrete *k*-valued functions depending on *n* variables be given, i.e. the set of functions defined on the set of all vertices of the Descartes product $S^n=S_1 \times S_2 \times \ldots \times S_n$ where is S_i, $i=\overline{1, n}$ an arbitrary collection of real numbers $i=\overline{0, n}$ and taking values from the set $\{0,1, \ldots, k-1\}$. In this consideration, there is a one-to-one correspondence between multivalued discrete functions, depending on the *n* arguments and subsets $N_f \subseteq S^n$. The function $f(x_1, x_2, \ldots, x_n)$ and the subset $N_f \subseteq S^n$ have a one-to-one correspondence in the case

$$f(x)=\begin{cases}\gamma, & if \;\; x \in N_f \\ 0, & if \;\; x \in S^n \setminus N_f\end{cases},$$

here $\gamma \in E_f \subseteq\{0,1, \ldots, k-1\}$.

It is assumed that the set E_f splits the function $f(x_1, x_2, \ldots, x_n)$ into the set of subfunctions $f_{\gamma_1}(\tilde{x}), f_{\gamma_2}(\tilde{x}), \ldots, f_{\gamma_m}(\tilde{x})$, and the set N_f into pairwise disjoint subsets $N_{f_{\gamma_1}}, N_{f_{\gamma_2}}, \ldots, N_{f_{\gamma_m}}$, where $m=|E_f|$,

$$f_{\gamma_i}(\tilde{x})=\begin{cases}\gamma_i, & if \;\; f(x)=\gamma_i \\ 0, & if \;\; f(x) \neq \gamma_i\end{cases},$$

$$N_{f_{\gamma_i}}=\{\tilde{\alpha}:(\tilde{\alpha} \in S^n) \wedge(f(\tilde{\alpha})=\gamma_i)\}, \; (i=\overline{1, m}).$$

It is easy to see that

$$\mathrm{J}_{\mathrm{M}}(x)=\begin{cases}k-1, & if \;\; x \in \mathrm{M} \\ \;\;0, & if \;\; x \notin \mathrm{M}\end{cases},$$

where $\mathrm{M} \subseteq E_k=\{0,1,2, \ldots, \mathrm{k}-1\}$.

An elementary function (e.f.) is the expression [16]

$$\mathfrak{A}=\min [J_{M_1}(x_1), J_{M_2}(x_2), \ldots, J_{M_n}(x_n), \gamma],$$

where

$$\varnothing \neq \mathrm{M}_j \subseteq E_k, \; (j=\overline{1, n}).$$

In the further presentation, for the sake of convenience, the formula $\max [\mathfrak{A}_1, \mathfrak{A}_2, \ldots, \mathfrak{A}_m]$ will be conventionally denoted as $\mathfrak{A}_1 \vee \mathfrak{A}_2 \vee \ldots \vee \mathfrak{A}_m=\bigvee_{i=1}^m \mathfrak{A}_i$: if $\mathfrak{A}_i$ there is an analogue of an e.f., then the indicated formula will be called the canonical normal form (c.n.f.).

The area of truth e.f. $\mathfrak{A}$ we will call the set $N_{\mathfrak{A}}$, in which the e.f. $\mathfrak{A}$ equals γ. It is easy to see that the domain $N_{\mathfrak{A}}=\prod_{j=1}^n M_j$ is a ball (a subset of a set S^n) of a Descartes product S^n. With such a geometric consideration of the e.f. matches the ball $N_{\mathfrak{A}}$, of the Descartes product S^n.

The rank of e.f. $\mathfrak{A}$ let's call the number

$$r(\mathfrak{A})=\sum_{j=1}^n(k-|\mathrm{M}_j|)=kn-\sum_{j=1}^n|\mathrm{M}_j|.$$

The formula $\mathfrak{M}=\bigvee_{i=1}^t \mathfrak{A}_i$, where everything $\mathfrak{A}_i$, $(i=\overline{1, t})$ is e.f., will be called the canonical normal form (c.n.f.).

Note that each multivalued discrete function $f(x_1, x_2, \ldots, x_n)$ corresponds to a non-empty class of c.n.f. that implement this function. The set of all balls corresponding to the e.f. a certain c.n.f. from this class, defines the coverage N_f of balls from S^n. Whence it follows that subsets $\mathrm{M} \subseteq E_k^n$ can be specified using c.n.f [17, 18].

Let $I=\{N_{\mathfrak{A}}\}$, there be some subset of balls from S^n.

A ball $N_B \in I$ is called maximal with respect to M if it does not exist in I. the ball $N_{\mathfrak{A}}$, such that $N_{\mathfrak{A}} \neq N_B$ and $N_{\mathfrak{A}} \supseteq N_B$.

To represent the function $f(x_1, x_2, \ldots, x_n)$ in the form of c.n.f. we considered its quasi-Boolean representation: $f=f_{\gamma_1} \vee f_{\gamma_2} \vee \ldots \vee f_{\gamma_m}$ moreover $\gamma_1<\gamma_2<\ldots<\gamma_m$.

It is easy to see that for the same function $f(\tilde{x})$ can have several equal quasi-Bullian representations. Indeed,

$$\begin{aligned} f &= f_{\gamma_1} \vee f_{\gamma_2} \vee \ldots \vee f_{\gamma_m} = \\ &= f^*=f_{\gamma_1}^* \vee f_{\gamma_2}^* \vee \ldots \vee f_{\gamma_m}^*, \end{aligned}$$

where

$$N_{f_{\gamma_i}^*}=N_{f_{\gamma_i}} \bigcup Q_i,$$

$$Q_i \subseteq \bigcup_{j>i} N_{f_{\gamma_j}}, \left(i=\overline{1, m}\right).$$

We will consider only one "maximum" view

$$f' = f'_{\gamma_1} \vee f'_{\gamma_2} \vee \ldots \vee f'_{\gamma_m},$$

where

$$N_{f'_{\gamma_i}} = \bigcup_{j=1}^{n} N_{f_{\gamma_j}}, \ \left(i=\overline{1,m}\right).$$

Consider all maximal balls $N_{B_j^i}, \left(i=\overline{1, m}\right)$, included in $N_{f'_{\gamma_i}}$, intersecting with $N_{f_{\gamma_i}}$ and such that the value B_j^i is equal γ_i to $N_{B_j^i}, \left(i=\overline{1, m}\right)$.

$\mathfrak{M} = \bigvee_{i=1}^{m} \bigvee_{j=1}^{I_i} B_j^i$ will be called the reduced canonical normal form of the function $f\left(\tilde{x}\right)$.

A covering of a set N_f with the maximal balls is said to be irreducible if, after removing any of the balls included in it, it will not be a covering.

C.n.f., realizing a function f is called dead-end if it corresponds to an irreducible cover of the set N_f.

Consider a multivalued function $F\left(x_1, x_2, \ldots, x_n\right)$ defined on

$$M \subseteq S^n : F\left(\tilde{x}\right) = \gamma_j, \text{ if } \tilde{x} \in M_j, \left(j=\overline{1, m}\right),$$

$$m<k, \ \gamma_j \in E_k, \ M = \bigcup_{i=0}^{m} M_i \text{ and } M_i \bigcap M_j = \varnothing \text{ for}$$

$\left(i \neq j, i, j = \overline{0, m}\right)$. Moreover $\gamma_1 < \gamma_2 < \ldots < \gamma_m, \ \gamma_0 = 0.$

Hence $F\left(x_1, x_2, \ldots, x_n\right)$ is defined by specifying pairwise disjoint sets $\mathrm{M}_0, \mathrm{M}_1, \ldots, \mathrm{M}_{\mathrm{m}}$. The function $F\left(\tilde{x}\right)$ is partially defined on the entire set S^n. There are various extensions in the class of multivalued discrete functions $F\left(\tilde{x}\right)$ that are not equivalent to each other.

The task of this work is to find the simplest, in a sense, additional definitions.

For $F\left(\tilde{x}\right)$, select all maximal balls $N_{B_j^i}$, $\left(i=\overline{1, m}, \ j=\overline{0, I_i}\right)$, contained in $E_k^n \setminus \bigcup_{v=0}^{i-1} M_v$, intersecting with M_i such that the value B_j^i is γ_i.

C.n.f. $\mathfrak{M} = \bigvee_{i=1}^{m} \bigvee_{j=1}^{I_i} B_j^i$ will be called the abbreviated normal form for $F\left(\tilde{x}\right)$. It is easy to see that c.n.f. $\mathfrak{M}_{\Sigma\mathrm{T}F}$ is uniquely determined by the function $F\left(\tilde{x}\right)$.

Let us now indicate the points at which, when the values of the function change $F\left(\tilde{x}\right)$ (transition to $F'\left(\tilde{x}\right)$), the values change $\mathfrak{M}_{\Sigma\mathrm{T}F}$ (transition to $\mathfrak{M}_{\Sigma\mathrm{T}F'}$).

Consider a multivalued discrete function $F\left(x_1, x_2, \ldots, x_n\right)$ defined on $M \subseteq S^n$:

$$F\left(\tilde{x}\right) = \gamma_j, \text{ if } \tilde{x} \in M_j, \left(j=\overline{0, k-1}\right), \tag{2}$$

where $M = \bigcup_{i=0}^{m} M_i$ and $M_i \bigcap M_j = \varnothing \ i \neq j.$

We denote by $\left\{\pi\right\}$ the set of all permutations of the set

$$\left\{0, 1, \ldots, k-1\right\} : \pi = \left(i_0, i_1, \ldots, i_{k-1}\right).$$

The functions

$$F_\pi\left(\tilde{x}\right) = i_j, \text{ if } \tilde{x} \in M_j, \left(j=\overline{0, k-1}\right),$$

we call π is a permutation of $F\left(\tilde{x}\right)$.

We will say that a point $\tilde{\alpha} \in E_k^n \setminus \mathrm{M}$ preserves the encoding (code) of a set $M_j \subseteq M, \ j=\overline{0, k-1}$ with respect to a permutation π, if $\mathfrak{M}_{\Sigma\mathrm{TF}}\left(\tilde{\alpha}\right) = j$ and $\mathfrak{M}_{\Sigma\mathrm{TF}_\pi}\left(\tilde{\alpha}\right) = i_j$.

A point $\tilde{\alpha} \in E_k^n \setminus \mathrm{M}$ is called a point that preserves the encoding of the set M_j, if it preserves the encoding M_j with respect to any permutation $\pi \in \left\{\pi\right\}$.

Let

$$\mathfrak{M} = \bigvee_{\pi : \pi \in (\pi)} \mathfrak{M}_{\Sigma\mathrm{TF}_\pi} \ .$$

Theorem. The point $\tilde{\alpha} \in E_k^n \setminus \bigcup_{i=0}^{k-1} \mathrm{M}_i$ preserves the code of the set M_j, then and only if, when:

1) in c.n.f. $\mathfrak{M}$ there is an e.f. $\mathfrak{A}$ such that

$$N_{\mathfrak{A}} \bigcap M_j \neq \varnothing, \ \tilde{\alpha} \in N_{\mathfrak{A}}, \ N_{\mathfrak{A}} \bigcap M_i = \varnothing,$$

where $\left(i = 0, \ldots, j-1, j+1, \ldots, k-1\right)$;

2) each ball $N_{\mathfrak{A}}$ where e.f. $\mathfrak{A}$ included in c.n.f.

$\mathfrak{M}$.

Has common points with M_i, $(i \neq j)$ and contains a point M_j.

Proof. Necessity. Let the point $\tilde{\alpha}$ preserves the code of the set M_j Then for the encoding $\pi = \{i_0, i_1, \ldots, i_{j-1}, i_{k-1}\}$ of c.n.f. $\mathfrak{M}_{\Sigma TF_\pi}$ there is an e.f. $\mathfrak{A}$ such that $\mathfrak{A}(\tilde{\alpha}) = k-1$.

By definition,

$$N_{\mathfrak{A}} \subseteq E_k^n \setminus \bigcup_{\substack{i=0 \\ i \neq j}}^{k-1} \mathrm{M}_i,$$

i.e. the interval $N_{\mathfrak{A}}$ does not intersect with any set M_i, $(i \neq j)$. The first condition is proved.

Let us now consider a permutation π in which $i_j = 0$. Suppose that in the c.n.f. $\mathfrak{M}_{\Sigma TF_\pi}$ there is an e.f. $\mathfrak{A}$ such that the ball $N_{\mathfrak{A}}$ has a common point with M_l, $(l \neq j)$ and contains M_j of the point $\tilde{\alpha}$.

Let us assume that this ball contains no points of the set M_j. It is easy to see that in this case $\mathfrak{M}_{\Sigma \mathrm{TF}_\pi}(\tilde{\alpha}) \neq 0$ and the permutation π also does not preserve the code of the set M_j. We have proved the second necessary condition for the preservation of the point $\tilde{\alpha}$ of the set M_j.

Sufficiency. Suppose the conditions of the theorem are true and for each ball $N_{\mathfrak{A}}(\mathrm{M}_l, \tilde{\alpha}), (l \neq j)$, having a point $\tilde{\alpha}$, intersecting the sets M_i and M_j, e.f. $\mathfrak{A}$ - belongs to $\mathfrak{M}_{\Sigma TF_\pi}$. Moreover, in any encoding $\pi = \{i_0, i_1, \ldots, i_{k-1}\} \in \{\pi\}$ e.f. $\mathfrak{A}(\tilde{\gamma}) = i_l$, for $\tilde{\gamma} \in N_\eta(\mathrm{M}_l, \tilde{\alpha})$. We denote by N_B, the ball such that $N_B \cap \mathrm{M}_j \neq \varnothing$, $N_B \cap \mathrm{M}_l = \varnothing$, $(i \neq l)$, $\tilde{\alpha} \in N_B$ for any permutation $\pi = \{i_0, i_1, \ldots, i_{k-1}\}$ of the e.f. $\mathfrak{A}(\tilde{\gamma}) = i_j$ for $\gamma \in N_B$, $B \in \mathfrak{M}_{\Sigma TF_\pi}$.

If in some encoding π the value of the function F_π on the sets of the set M_I less than the values F_π at the M_j, that is $i_j > i_I$, then the point $\tilde{\alpha}$ will be covered by intervals $N_{\mathfrak{A}}(\mathrm{M}_l, \tilde{\alpha})$, but will also be covered by the interval N_B. Since the external operation in c.n.f. is to take a maximum, then $\mathfrak{M}_{\Sigma TF_\pi}(\tilde{\alpha}) = i_j$.

If $i_j > i_l$, then $\mathfrak{M}_{\Sigma TF_\pi}(\tilde{\delta}) = i_l$ for $\tilde{\delta} \in N_{\mathfrak{A}}(\mathrm{M}_l, \tilde{\alpha})$.

But in the interval $N_{\mathfrak{A}}(\mathrm{M}_l, \tilde{\alpha})$, there exists a point $\tilde{\gamma}$ belonging to M_j, such that $\mathfrak{M}_{\Sigma TF_\pi}(\tilde{\gamma}) = i_j$, and by hypothesis $i_j > i_l$. Consequently, the function defined by us at a point $\tilde{\gamma}$ is not equal i_j. Therefore, each such interval $N_{\mathfrak{A}}(\mathrm{M}_l, \tilde{\alpha})$, cannot participate in constructing an extension.

The sufficiency of the theorem is proved.

ALGORITHM FOR THE SYNTHESIS OF A SET INVARIANT WITH RESPECT TO ENCODING AND CONTINUATION

Based on the results of the theorem, we construct an algorithm for synthesizing a set of vertices of the Descartes product

$$S^n = S_1 \times S_2 \times \ldots \times S_n \text{ of sets } \mathrm{M}_j, \left(j = \overline{0,\, k-1}\right)$$

preserving the encoding when obtaining optimal extensions of a partially defined function $F(\tilde{x})$.

Let M_j be the set of all vertices $\tilde{\alpha} \in S^n$, of the coding-preserving set M_j, $\left(j = \overline{0,\, k-1}\right)$ and $\tilde{M} = \bigcup_{i=0}^{k-1} \tilde{M}_i$.

We consider the algorithm for the synthesis of sets M_j. We describe it as follows:

1. For any M_j we find $\{N_j\}$ the set of balls $N_{\mathfrak{A}}$ such that $N_{\mathfrak{A}} \cap \mathrm{M}_j \neq \varnothing$ and $\mathfrak{A} \in \mathfrak{M}$.
2. For $\{N_j\}$, $\left(j = \overline{0,\, k-1}\right)$, construct the set of all balls $\{N_j\}'$ whose intersection with M_i, $\left(i = \overline{0,\, k-1}\right)$ are empty.
3. If $\{N_j\} \neq \varnothing$, then in $\bigcup_{N : N \in \{N_j\}'} N$ we define points $\tilde{\alpha}$, that are not included M_j and are such that there is no ball $N_{\mathfrak{A}}$ that is contained $\tilde{\alpha} : N_{\mathfrak{A}} \cap \mathrm{M}_j \neq \varnothing$, in $\bigcup_{N : N \in \{N_j\}'} N$ and e.f. $\mathfrak{A} \in \mathfrak{M}$. It is easy to see that the multitude of all such $\tilde{\alpha}$, forms M_j, $\left(j = \overline{0,\, k-1}\right)$.
4. Combining of the sets $\tilde{\mathrm{M}}_0, \tilde{\mathrm{M}}_1, \ldots, \tilde{\mathrm{M}}_{k-1}$ we obtain the set $\tilde{\mathrm{M}}$.

DESCRIPTION OF THE PROGRAM FOR THE SYNTHESIS OF AN INVARIANT SET.

For a given set of vertices, divided into classes, the program constructs an invariant continuation for each class.

There are eight procedures in her body. A diagram of the interaction is shown in Figure 1.

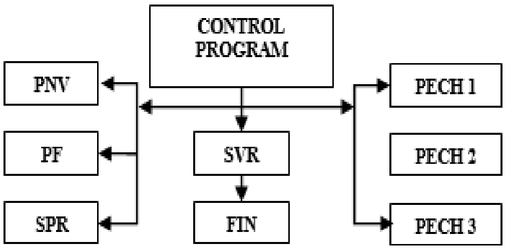


Figure 1

The SPR procedure expands the specified vertex to an interval of dimension (n-1). The PF (P) procedure is used to find the intersection of intervals; P-sign (if for one variable P = 0, then the intervals intersect; if for all variables P = 0, then the intervals intersect). The PNV procedure is used to search for such vertices that intersect with the obtained interval of dimension (n-1). The SVR procedure for a given interval synthesizes the vertices included in this interval. FIN builds index working sets. The procedures PECH 1 (A_1, A_2, PM), PECH 2 (A_1, A_2, PM), PECH 3 (A_1, A_2, PM) are used to print one-, two- and three-dimensional arrays, respectively, where A_1, A_2 are boundary values indices, PM is the array to be printed.

INSTRUCTIONS FOR THE PROGRAM

The following values are the initial values for the program: CK - number of classes; K is the significance of the function; N is the number of variables; CB is the number of vertices in all classes; R_1, R_2, R_3, R_4, R_5 - dimensions of arrays used for intermediate and final results; $R_1 = N^{R_2}$; R_2 = max (N, CB-1); R_3 is the number of extensions in MR; R_4 - number of vertices in extensions from MR; R_5 is the number of vertices in the invariant continuation; AK (CBxN) - a table, each row of which corresponds to one vertex of the cube E_k^n, the *i*-th element of the row is equal to $2^{k-\alpha-1}$, where α - is the value of the *i*-th variable; CVK (1x (CK + 1)) - vector, *i* + 1 - th element of which is equal to the number of the last vertex of the *i* - th class; Q(8x8) is the matrix used to check for intersection of intervals:

$$Q = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 2 & 0 & 0 & 2 & 2 \\ 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 4 & 4 & 4 & 4 \\ 0 & 1 & 0 & 1 & 4 & 5 & 4 & 5 \\ 0 & 0 & 2 & 2 & 4 & 4 & 6 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{vmatrix}$$

The end results are arrays *IV* (1x (CK + 1)) - a vector, the (*i* + 1) th element of which is equal to the number of the last invariant vertex of the i-th class, INP (1xCIV (CK)) is a table, each row of which corresponds to an invariant top.

The block diagram is shown in Figure 2.

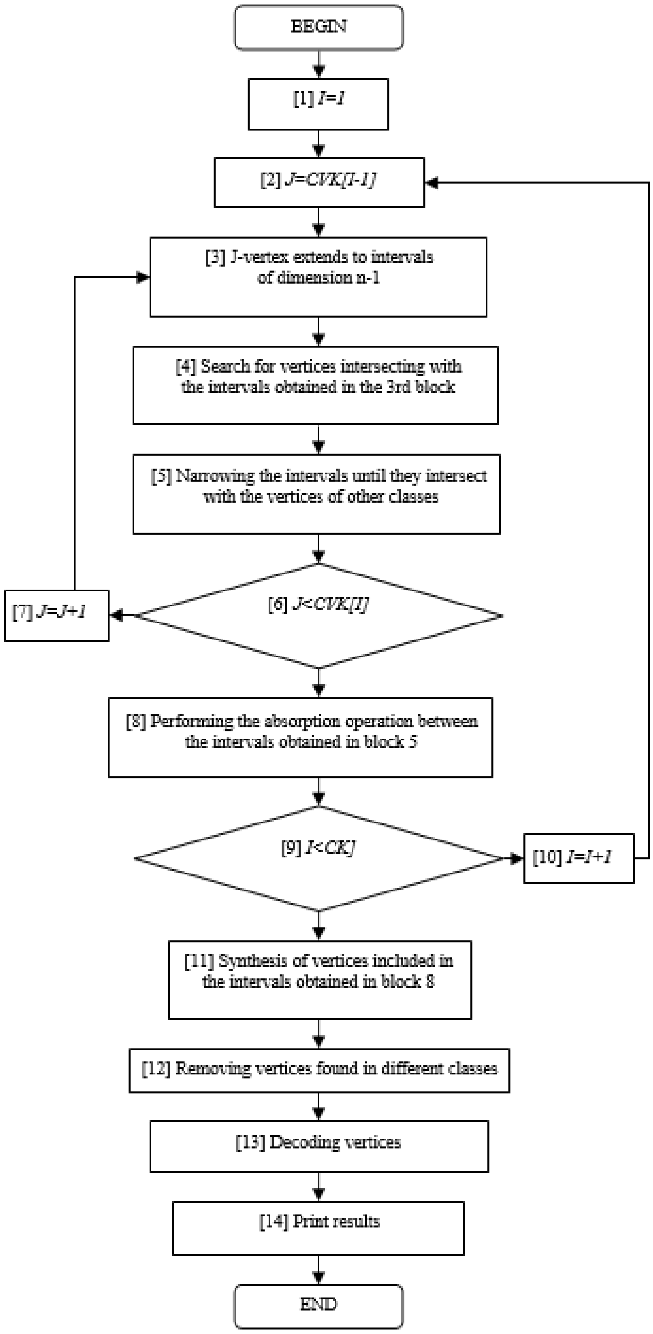

BEGIN
[1] I=1
[2] J=CVK[I-1]
[3] J-vertex extends to intervals of dimension n-1
[4] Search for vertices intersecting with the intervals obtained in the 3rd block
[5] Narrowing the intervals until they intersect with the vertices of other classes
[6] J<CVK[I]
[7] J=J+1
[8] Performing the absorption operation between the intervals obtained in block 5
[9] I<CK]
[10] I=I+1
[11] Synthesis of vertices included in the intervals obtained in block 8
[12] Removing vertices found in different classes
[13] Decoding vertices
[14] Print results
END

Figure 2

Test case. As an example, let's take five vertices of the cube E_3^3, divided into three classes. The initial data are the following: K = 3, N = 3, CK = 3, CB = 5, R1 = 27, R2 = 4, R3 = 100, R5 = 20,

$$A = \begin{vmatrix} 1 & 1 & 2 \\ 2 & 2 & 4 \\ 1 & 4 & 2 \\ 4 & 2 & 2 \\ 4 & 1 & 1 \end{vmatrix},$$

matrix Q and

$$CVK = \begin{vmatrix} 1 \\ 1 \\ 3 \\ 5 \end{vmatrix}.$$

The end result will be

$$CIV = \begin{vmatrix} 0 & 1 & 5 & 8 \end{vmatrix},$$

$$INP = \begin{vmatrix} 2 & 2 & 1 \\ 1 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \\ 0 & 1 & 2 \end{vmatrix}.$$

The counting time of the test case is 15 seconds.

CONCLUSION.

The article investigates partially defined multivalued functions and the problem of logical continuation of such functions in the class of canonical normal forms. A theorem is proved on the invariant continuation of multivalued functions in the class of disjunctive norms, which does not depend on the adopted coding. An algorithm for the formation of a set of invariant points, independent of the adopted encoding, and software for such an invariant continuation are constructed.

REFERENCES

[1] Zhuravlev Yu.I. On a class of not everywhere defined functions of the algebra of logic, In: Discrete Analysis, Issue 2.-Novosibirsk: IM SB AS USSR, 1964, p.23-27.

[2] Nurlibaev A.N. On normal forms k - valued logic// Collected Works on Mathematical Cybernetics. Issue. I. M., Publishing House of the Computing Center of the Academy of Sciences of the USSR Moscow, 1976.

[3] Yablonsky S.V. Functional constructions of the k - valued logic// Proceedings of Mat. Inst. Steklov 51, S. 5-142, 1958.

[4] I.H.Normatov: Principle of independence of continuation of functions multivalued logic from coding, Journal of Physics, (2018), 1210.

[5] Kabulov A.V. and Normatov I.H. About problems of decoding and searching for the maximum upper zero of discrete monotone functions// – Journal of Physics Conference Series 1260(10):102006·August 2019 DOI: 10.1088/1742-6596/1260/10/102006 pp. – 1–7.

[6] A.V.Kabulov, E.Urunbaev, I.H Normatov and A.O.Ashurov: Logical method for constructing the optimal corrector of fuzzy heuristic algorithms, 2019 International Conference on Information Science and Communications Technologies (ICISCT), (2019), 1 – 4.

[7] A.V.Kabulov, I.H.Normatov and A.O.Ashurov: Computational methods of minimization of multiple functions, Journal of Physics, (2019), 1260.

[8] A.V.Kabulov, E.Urunbaev, I.H Normatov and A.O.Ashurov: Synthesis methods of optimal discrete corrective functions, Advances in Mathematics: Scientific Journal 9, 9(2020), 6467 – 6482.

[9] A.V.Kabulov, I.H.Normatov SH. Boltaev and I.Saymanov: Logic method of classification of objects with non-joining classes, Advances in Mathematics: Scientific Journal 9, 10(2020), 1857 – 8365.

[10] I.H.Normatov and E.Kamolov: Development of an algorithm for optimizing the technological process of kaolin enrichment, IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, (2020), 1 – 4.

[11] A.V.Kabulov, I.H Normatov, A.Seytov and A.Kudaybergenov: Optimal Management of Water Resources in Large Main Canals with Cascade Pumping Stations, IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, (2020), 1 – 4.

[12] A.V.Kabulov, I.H Normatov and A.Karimov: Algorithmization control of complex systems based on functioning tables, Journal of Physics Conference Series 6(2020), 1 – 9.

[13] Kabulov, A.V., Normatov I.H., Karimov, A.A., Navruzov, E.R. Constructing control models of complex systems in the language of functioning tables, European Journal of Molecular and Clinical Medicine, 2020, 7(2), стр. 758–771

[14] Kabulov A.V., Normatov I.H., Urunbaev E., Ashurov A.: About the problem of minimal tests searching, Advances in Mathematics: Scientific Journal, 2020, 9(12), стр. 10419–10430

[15] Kabulov A.V., Normatov I.H., Karimov A.A., Navruzov E.R.: Algorithm of constructing control models of complex systems in the language of functioning tables, Advances in Mathematics: Scientific Journal, 2020, 9(12), стр. 10397–10417

[16] Katerinochkina N. N. Search for the maximum upper zero of a monotone function of the logic algebra. DAN SSSR, No. 3, 1975, p. 224.

[17] N. A. Solovyov Tests. - Novosibirsk. The science. 1978. P. 187.

[18] V. Gogolev Some estimates of disjunctive normal forms of logic algebra functions. In sat. Problems of Cybernetics. Moscow, Russia. The science.1967. issue 19, pp. 75-94.

The Effects of Electrode Physical Parameters on the Statistical Life Models of Li-Ion Battery

Talal Mouais
Department of Electrical and Computer Engineering
Effat University
Jeddah, Saudi Arabia
tamouis@effat.edu.sa

Omar A. Kittaneh
Department of Science, Math, and Technology
Effat University
Jeddah, Saudi Arabia
okitanneh@effatuniversity.edu.sa

Mohammed Abdulmajid
Department of Electrical and Computer Engineering
Effat University
Jeddah, Saudi Arabia
moabdulmajid@effatuniversity.edu.sa

Abstract- This paper analyzes the impact of two different electrode types from different manufacturers on the statistical lifetime models of Li-ion batteries. The analysis includes some standard goodness of fit tests and analysis of variance. It was found that batteries with a thicker electrode and different physical parameters will have different battery lifetime models.

Index Terms— accelerated life test; Li-ion battery, the goodness of fit tests; reliability; Analysis of variance (ANOVA) test; Weibull model; Lognormal model; Normal model

I. INTRODUCTION

Electrodes are the most significant components in the Li-ion battery. It affects the capacity, power density, and energy density of the battery. Mei found in [1] that the electrode's physical parameters such as thickness and volume play an important role in the heat generation rate and the energy density of the battery. Furthermore, he found that the particle size of the electrode impacts the energy density and power density.

Zhao found in [2] that thicker electrodes can increase the proportion of active materials and increase energy. However, increasing the thickness could negatively affect the thermal and electrochemical performances of Li-ion batteries. Besides, he found that batteries with a thicker electrode have more intensive and uneven temperature responses across the electrochemical cell when discharging under the same rate. This can lead to the depletion of active materials and faster capacity fading of the entire battery. Moreover, he concluded that a Li-ion battery with a thicker electrode is more sensitive to increased discharge rate. This is due to the ohmic heat generated, which led to the deterioration of the Li-ion battery's health

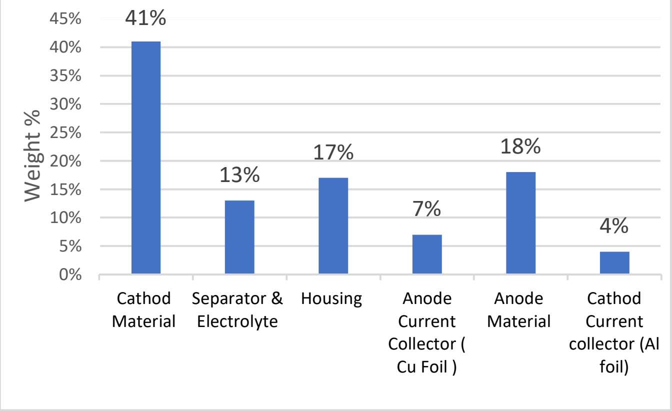


Figure 1. Show the weight percentage occupied by the component of a commercial Li-ion battery. Adapted from [3]

Figure 1 shows the breakdown of the weight percentage by the components of a commercial Li-ion battery. It is very clear from the figure that the electrodes comprise the most significant material components of the Li-ion battery.

Due to the significant impact of the electrodes on the Li-ion battery's performance, and since it comprises the largest and the heaviest material components of Li-ion battery, investigating the effects of the physical parameters of the electrodes on the lifetime models of Li-ion battery is crucial.

There are mainly three methods to model Li-ion batteries lifetime: Physics-based, statistical, and data-driven methods. This paper will focus on the statistical method due to its simplicity and convenience, which would help in further analysis. The statistical approach used in this work is based on trying to fit the experimental data of lifetimes of two types of batteries, B1 and B2, obtained from [4] by some possible parametric statistical models which might not be normally distributed as traditionally assumed in [4]. The three models examined are weibull, lognormal and normal distributions. In fact, this work favors the lognormal as the best fitting model for the B1 type but rejects all for B2. This would be achieved by implementing some standard statistical goodness of fit tests and supported by the analysis of variance technique that was investigated by [4], which will be performed in this work in more detail.

The paper is organized as follows. Section II introduces the experimental data of the two types of batteries obtained from [4]. Section III describes the statistical tools used in analyzing the experimental data. Section IV provides the results of the statistical techniques described in Section III. Lastly, the paper is concluded in section V.

II. DETAILS OF THE EXPERIMENTAL DATA

The experiment in [4] was conducted using the following two cells with the below parameters

TABLE 1: THE PARAMETERS OF THE EXPERIMENT IN [4] CONDUCTED

	B1	B2
Nominal Capacity(Ah)	1.1 Ah	1.35 Ah
Length (mm)	50	50
Width	33.8	33.8
Thickness (mm)	5.4	6.7
Weight (g)	22	28

Arbin BT-2000 was used for testing the cycle life of batteries. The number of cycles to failure for batteries B1 & B2 are in Table2.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

TABLE 2. THE RESULTS OF THE EXPERIMENT CONDUCTED IN [4]

Battery Sample	Discharge condition	Number of cycles to failure
B1_1	0.5C	730
B1_2	0.5C	471
B1_3	0.5C	537
B1_4	0.5C	515
B1_5	1.0C	608
B1_6	1.0C	480
B1_7	1.0C	601
B1_8	1.0C	661
B2_1	0.5C	1336
B2_2	0.5C	764
B2_3	0.5C	803
B2_4	0.5C	1321
B2_5	1.0C	669
B2_6	1.0C	741
B2_7	1.0C	747
B2_8	1.0C	724

III. METHODOLOGY AND METHODS

For the convenience of the reader, this section, as shown in Table 3, lists three of the standard goodness of fit tests that are used to check the appropriateness of the three suggested probability distributions to the experimental data. The table also shows the analysis of variance test, which is used to test the hypothesis that the homogeneity of the mean lifetime between the two types of the batteries and when changing the stress from 0.5C to 1C.

Table 3. Methods used for analyzing the data

Test	Description
Kolmogorov-Smirnov (KS)	KS test [5] is powerful nonparametric goodness of fit test that finds the maximum vertical distance between the empirical and completely specified theoretical distributions. The test is not accurate when the parameters are estimated from data, as in our case. Thus, its modified version [6], referred to as the Lilliefors test, is used instead.
Anderson Darling (AD)	AD goodness of fit test [7] for any completely specified distribution, i.e., with known or estimated parameters considering the adjustment of the test statistic or its critical values. The test applies to the three distributions.
Jarque - Bera (JB)	JB test [8] is very efficient but usually applied after taking the approval of other two tests like KS and AD. The test is only applicable to normal and lognormal, but not Weibull.
Analysis of variance (ANOVA)	ANOVA [9] is used to decide whether there is a significant difference between the means of two independent random variables, given they are sufficiently normally distributed and have similar variances.

IV. RESULTS AND DISCUSSION

The goodness of fit tests:

The goodness of fit tests was conducted using the number of cycles to failure for each battery at a different discharge rate. The tests are applied for the three probability distributions, and the results are reported in Tables 4-9.

TABLES 4: GOODNESS OF FIT TESTS FOR THE DATA SET AT B1 (0.5 C)

	Test	TS	CV	PV	Result
Weibull	KS	0.358	0.385	0.09	Accept
	AD	0.5	0.696	0.19	Accept
	JB	NA			
Lognormal	KS	0.32	0.375	0.15	Accept
	AD	0.374	0.56	0.25	Accept
	JB	0.62	0.214	0.21	Accept
Normal	KS	0.341	0.375	0.11	Accept
	AD	0.431	0.557	0.15	Accept
	JB	0.713	0.852	0.12	Accept

TABLES 5: GOODNESS OF FIT TESTS FOR THE DATA SET AT B1 (1 C)

	Test	TS	CV	PV	Result
Weibull	KS	0.28	0.385	0.47	Accept
	AD	0.348	0.696	0.49	Accept
	JB	NA			
Lognormal	KS	0.335	0.375	0.12	Accept
	AD	0.384	0.557	0.23	Accept
	JB	0.56	0.852	0.31	Accept
Normal	KS	0.32	0.375	0.15	Accept
	AD	0.347	0.557	0.31	Accept
	JB	0.474	0.852	0.5	Accept

TABLES 6: GOODNESS OF FIT TESTS FOR THE DATA SET AT B1 (1C &0.5 C)

	Test	TS	CV	PV	Result
Weibull	KS	0.178	0.29	0.5	Accept
	AD	0.291	0.72	0.65	Accept
	JB	NA			
Lognormal	KS	0.146	0.29	0.5	Accept
	AD	0.212	0.67	0.82	Accept
	JB	0.541	2.12	0.5	Accept
Normal	KS	0.1632	0.29	0.5	Accept
	AD	0.235	0.67	0.74	Accept
	JB	0.571	2.12	0.5	Accept

It can be seen from Tables (4-6) that all models are accepted to fit the data B1 (0.5 C), B1 (1 C), and B1 (1 C and 0.5 C). B1 (0.5 C) prefers lognormal, B1 (1 C) prefers its companion

Weibull [10-12], whereas B1 (1 C and 0.5 C) prefers again lognormal. This suggests that the change in the stress from 0.5 C to 1 C does not change the lifetime model of the batteries of type B1.

TABLES 7: GOODNESS OF FIT TESTS FOR THE DATA SET AT B2 (0.5C)

	Test	TS	CV	PV	Result
Weibull	KS	0.332	0.385	0.15	Accept
	AD	0.645	0.696	0.07	Accept
	JB	NA			
Lognormal	KS	0.301	0.375	0.21	Accept
	AD	0.494	0.557	0.09	Accept
	JB	0.65	0.85	0.17	Accept
Normal	KS	0.3	0.375	0.21	Accept
	AD	0.5	0.56	0.08	Accept
	JB	0.66	0.85	0.16	Accept

TABLES 8: GOODNESS OF FIT TESTS FOR THE DATA SET AT B2 (1C)

	Test	TS	CV	PV	Result
Weibull	KS	0.261	0.385	0.5	Accept
	AD	0.414	0.696	0.33	Accept
	JB	NA			
Lognormal	KS	0.298	0.375	0.22	Accept
	AD	0.419	0.557	0.17	Accept
	JB	0.704	0.852	0.13	Accept
Normal	KS	0.292	0.375	0.25	Accept
	AD	0.406	0.557	0.19	Accept
	JB	0.686	0.852	0.14	Accept

TABLES 9: GOODNESS OF FIT TESTS FOR THE DATA SET AT B2 (1C & 0.5C)

	Test	TS	CV	PV	Result
Weibull	KS	0.368	0.288	0.001	Reject
	AD	1.207	0.717	0.002	Reject
	JB	NA			
Lognormal	KS	0.343	0.288	0.007	Reject
	AD	1.059	0.667	0.003	Reject
	JB	1.608	2.116	0.080	Accept
Normal	KS	0.372	0.288	0.001	Reject
	AD	1.221	0.667	0.001	Reject
	JB	1.752	2.116	0.070	Accept

On the other hand, It can be seen from Tables (7-9) that all models are accepted to fit the data B2 (0.5 C), B2 (1 C) with preference to lognormal and Weibull, but all are rejected or weakly accepted to fit B2 (1 C and 0.5 C). This suggests that the change in the stress from 0.5 C to 1 C does change the lifetime model of the batteries of type B2.

Figures 2a, 2b, and 2c show the probability plot for Normal, Weibull, and lognormal distribution, respectively, for battery B1.

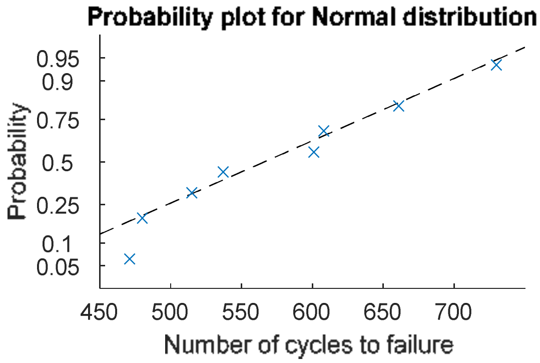


Figure 2a. Probability plot for Normal distribution for B1 at 0.5 C and 1 C.

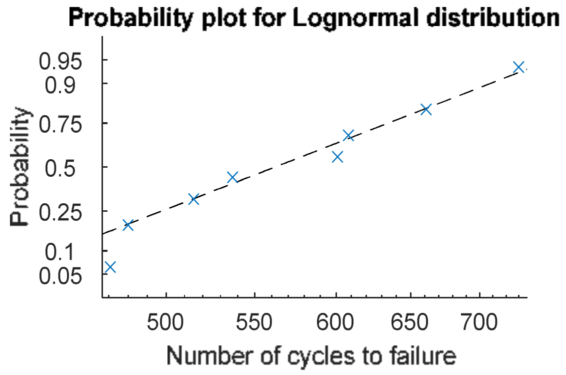


Figure 2b. Probability plot for Lognormal distribution for B1 at 0.5 C and 1 C.

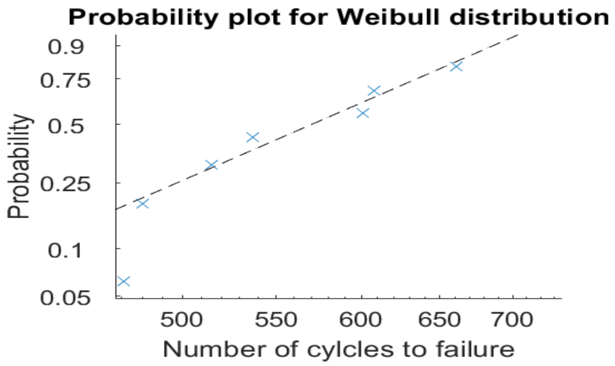


Figure 2c. Probability plot for Weibull distribution for B1 at 0.5 C and 1 C.

It is clear from Figures 2a, 2b, and 2c that almost all the data points are close or on the trend line of the three distributions. This indicates that the data fit the three models in agreement with the goodness of fit tests' results in Tables 4-6.

Figures 3a, 3b, and 3c show the probability plot for normal, Weibull, and lognormal distributions, respectively. For battery B2, where we can see that none of the three distributions is an acceptable fit to the data, which means that B2 (0.5 C) and B2 (1 C) have entirely different distributions as agreed with the goodness of fit test results shown in Tables 7-9.

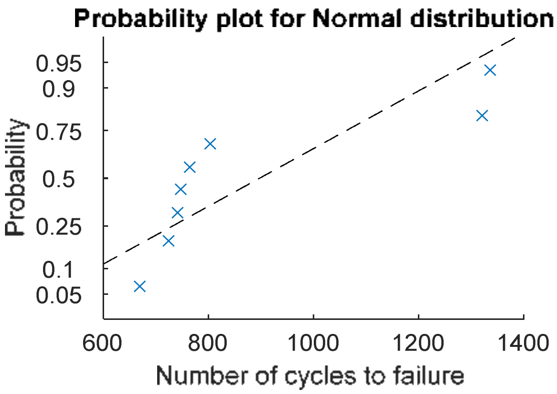


Figure 3a. Probability plot for Normal distribution for B2 at 0.5 C and 1 C.

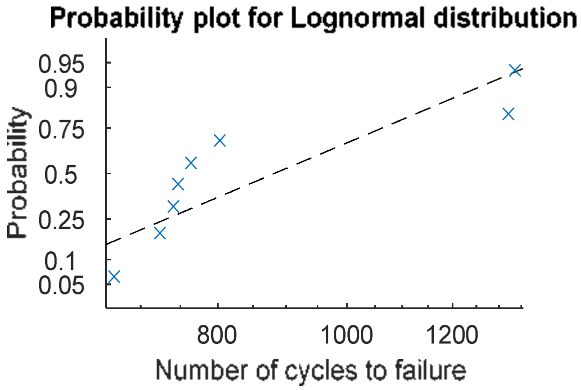


Figure 3b. Probability plot for Lognormal distribution for B2 at 0.5 C and 1 C.

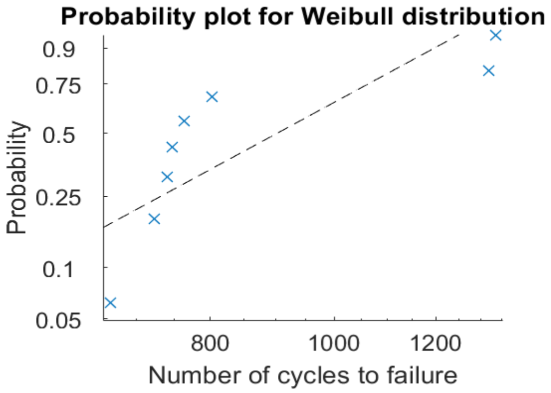


Figure 3c. Probability plot for Weibull distribution for B2 at 0.5 C and 1 C.

Analysis of variance (ANOVA) for the data:

One-way ANOVA is used to test whether two data sets have similar means at a certain appropriate level of significance; here, we choose it to be 0.05. This technique requires that the two data sets be independent, normally distributed, and acceptably similar standard deviations. B1 (0.5 C) and B1 (1 C) data sets satisfy the above assumptions because they are normally distributed as clear from Tables 4 and 5, independent, and have standard deviations of 115 and 77 cycles, which can be confirmed using Levene's test of equality of variances [13]. Indeed, it can be checked that the p-value of Levene's test for this case is equal to 0.5234. Table 10 below shows ANOVA results applied on B1 (0.5 C) and B1 (1 C).

TABLES 10: ANALYSIS OF VARIANCE (ANOVA) FOR B1 (0.5 C) AND B1 (1 C).

	SS	df	MS	F	P-value	F dist
Between Groups	1176.13	1	1176.1	0.1240	0.737	5.9873
Within Groups	56893.8	6	9482.3			
Total	58069.9	7				

Since the p-value in Table 10 is 0.7367>>0.05, where 0.05 is the level of significance, we don't reject the null hypothesis that means values are the same at different discharge rates. This indicates that the mean lifetime of B1 (0.5C) and B1 (1C) are the same. Therefore, for B1 batteries changing the stress from 0.5 to 1 does not change its mean lifetime, which entirely agrees with the goodness of fit tests in Tables 4-6.

On the other hand, for battery type B2, ANOVA is not applicable because the standard deviations are significantly different, which can be again checked by Levene's test. The p-value, in this case, is equal to 0.0054. It is easy to check that the standard deviations of B2 (0.5 C) and B2 (1 C) are equal to 315 and 36 cycles, respectively. In this case, we can look at the mean values of B2 (0.5 C) and B2 (1 C), which are 1056 and 720cycles, respectively, which are very different, whereas the mean values of B1 (0.5 C) and B1 (1 C) are very close of 563 and 588 cycles, respectively.

However, the sample size used in this study is relatively small, where obtaining complete data for the lifetime of such batteries is very difficult due to time and cost limitations. Therefore, it is necessary to increase the sample size and consider efficient censoring [14-16] and truncating [17-18] schemes to overcome this problem.

V. CONCLUSION AND FINDINGS

After conducting the statistical analysis for B1, it is found that B1 (0.5 C) better fits lognormal, and B1 (1 C) better fits Weibull, whereas B1 (0.5 C and 1 C) prefers the lognormal distribution. In fact, Levene's test of equality of variances confirms that the three data sets have similar standard deviations, and ANOVA proves that their means are also very similar. All these standard statistical evidence proves that the B1 battery for 0.5 C and 1 C has the same lifetime.

On the other hand, B2 is sensitive to a high rate of discharge due to the contribution of the ohmic heat generated. By looking at the statistical analysis of B2, it can be seen that B2 (0.5 C) better fits lognormal, and B2 (1 C) better fits Weibull, but B2 (0.5 C and 1 C) does not fit any model. Also, Levene's test shows that these two data sets have very different standard deviations as the standard deviation of B2 (0.5 C) is almost ten times that of B2 (1 C). This violates the main condition of ANOVA, yet, simple calculations for the sample means would help at this stage, which indicates that the mean lifetime significantly decreases when the stress increases from 0.5 C to 1 C.

REFERENCES:

[1] Mei, W., Chen, H., Sun, J., & Wang, Q. (2019). The effect of electrode design parameters on battery performance and optimization of electrode thickness based on the electrochemical–thermal coupling model. Sustainable Energy & Fuels, 3(1), 148–165.

[2] Zhao, R., Liu, J., & Gu, J. (2015). The effects of electrode thickness on the electrochemical and thermal characteristics of lithium ion battery. Applied Energy, 139, 220–229.

[3] Pender, J., Jha, G., Youn, D., Ziegler, J., Andoni, I., Choi, E., Heller, A., Dunn, B., Weiss, P., Penner, R., & Mullins, C. (2020). Electrode Degradation in Lithium-Ion Batteries. ACS Nano, 14(2), 1243.

[4] Williard, N., Wei He, Osterman, M., & Pecht, M. (2012). Reliability and failure analysis of Lithium Ion batteries for electronic systems. 2012 13th International Conference on Electronic Packaging Technology & High Density Packaging, 1051–1055.

[5] Massey, F.J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. 46(253), 68–78.

[6] Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov Tests for Normality with Mean and Variance Unknown. Journal of the American Statistical Association, 62, 399-402.

[7] Anderson, T., Darling, D. (1954). A Test of Goodness of Fit. Journal of the American Statistical Association, 49(268), 765-769.

[8] Jarque, C., Bera, A. (1987). A Test for Normality of Observations and Regression Residuals. International Statistical Review / Revue Internationale de Statistique, 55(2), 163.

[9] Montgomery, D.C. (2017). Design and analysis of experiments. John Wiley & Sons.

[10] Kittaneh, O.A., Shehata, M., Majid, M.A. (2018). An efficient censoring scheme for lifetime of connected solid-state lighting based on entropy measures. (L&T), 15th IEEE Conference.

[11] Kittaneh, O.A., Beltagy, M. (2017). Efficiency estimation of type-I censored sample from Weibull distribution based on sup-entropy Commun. Stat. Comput. Simul., 46, 2678-2688.

[12] Kittaneh, O.A., Majid, M. A. (2019). Comparison of two-lifetime models of solid-state lighting based on sup-entropy. Heliyon, 5(10), e02551.

[13] Derrick, B., Ruck, A., Toher, D., White, P. (2018). Tests for equality of variances between two samples which contain both paired observations and independent observations. Journal of Applied Quantitative Methods. 13(2), 36–47.

[14] Kittaneh, O.A., Heba, A.M., Sara, H., Majid, M.A. (2021). On Censoring the Normal Distribution, to appear in IMA Journal of Mathematical Control and Information, recently accepted.

[15] Kittaneh, O. (2012). Deriving the efficiency function for type I censored sample from exponential distribution using sup-entropy. Journal of Statistics, 19(1), 43–53.

[16] Kittaneh, O.A., Akbar, M. (2014). Deriving the efficiency function for type I censored sample from Pareto distribution using sup-entropy. IMA Journal of Mathematical Control and Information, 33 (2), 231–237.

[17] Omar Abd Al-Rahman, I.K. (2015). A measure of discrimination between two double truncated distributions. Communications in Statistics-Theory and Methods, 44(9), 1797–1805.

[18] Ebrahimi, N., Soofi, E. (1990). Relative information loss under Type II censored exponential data. Biometrika, 77(2), 429-435.

[19] Di Crescenzo, A., Longobardi, M. (2004). A measure of discrimination between past lifetime distributions. Statistics & Probability Letters, 67(2), 173-182.

Smart Street Light Management System with Automatic Brightness Adjustment Using Bolt IoT Platform

Sk Mahammad Sorif
Dept. of Electrical Engineering
Aliah University, Kolkata-700160
West Bengal, India
sorif9933@gmail.com

Dipanjan Saha
Dept. of Electrical Engineering
Aliah University, Kolkata-700160
West Bengal, India
dipanjansaha019@gmail.com

Pallav Dutta
Dept. of Electrical Engineering
Aliah University, Kolkata-700160
West Bengal, India
pallav.dutta@aol.com

***Abstract*— This paper presents a streetlamp control system based on the Bolt IoT platform. The aim of this project is conservation of energy by reducing electricity wastage and to minimize the manpower. The scheme utilizes Light Emitting Diodes (LED) that doesn't take huge amount of power and being directional light sources, it can radiate light in specific direction thereby improving the efficiency of the streetlamps. Using LDR with LED lights the intensity can be controlled. IR sensors utilized on one roadside send signals for the LEDs to get glowing for the next specific section of the road after sensing the density and movement of vehicles. The proposed work has achieved a better performance compared to the existing systems.**

***Keywords*— Bolt IoT platform, Internet of Things (IOT), Light intensity, Power saving, Streetlight**

I. Introduction

IOT is the interface of physical devices which permits the devices to contact with each other and make the devices sensed and controlled remotely. These advanced automation and analytics system use artificial intelligence technology to give automated and advanced products and services. IOT based systems permit better transparency, control, and great performance. [1].

IOT has different automation applications like smart parking, smart home, smart roads, smart lighting and so on. Streetlamps are the crucial requirement in present time of transportation for safety purposes and keeping away accidents during night. Despite that in the present occupied life nobody tries to turn it off/on when not needed. This project gives solution to this by reducing manpower and conserving the energy. The current manual streetlamp system has a few problems like timing problem, maintenance issues and connectivity issues. These problems can be eliminated by IOT technology [2]. This system depends on smart and weather adaptive automated street lighting and management [3]. Automation simplifies different issues on the planet economy just as in daily life [4].

The LED (Light Emitting Diode) is viewed as a next generation light source since not exclusively is it energy efficient however it has the long life needed for the illumination of outdoors, workplaces and homes. Recently, the research focus has moved to the plan of an intelligent LED lighting system on account of the fact that the LED is effortlessly combined with different electronics, for example, micro controllers, wired and/or wireless transmission devices and a variety of sensors to execute intelligent lighting system [5]. It utilizes latest innovation in LED light source to replace traditional streetlights such as HID lamps or High-Pressure Sodium lights. The LED lights are embraced in view of its different advantages over existing innovations like power saving because of increased current luminous efficiency, high colour rendering index, reduced maintenance cost, accelerated start-up, and durability [6]. These days flexibility of streetlamp system is being highly challenged. Most part of the control runs in a manual arrangement though some are automated depend on their surrounding parameters. Taking care of remote area location is the main problem. Manual mistakes can develop energy wastage and the performance of the system become low [7]. The point of this paper is to automate the streetlamps to increase the productivity and precision of the system in a practical way and also allows wireless accessibility and control over the system [8]. The main purpose of the system is the energy preservation for the fact that the assets like hydro, coal, thermal that we depend upon are not renewable energy, so presenting power saving components like LDR and LEDs can illuminate a huge region with high-intensity light whenever required [9].

The whole system is based on the Bolt IoT platform. It gives Wireless Fidelity (Wi-Fi) capabilities and cloud connectivity with the sensors in the system. One of the main purposes of connecting all devices to the web and cloud is to have wireless access of those devices. This system helps the user to control the devices from distant areas over the web, cloud or local area network (LAN). But in the LAN type connection range of the system get reduced. If the connection with the devices is over web or cloud then one can control those devices from anywhere on the planet with active internet from both system and user side. Utilizing Bolt IoT we can reduce code size about 80% and it can be run 10 times faster [10].

During daytime there is no need of street lamps so the LDR keeps the streetlamp off until the light level is low or the light frequency is low and the LDR resistance is high. This keeps current from going to the base of the transistors. Due to this reason the lights do not glow. When sufficiently high frequency light falls on the equipment, the semiconductor soak up photons and give bound electrons adequate energy to fence into the conduction band. The following free electron (and its hole partner) conduct electricity, thus dropping resistance. It supports client server operation where a single user can control the overall system [7]. It decreases heat and carbon dioxide emissions [1]. IOT based streetlamps automation is a practical and eco-friendly

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

technique which additionally take out the issues in disposal of incandescent lights and power saving [11].

II. LITERATURE SURVEY

In the present years, many efforts have been taken to automate the existing streetlamp system. For any smart streetlight system, it should work in an efficient method to maximize the quality and productivity. So, by implementing a more dependable system can remove a significant road lighting cost and reduce human effort as well. However, many methods are still operating with conventional light sources, it may reduce the human effort but the light pollution and energy wastage still exist.

F. Dheena et al. [12] in the year 2017 proposed a paper on smart streetlight management system using Arduino (ESP8266EX Wi-Fi Module), LDR, Relay, DHT11 sensor. In this system LDR detects when adequate light is not available, then with the help of relay the LED lights are turned ON or OFF. The advantage of this system is that it is simple and it can reduce manual work. It can also save energy. But the initial installation cost and maintenance are high.

The paper proposed by Jeetendra Swami et al. [13] in the year 2019 is quite same like the previous one, but here they used sensors also for vehicle detection. Still the installation cost and maintenance is high for this system.

Nabil Ouerhani et al. [14] in the year 2016 published their paper with respect to streetlamp controlling utilizing Zigbee remote module. They included microcontroller, LDR, and a transmission module. Zigbee permits wireless communication with the light module. The system consists of two LDR sensors to analyze the day-night variations and light health conditions. The outcomes from the LDR are moved to the microcontroller after processing the data and further into the transmission module. The wireless Zigbee sends the information to the control centre to monitor and operate each streetlamp. The system utilizes Zigbee network and the range of Zigbee is exceptionally short. Similar type of work also happened by M. Caroline Viola Stella Mary et al. [15] in the year 2018.

Siddaarthan Chintra Suseendran et al. [16] in the year 2018 implement a streetlamp control system using Raspberry Pi 3 model. They used python code in such a manner that intensity of light is controlled with the help of LDR. It also used Zigbee remote module. This system reduces power consumption but the design is complex.

Archibong, Ekaette Ifiok et al. [17] implemented a traffic flow-based streetlamp control system powered by solar panels. They utilized Arduino Wi-Fi module (ESP2866MOD), IOT module (Particle Electron 3G), Flying Fish infrared sensor (IR) module with adjustable distance, ULN2003 driver IC, LED array,18650 cylindrical lithium-ion batteries rated 3.7V, 2200mAh, LDR, SARODA SP09-05 model solar PV module rated 18V, 20W, wireless router and a computer and energy consumption was reduced by three times. Sensors are given at either side of the streets to recognize the vehicle movements and to instruct the microcontroller to turn on and off the lights accordingly. The installation cost is relatively high for this system due to solar panel and all other components

The paper proposed by Omkar Rudrawar et al. [18] planned a system to control street light using power electronics. In this system the light strength is controlled by TRIAC by controlling the applied voltage. As this voltage is proportionate to the light intensity. The system is automatically turned ON/OFF according to the information of sunrise or sunset from reliable internet sources. This system can be used with the existing manual system so initial cost is low but the maintenance is a problem due to not knowing which part is defective.

The paper by M. Sahithi Prasanthi et al. [19] planned a system for controlling the traffic light by image processing. The system will identify vehicles through pictures instead of utilizing electronic sensors installed in the sidewalk. A camera will be placed near by traffic signal. It will capture picture sequentially. The image sequence will at that point be analysed utilizing digital image processing for vehicle identification, and according to traffic conditions on the road, light can be controlled.

III. HARDWARE ARCHITECTURE & PROPOSED SYSTEM

The system consists of LDR, IR sensor, Bolt Wi-Fi module, Bolt cloud, LED and few basic electronic components e.g. Breadboard, Resistor, Connecting wires etc. A single system is capable of controlling four lights. The required connection for the proposed system is shown in Figure 1. The BOLT WIFI MODULE board is powered by using USB cable. Here the two sensors IR sensor and Light dependent resistor (LDR) sensor are interfaced to the board. The respective ground and VCC pin from the IR

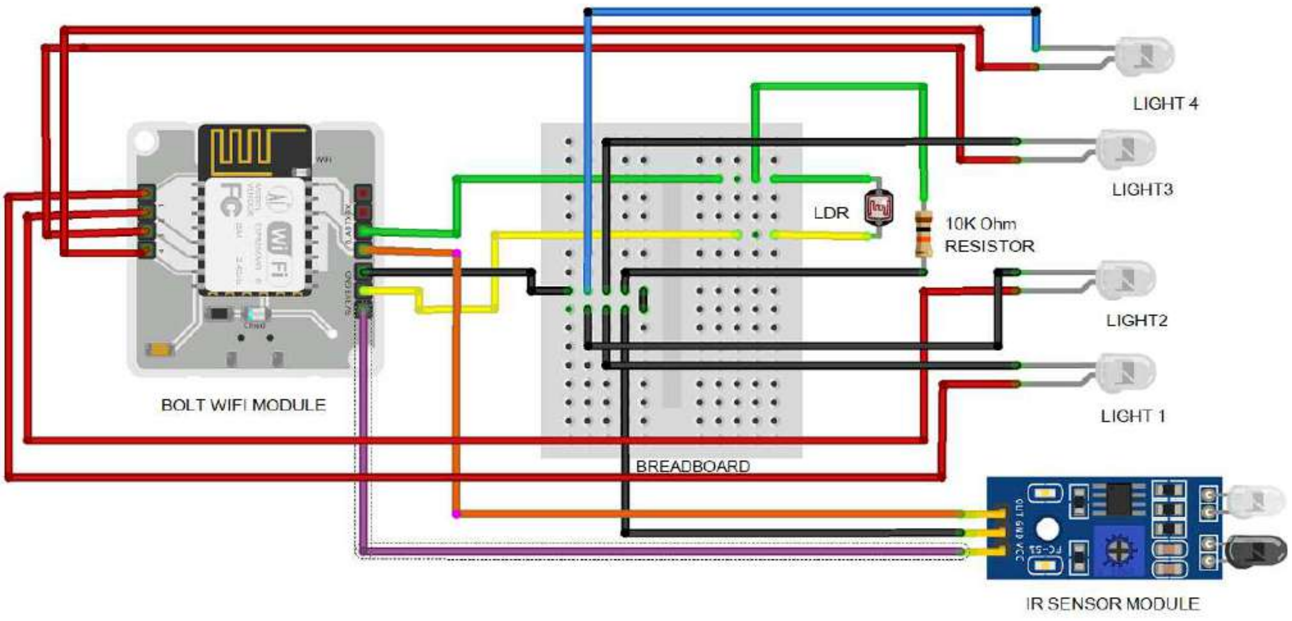


Figure 1: Circuit and connection diagram

sensor connected to GND and 5V pins of BOLT module respectively and output pin is connected to GPIO PIN 0. One of the terminals of LDR sensor is given to analog read pin A0 of BOLT module and another terminal is in 3v3 supply of bolt. A 10K resistor is connected to A0 pin with LDR and another leg of resistor is in ground. And all the positive pin of LED is connected in 1,2,3,4 GPIO pin of bolt module respectively and negative pin are connected to ground of bolt module.

For connecting bolt Wi-Fi module to bolt cloud through Message Queue Telemetry Transport (MQTT) protocol Bolt module required a local Wi-Fi Hotspot and we can access bolt cloud through mobile/tablet/laptop through HTTP & HTTPS protocols. Through cloud dashboard we can fetch data from all simultaneous systems which can control four lights individually. In Oracle Virtual Box workstation on UBUNTU server we write the python code for automatic street light controller. UBUNTU server connects bolt module through bolt cloud API key. By python coding on UBUNTU, we can control our proposed system.

The circuit and connection diagram of this system is shown in Figure.1.

A. Bolt Wi-Fi Module

The whole system is based on the Bolt IoT platform. Bolt IoT platform is an integrated IoT platform that gives Wireless Fidelity (Wi-Fi) capabilities and cloud connectivity to the sensors and actuators used in the system. This platform is based on Espressif-8266 Wi-Fi (ESP-8266 Wi-Fi) module. Machine Learning (ML) algorithms can be easily integrated with Bolt IoT projects for detecting or predicting anomalies in the sensor values. Some of the reasons to choose Bolt IoT are its ability to reduce code by about 80%- and 10-times faster time of deployment. Table I portrait the specifications of the Bolt IoT platform used [20] and Figure 2 depicts the view of Bolt IoT module.

Table I: Specifications of Bolt IoT platform

Parameters	Details
Processing/ Connectivity Module	ESP8266 (Customizable)
MCU	32-bit RISC CPU: Tensilica Xtensa LX106
Power	5V/1A DC - Micro-USB port/ 5V & GND pins
Voltage (Operating)	3.3V
Clock Frequency (CPU)	80 MHz
Internal Memory (MCU)	96KB of data RAM, 64 KB of instruction RAM
External Memory (MCU)	4 MB Flash memory [QSPI]
GPIO pins	5 Digital pins
ADC	1 pin 10-bit ADC
PWM	All 5 Digital pins capable of PWM
Dimension	35mm x 35mm
Boot Time	< 1 second

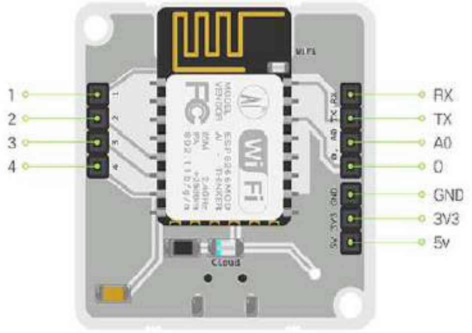


Figure 2: View of Bolt Wi-Fi Module

There are three types of connectivity in Bolt IoT. Details are given in Table II.

Table II: Types of connectivity in Bolt IoT

Wi-Fi	802.11 b/g/n Automatic AP mode when not connected to Wi-Fi WPA/WEP/WPA2 authentication 2.4 GHz Wi-Fi
UART	8-N-1 3.3V TTL UART [using GND, RX, TX pins] [2400/ 480/ 9600/ 19200 bitrates]
Cloud	Optional: Custom cloud using Bolt APIs

B. *LDR*

An LDR is a light sensitive electronics device whose resistivity is a component of the incident electromagnetic radiation. They are also known as photo conductive cells, photo conductors or simply photocells. LDRs are created with the help of semiconductor materials which have high resistance. So, when the photons fall on the gadget, the electrons in the valence band of the semiconductor material are aroused to the conduction band. LDR is used in this circuit as a darkness detector.

C. *IR Sensor*

An infrared sensor is an electronic instrument that is utilized to detect certain characteristics of its surroundings by either transmitting or detecting infrared radiation. It is additionally can measure heat of an object and detect motion. Infrared waves are not noticeable to the natural eye. In the electromagnetic range, infrared radiation is the region having frequencies longer than visible light frequencies, yet more limited than microwaves. The infrared regions roughly differentiated from 0.75 to 1000µm.

In this proposed system IR sensor is installed at the starting point of the system and then sequentially one by one four street lights are installed e.g., Light 1,2,3 & 4 as shown in Figure 3.

The LDR (Light Dependent Resistor) is used to sense the amount of light or darkness in the environment in order to switch ON/OFF the lights. In day time when sufficient light is present, street lights are OFF. The system waits for the ambient light to fall which is detected by the LDR; system gets activated and keeps all the street light (LED) at

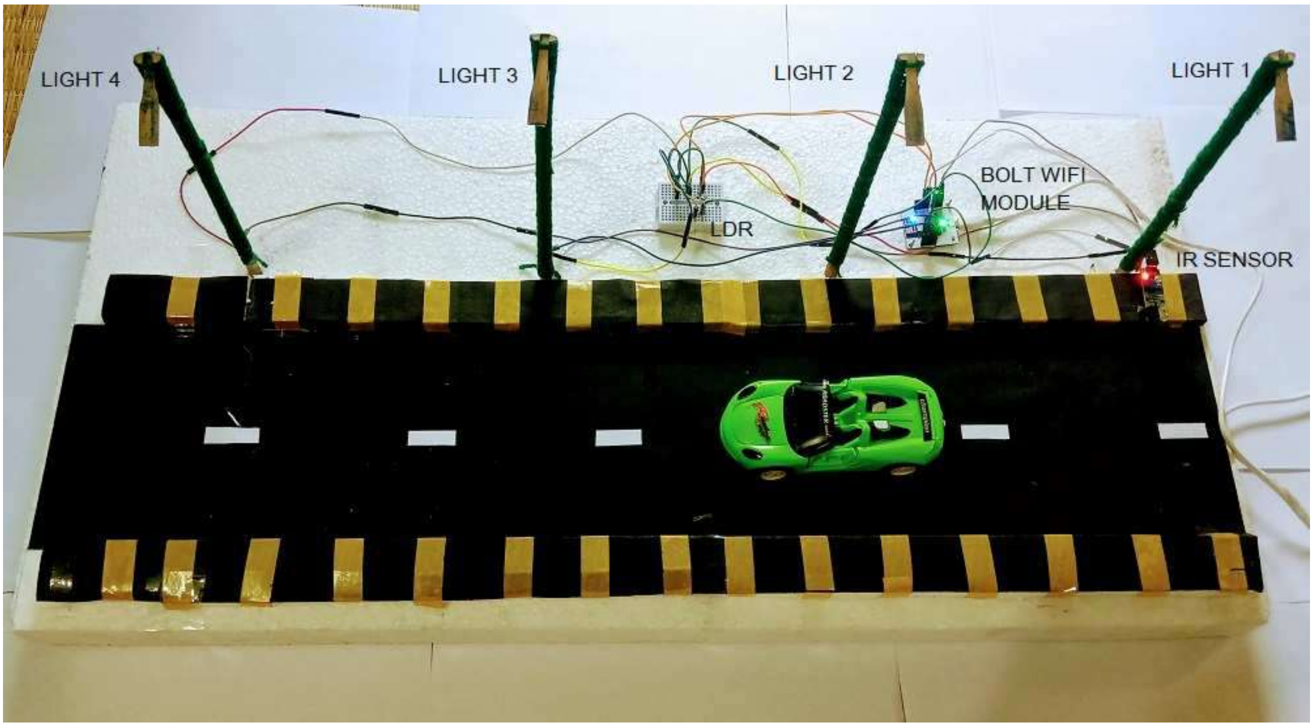


Figure 3: Prototype of the Proposed System

40% brightness and an alert SMS is sent to the authority through twilio.com [21].

Whenever the system detects an object by IR sensor, all the LEDs are powered to 100% brightness. When objects move out of the system reach all the street light are reset to 40%.

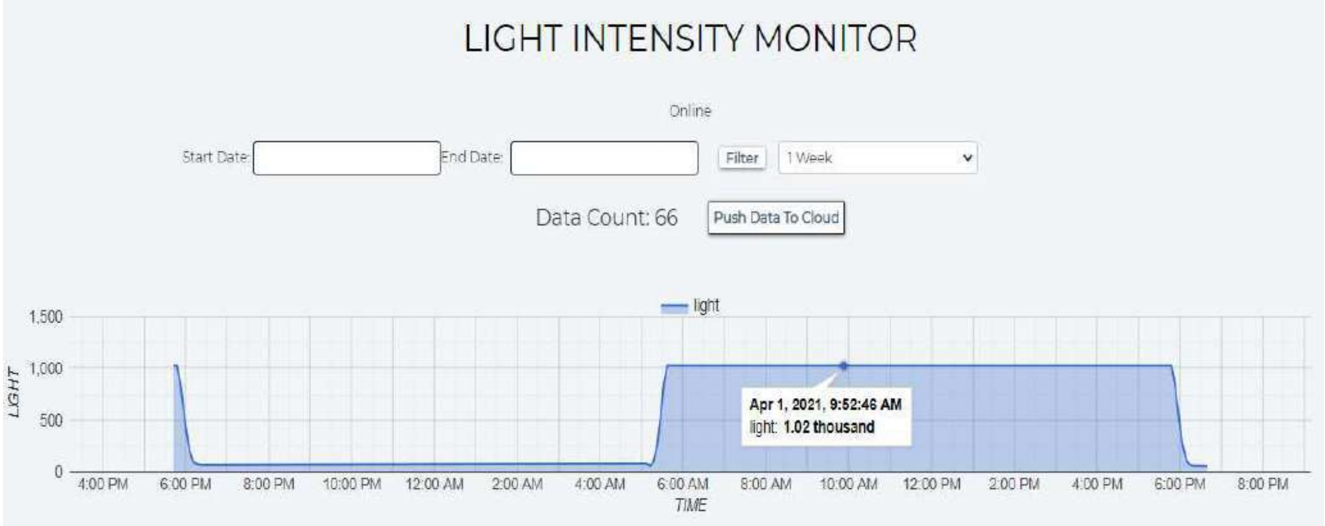
Afterwards when sufficient light is available the LDR sense and turned off LED automatically, and send an SMS to the authority that street light is turned OFF. Each of the street light can also be individually turned ON/OFF manually via login in bolt cloud dash board or using BOLT IOT mobile application.

IV. RESULT & DISCUSSION

This IOT based automated streetlamp system is extremely cost effective. The project aim is the conservation of energy. It can likewise terminate the CO2 outflows and light pollution. The method does not require manpower and frequent check rather the system status can be thoroughly monitored through updated data stored in bolt cloud. In cloud we can monitor and analyze intensity value of light over the day. The Figure 4 shows the values.

The system checks for the evening to fall which is recognized by the LDR; system gets triggered and keeps all the streetlamp (LED) at 40% of its peak brightness. And send an alert to the authority via SMS that street light is on as shown in Figure 5.

Whenever the system detects an object (human, vehicle) through IR sensor the associated LEDs are powered to its 100% brightness. After objects moves out of the system reach all the street light are reset to 40% of its peak brightness. After that when intensity value of the ambient light become high which implies day the street light are turn off automatically and send an alert to the authority via SMS that street light is OFF as shown in Figure 6.



LIGHT INTENSITY MONITOR

Online

Start Date: End Date: 1 Week

Data Count: 66

Figure 4: Light intensity data from LDR

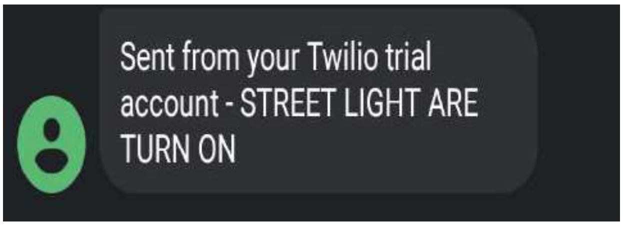

Sent from your Twilio trial account - STREET LIGHT ARE TURN ON

Figure 5: Street light turn ON SMS alert

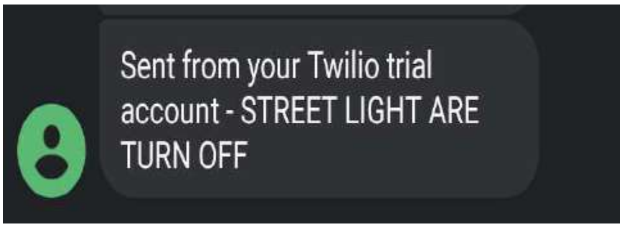

Sent from your Twilio trial account - STREET LIGHT ARE TURN OFF

Figure 6: Street light turn OFF SMS alert

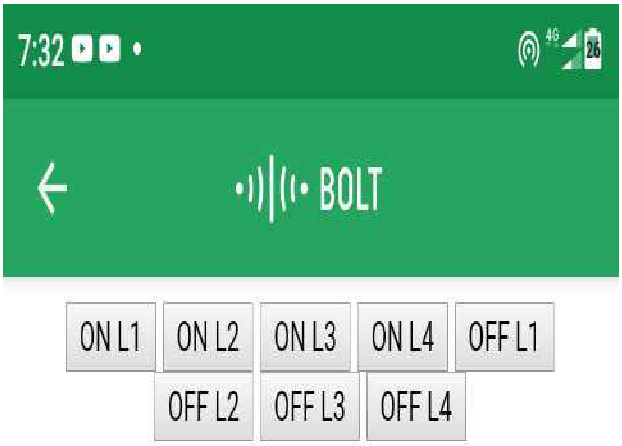


Figure 7: Manual street light control switch in Bolt IoT mobile application

The control of individual street lights is also possible through login in Bolt cloud dash board. Figure 7 shows the individual street light control switch in Bolt IoT mobile application.

V. CONCLUSION

A great portion of energy can be saved by replacing sodium vapor lamps by LED. By adding smart switching to LED streetlamps, it is turned ON/OFF automatically. Owing to the intelligence offered by this proposed system the brightness of the street lights can also be controlled through sensing the natural light or traffic flow. This prevents unnecessary wastage of electrical energy further.

The LED dimming feature is joined with the different sensors that give the distance information from any objects and the surrounding light intensity around the LED light. The intensity of the LED is automatically controlled. The LED dimming feature can operate with a wired or wireless system to realize universal LED lighting systems. It gives an effective and intelligent automatic streetlamp control system with the assistance of LDR. There is a message system to alert and communicate the status of the lights. In this proposed system the street lights send message when it will turn OFF/ON through twillo.com. By this if any day street lights are not turned off/on, system do not give an alert so authority can take measure upon it. Here one IR sensor is used for controlling four street lamps so it reduced the cost of the system. For any emergency condition we can manually operate each of the street lights through web interface or mobile application, which make the system more reliable. It can decrease the energy utilization and maintenance cost. It tends to be applied in urban as well as rural areas. The system is expandable and absolutely adaptable to the requirements of the user. It establishes a safe environment with maximum intensity light at whatever point required. This system can report street lamp failure, which make the maintenance of street lamp simpler and less formidable.

The existing bulbs or lights are also may be linked to this low-cost street light management system. The lights will make good use of the Internet of Things (IoT) connectivity to not only prevent the power wastage but also for better management and fault detection.

The need of the system is to reduce energy consumption, decrease the maintenance cost and to expand the lifespan of the system. By allowing this method the energy can be used more efficiently in smart street lightning system. This smart and relatively low cost IoT system can also reduce the CO2 emissions and help to protect the environment.

REFERENCES

[1] J.Arthi, W.Lydiapreethi, and B. Gunasundari,"IOT Based Smart LED Street Lighting System" IJRTI | Volume 2, Issue 4 | ISSN: 2456-3315.

[2] SayaliArkade, Akshada Mohite, Rutuj, Vikas, "IoT Based Street Lights For Smart City" International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue XII, December 2016.

[3] A. K. Tripathy, A. K. Mishra and T. K. Das, "Smart lighting: Intelligent and weather adaptive lighting in street lights using IOT," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala, India, 2017, pp. 1236-1239, doi: 10.1109/ICICICT1.2017.8342746.

[4] M. Saifuzzaman, N. N. Moon and F. N. Nur, "IoT based street lighting and traffic management system," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, Bangladesh, 2017, pp. 121-124, doi: 10.1109/R10-HTC.2017.8288921.

[5] A. Suzdalenko and I. Galkin, "Choice of power and control hardware for smart LED luminary," in Proceeding of the 12th IEEE Biennial Baltic Electronics Conference, Tallinn, Estonia, pp. 331-334, 2010.

[6] Ritisha Salvi, Shraddha Margaj, Kavita Mate1, Bhakti Aher, Smart Street Light Using Arduino Uno Microcontroller" International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 5, Issue 3, March 2017.

[7] Monika R. Kodali and S. Yerroju, "Energy efficient smart street light," 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Tumkur, India, 2017, pp. 190-193, doi: 10.1109/ICATCCT.2017.8389131.

[8] B. Abinaya, S. Gurupriya and M. Pooja, "Iot based smart and adaptive lighting in street lights," 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), Chennai, 2017, pp. 195-198, doi: 10.1109/ICCCT2.2017.7972267.

[9] Dr.A.S.C.S.SASTRY, K.A.S.K. Bhargav, K. Surya Pavan, M.Narendra "Smart Street Light System using IoT" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 07 Issue: 03 | Mar 2020.

[10] S. Rehan and R. Singh, "Industrial and Home Automation, Control, Safety and Security System using Bolt IoT Platform," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 787-793, doi: 10.1109/ICOSEC49089.2020.9215345.

[11] Parkash, Prabu V, Dandu Rajendra, "Internet of Things Based Intelligent Street Lighting System for Smart City" International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 5, May 2016.

[12] P. P. F. Dheena, G. S. Raj, G. Dutt and S. V. Jinny, "IOT based smart street light management system," 2017 IEEE International Conference on Circuits and Systems (ICCS), Thiruvananthapuram, India, 2017, pp. 368-371, doi: 10.1109/ICCS1.2017.8326023.

[13] Jeetendra Swami, Himanshu Patel, Krishna Patel and Ms.Nisha Bhalse, " IoT Based Street Light Automation System" International Journal of Trend in Research and Development, Volume 6(2), ISSN: 2394-9333.

[14] N. Ouerhani, N. Pazos, M. Aeberli and M. Muller, "IoT-based dynamic street light control for smart cities use cases," 2016 International Symposium on Networks, Computers and Communications (ISNCC), Yasmine Hammamet, Tunisia, 2016, pp. 1-5, doi: 10.1109/ISNCC.2016.7746112.

[15] M. C. V. S. Mary, G. P. Devaraj, T. A. Theepak, D. J. Pushparaj and J. M. Esther, "Intelligent Energy Efficient Street Light Controlling System based on IoT for Smart City," 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 551-554, doi: 10.1109/ICSSIT.2018.8748324.

[16] S. C. Suseendran, K. B. Nanda, J. Andrew and M. S. Bennet Praba, "Smart Street lighting System," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 630-633, doi: 10.1109/CESYS.2018.8723949.

[17] E. I. Archibong, S. Ozuomba and E. Ekott, "Internet of Things (IoT)-based, Solar Powered Street Light System with Anti-vandalisation Mechanism," 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), Ayobo, Nigeria, 2020, pp. 1-6, doi: 10.1109/ICMCECS47690.2020.240867.

[18] O. Rudrawar, S. Daga, J. R. Chadha and P. S. Kulkami, "Smart street lighting system with light intensity control using power electronics," 2018 Technologies for Smart-City Energy Security and Power (ICSESP), Bhubaneswar, India, 2018, pp. 1-5, doi: 10.1109/ICSESP.2018.8376692.

[19] M. S. Prasanthi et al., "IOT Based Streetlight Management," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 1796-1801, doi: 10.1109/RTEICT42901.2018.9012380.

[20] B. I. T . (I. P. Limited), "IoT Platform," BOLT. [Online]. Available: https://www.boltiot.com/techspecs. [Accessed: 27-feb-2021].

[21] "Communication APIs for SMS, Voice, Video and Authentication," Twilio. [Online]. Available: https://www.twilio.com/. [Accessed: 27-feb-2021]

MLP for Spatio-Temporal Traffic Volume Forecasting

1st Asimina Dimara
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece, 57001
Email: adimara@iti.gr

2nd Dimitrios Triantafyllidis
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece, 57001
Email: triantafd@iti.gr

3rd Stelios Krinidis
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece, 57001
Email: krinidis@iti.gr

4th Konstantinos Kitsikoudis
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece, 57001
Email: kzkitsik@iti.gr

5th Dimosthenis Ioannidis
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece, 57001
Email: djoannid@iti.gr

6th Efthalia Valkouma
Egnatia Odos S.A
Thessaloniki, Greece
Email: thvalkou@egnatia.gr

7th Stilianos Skarvelakis
Egnatia Odos S.A
Thessaloniki, Greece
Email: sskarvel@egnatia.gr

8th Stavros Antipas
Patras Port Authority S.A
Patras, Greece, 26333
Email: santipas@patrasport.gr

9th Dimitrios Tzovaras
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece, 57001
Email: Dimitrios.Tzovaras@iti.gr

***Abstract*—Avoiding traffic congestion phenomena is an important aspect of efficient transportation infrastructure (e.g. toll roads) management. Traffic congestion phenomena can be avoided by forecasting volume traffic data. This paper aims to analyze how specific factors and parameters affect the behavior of traffic volume, like vehicle category, date, and weather data at a given timestamp and place. Moreover, the procedure of data prepossessing is presented to produce a cleaner data set that gives more fundamental information. Subsequently, spatio-temporal toll road prediction is achieved through a multi-layer perceptron. Finally, the proposed low-cost method is evaluated using real-life data from a toll plaza, while the experimental results show the efficiency of the proposed method.**

***Index Terms*—Traffic volume, spatio-temporal traffic, toll road traffic, traffic forecasting, data driven.**

I. Introduction

Traffic congestion not only leads to vehicular queuing and longer driving times but also results in waste fuel, increased carbon dioxide emissions and air pollution [1]. The first step towards solving traffic congestion is an integral approach which will enable traffic regulation through traffic volume management. Traffic management aims at safe travel and bottleneck-free roads and is part of the Intelligent Transportation Systems (ITS) [2]. A main aspect of ITS is to develop models that can be deployed to estimate, simulate and/or forecast traffic volume, traffic congestion and/or traffic density.

Especially, both the users and administrators of a toll road want to gain insight into the traffic volume at a specific toll plaza [3]. The existing literature referring to traffic forecasting is voluminous [4] and is implemented for many different aspects like travel time prediction, traffic volume prediction and waiting time in toll roads [5]. In [6] the authors state that using an artificial neural network (ANN) to predict traffic volume provides a lower error in comparison with other models (e.g. regression models). Furthermore, many of the existing models use cameras to detect traffic volume [7] and they are efficient but high-cost methods.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Moreover, a fusion of ANN method and support vector machines is used to estimate traffic using many parameters like road vehicle counts, socio-economic data, geometry of the road and other parameters [8]. These type of models use a combination of road data and economic data to predict traffic [9] and are found to be accurate enough but harvesting all this type of data may prove a thorny task. Nonetheless, there is less literature about predicting volume traffic in toll plazas. Probably, because each local toll road is affected by multi-and various parameters like the local climate, distance from big cities, socio-economic data, and pavement quality. Therefore, even if a suggested method is accurate enough, it must be rebuilt for a specific toll plaza.

As a result, this paper suggests a low-cost method to forecast the traffic volume while using the least possible data that can be effortlessly retrieved like the local weather and number of vehicles passing daily through the toll road. Moreover, an effort is made to gain awareness of how this information affects volume traffic and employ the greatest possible advantage from it. Initially, the suggested methodology, data wrangling and preprocessing is addressed in Section II. Subsequently, the results are presented in Section III. Finally, conclusions are drawn in the final section.

II. Methodology

To achieve an increased performance some additional features that may affect the traffic volume have been integrated into the traffic forecasting using the proposed model [10]. In this paper, traffic data has been enhanced with temporal features (e.g., month, hours, day, and holiday periods) and weather features (e.g. temperature, cloud coverage and weather types (e.g sunny, foggy, rainy, snowy)). Specifically, temporal features are used to capture how seasonality may affect the traffic volume.

On the other hand, weather variables may not be directly linked to traffic volume forecasting, but they can affect it indirectly. Non expected weather conditions may disrupt holiday travel plans, while extreme weather conditions may lead to unexpected delays or postponement of business truck schedules. All the aforementioned circumstances that lead to schedule cancellations or postponements affect traffic, as a result necessitating the knowledge of weather conditions for more accurate predictions.

A. Data Wrangling

To explain the process with thoroughness, some further information about the data is provided. The vehicles are classified into four categories depending on the vehicle type as follows:

$$category = \begin{cases} CAT1, & \text{Motorcycle, Tricycle Vehicles} \\ CAT2, & \text{Passenger cars} \\ CAT3, & \text{Trucks, buses with} < 4 \text{ axes} \\ CAT4, & \text{Trucks, buses with} \geq 4 \text{ axes} \end{cases} \tag{1}$$

The toll road fees correspond to the four different vehicle categories, therefore a toll system can estimate the total number of vehicles for each category directly from the daily total sum of fees. As mentioned, the data set holds traffic information in a daily time series structure.

As it may be observed in Fig. 1 the number of vehicles that belong to CAT1 and CAT2 fluctuate similarly referring to months and are characterized by a traffic increase during the summer months (June, July, August). On the other hand, the number of vehicles that belong to CAT3 and CAT4 (Fig. 2) have similar patterns, with very small variations without following a certain motif during the year. This could indicate that CAT1 and CAT2 could be grouped together, as well as CAT3 and CAT4.

As a result, the problem could be split into 2 forecasting problems, forecasting of vehicles that belong to light vehicles (CAT1 and CAT2) and prediction of vehicles that belong to heavy vehicles (CAT3 and CAT4).

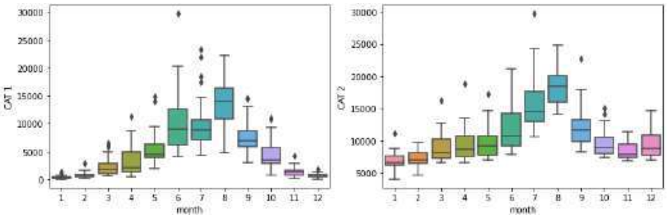


Fig. 1. CAT1, CAT2 per month

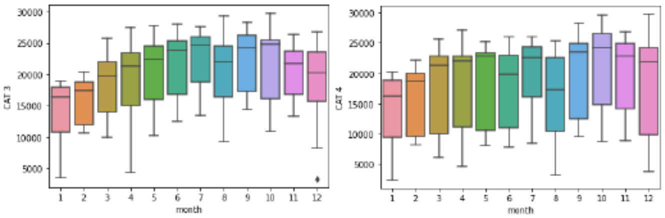


Fig. 2. CAT3, CAT4 per month

To further corroborate the previous assumption Fig. 3 and Fig. 4 reveals the same patterns in traffic volume for toll plazas for CAT1 and CAT2, and for CAT3 and CAT4. Specifically, on weekends more light vehicles seem to pass through the tolls than on weekdays, in contrast to heavy vehicles where more vehicles pass through the tolls on weekdays. Similar behaviors between the categories may also be noticed in Fig. 5 and Fig. 6.

It should be mentioned that for holiday periods the two grouped categories follow the same assumption as depicted in Fig. 7 and in Fig.8. Consequently, the problem may be split into light vehicles volume forecasting and heavy vehicles volume forecasting.

B. Data preprocessing

1) Impute missing values: The number of vehicles never have missing values as they derive from the fees. On the

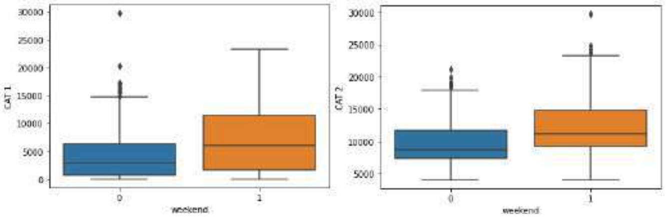


Fig. 3. CAT1, CAT2 per weekend

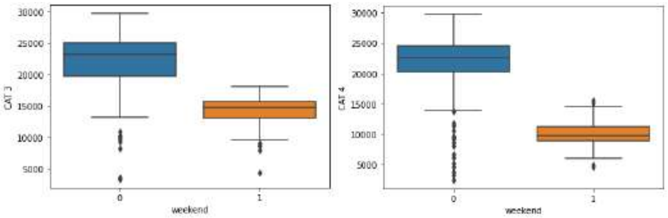


Fig. 4. CAT3, CAT4 per weekend

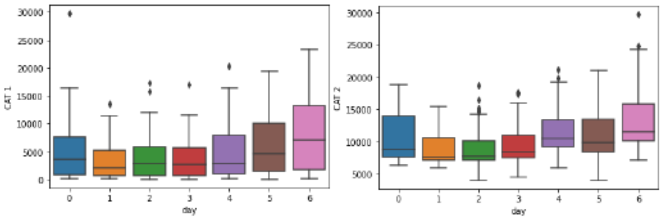


Fig. 5. CAT1, CAT2 per weekday

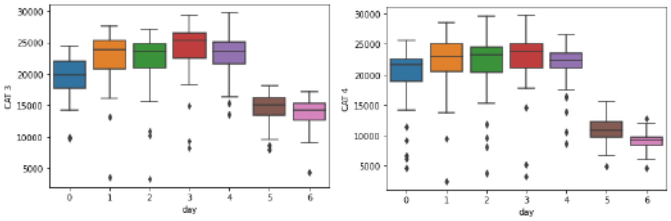


Fig. 6. CAT3, CAT4 per weekday

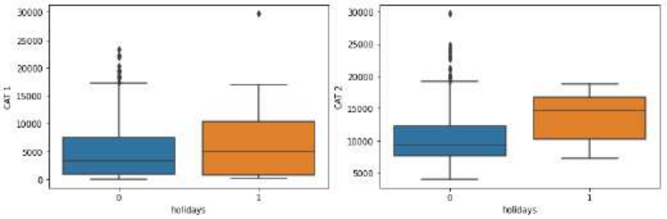


Fig. 7. CAT1, CAT2 per holidays

other hand missing data may appear in the weather variables. Every type of variable requires a different imputing method to be applied. For continuous values (e.g temperature), missing values are filled using the linear interpolation method [11], while missing data for categorical values (e.g weather state) is replaced with the most frequent values within the data set [12].

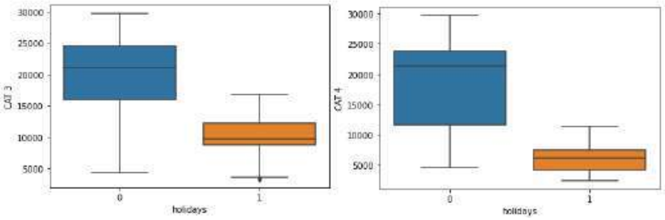


Fig. 8. CAT3, CAT4 per holidays

2) Feature Encoding: Label encoding [13] is applied to the weather state feature to convert the possible weather states into a numeric form as described below:

$$weather\ state = \begin{cases} 0, & \text{Sunny} \\ 1, & \text{Foggy} \\ 2, & \text{Rainy} \\ 3, & \text{Snowy} \end{cases} \tag{2}$$

3) Outliers: Any outliers found were replaced by the average traffic of the current month for each category. In Figures 1 to 8 some outliers are observed to appear throughout the dataset. There are various methods to detect an outlier, however, within this paper, an empirical method for outlier dection was deployed, Z-score method [14], which describes the position of a record's distance (d) from the average (μ) measured in standard deviation unit (σ) and is defined as follows [14]:

$$Z_{score}(d) = \frac{d-\mu}{\sigma} \tag{3}$$

Any data that lie beyond 3σ from the average are considered as outliers and are replaced by the average traffic of the corresponding month.

4) Data modeling: To obtain insights about future traffic volume, past traffic data, temporal and weather variables were used. In addition, to achieve a better model performance, the method proposed in this paper also uses temporal and weather data from the nested period that are retrieved from Numerical Weather Prediction (NWP) models [15]. Specifically, for every record, data from n past days from current day (t) is selected and next day temporal and weather features in order to forecast the day ahead traffic volume ($t+1$). The data modeling as described above is depicted in Fig. 9.

C. Traffic volume forecasting

Within this paper, a MLP model has been developed that implements the forecasting process, which is a conventional feed-forward Neural Network that contains more than one layer [16]. Some worth noting characteristics are that the first layer has as many nodes as the input features, while the number of the last layers' nodes depends on the machine learning task. For regression tasks, the last layer usually contains only one node. In the hidden or intermediate layers, the number of nodes may vary depending on the complexity

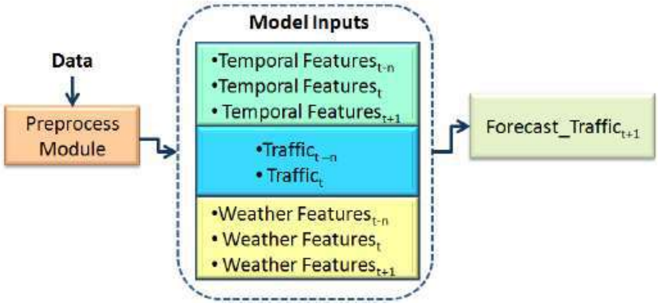


Fig. 9. Data transformation

of the problem or the source domain (e.g, images require a more complex architecture).

During training process, given a fully processed data set with N records, which can be defined as $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, each $\mathbf{y}_i = \{y_1, \ldots, y_N\}$ is the ground truth number of vehicles and $\mathbf{x}_i = \{x_{i1}, \ldots, x_{im}\}$ represent the m input features as calculated in (4). As Fig. 9 implies traffic data from current and past n days are added as input. Temporal Features encapsulate 4 different features (month, day, weekend and holidays), while weather features contain 3 different features (temperature, cloud coverage, and weather state). For these contextual features, not only current and past n days are fed as input but also the day's ahead values that are retrieved from NPW models. To sum up, the number of input features a day's ahead prediction requires considering data from past n days can be calculated as follows:

$$\begin{aligned} Features = m = {} & (n+1) * 1[traffic\ data] \\ & + (n+2) * 4[temporal\ features] \\ & + (n+2) * 3[weather\ features] \\ & = 7 * (n+2) + n + 1 \\ & = 8 * n + 15 \end{aligned} \tag{4}$$

As the number of hidden layers and nodes per layer is increased the MLP becomes more complex and is more susceptible to overfitting. To measure the model's performance, a loss function is defined. The loss function the MLP Regressor uses to evaluate the model in each iteration is the mean squared error (MSE) [17]. The divergence between the predicted number of vehicles $\hat{y}_i$ and the actual number of vehicles y_i, calculated for N records is expressed with the following equation [17]:

$$Loss(y_i, \hat{y}_i) = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{5}$$

Moreover, during the training procedure the weights are updated using an optimizer in order that mitigate the loss expressed by the loss function through an iterative process. In this work, Adam optimizer is used [18].

The MLP model is trained utilizing the aforementioned data. After the training phase the MLP forecasts the daily volume traffic. The retrieved data is pre-processed before fed in the MLP. Progressively, the MLP model is retrained every week with the new traffic and weather data in order to include the complete traffic volume and weather changes while improving itself through time or even capturing steep changes of the weather. A flow chart of the overall traffic forecasting model that was described is depicted in Fig. 10.

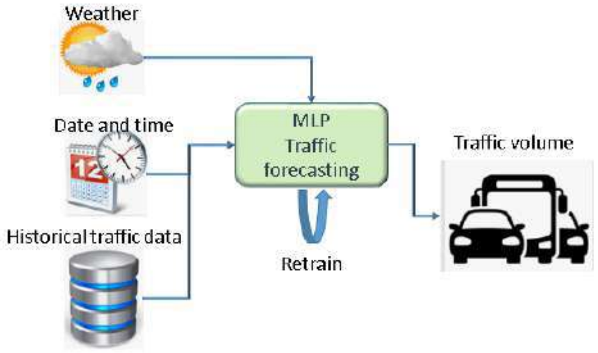


Fig. 10. Flow chart of the overall traffic forecasting model

III. Experimental Results

A. Use case

This section elaborates results obtained from the conducted experiments that took place at the toll plaza in Analipsi (Fig. 11), that is located between Kavala and Thessaloniki, Greece. The traffic data that were collected from the toll's implementation system were retrieved and used as input to the neural network that was developed within this work in order to forecast the day's ahead traffic volume. Furthermore, weather variables were retrieved based on the toll's exact location.

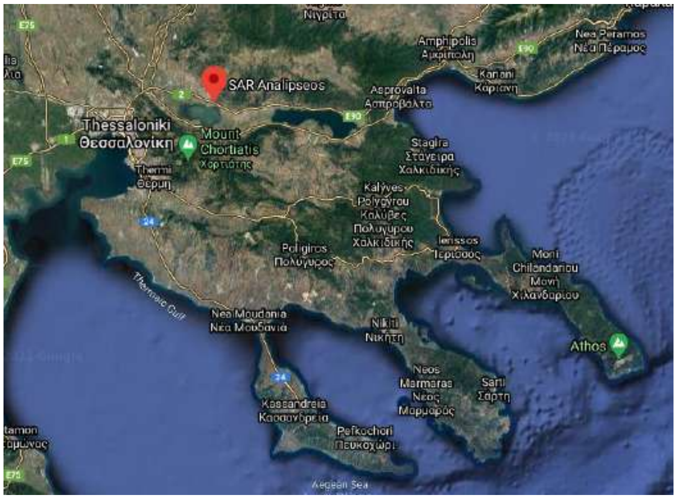
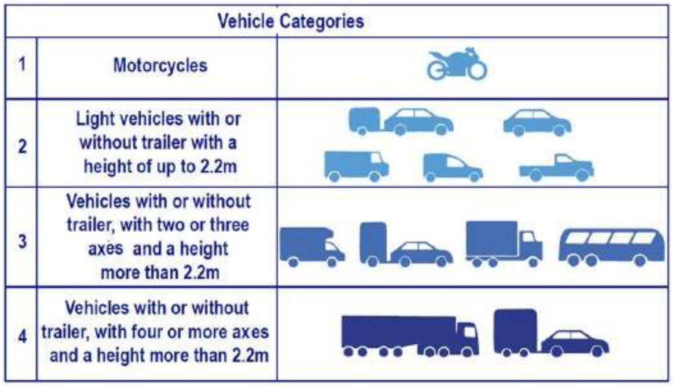


Fig. 11. Analipsi toll plaza map

The toll rates in Greece for various vehicles are as depicted in Fig. 12. It may be observed that the categories are the same as mentioned above for the CAT1, CAT2, CAT3, and CAT4, respectively.



Vehicle Categories		
1	Motorcycles	
2	Light vehicles with or without trailer with a height of up to 2.2m	
3	Vehicles with or without trailer, with two or three axes and a height more than 2.2m	
4	Vehicles with or without trailer, with four or more axes and a height more than 2.2m	

Fig. 12. Toll rate vehicle categories

B. MLP architecture

The lack of existence of a theory-based principle in order to define the optimal topology for a neural network often leads to trial and error methods for the definition of the network architecture [19]. However some broadly accurate practise-based guides exist that suggest the number of layers and neurons for a neural network [19]. As stated by Hecht Nielsen (1987) [20], the hidden layers' neurons should contain are:

$$nodes = 2m + 1, \tag{6}$$

where m is the number of input features (input nodes).

According to Huang [21] a neural network with two hidden layers should have a specific number of neurons. In the first hidden layer, the number of neurons that are suggested are:

$$nodes = \sqrt{(m+2)N} + 2\sqrt{\frac{N}{m+2}}, \tag{7}$$

while the sufficient number of neurons for the second hidden layer should be obtained from:

$$nodes = \sqrt{(m+2)N} \tag{8}$$

It is also worth mentioning that tuning the number of neurons using the power of two increases the speed of finding a layer size for a decent performance and leads to faster computational solutions [22]. Therefore, in the current work another empirical method was used defining the number of nodes as a power of two:

$$nodes = 2^k, \tag{9}$$

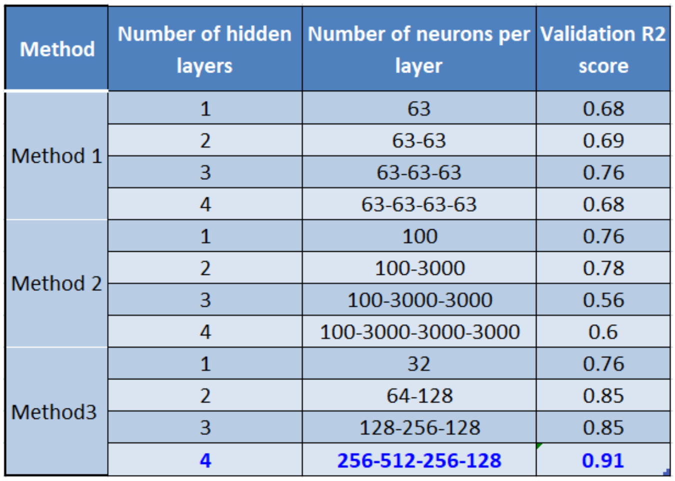
where $k \in \mathbb{N}$.

All above methods were tested for their performance through a trial and error procedure and the network with the highest performance was selected as the most appropriate. The training samples were approximately $N = 280$ while the input features considering the two previous days for the forecasting process are $m = 31$ (4).

$$m = 8 * n + 15 \overset{n=2}{=} 31 \tag{10}$$

According to Nielsen's method ([20] and (4)) each layer should contain $nodes = 63$, while using Huang's method [21] (Method 2) the first layer should contain $nodes \approx 100$, while the second layer should contain $nodes \approx 3000$.

The results obtained by the tuning procedure during the trial and error phase are presented in Fig. 13. Generally, the model seems to perform better having more than one layer, but utilizing Huang's method [21] (Method 2) the performance drastically decreases. As it may be observed in Fig. 13 Method 1 (Nielsen [20]) reveals poor results during all tests, while (9) shows better performance. Taken into account the overall performance of each network topology, method 3 (power of two [22]) with four hidden layers was selected as the most appropriate.



Method	Number of hidden layers	Number of neurons per layer	Validation R2 score
Method 1	1	63	0.68
	2	63-63	0.69
	3	63-63-63	0.76
	4	63-63-63-63	0.68
Method 2	1	100	0.76
	2	100-3000	0.78
	3	100-3000-3000	0.56
	4	100-3000-3000-3000	0.6
Method3	1	32	0.76
	2	64-128	0.85
	3	128-256-128	0.85
	4	**256-512-256-128**	**0.91**

Fig. 13. Tuning nodes per hidden layer

The overall MLP architecture that was selected is as depicted in Fig. 14.

C. MLP performance

In Fig. 15 and Fig. 16 the performance of both MLP models is depicted for heavy and light vehicles, respectively. The initial data set contains daily data during the period 2019 (January 1 - December 31) considering the total number of vehicles (for each category) that pass through the tolls. The data set was split to 75% for training and 25% for test. Specifically, the data samples that were used for training were selected by randomly getting three weeks from each month while the remaining week from each was used for testing. The metrics used to evaluate the MLP performance are Mean absolute Error (MAE) [23] and Mean Absolute Percentage Error (MAPE) [24]:

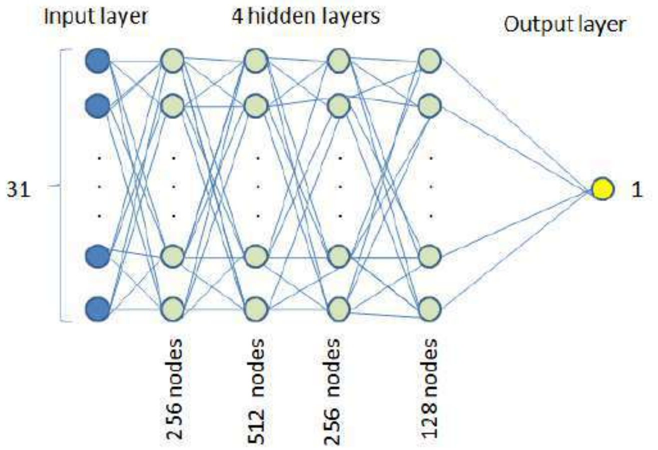


Fig. 14. MLP architecture

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |y_i - \hat{y_i}| \tag{11}$$

$$MAPE = \frac{1}{N} \cdot \sum_{i=1}^{N} |\frac{y_i - \hat{y_i}}{y_i}| \tag{12}$$

An insubstantial deviation may be observed between the number of predicted and actual vehicles (Fig. 15 and Fig. 16). Furthermore, the predicted pattern seems to capture the actual fluctuations even in peak traffic volume moments. Specifically, the mean absolute percentage error in both cases is approximately 8.85%. Considering that 100 vehicles will pass through a specific toll plaza a MAPE score of this value implies that the forecasting model deviates by only 9 vehicles from the actual value.

In terms of performance, considering the nature of the data (total vehicles per day), the lack of intermediate values caused by the given time interval (aggregated daily values), is very reasonable and reveals decent credibility. To summarize, the selected architecture for the MLP proved to have an acceptable performance and is considered as appropriate for the forecasting process.

Conclusions

The proposed method is a tool for traffic volume forecasting at a specific toll plaza. Available data are analyzed revealing hidden information of the associated data, while the retrieved data is preprocessed to better fit the model. Traffic volume patterns during the experimental tests are accurate compared to the expected traffic volume.

The proposed model is a non-intrusive and low-cost traffic volume forecasting tool that informs users and administrators about the traffic volume of an exact location while indicating potential traffic congestion. Moreover, this traffic forecasting tool may be utilized in real-time traffic monitoring and management if data are retrieved real-time. Finally, it has the potential of enhancing the management of road networks while reducing traffic congestion.

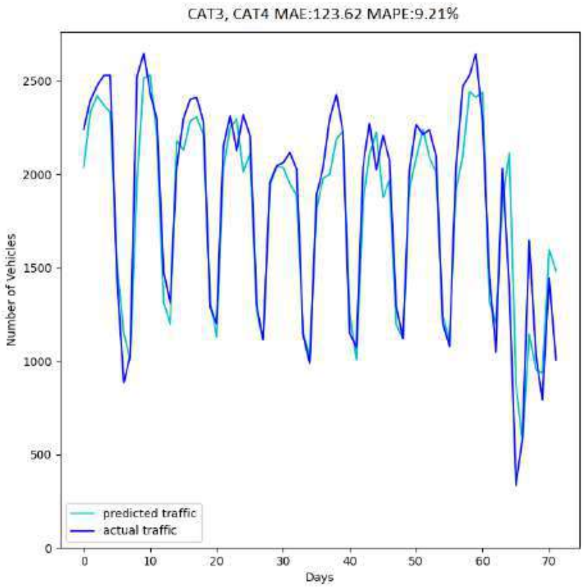


Fig. 15. Heavy Vehicles MLP performance

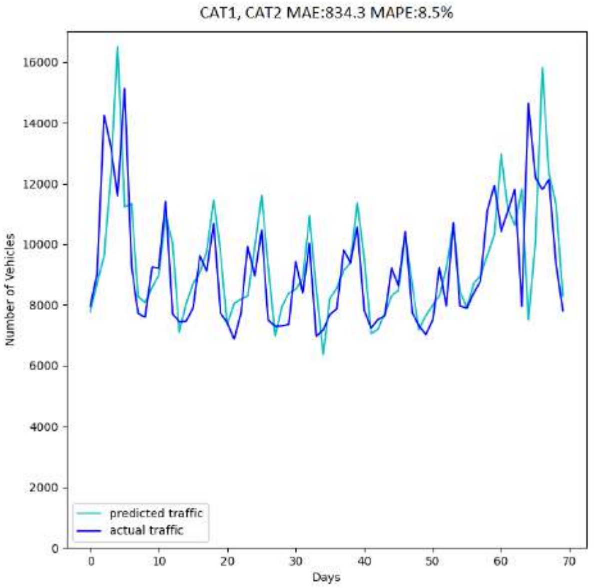


Fig. 16. Light Vehicles MLP performance

IV. Acknowledgments

This work is partially supported by the PERFFECT project co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code:T1EDK-03368) and by the SELAS project co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code:T1EDK-03547).

References

[1] G. Li, W. Lai, X. Sui, X. Li, X. Qu, T. Zhang, and Y. Li, "Influence of traffic congestion on driver behavior in post-congestion driving," *Accident Analysis & Prevention*, vol. 141, p. 105508, 2020.

[2] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3529–3536.

[3] S. B. Seo, P. Yadav, and D. Singh, "Lora based architecture for smart town traffic management system," *Multimedia Tools and Applications*, pp. 1–16, 2020.

[4] G. Solvoll, T. A. Mathisen, and M. Welde, "Forecasting air traffic demand for major infrastructure changes," *Research in Transportation Economics*, vol. 82, p. 100873, 2020.

[5] D. Triantafyllou, N. Kotoulas, S. Krinidis, D. Ioannidis, and D. Tzovaras, "Large vehicle recognition and classification for traffic management and flow optimization in narrow roads," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2017, pp. 1–4.

[6] R. K. Duraku and R. Ramadani, "Development of traffic volume forecasting using multiple regression analysis and artificial neural network," *Civil Engineering Journal*, vol. 5, no. 8, pp. 1698–1713, 2019.

[7] Z. Kadim, K. M. Johari, D. F. Samaon, Y. S. Li, and H. W. Hon, "Real-time deep-learning based traffic volume count for high-traffic urban arterial roads," in *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE, 2020, pp. 53–58.

[8] S. M. Khan, S. Islam, M. Z. Khan, K. Dey, M. Chowdhury, N. Huynh, and M. Torkjazi, "Development of statewide annual average daily traffic estimation model from short-term counts: A comparative study for south carolina," *Transportation Research Record*, vol. 2672, no. 43, pp. 55–64, 2018.

[9] M. Fu, J. A. Kelly, and J. P. Clinch, "Estimating annual average daily traffic and transport emissions for a national road network: A bottom-up methodology for both nationally-aggregated and spatially-disaggregated results," *Journal of Transport Geography*, vol. 58, pp. 186–195, 2017.

[16] J. Ilonen, J.-K. Kamarainen, and J. Lampinen, "Differential evolution training algorithm for feed-forward neural networks," *Neural Processing Letters*, vol. 17, no. 1, pp. 93–105, 2003.

[10] A. Dimara, D. Triantafyllidis, S. Krinidis, K. Kitsikoudis, D. Ioannidis, S. Antipas, and D. Tzovaras, "Fusing birch with g. boosting for improving temporal traffic congestion tailored to port gates: Case study in patras, greece," in *2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*. IEEE, 2020, pp. 1–5.

[11] Y. Yuan, C. Zhang, Y. Wang, C. Liu, J. Ji, and C. Feng, "Linear interpolation process and its influence on the secondary equipment in substations," in *2017 China International Electrical and Energy Conference (CIEEC)*, 2017, pp. 205–209.

[12] P. Schober and T. R. Vetter, "Missing data and imputation methods," *Anesthesia and analgesia*, vol. 131, no. 5, p. 1419, 2020.

[13] X. Liu, S. Wang, X. Zhang, X. You, J. Wu, and D. Dou, "Label-guided learning for text classification," *arXiv preprint arXiv:2002.10772*, 2020.

[14] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[15] D. Cho, C. Yoo, J. Im, and D.-H. Cha, "Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas," *Earth and Space Science*, vol. 7, no. 4, p. e2019EA000740, 2020.

[17] J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[18] S. Bock and M. Weiß, "A proof of local convergence for the adam optimizer," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[19] A. Dimara and C.-N. Anagnostopoulos, "Data based stock portfolio construction using computational intelligence," in *International Conference on Internet Science*. Springer, 2017, pp. 76–94.

[20] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," in *Proceedings of the international conference on Neural Networks*, vol. 3. IEEE Press New York, 1987, pp. 11–14.

[21] G.-B. Huang and H. A. Babri, "General approximation theorem on feedforward networks," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.*, vol. 2. IEEE, 1997, pp. 698–702.

[22] R. K. Jain, "Rajeswari ponnuru (intel), ajit kumar p.(intel), and ravi keron n.(intel). cifar-10 classification using intel optimization for tensorflow," 2017.

[23] W. Wang and Y. Lu, "Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model," in *IOP conference series: materials science and engineering*, vol. 324, no. 1. IOP Publishing, 2018, p. 012049.

[24] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, 2016.

Approaches in Determining Software Development Methods for Organizations: A Systematic Literature Review

Fahmi Alaydrus
Faculty of Computer Science
University of Indonesia
Jakarta, Indonesia
fahmi02@ui.ac.id

Teguh Raharjo
Faculty of Computer Science
University of Indonesia
Jakarta, Indonesia
teguhr2000@gmail.com

Bob Hardian
Faculty of Computer Science
University of Indonesia
Jakarta, Indonesia
hardian@cs.ui.ac.id

Adi Prasetyo
Faculty of Computer Science
University of Indonesia
Jakarta, Indonesia
adip12@cs.ui.ac.id

***Abstract*— Harvard Business Review (HBR) reported 27% of projects were over budget, McKinsey & Company reported 66% of projects were over budget, and the Project Management Institute (PMI) reported more than 40% of projects were over budget. The causes are losing focus, execution issues, content issues, and skill issues. The failure of the software project falls into the process domain category. Undefined processes can lead to low-quality results regarding unexpected outcomes, schedule, and budget. In short, the project is poorly executed. A software development method is needed to execute the project correctly, so the quality is guaranteed, on time, and within budget. This study aims to summarize the approaches used to determine software development methods for organizations and how to use them. A systematic literature review technique was conducted and obtained 16 relevant papers. Ten approaches for determining software development methods were found. Instead of selecting a suitable software development method, the organization could create a software development method by tailoring several deemed appropriate methods to its needs.**

Keywords*— *Software Development Method, Systematic Literature Review, Approaches Selecting Software Development Methods.

I. Introduction

Harvard Business Review reports that 27% of projects usually go over budget [1]. In addition, at least one in six IT projects is 200% over-cost and 70% over-schedule [1]. IT projects performance has an average cost overrun of 66% [2]. Over 40% of projects were completed over budget [3]. The following are problem groups for most project failures: (1) Losing focus due to unclear objectives and lack of business focus [2][3]. (2) Execution issues related to unrealistic schedules and reactive planning [2]. (3) Content issues such as changing requirements and technical complexity [2]. (4) Skill issues such as unaligned teams, lack of skill, and poor communication [2][3]. The causes of project failure groups ratio can be seen in Fig 1.

According to Marchewka, that project failure issues are part of the software failure category's process domain [4]. The process domain is a set of project management and product development processes such as defining project goals and objectives and developing and implementing realistic project plans [4]. Undefined processes can lead to low-quality results in terms of not expected value, outcome, meeting schedule, budget, or quality objectives [4]. Usually, requirements that are not well-defined lead to additional work or a product, process, or system that stakeholders did not ask for or do not need [4]. In short, the project is poorly executed [4].

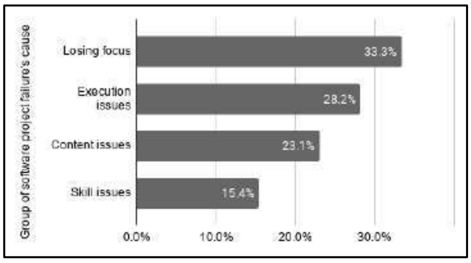


Fig. 1. Group of software project failure causes [2]

One of the project success factors is the proven methodology and tools used [2]. A software project requires a development method to control the people involved, manage the project processes, determine the technology to be prepared, and implement the organizational rules that enable developing complex software on time with quality, of course, to avoid excess budget [5]. The organization can execute software development projects without any development method, but the consequences are often more expensive and less reliable than they should be [6]. The right software development method can accommodate the needs of the project being carried out [8]. Software project needs a development methods, so that the process is structured, projects can be successful with guaranteed quality, work on time, and within budget [7].

The organization can take several approaches to determine software development methods, so the software development method complies with the organization's conditions and needs [5]. This study aims to summarize what approaches the organization can use to define software development methods and how it uses. Many previous studies have discussed how to use an approach for determining software development methods, but no research has been conducted to summarize the approach that the organization can use for its determination. This research also provides insights for determining software development methods that comply with the organization's context, conditions, and needs.

II. Literature Study

A. Software Development Method

A software development method is the approach or method used in the software development process [6].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Sometimes software development methods can be referred as process models or software development methods [5], software development methodologies [8], software development strategies and methods [4][6]. Software development methods are divided into two classification, namely plan-driven and agile [5]. Some samples of software development methods are Waterfall, Incremental and Iterative Development, Prototype, Spiral, Extreme Programming, and Scrum [5][6][8].

B. *The criteria for determining the appropriate software development method for an organization*

The appropriate software development method for an organization is nominated based on organizational conditions, needs, and problems related to software development [5]. There are criteria and sub-criteria used to analyze conditions, needs, and problems related to application development to determine the appropriate development method [9].

TABLE I. CRITERIA AND SUB CRITERIA FOR DETERMINING SOFTWARE DEVELOPMENT METHODS

Criteria	Sample of sub-criteria
Personnel	Personnel substitutions, size of team, willingness of the team to change, team diversity, distribution team.
Requirements	Changing needs on an ongoing basis, needs that are not clear, standards of needs, necessities to meet needs.
Application	Risk level, complexity, hardware architecture, application type, application size, application performance, relationship with existing or planned systems, configuration and documentation management, development stage, products quality.
Organization	Organizational maturity, organization current practices, technical support, organization standards, commitment management, culture, organizational stability, organizational facilities, organization size.
Operational	User resistance, end user commitment, user engagement, number of users, user rotation, prerequisites to be met such as policies, regulations.
Business	Potential project losses, dependence on other projects, dependency, financial considerations, marketing activities, user satisfaction.
Technology	Availability of tools, the use of technology, development tools experience, adoption of new technologies.
Process	Process flexibility, process simplification, learning process.

III. RESEARCH METHODOLOGY

The systematic literature review (SLR) method is used to answer these research questions by comprehensively summarizing the related research that has been done before [10], and combine with existing theories. There are 3 processes carried out in the SLR method based on Kitchenham, namely planning, implementation, and reporting [10].

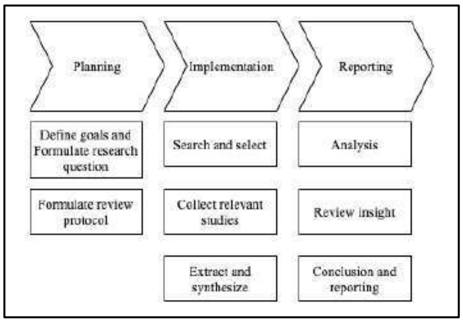


Fig. 2. SLR Methods Stages

A. *SLR Planning Stage*

The first thing to do in the SLR planning stage is to set goals and formulate research questions. Specify the research questions (RQ) was conducted to keep the review focused. It was designed by Population, Intervention, Comparison, Outcomes, and Context (PICOC) formula [11], shown in Table II.

TABLE II. RESEARCH QUESTION'S CRITERIA

Population	Software, software development, software process, software development life cycle
Intervention	Determining approach, technique, tools
Comparison	None
Outcomes	Approach, technique, or tool are used for determining software development method in organization
Context	Studies in organization and academia, software project management

The following are the research question and its motivation discussed in this literature review are shown in Table III.

TABLE III. RESEARCH QUESTIONS

ID	Questions	Motivation
RQ I	What methods can be used to determine the appropriate software development method for the organization?	Identify the most used approach for determining software development method in organization
RQ II	How to use the approach in determining software development method?	Identify how to use the approach to determine software development methods

Thereafter, protocol formulations are performed. After that, the search protocol formulation was carried out. The keywords used for the search are as follows

("software life cycle" OR "software process" OR sdlc OR "software development" OR "software development method*") AND (select* OR implement* OR determin* OR adopt* OR best OR creat* OR decision)

Filtering of literature searches is carried out so that research remains relevant. Following are the filter criteria shown in Table IV.

TABLE IV. CRITERIA OF FILTERING LITERATURE

ID	Criteria	Type
IN1	Publications in the period 2015-2021	Inclusion
IN2	Publications in English	Inclusion
IN3	Publications in the form of journals and proceedings	Inclusion
IN4	Publications must be complete and downloadable full-text version	Inclusion
EX1	Research are related studies are not opinions, books, presentations, and articles	Exclusion
EX2	The topic does not discuss other than the scope of selecting software development and modeling methods	Exclusion

B. *SLR Implementation Stage*

The second stage, namely SLR implementation, is a search and selection of relevant publications to be used as literature in this study were conducted according to the keywords that had been planned in the first stage. The databases used for searching literature are ACM Digital, IEEE, Pro Quest, Science Direct, Scopus, Springer Link, and Wiley Online Library. The SLR process is shown in Fig 3.

Firstly, a search is performed using predetermined keywords. The inclusions in this search are publications in English (IN2), and publications must be journals or proceedings (IN3). The exclusion of this search is publications in opinions, books, presentations or articles (EX1). From the results of the search and inclusion, there were 756 publications. Secondly, filtering the results by displaying publications in the range of time from 2015-2021. There are 426 publications obtained. Thirdly, conducting a review with relevant publications related to selecting software development methods and modelling (IN4) and downloadable publications (IN5). Publications that are not related to the topic of selecting software development methods and modelling will be excluded. From the third stage, 83 publications were obtained. In the fourth stage, a publication is reviewed by reading its contents. Sixteen publications can be used in this study.

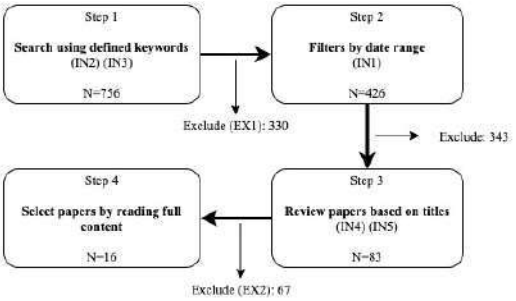


Fig. 3. SLR Process

The results from the extraction and synthesis of publications that will be used as references in this study can be seen in Table V.

TABLE V. THE RESULTS OF THE SELECTED PAPER

Database	Total Papers
ACM Digital	4
IEEE	4
ProQuest	1
Scopus	2
Science Direct	2
Springer Link	2
Wiley Online Library	1
Total	**16**

Publication quality assessment was required for further study selection after inclusion and exclusion criteria [11]. Several questions were developed to evaluate quality of studies shown in Table VI.

TABLE VI. QUALITY ASSESSMENT QUESTIONS

QA1	The discussion in publications is in the scope of software development selection and modeling methods
QA2	Population of study is adequate for data analysis
QA3	Study uses adequate methodology

Each question has following value: (1) not good, (2) adequate, and (3) good. The maximum score is 9 and the least value to pass quality value is 4.5. Table VII below show the quality assessment results.

TABLE VII. QUALITY ASSESSMENT RESULTS

Reference	Q1	Q2	Q3	Total
[12]	3	3	3	**9**
[13]	3	3	3	**9**
[14]	3	2	3	**8**
[15]	3	2	2	**7**
[16]	3	2	2	**7**
[17]	2	2	3	**7**
[18]	3	3	3	**9**
[19]	3	3	3	**9**
[20]	3	3	3	**9**
[21]	3	3	3	**9**
[22]	2	1	2	**5**
[23]	3	2	2	**7**
[24]	3	3	3	**9**
[25]	3	3	3	**9**
[26]	3	2	2	**7**
[27]	3	2	3	**8**

All the final study score was above 4.5. It means those studies eligible to be used for data extraction and data synthesizes.

C. SLR Reporting Stage

The last stage, namely SLR reporting, is selected publications as references will be analyzed comprehensively to find the approach that can be used to determine the software development method for the organization and how to use that approaches. Can be seen in Table VIII, is a list of publications used in this study.

TABLE VIII. THE PUBLICATION TO BE USED

Title	Year	Reference
Systematic Approach for Mapping Software Development Methods to the Essence Framework	2016	[12]
Analysis of software development method selection: a case of a private financial institution	2019	[13]
An Excursion to Software Development Life Cycle Models- an Old to Ever-growing Models	2016	[14]
Exploring the Use of the Cynefin Framework to Inform Software Development Approach Decisions	2015	[15]
Agile manifesto and practices selection for tailoring software development: A systematic literature review	2018	[16]
An Approach for Systematic Planning of Project Management Methods and Project Processes in Product Development	2020	[17]
Tailoring Agile-Based Software Development Processes	2019	[18]
Agile methods tailoring – A systematic literature review	2015	[19]
Suitability of existing Software development Life Cycle (SDLC) in context of Mobile Application Development Life Cycle (MADLC)	2016	[20]
The Use of Analytic Hierarchy Process for Software Development Method Selection: A Perspective of e-Government in Indonesia	2017	[21]
Adopting scrum framework in a software development of payroll information system	2020	[22]
Decision Support Framework for the Adoption of Software Development Methodologies	2019	[23]

Title	Year	Reference
Selection of software development model using TOPSIS methodology	2018	[24]
Software process selection system based on multicriteria decision making	2020	[25]
An integrated approach to formulate a value-based software process tailoring framework	2016	[26]
SDLC Model Selection Tool and Risk Incorporation	2017	[27]

IV. Results and Discussion

The final stage of SLR is reporting. Sixteen literatures have been reviewed, a list of the approach was carried out to determine the method of software development and the determining factors in using this approach.

A. Approaches in Determining Software Development Methods

Answering the question RQ1, "what methods can be used to determine the appropriate software development method for the organization?", based on the literature review, that found 10 approaches. The explanation of these approaches are follows.

Comparative Analysis. This approach is obtained from four literatures [14][20][22][27]. Comparative analysis is carried out based on the calculation of specific situations, objectives, problems, and limitations encountered in project implementation [14][20][22][27]. After understand the conditions of the project, comparisons are made between the characteristics of the project and the existing software development methods [14]. Several criteria are used such as cost, clarity of requirements, documentation, project size, flexibility, and etc [14][20].

Multi-criteria Decision Making Technique. This approach is obtained from four literatures [13][24][25][21]. Multi-criteria decision-making (MCDM) is a sub-discipline of operations research that explicitly considers several conflicting criteria in decision-making [25]. There are several techniques found from literature studies including Analytic Hierarchy Process (AHP) [13][21], Fuzzy AHP [25], and TOPSIS [24].

Intentional Modelling Framework. This approach is obtained from one literature [16]. This framework belongs to the tailoring category of practices from agile software development methods [16]. The choice of practice based on the agile value such as customer-centered, software that works, collaboration, simplicity, communication, nature, learning, pragmatism and adaptability [16].

Goal-oriented Meta-model. This approach is obtained from one literature [16]. This framework belongs to the tailoring category of practices from software development methods [16]. The choice of practice based on the objectives of each activity. Each activity is broken down into the smallest practical parts, then the tailored practice must be understood by the development team [16].

Contingency Factors. This approach is obtained from one literature [19]. These approach deals with customizing software development methods by selecting several methods that will be available to the organization and making the selection of methods based on the development context such as degree of uncertainty, impact and structure of project. However, the development team must understand the chosen execution method [19].

Method Engineering. This approach is obtained from one literature [19]. This approach focuses on the methods of each activity in software development that occurs in the organization. This approach starts with understanding the environment and characteristics of the project, and there is a fragments repository that contains practical collections, will be combined for each project implementation [19].

Cynefin Framework. This approach is obtained from two literatures [15][17]. Cynefin is a decision framework that recognizes the causal differences that exist between different types of systems and proposes new mechanisms for understanding the level of complexity when decisions are made [15]. There are 5 classifications in the cynefin framework, namely complex, complicated, chaotic, simple, disorder [15].

Stacey Matrix. This approach is obtained from one literature [17]. The Stacey Matrix is designed to help understand the factors that contribute to complexity and select the best management action to address different levels of complexity [17]. The basis of the matrix is two dimensions: agreement and certainty [17].

Value-based Software Process Tailoring Framework. This approach is obtained from two literatures [18] [26]. The framework is a combination of value-based factors, MoSCoW rules, Quality Functional Deployment (QFD), Activity-Based Costing (ABC), Priority Map, Value Index and Value Graph [26]. The purpose of this framework is to provide a systematic method using an integrated approach that is able to reduce subjectivity in decision making and satisfy stakeholders [26], also more efficient and outcome-based [18].

Essence Framework. This approach is obtained from one literature [12]. The Essence Framework is a common ground defines the Language and Kernel for modeling software development methods [12] [28] [29]. The language is a domain-specific languages for defining the Kernel, Practices, and Method [28]. The kernel is a conceptual model of software development domain [30]. In short, this framework is a composition of practices. There are similarities, or kernels, shared between all of these methods and practices [12].

B. How to Use of The Approach in Determining Software Development Method

To answer the RQ2 question, "How to use an approach in determining software development methods?", The explanation for each approach from the previous section are as follows:

Comparative Analysis. (1) understand the organization's condition and situation from the aspects of personnel, needs, applications, organization, operations, business, technology, and processes [14]. (2) compile a list of available software development methods and their characteristics, then compare project and organizational characteristics with software development methods [14]. (3) select a software development method based on the highest compatibility level [20].

Multi-criteria Decision Making Technique. The AHP technique, the steps are taken as follows. (1) determining the criteria, sub-criteria, and alternative methods of software development. (2) selecting respondents to rank their importance using the Saaty Scale Preference. (3) calculating to get the weight of each criterion, sub-criteria, and alternative. (4) calculating the consistency ratio to ensure the consistency level of the answers. If the consistency ratio value is below 10%, repeat step number 3. (5) sort the alternatives based on the highest weight [13][21]. The Fuzzy AHP technique, the steps taken are as follows. (1) define linguistic value with corresponding fuzzy numbers. (2) defuzzification of the decision matrix. (3) define criteria, sub-criteria, and alternatives. (4) calculate matrices to get weighted numbers. (5) calculate consistency ratio. (6) sorting weight from alternatives. [25]. The TOPSIS technique, the steps are taken as follows. (1) determine the number of alternatives, criteria, and decision-makers involvement. (2) construct decision matrices. (3) standardize the decision matrix. (4) calculate weighted standardized decision matrix. (5) determine ideal solution and negative ideal solution. (6) find relative closeness to the ideal solution. (7) sorting the alternatives in descending [24].

Intentional Modelling Framework. (1) determine the goals of the project [16]. (2) determine the activities that must be done to achieve these goals [16]. (3) choose practices that are in accordance with the activities to be carried out [16]. (4) check the vulnerability of each selected practice, if there is a vulnerability, then solve the vulnerability by looking for or linking with other practices [16]. (5) apply the practice [16].

Goal-oriented Meta-model. (1) determine the objectives of each activity, then arrange or classify the elements needed [16]. (2) combine or generalize similar elements [16]. (3) do coupling process, installing independent elements with elements that are closely related to each other [16]. (4) the independent elements that have been linked are prioritized according to the appropriate criteria [16]. Ordering priorities can be done by using AHP's formal decision-making techniques [16].

Contingency Factors. (1) summarize all the methods available in the organization [19]. (2) determine the project development context to be carried out, such as the level of uncertainty, the impact of the project, and the project structure [19]. (3) prepare the criteria needed for each project [19]. (4) select the methods that are available or have been used in the organization to organize them into a series of development activities [19]. That way, the entire mechanism of the method has been understood by the team in the organization [19].

Method Engineering. (1) ensure that the organization has a repository of fragments, collection of methods, techniques, or tools used in developing software [19]. (2) understand the project environment such as goals, team, internal and external environment, and project's characteristics [19]. (3) select the fragment from the repository and validate the characteristics of the projects [19]. (4) combine the fragments for implementation in the project [19]. (5) adapt and measure the performance of the project implementation against the compiled collection of fragments [19].

Cynefin Framework. (1) mapping project criteria into Cynefin's domain to help decision makers understand the types of projects to be executed [17]. (2) mapping the criteria into the simple domain if the practice is easiest and has the least causal relationship [17]. (3) mapping criteria into a complicated domain if practices have clear relationships between explainable causes and effects [17]. (4) mapping the criteria into a complex domain if the domain causes the difficulty for process improvement [17]. (5) mapping the criteria into the chaotic domain if the criteria for the project to be carried out are the most difficult and have chaotic causes [17]. (6) mapping the criteria into disorder domain if all perspectives contradict one another [17]. (7) once the mapping is done, a software development method can be selected according to the characteristics of the project [17].

Stacey Matrix. Determining the software development method with this approach is the same as the Cynefin Framework[17]. However, the mapping is entered into a Stacey matrix with measurement parameters that are carried out by determining the level of uncertainty and agreement of a project. [17]. Once mapped, validate the characteristics of the existing software development methods for selection [17].

Value-based Software Process Tailoring Framework. (1) identify the value factors that are most appropriate for the needs of the project [18][26]. (2) do an assessment of the factors that have been determined to rank priorities is carried out using the MoSCow rules to guide the assessment process [26]. (3) assess the relationship between the identified value factors and the activities of the software development process [26]. (3) estimate the cost of each process and task in making the software to completion [26]. (4) if it is deemed unsuitable, select, eliminate, or combine related activities based on the identified value factors, then estimate the cost again until the appropriate value is obtained [26]. Experienced practitioners is needed for using this approach [18][26].

Essence Framework. (1) understand the concept of the kernel and essence language such as Alpha, Activity Space, Work Product, and Activity [12][28][30]. (3) identify the Alpha Kernel that is relevant to the activities carried out [12]. (3) outline the existing software development practices and be understood by the team [12]. (4) add a sub-alpha containing the activity derived from alpha [12]. (5) define alpha states and checkpoints for each alpha activity [12]. (6) add a work product, which contains the output of the activity or defined alpha. (7) define activities based on the expected output or work product [12][28]. (8) identify the relevant kernel activity space to put a series of activities [12][28]. (8) connect activities with kernel activity spaces [28].

TABLE IX. LIST OF APPROACHES DETERMINING SOFTWARE DEVELOPMENT METHODS

No	Classification	Approaches	References
1	Selecting	Comparative Analysis	[22], [14], [20], [27]
2		Multi-criteria Decision Making Technique	[24], [13], [21], [25]
3		Cynefin Framework	[17]
4		Stacey Matrix	[17]
5	Tailoring	Intentional Modelling Framework	[16]
6		Goal-oriented Meta-model	[16]
7		Contingency Factors	[19]
8		Method Engineering	[19]
9		Value-based Software Process Tailoring Framework	[18] [26]
10		Essence Framework	[12]

V. Conclusions

This research was conducted by a systematic literature review. In general, there are two classifications for determining software development methods for organizations, namely selecting, and tailoring. Sixteen literature reviewed, there are 10 approaches to determine software development methods including Comparative Analysis, Multi-criteria

Decision Making Techniques, Intentional Modeling Frameworks, Goals-oriented Meta-models, Contingency Factors, Method Engineering, Cynefin Framework, Stacey Matrix, Value -based Processes Tailoring Framework, and the Essence Framework. To determine the most suitable software development method for the organization, not only select an existing software development method, but also create a software development method are possible that fits the context.

VI. Future Work and Limitation

This study's limitation is the used criteria, keywords, and databases have not reached all related research that has ever existed. Furthermore, this research can be continued by evaluating the effectiveness of this approach in determining the organization's most suitable software development method.

References

[1] B. Flyvbjerg and A. Budzier, “Why Your IT Project May Be Riskier than You Think,” *Harv. Bus. Rev.*, 2011.

[2] M. Bloch, S. Blumberg, and J. Laartz, “Delivering large-scale IT projects on time, on budget, and on value,” *McKinsey Co.*, 2012.

[3] PMI, “Success Rates Rise: Transforming the high cost of low performance,” *Pulse Prof. - 9th Glob. Proj. Manag. Surv.*, 2017.

[4] J. T. Marchewka, *Information Technology Project Management: Providing Measurable Organizational Value*. 2015.

[5] I. Sommerville, *Software Engineering 10th edition*. 2016.

[6] R. Pressman and B. Maxim, *Software Engineering: A Practitioner's Approach, 8th Edition*. 2014.

[7] R. Kneuper, *Software Processes and Life Cycle Models: An Introduction to Modelling, Using and Managing Agile, Plan-Driven and Hybrid Processes*. 2018.

[8] R. Roth, B. H. Wixom, and A. Dennis, *Systems Analysis and Design: An Object-Oriented Approach with UML 5th Edition*. 2012.

[9] P. Clarke and R. V. O'Connor, “The situational factors that affect the software development process: Towards a comprehensive reference framework,” *Inf. Softw. Technol.*, 2011.

[10] B. Kitchenham *et al.*, “Systematic literature reviews in software engineering-A tertiary study,” *Inf. Softw. Technol.*, 2009.

[11] B. Kitchenham, “Procedures for Performing Systematic Reviews,” 2004.

[12] G. Giray, E. Tüzün, B. Tekinerdogan, and Y. Macit, “Systematic approach for mapping software development methods to the essence framework,” *Proc. - 5th Int. Work. Theory-Oriented Softw. Eng. TOSE 2016*, 2016.

[13] E. E. Surbakti, B. Purwandari, I. Solichah, and L. Kumaralalita, “Analysis of software development method selection: A case of a private financial institution,” *ACM Int. Conf. Proceeding Ser.*, 2019.

[14] U. S. Shah, D. C. Jinwala, and S. J. Patel, “An Excursion to Software Development Life Cycle Models,” *ACM SIGSOFT Softw. Eng. Notes*, 2016.

[15] R. V. O'Connor and M. Lepmets, “Exploring the use of the cynefin framework to inform software development approach decisions,” *ACM Int. Conf. Proceeding Ser.*, vol. 24-26-Augu, pp. 97–101, 2015.

[16] S. Kiv, S. Heng, M. Kolp, and Y. Wautelet, *Agile manifesto and practices selection for tailoring software development: A systematic literature review*. Springer International Publishing, 2018.

[17] J. Baschin, T. Huth, and T. Vietor, “An approach for systematic planning of project management methods and project processes in product development,” *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, 2020.

[18] R. Akbar, “Tailoring Agile-Based Software Development Processes,” *IEEE Access*, 2019.

[19] A. S. Campanelli and F. S. Parreiras, “Agile methods tailoring - A systematic literature review,” *J. Syst. Softw.*, vol. 110, 2015.

[20] A. Kaur and K. Kaur, “Suitability of Existing Software Development Life Cycle (SDLC) in Context of Mobile Application Development Life Cycle (MADLC),” *Int. J. Comput. Appl.*, 2015.

[21] M. Helingo, B. Purwandari, R. Satria, and I. Solichah, “The Use of Analytic Hierarchy Process for Software Development Method Selection: A Perspective of e-Government in Indonesia,” *Procedia Comput. Sci.*, 2017.

[22] B. G. Sudarsono, Fransiscus, H. Hartono, D. Y. Bernanda, and J. F. Andry, “Adopting scrum framework in a software development of payroll information system,” *Int. J. Adv. Trends Comput. Sci. Eng.*, 2020.

[23] L. Simelane and T. Zuva, “Decision Support Framework for the Adoption of Software Development Methodologies,” *2019 Int. Multidiscip. Inf. Technol. Eng. Conf.*, 2019.

[24] D. Gaur and S. Aggarwal, *Selection of software development model using TOPSIS methodology*, vol. 847. Springer Singapore, 2019.

[25] P. Pandey and R. Litoriya, “Software process selection system based on multicriteria decision making,” *J. Softw. Evol. Process*, 2020.

[26] N. Azura Zakaria, S. Ibrahim, and M. Naz'ri Mahrin, “An integrated approach to formulate a value-based software process tailoring framework,” 2016.

[27] P. Agarwal, A. Singhal, and A. Garg, “SDLC Model Selection Tool and Risk Incorporation,” *Int. J. Comput. Appl.*, 2017.

[28] Object Management Group, “Essence – Kernel and Language for Software Engineering Methods,” no. Version 1.2, 2018.

[29] I. Jacobson, H. Lawson, P.-W. Ng, P. E. McMahon, and M. Goedicke, *The Essentials of Modern Software Engineering: Free the Practices from the Method Prisons!* ACM Digital and Morgan & Claypool, 2019.

[30] I. Jacobson, P.-W. Ng, P. E. McMahon, I. Spence, and S. Lidman, “The Essence of Software Engineering - The SEMAT Kernel,” *Commun. ACM*, 2012.

Application of Machine Learning Techniques to the Prediction of Student Success

Eluwumi Buraimoh
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
2287804@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

Kershree Padayachee
Centre for Learning, Teaching and Development
The University of the Witwatersrand
Johannesburg, South Africa
kershree.padayachee@wits.ac.za

***Abstract*—This study presents six machine learning models in the prediction of student success in a technology-mediated environment. Student behavioral attributes with a learning management environment have proven to be a significant determinant in forecasting students' performance. This study attempts to provide the model with optimum accuracy to determine students who need assistance to improve their educational performances and other learning outcomes. We examined the impacts of SMOTE data re-sampling and the effect of attribute selection in this study. The models' performances were enhanced with the re-sampling method as the imbalanced dataset was identified to have performed poorly. Attribute Selection with the top ten attributes and 10-fold cross-validation offer best performances. The six predictive models utilized in this study are Linear Discriminant Analysis, Logistic Regression, Classification and Regression Tree, K-Nearest Neighbour, Naïve Bayes Classifier, and Support Vector Machines. Classification and Regression Tree model and Linear Regression had the best accuracy score of 0.86 after 10-fold cross-validation and top ten attribute selection. This study concludes that student behavioral attributes are useful predictors of student success.**

***Keywords*—Blended-Learning, Data Re-sampling, Machine Learning Models, Information Gain, Feature Selection, SMOTE**

I. Introduction

The machine learning technique is one of the main methods used in studying student performance or success, aside from statistical analysis and data mining. Academic performance is a daunting challenge for tertiary education institutions across the globe. [1] and [2] described data analytics as a tool for identifying students who are struggling educationally and enhancing throughput in various educational institutions.

This study falls under the category of Educational Data Mining (EDM). EDM is a subdivision of data mining that specializes in designing, evaluating, and implementing various automatic tools for measuring vast amounts of data from academic environments [3]. This study investigates student success through student behavioral attributes in the learning management environment. In any learning environment, student engagement is a crucial indicator for assessing a student's success or failure [4]. The channels of education delivery include traditional classroom, online-learning, blended-learning, and others. The Learning Management System (LMS) is a learning platform that allows instructors and learners to communicate without having to meet in person [5]. The global adoption of LMS platforms in learning is increasing by the day as several factors have warranted this acceptance. The reason for the adoption of LMS include the convenience of learning at student pace, improvement of cost-efficiency for the institutions and full coverage of a large number of students [6]. The blended-learning, which is interchangeably called hybrid-learning, is an infusion of both standard classroom and technology-aided settings [7]. [8] provided objectives of blended-learning as an effective learning process, student-teacher physical contact, academic performance enhancement, and learner's freedom. However, the reported rate of failure in blended-learning in the undergraduate programs has dramatically increased in current time [9]. Research into the determinant factors for a boost in student success in this blended-learning environment will increase throughput in tertiary education across the globe as the present COVID-19 pandemic necessitated the adoption of one form of online or the other. This research provides machine learning models with optimal performance in predicting student success in undergraduate as a form detective tool in assisting students from not dropping out of the blended-learning course and increasing academic outcomes.

In this paper, Section II discusses the related work on the prediction of the success of students with the application of machine learning. Section III highlights the research design, including data collection and pre-processing, feature selection, machine learning techniques, and evaluation metrics. Section IV outlines the results. Section V concludes the paper.

II. Related Work

Student success prediction is a crucial challenge in the technology-aided settings [5]. Past researches have provided various factors that influence student success. Some the factors are the student demographic information such as gender [10], previous academic performance [11] and interactions with the learning environment [6]. Therefore, this research investigates the influencing power of the student's interaction with the Learning Management System. Most scholars proposed that

978-1-6654-4067-7/21/$31.00 © 2021 IEEE

the engagement/interaction on LMS has a positive correlation with student success [6], [12]. [13] studied student performance on final examination grade in an undergraduate program using Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Sequential Minimal Optimisation and Neural Network. The Logistic Regression performance was the best among the algorithms used in the study with an accuracy value of 66%. The authors in [14] predicted student academic performance in the virtual learning environment for four categories. Artificial Neural Network outperformed Logistic Regression and Support Vector Machine algorithm with a classification accuracy of 84% to 93%. [15] used the Recurrent Neural Network (RNN) in their in-depth knowledge and engagement study. The authors achieved an accuracy value of 88.3% for RNN in their research. [5] investigated student learning performance from LMS data using Support Vector Machine, Linear Discriminant Analysis, Random Forest, K-Nearest Neighbour, and Classification and Regression Tree (CART). The result showed the performance of random forest as the best with an accuracy value of 90%. [5] further suggested the class imbalance problem's solution in future work. [16] applied the Logistic Regression predictive model in their study on variables (degree of engagement, degree of prestige, degree of visibility, student access amount, management system by subject, experience, age, and gender) and got an accuracy value of 87.53%. [12] investigated the academic success of students based on their learning management network activities. The predictive models used in the analysis were Artificial Neural Network, Decision Tree, and Naïve Bayes with bagging boosting and ensemble techniques. The highest accuracy value of 82% was obtained from the Decision Tree classifier.

Understanding data sources in prediction are essential as it lays the groundwork for future research that has yet to be pursued. It also cuts down computation times on feature extraction [17]. Many studies have used qualitative data (surveys), quantitative data (student behavioral data from online learning activities), and others used combinations of both data types in the prediction of student performance [4]. [6] in their research used publicly available data from the Open University of the United Kingdom to investigate the student's performance. Four online courses from the Moodle LMS log-file data of the undergraduate students at Tel Aviv University, Israel, were utilized by [18] used.

Captured images from videos were developed and used for engagement recognition in relation to student performance by [19]. [20] investigated the connection between student engagement and academic success in a technology-mediated platform using the LMS data from a North American university undergraduate science course. [21] explored interview questions for student engagement challenges in e-learning platforms at different Saudi universities and their relationship with student performance. [16] applied Moodle LMS data from graduate courses in public management course at the university in Brazil from 2014 to 2015. [6] used four variables to analyze student performance in online learning: initial evaluation results, the highest level of education, final test score, and clicks on the learning site. Academic achievement was strongly correlated with student clicks on nine online learning sites, final grades, and evaluation performance. Students click on forumng and oucontent were also found to be influential in predicting student performance. Student participation and final exam grade were positively impacted by forum conversation and access to course content. [20] examined the connection between student engagement and performance in a technology-mediated setting using nine engagement metrics and a cluster analysis derived through students' activity records. According to their findings, student characteristics such as frequency of logins, material read, and the number of forum read affected quiz results, resulting in a high final course score. Because of the positive association between participation and results, [20] suggested that student engagement may be a determinant of academic success.

III. Methodology

In this study, we seek to predict the success using the student behavioral patterns/activities on learning management. Six machine learning predictive models will be trained with K-fold cross-validation after feature selection using the top five and ten features after applying the information gain filter on the dataset obtained from LMS. The confusion matrix, accuracy, precision, recall, and f1-score will be used to evaluate the models' performance in this study to ascertain the model with the optimum performance.

A. Data Acquisition and Pre-processing

Students dataset containing the demographic, behavioral, and academic records were obtained from Kalboard 360 LMS. The data has 480 records of 305 male and 175 female students in an institution [12]. The data consists of 16 numerical and categorical attributes of students gathered over two semesters(first and second). The target variables are low, medium, and high represented below:

$$Grade = \begin{cases} 0-69, & \text{Low} \\ 70-89, & \text{Medium} \\ 90-100, & \text{High} \end{cases}$$

The dataset's class distribution is 127, 211, and 142 for low, medium, and high classes. To fix the class imbalanced distribution challenge, we will be employing Synthetic Minority Oversampling Technique (SMOTE) to avoid dominance from the majority class and improve the models' performance. The SMOTE works by generating new instances from the minority group prior to training the model [22].

B. Attribute Selection

To predict student success in a blended-learning environment, we explored the Information Gain evaluation filter to determine the most contributing attribute. For attribute selection, entropy is used to measure the value of attributes in descending order. We will be selecting the attributes with high entropy for dimensionality reduction and improvement in model performance [23]. Information gain value is represented mathematically as 0≤e≤1. This means that the value spans from 0 to 1.

C. Classification Models

In this study, two linear (Linear Discriminant Analysis and Logistic Regression) and four non-linear (Classification and Regression Tree, K-Nearest Neighbor, Naïve Bayes and Support Vector Machines) supervised machine learning models will be trained to predict the success of the student in a blended-learning setting.

a) Linear Discriminant Analysis: The Linear Discriminant Analysis (LDA) is a predictive model used to model the differences in classes. The discrimination is done by comparing the means of attributes. LDA is used for dimensionality reduction and the minimization of the possibility of misclassifying cases. The structure of LDA used in this paper is from [24].

b) Logistic Regression: The Logistic Regression (LR) is used for predictive exploration. LR is also used to forecast categorical dependent attributes, mostly with the support of predictor attributes. The LR is focused on the estimate of the greatest probability, and the estimate must be most likely. The architecture of the LR used in this study follows [25].

c) Classification and Regression Tree: The Classification and Regression Trees (CART) build a framework from the training set. The division points are selected rapaciously by comparing each attribute and the significance of each attribute in the training set to reduce the loss function [24]. The application of the CART model in this paper is from [24].

d) K-Nearest Neighbor: The K-Nearest Neighbor (KNN) identifies the centroid sample in the training dataset for new samples. The overall average sample is considered here as forecast from the centroid closest neighbor [5]. The distance measure used is the Euclidean distance. KNN is a simple model but very useful in prediction.

e) The Naïve Bayes Classifier: The Naïve Bayes Classifier (NBC) is the most proactive and logical learning algorithm for most classification problems. NBC is based on Bayes' theory of strong assumptions of independence within attributes using a Bayesian framework [1], [6], [24]. The execution of NBC in this study is gotten from [24]

f) Support Vector Machines: The Support Vector Machines(SVM) is an efficient, strong, and reliable predictive model that is identified by a separating hyperplane. SVM generates a decision boundary that is used for prediction. SVM works by determining the closest data dimensions called support vectors to the inference segregation in the training dataset and separates the current test variable through the use of the functional margin [26]. The implementation of the SVM used in this paper comes from [26]

D. Performance Metric

The efficiency of the six classification models used in this study will be assessed through performance metrics such as the confusion matrix, accuracy, recall, precision, and F1 score after K-fold cross-validation, where k=10. These performance metrics have been widely used in previous research.

Table I: Information Gain and Attributes Categorisation for the Set of Attributes in the Prediction of Student Success

Rank	Entropy	Attribute	Attribute Categorisation
1	0.46	Visited Resources	Behavioural
2	0.4	Student Absence Days	
3	0.37	Raised Hands	
4	0.26	Announcement View	
5	0.15	Parent Answering Survey	
6	0.13	Nationality	Demographic
7	0.13	Relation	
8	0.12	Place of Birth	
9	0.11	Discussion	Behavioural
10	0.10	Parent School Satisfaction	
11	0.07	Topic	Academic
12	0.05	Gender	Demographic
13	0.04	Grade Id	Academic
14	0.01	Semester	
15	0.01	Stage Id	
16	0.00	Section Id	

- **Confusion Matrix**: The value of the information provided by a predictive model about expected and actual class labels is held in the confusion matrix. Our models are evaluated using the information in the matrix.
- **Accuracy**: The accuracy score is a common metric for evaluating classification models. It's calculated as the number of precise predictions dependent on the total number of predictions.
- **Recall**: The number of accurate positive predictions and the ratio of the total number of positives are referred to as recall. This is also known as the true positive rate.
- **Precision**: The number of accurate positive predictions as a percentage of the total number of positive predictions is known as precision.
- **F1 score**: The F1 score represents the average of recall and precision. It serves as a red flag of incorrectly classified performance.

IV. Results and Analysis

The results of the experiments performed with the six predictive models are presented here. After applying the information gain attribute evaluation in InformationG, ten attributes were the most contributing attributes in predicting student success. The top five attributes are categorized under student behavioral attributes with the LMS, as shown in InformationG. The two separate experiments were performed using the top 5 and 10 attributes for attribute selection to reduce the dimension and improve models' performances. K-fold (5 and 10) cross-validations were utilized for training the models after resampling the data with the SMOTE technique in this study.

The Information Gain results show the order of the importance of the features in the prediction of student success in descending order. Figure 1 also gives the graphical representation of the features in order of entropy value.

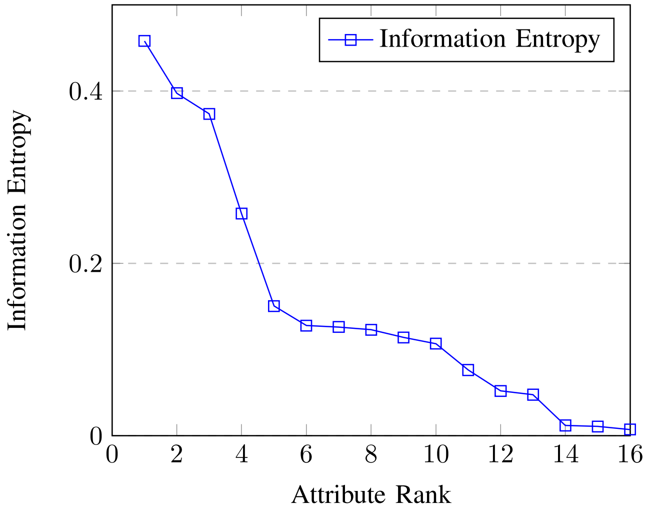


Figure 1: A Chart representation of the Information Gain for Set of Attributes to Predict Student Success

		Predicted		
		Low	Medium	High
Actual	Low	98%	2%	0%
	Medium	19%	62%	19%
	High	0%	10%	90%

Table II: Confusion Matrix showing the performance of **Linear Discriminant Analysis** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	98%	2%	0%
	Medium	14%	67%	19%
	High	0%	10%	90%

Table III: Confusion Matrix showing the performance of **Logistic Regression** Model after 10-fold cross-validation using the top 10 attributes

A. *Prediction Models*

This segment of the paper presents the results obtained from six models used in this study. The confusion matrices in Tables II, III, IV, V, VI and VII highlight the performances of the predictive models. The confusion matrices are determined from the testing dataset with the SMOTE-balanced dataset after the 10-fold cross-validation. We obtained the best model performances from the balanced dataset in contrast to the imbalanced dataset performance. tab: 10-fold comparison and tab: 5-fold comparison also compared the performances of the application of the 10 and 5 fold cross-validation on the balanced and imbalanced dataset. The performance of the 10-fold cross-validation outweighs that of 5-fold in the balanced dataset, while 10 and 5 fold cross-validate make no difference in the models' performances in the imbalanced dataset. From tab: Linear, the Linear Discriminant Analysis obtained an accuracy score of **0.84** after attribute selection of the top 10 attributes and 10-fold cross-validation. The percentage of the classification of the classes are 98, 62 and 90 for the low, medium, and high, respectively. The accuracy score for the Logistic Regression model is **0.86** and tab: Logistic shows the performance of Logistic Regression for low-class as 98%, medium as 67%, and high as 90%. tab: CART represents the Classification and Regression Tree model's performance with an accuracy score of **0.86**. The percentage of the low, medium, and high classes is 95, 72, and 88, respectively. The K-Nearest Neighbour accuracy score is **0.81** which was obtained from the confusion matrix in tab: KNeighbour with the low, medium, and high classes percentage as 92, 56, and 90. The dominant class in this classification is low-class. The Naïve Bayes Classifier performance as represented in tab: Naïve gives the accuracy score of **0.82** where the classes are rightly classified in the percentage of 88, 61, and 92 for low, medium, and high. The performance of Support Vector Machines as illustrated in tab: Support presents an accuracy score of **0.72** achieved from the classification of 90%, 36%, and 84% for low, medium, and high classes. The best accuracy were achieved from the **Logistic Regression** and **Classification and Regression Tree** with an accuracy score of **0.86**. The poorest performance was obtained from the **Support Vector Machines** with an accuracy score of **0.72**. In summary, the performances of the six models used in this study are represented in tab: Summary where the Classification and Regression Tree outperformed the five other models with an accuracy value of 0.86, precision value of 0.86, recall value of 0.86, F1-score value of 0.86 and Area Under Curve (AUC) value of 0.97. The other models in order of their performances are Logistic Regression, Linear Discriminant Analysis, Naïve Bayes Classifier, K-Nearest Neighbour, and Support Vector Machines.

V. Conclusion

This paper gives a framework for predicting student success in a blended-learning course to assist vulnerable students who are prone to fail or withdraw from the program. The result represented in InformationG shows that the behavioral and demographic attributes are the most contributing predictors in forecasting student performance. The academic attributes have no tangible contribution to attribute importance. The

		Predicted		
		Low	Medium	High
Actual	Low	95%	5%	0%
	Medium	6%	72%	22%
	High	0%	12%	88%

Table IV: Confusion Matrix showing the performance of **Classification and Regression Tree** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	92%	5%	3%
	Medium	28%	56%	16%
	High	2%	8%	90%

Table V: Confusion Matrix showing the performance of **K-Nearest Neighbour** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	88%	12%	0%
	Medium	11%	61%	28%
	High	0%	8%	92%

Table VI: Confusion Matrix showing the performance of **Naïve Bayes** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	90%	8%	2%
	Medium	36%	36%	28%
	High	2%	14%	84%

Table VII: Confusion Matrix showing the performance of **Support Vector Machines** Model after **10-fold cross-validation** using the **top 10 attributes**

Predictive Model	**SMOTE + FS**		**Raw_data +FS**	
	5 Features	**10 Features**	**5 Features**	**10 Features**
LDA	0.82	0.84	**0.75**	**0.74**
LR	**0.84**	**0.86**	0.74	**0.74**
CART	0.82	**0.86**	0.69	0.73
KNN	0.75	0.81	0.58	0.60
NBC	0.82	0.82	0.71	0.67
SVM	0.70	0.72	0.56	0.60

Table VIII: Comparison of the Models' Performances in terms of the Accuracy score for balanced (SMOTE) and Imbalanced (raw) data after **10-Fold Cross-Validation** using 5 and 10 Features.

Predictive Model	**SMOTE + FS**		**Raw_data +FS**	
	5 Features	**10 Features**	**5 Features**	**10 Features**
LDA	0.81	0.83	**0.75**	0.74
LR	0.83	**0.84**	**0.74**	**0.74**
CART	**0.86**	**0.84**	0.69	0.73
KNN	0.74	0.76	0.58	0.60
NBC	0.83	0.82	0.71	0.67
SVM	0.70	0.72	0.56	0.60

Table IX: Comparison of the Models' Performances in terms of the Accuracy score for balanced (SMOTE) and Imbalanced (raw) data after **5-Fold Cross-Validation** using 5 and 10 Features.

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
LDA	0.84	0.84	0.84	0.83	0.94
LR	**0.86**	0.85	**0.86**	0.85	0.94
CART	**0.86**	**0.86**	**0.86**	**0.86**	**0.97**
KNN	0.81	0.81	0.81	0.80	0.92
NB	0.82	0.81	0.82	0.81	0.92
SVM	0.72	0.70	0.72	0.70	0.90

Table X: The Summary of the Performance Metrics of the Predictive Models' Performances after 10-fold Cross-Validation on the SMOTE balanced dataset and top 10 Attribute Selection.

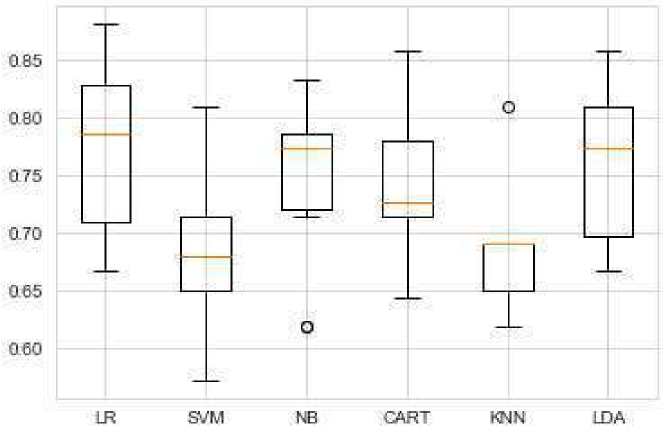


Figure 2: Algorithm Comparison Performance by Training Set

imbalanced data resulted in the models' poor performances in this study; hence, SMOTE method re-sampling technique for balancing to equal count. The application of SMOTE method and the attribute selection using the top 10 features recorded high-performances in the models' performances as expressed in tab: 10-fold comparison. The 5-fold cross-validation results in tab: 5-fold comparison were not as good as that of 10-fold cross-validation in tab: 10-fold comparison except for CART and NBC with the top 5 features. It is also important to note that the 5 or 10 fold cross-validation results are the same for the imbalanced data with 5 and 10 top attributes as shown in tab: 5-fold comparison and tab: 10-fold comparison. In fig:Algorithm Comparison, the LR and LDA performances were the best at the training phase. From tab: 10-fold comparison, the accuracy scores of CART and LR were the best, with a score of 0.86. The other accuracy scores are 0.84, 0.82, 0.81, and 0.72 for LDA, NBC, KNN, and SVM respectively. On critical analysis of the results illustrated in tab: Summary, the CART performance for classification of student success band was the best with an accuracy score of 0.86, precision score of 0.86, recall of 0.86, and an AUC of 0.97. The other models in order of performances are LR, LDA, NBC, KNN, and SVM. We observed that the class's misclassification rate in medium-class was higher than any other classes for imbalanced and SMOTE balanced data. The most rightly predicted class is low, followed by high class. In summary, the results obtained from this study show that machine learning techniques are efficient in identifying student performance on time for possible aid to prevent failure in their courses. The behavioral and demographic attributes are also essential in student performance classification. The constraint of this study is the type of data used. A more robust dataset in terms of the number of attributes would have given a holistic view of other essential attributes needed to classify the student performance. The findings of this study are solely dependent on the data utilized in the research. In the future study, we intend to investigate the reasons for the medium class's low classification for both imbalanced and balanced datasets. This study's machine learning models will also be extended to data from the university repository and not freely available dataset online used in this study. This paper's contribution is the presentation of the critical behavioral attributes for timely identification of students who are prone to withdraw from their courses for assistance by the institutions or administrators in an expeditious manner. This study concludes by highlighting that students' behavioral activities with the LMS are positive predictors to detect the students' performance.

Acknowledgment

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant numbers: 121835).

References

[1] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *In Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT '20), September 14-16, 2020, Cape Town, South Africa. ACM, New York, NY, USA, 10 pages.* ACM, 2020.

[2] A. D. Kumar, R. P. Selvam, and K. S. Kumar, "Review on prediction algorithms in educational data mining," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 8, pp. 531–537, 2018.

[3] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.

[4] M. Hu and H. Li, "Student engagement in online learning: A review," in *2017 International Symposium on Educational Technology (ISET)*. IEEE, 2017, pp. 39–43.

[5] A. Dutt and M. A. Ismail, "Can we predict student learning performance from lms data? a classification approach," in *3rd International Conference on Current Issues in Education (ICCIE 2018)*. Atlantis Press, 2019, pp. 24–29.

[6] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[7] W. W. Porter, C. R. Graham, K. A. Spring, and K. R. Welch, "Blended learning in higher education: Institutional adoption and implementation," *Computers & Education*, vol. 75, pp. 185–195, 2014.

[8] R. T. Osguthorpe and C. R. Graham, "Blended learning environments: Definitions and directions," *Quarterly review of distance education*, vol. 4, no. 3, pp. 227–33, 2003.

[9] T. Abed, R. Ajoodha, and A. Jadhav, "A prediction model to improve student placement at a south african higher education institution," in *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 2020, pp. 1–6.

[10] Z. Cai, X. Fan, and J. Du, "Gender and attitudes toward technology use: A meta-analysis," *Computers & Education*, vol. 105, pp. 1–13, 2017.

[11] C. J. Asarta and J. R. Schmidt, "Comparing student performance in blended and traditional courses: Does prior academic achievement matter?" *The Internet and Higher Education*, vol. 32, pp. 29–38, 2017.

[12] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016.

[13] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3. IOP Publishing, 2020, p. 032019.

[14] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," *Computers in Human Behavior*, vol. 104, p. 106189, 2020.

[15] K. Mongkhonvanit, K. Kanopka, and D. Lang, "Deep knowledge tracing and engagement with moocs," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 340–342.

[16] J. C. S. Silva, J. L. Ramos, R. L. Rodrigues, A. S. Gomes, F. d. F. de Souza, and A. M. A. Maciel, "An edm approach to the analysis of students' engagement in online courses from constructs of the transactional distance," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2016, pp. 230–231.

[17] J. Gardner and C. Brooks, "Student success prediction in moocs," *User Modeling and User-Adapted Interaction*, vol. 28, no. 2, pp. 127–203, 2018.

[18] T. Soffer and A. Cohen, "Students' engagement characteristics predict success and completion of online courses," *Journal of Computer Assisted Learning*, vol. 35, no. 3, pp. 378–389, 2019.

[19] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[20] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Relationship between student engagement and performance in e-learning environment using association rules," in *2018 IEEE World Engineering Education Conference (EDUNINE)*. IEEE, 2018, pp. 1–6.

[21] M. A. Alsubhi, N. S. Ashaari, and T. S. M. T. Wook, "The challenge of increasing student engagement in e-learning platforms," in *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*. IEEE, 2019, pp. 266–271.

[22] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707*, 2013.

[23] R. Ajoodha, S. Dukhan, and A. Jadhav, "Data-driven student support for academic success by developing student skill profiles," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. IEEE, 2020, pp. 1–8.

[24] J. Brownlee, "Machine learning mastery with python," *Machine Learning Mastery Pty Ltd*, pp. 100–120, 2016.

[25] S. Sperandei, "Understanding logistic regression analysis," *Biochemia medica: Biochemia medica*, vol. 24, no. 1, pp. 12–18, 2014.

[26] G. P. S. Manu, "Classifying educational data using support vector machines: A supervised data mining technique," *Indian Journal of Science and Technology*, vol. 9, p. 34, 2016.

RCNX: Residual Capsule NeXt

Arjun Narukkanchira Anilkumar
IOT Collaboratory IUPUI
Department of Electrical and Computer Engineering
Purdue School of Engineering and Technology
Indianapolis, USA
arjuna@iupui.edu

Mohamed El-Sharkawy
IOT Collaboratory IUPUI
Department of Electrical and Computer Engineering
Purdue School of Engineering and Technology
Indianapolis, USA
melshark@iupui.edu

***Abstract*— Contrary to the popular Convolutional Neural Network (CNN), which depends on the shift-invariance in the image, Capsule Networks depends on hierarchical model relations. This aspect of Capsule Networks keeps them in the machine learning domain despite their enormous size with only comparable accuracy to the CNNs. Capsules utilize an intricate algorithm to create a route by agreement, leading to their size and uniqueness. Recent developments in Capsule Networks have contributed to mitigating the problem of size and improving accuracy. We focus on modifying one of the Capsule Network, Residual Capsule Network (RCN), to a comparable size to modern CNNs, thus restating Capsule Network's importance. In this paper, Residual Capsule NeXt (RCNX) is proposed as an effective and more advanced version of RCN with a size of 1.5M parameters and an unprecedented improvement in the network's accuracy on the CIFAR-10 dataset to 89.3%. This accuracy and size exceed the famous embedded CNN model MobileNetV3.**

Keywords—Convolution Neural Networks, Capsule Network, Residual Capsule Network, ResNeXt, CIFAR-10, Residual Capsule NeXt, RCNX

I. Introduction

In recent times, CNNs are prevalently convenient for image recognition in the Machine learning domain. Before the arrival of machine learning, any signal processing for an image used convolution, which led to the intuitive development of CNNs. Recently machine learning algorithms are moving towards an approximation of human vision[2]. One of the earliest successful attempts was AlexNet. The neural arrangements of the AlexNet had similarities to the human visual pathway[2], [3]. The model replicated the interlaced Optic nerves in the human visual pathway.

Capsule Networks[1] are different in many ways when compared to CNN. Capsule Networks took a forward step in the improvement of end layers of image recognition. This network is first proposed by one of the authors of AlexNet, i.e., Geoffrey Hinton[1]. Capsule Networks implement the concept of capsules for each characteristic detected in an image. These capsules indeed required a new way of routing between nodes to bring the idea of routing by agreement. All these combined, we arrive at an excellent model to recognize handwritten digits. Furthermore, Capsule Networks can recreate the different handwritten digits in different styles, a fascinating technique known as the reconstruction[1].

Developments are brought into the Capsule network to advance its application. These advancements helped Capsule networks to improve from simple handwritten digit classification to fully functional object recognition in an image. We focus on one such model, which improved the Capsule network's application to become an excellent image recognition tool, namely, Residual Capsule Network[4].

In this paper, a new machine learning model, Residual Capsule Next (RCNX), is proposed, which aims to improve contemporary architecture Residual Capsule Network (RCN)[4]. By utilizing well-established ideas that improve model size and accuracy, we transform RCN into a machine learning model for embedded systems. The paper focuses on reducing the model size tremendously by including modifications that have been proven in different Capsule Network models while improving the model's accuracy. The complex modifications proposed in this paper include replacing initial convolutional layers, change in architecture, improvement in routing by agreement algorithm, including a complex reconstruction network, activation functions, and tuning the model to be excellent in every layer.

II. Background

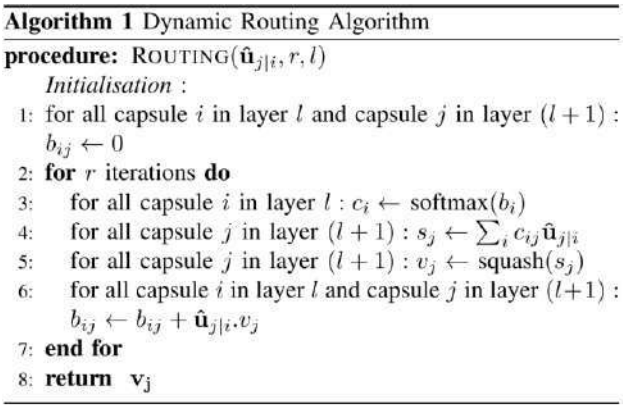
A. Convolutional Neural Networks or CNNs

Space invariant artificial neural networks, commonly known as CNNs, have their name due to the inclusion of existing convolution algorithms and their shift-invariant function[5]. CNN derives its understanding from the similarities and differences from region to region in an image. They feed these scalar-feature detectors output through multiple layers of feature detectors and pooling layers and embedded in them to reduce the complexity while capturing the assumed essential data[4].

B. Capsule Network

Acknowledging limitations of the above-mentioned scalar-feature detectors and the loss of information due to pooling layers led to Capsule Networks' invention[1]. Capsule networks substituted the shift-invariant scalar features with an equivariant vectored-feature detector that brought forth this new branch on neural networks[1], [4]. The pooling layer's loss of information is replaced with the invention of the 'routing-by-agreement' algorithm.

Capsules are a collection of neurons that singly activate based on numerous aspects of an object, such as size, position, and hue[4]. Every capsule excites depending on a single significant aspect of the object. The output of these feature detectors is vectors encoded with each aspect's probabilities that the network knows[1], [4]. Routing by agreement improves the complexity of the network while the increment in size is tolerable. By combining the probabilities vector yielded by each capsule, it is possible to estimate the network's prediction reasonably. The initial algorithm that the capsule networks used is the Dynamic Routing Algorithm[1]. The Dynamic Routing Algorithm is as per Fig. 1[1].



```
Algorithm 1 Dynamic Routing Algorithm
procedure: ROUTING($\hat{\mathbf{u}}_{j|i}, r, l$)
  Initialisation :
  for all capsule $i$ in layer $l$ and capsule $j$ in layer $(l+1)$ :
  $b_{ij} \leftarrow 0$
  for $r$ iterations do
    for all capsule $i$ in layer $l$ : $c_i \leftarrow \mathrm{softmax}(b_i)$
    for all capsule $j$ in layer $(l+1)$ : $s_j \leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$
    for all capsule $j$ in layer $(l+1)$ : $v_j \leftarrow \mathrm{squash}(s_j)$
    for all capsule $i$ in layer $l$ and capsule $j$ in layer $(l+1)$ :
    $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.v_j$
  end for
  return $\mathbf{v_j}$
```

Fig. 1. Algorithm 1: Dynamic Routing Algorithm[1]

The squashing function used in Algorithm 1 is [1]:

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \qquad (1)$$

C. *Residual Capsule Network (RCN)*

RCN is cumulative of architectures of ResNet and Capsule Network[4]. The abstract of RCN architecture is a redundant intricate layer of ResNet CNN applied on the initial Capsule Networks section. These repetitive layers allow the network to create a deeper feature extraction process before passing it on to Capsules[4]. The Capsule Network consists of a light convolutional layer before the Capsules. This simple convolution is replaced with 8 layers of redundant ResNet convolutions, which improves the complexity tremendously and boosts the Capsule Network's ability, thereby increasing the model's accuracy[4].

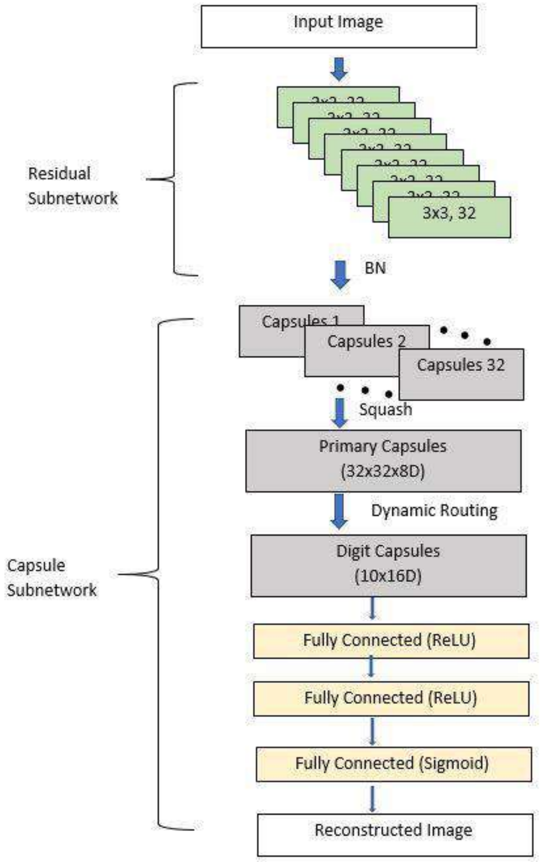


Fig. 2. Residual Capsule Network[4]

This modification is consistent with Capsule Network's view as no loss of information or embedded pooling layers. This RCN produced 84.16% accuracy with 11.86 M parameters for the CIFAR-10 dataset is significantly better than the 89.4% with the seven ensemble Capsule Network model with 101.5M parameters[4].

Considering the redundancy of the ResNet convolutions and the inefficiency of ResNet in comparison to ResNeXt[6], it is only fair to observe there is room for improvement. Although there has been an improvement in the routing-by-agreement algorithms, the RCN authors have not included any such changes to the later networks' capsules. Keeping in mind these variations, we propose changes in the initial convolutional layers and routing-by-agreement algorithm.

D. *ResNext and Cardinality*

ResNeXt is a neural network that brought improvements to ResNet. It is a homogeneous network with the ability to compress the conventional ResNet substantially[6]. "Cardinality" is the additional dimension that is provided to the network to move forward from the fundamental channels and kernels of convolutions in ResNet[6]. Cardinality delineates the extent of transformations. The implementation of Cardinality in ResNet architecture leads to ResNeXt.

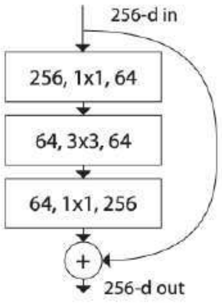


Fig. 3. ResNet convolution.

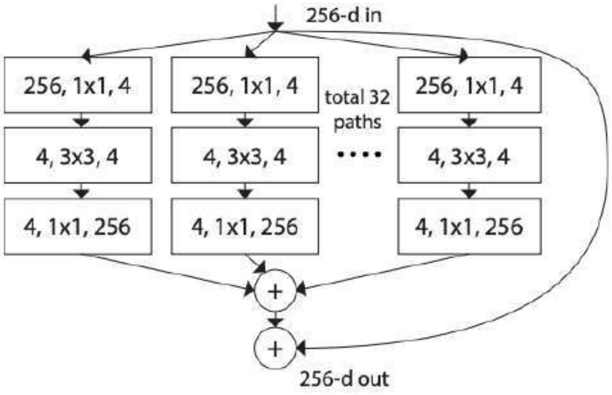
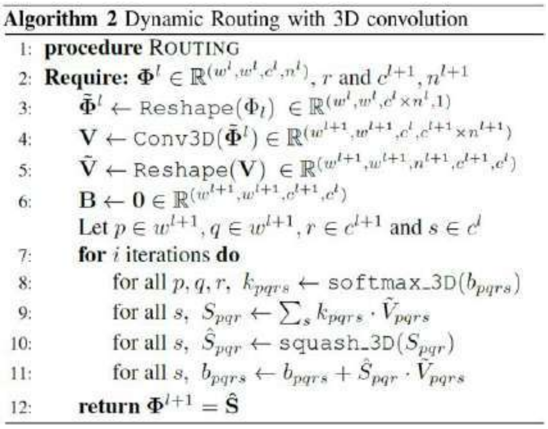


Fig. 4. ResNeXt convolution.

E. *DeepCaps*

Compared to various routing-by-agreement algorithms, the performance boost proved in the DeepCaps Network with the inclusion of Dynamic Routing with 3D convolution stands out. This significance to DeepCaps routing-by-agreement algorithm is due to the algorithm's simplicity in theory and its application. The following Fig. 5. contains the algorithm of Dynamic Routing with 3D convolution[7].



```
Algorithm 2 Dynamic Routing with 3D convolution
procedure ROUTING
Require: $\Phi^l \in \mathbb{R}^{(w^l,w^l,c^l,n^l)}$, $r$ and $c^{l+1}, n^{l+1}$
    $\tilde{\Phi}^l \leftarrow$ Reshape($\Phi_l$) $\in \mathbb{R}^{(w^l,w^l,c^l \times n^l,1)}$
    $\mathbf{V} \leftarrow$ Conv3D($\tilde{\Phi}^l$) $\in \mathbb{R}^{(w^{l+1},w^{l+1},c^l,c^{l+1} \times n^{l+1})}$
    $\tilde{\mathbf{V}} \leftarrow$ Reshape($\mathbf{V}$) $\in \mathbb{R}^{(w^{l+1},w^{l+1},n^{l+1},c^{l+1},c^l)}$
    $\mathbf{B} \leftarrow \mathbf{0} \in \mathbb{R}^{(w^{l+1},w^{l+1},c^{l+1},c^l)}$
    Let $p \in w^{l+1}, q \in w^{l+1}, r \in c^{l+1}$ and $s \in c^l$
    for $i$ iterations do
        for all $p, q, r$,  $k_{pqrs} \leftarrow$ softmax_3D($b_{pqrs}$)
        for all $s$,  $S_{pqr} \leftarrow \sum_s k_{pqrs} \cdot \tilde{V}_{pqrs}$
        for all $s$,  $\hat{S}_{pqr} \leftarrow$ squash_3D($S_{pqr}$)
        for all $s$,  $b_{pqrs} \leftarrow b_{pqrs} + \hat{S}_{pqr} \cdot \tilde{V}_{pqrs}$
    return $\Phi^{l+1} = \hat{\mathbf{S}}$
```

Fig. 5. Algorithm 2: Dynamic Routing with 3D convolution[7]

F. *3-Level Residual Capsule Network*

Inspired by the You Only Look Once (YOLO) network model, the RCN authors incorporated 3-staged feature detectors to bring a compound effect of various views of the image throughout the network to the RCN model[8]. This novel model, 3-Level Residual Capsule Network, is an exemplary work in utilizing Capsule Network's potential in deep stages. Nevertheless, an increase of complexity traces to an upsurge in the parameters, and as the network is a concatenation of three sets of 8-Layered ResNets. This modification increased accuracy to 86.42% while compensating for the increase in parameter via reducing intricacy in the reconstruction and maintaining a single layer output instead of RGB reconstruction[8].

From the 3-Level Residual Capsule Network, we note the contraction in the reconstruction network is deteriorating the performance. Also, the 3-Staged model is a significant factor of improved accuracy.

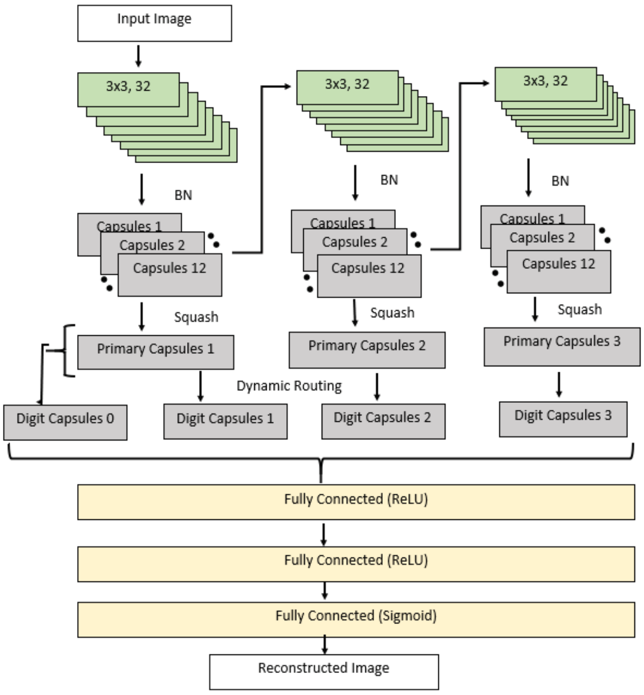


Fig. 6. 3-Level Residual Capsule Network [8]

III. PROPOSED RCNX: RESIDUAL CAPSULE NEXT

The following sections reveal the underlying enhancements included to RCN and model architecture of the proposed RCNX: Residual Capsule NeXt. Fig. 7. demonstrates the new architecture of RCNX.

A. *Primary Res Next Layers*

To accomplish superior accuracy, convolution layers of Capsule Network should be made capable of compound feature extraction. Even though Capsule Network performs based on routed capsules, the preliminary layers are convolutions dependent[8], [4]. Redundant layers of convolution in the initial stages of Capsule Network give us better multiplex feature extraction, and this provides the model with a good starting point.

The RCN considers eight layers of the residual network without any inclusion of bottleneck design or any different cardinality to improve the complexity and reduce parameters[4]. As mentioned previously, this paper proposes modifications in such features which is incomplete in the RCN network. We brought considerable compression and improvement to the RCN network. RCNX standouts with the inclusion of the new cardinality to RCN. As explained in the background, ResNeXt architecture reduced the number of parameters and, at the same time, improved the accuracy of the model's image classification properties.

With the missing structural advantages, RCN lacks the complexity to uncover a deeper understanding of the images, which can be the reason for not developing higher accuracy despite eight repetitive layers of ResNets. Since we include the required complexity using ResNeXt, we eliminated the redundant layers. ResNeXt structure embedded in the RCNX is with variable cardinality before reaching the capsules. While including the 3-level staged architecture, we also find the network improves learning of the image for different capsules to learn from various views of the image.

B. *3-Level ResNeXts before Capsule Network*

With the incorporation of various views, the proposed RCNX can learn intricate features and can do it fast. Allowing the training network to go through each level repetitively due to the intricate structure brings the best of RCNX with minimal effort.

In the 3-Level structure, the ResNeXt models with four and two cardinalities are included, with filters of the same size, i.e., 32. This varying cardinality brings variable total filter lengths for different capsules. Primary capsules receiving three different views to the image were structured to produce probability vectors of varying lengths, including flexibility in designing various features with various dimensions. We use the dimensions of 8, 24, 32, and 8 across four primary capsules.

C. *Efficient Capsules with '3D convolution-based dynamic routing.'*

Using 3D convolution-based dynamic routing, DeepCaps authors modified and improved routing by agreement algorithm for capsule networks[7]. This 3D convolution helps to trim the network in size. The convolution is considering that neighbouring neurons produce a similar pose, and this can be a cluster[7].

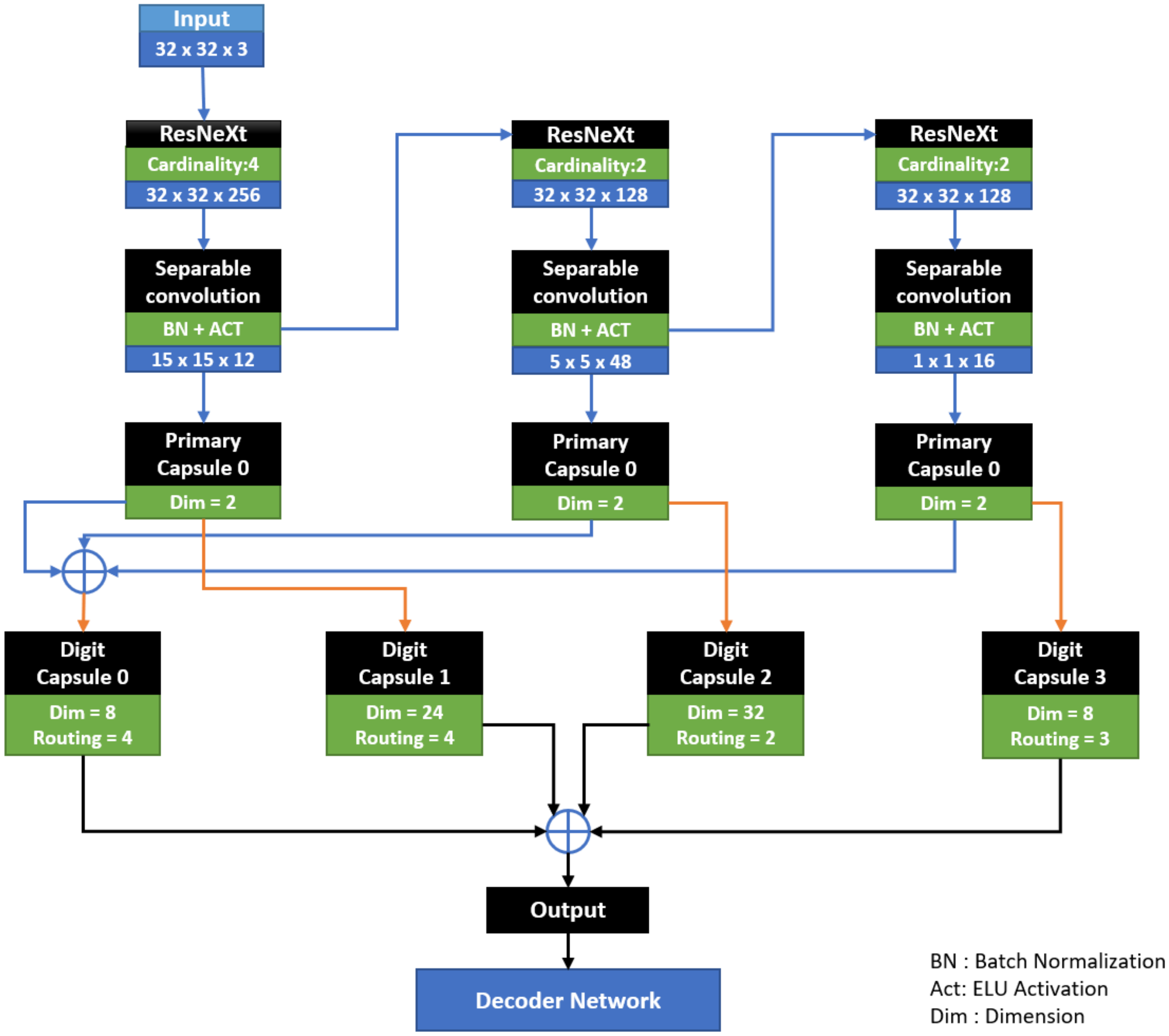


Fig. 7. Proposed RCNX: Residual Capsule NeXt Architecture

Using the 3D convolution-based dynamic routing algorithm, we tremendously reduced the repeated routing from layer ‘L’ to layer ‘L+1’. This reduction amount to a reduction factor of $c * (w^L w^{L+1})^2$ parameters in each capsule, where c represents the number of channels, and w^L represents the width of layer L[7].

The Capsule network of the RCNX is optimized to perform iterations that give the best performance and accuracy. We achieve optimization with Neural Network Intelligence (NNI)[9], and thus we use routings of 4, 4, 2, and 3 for each capsule network to be effective.

D. 3D reconstruction by decoder network

At the end of RCN, the reconstruction network does not comprise 3-dimensional inverse rendering but is limited to a 2-dimensional reconstruction. We include this 3-dimensional reconstruction as this is important for the Capsule Networks, in general, to learn quickly, and thereby creating the model to be more involved by integrating class independent decoder.

E. Elu activation function

Activation functions provide image classification models non-linearity that help them learn mapping functions, contributing to training and performance[10]. After most convolution layers, we use activation functions, and activation functions are mainly an integral part of the ResNeXt layers. The activation function in RCN is ReLU[8]. This ReLU activation is replaced by the Exponential Linear Unit (ELU) activation function as it is understood via repeated trials and NNI hyperparameter search that ELU outperforms ReLU in the case of proposed RCNX architecture[9], [10]. ELU also avoids some cons of ReLU activation. ELU activation excludes the problem of dead ReLU[10]. It also produces negative outputs, creates better optimization of biases and weights, and prevents saturation by avoiding non-zero gradient errors.

F. Summary

In summary, a 32x32x3 image travels through the proposed Residual Capsule NeXt as follows. The 3-channel image is convolved with primary ResNeXt convolutional layers, which extracts in-depth features, and creates channels of size $c * f$, where c is the cardinality and f is the filter length. These networks generate channels of 256 and 128. These pass through a separable convolution before proceeding to primary capsules, thereby creating a deeper model of RCNX architecture.

The separable convolution changes the image height and filter sizes to 15x15x12, 5x5x48, and 1x1x16, which provide the proposed RCNX with a wide range of views of the images. These views at three levels traverse through Primary capsules that have only 2-dimensional capsule vectors as output. These, in a manner of merged layers, get connected to one Digit Capsule, and others are individually connected to separate Digit Capsules.

The digit capsules are of varying sizes yet longer dimensions than Primary Capsules with optimized routing numbers of merges to form an output. Further, we decode output to form an inverse rendering effect by the decoder network to the corresponding input. This decoder network only activates during training and is removed while testing the proposed RCNX.

IV. Training Setup

Lenovo Think System compute node with Intel Xenon Gold processors, with 128GB RAM, and NVIDIA Tesla V100 is used in training and inference of RCNX[11]. CIFAR-10 Dataset is used for training and inference. The model is trained for 35 epochs, with a batch size of 32, Nadam optimizer, and LR decay of 0.9.

V. Results

The proposed RCNX model is trained and tested against the CIFAR-10 benchmark[12]. RCNX delivered an accuracy of 89.31% during the test, and the evaluation model size is 1.58M parameters.

Table I

Performance of various network models on CIFAR10

Model Name	***No. of Parameters***	***Test Accuracy***
Proposed RCNX	**1.58 M**	**89.31%**
Residual Capsule Network V2	1.95 M	85.12%
Baseline Residual Capsule Network	11.8 M	84.16%
3-Level Residual Capsule Network	10.8 M	86.42%
CapsNet	101 M	89.40%
DC Net	11.8 M	82.63%
DC Net ++	13.4 M	89.71%
MobileNet V3	1.83 M	88.93%

Comparing results of the proposed RCNX with results of other capsule network models like baseline Residual Capsule Network, Capsule Network, DCNET, DCNET++, Residual Capsule Network V2, and 3-Level Residual Capsule Network as per Table I[1], [4], [8], [12], we can easily conclude that the proposed RCNX is producing unparalled results.

CIFAR-10 benchmark dataset contains 60,000 images of 10 classes with 6000 pictures per class. 10,000 images in these are for inference and remainder for training the neural network models[12].

VI. Conclusion

Here, a new architecture, RCNX: Residual Capsule NeXt, is introduced. Using ResNeXt convolutions, 3D convolution-based dynamically routed Capsules, architecture following 3-Level RCN, full image reconstruction, ELU activation function, and hyperparameters tuned with NNI, and we achieved a model that is efficient with an accuracy of 89.31% and a reduced model size of 1.58 M parameters when tested on CIFAR-10. Thus, the proposed RCNX is better than embedded models like MobileNetV3, and RCNX is the first capsule network to achieve this in image classification tasks.

Acknowledgement

The authors would like to thank Lilly Endowment Inc for their support to Indiana University Pervasive Technology Institute.

References

[1] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA., p. 11 [Online]. Available: https://papers.nips.cc/paper/2017/file/2cad8fa47bbef28 2badbb8de5374b894-Paper.pdf. [Accessed: 10-Jan-2021]

[2] V. Dragoi, “Visual Processing: Cortical Pathways (Section 2, Chapter 15) Neuroscience Online: An Electronic Textbook for the Neurosciences | Department of Neurobiology and Anatomy - The University of Texas Medical School at Houston,” 07-Oct-2020. [Online]. Available: https://nba.uth.tmc.edu/neuroscience/m/s2/chapter15.html. [Accessed: 10-Jan-2021]

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[4] S. B. S. Bhamidi and M. El-Sharkawy, “Residual Capsule Network,” in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2019, pp. 0557–0560, doi: 10.1109/UEMCON47517.2019.8993019.

[5] G. W. Lindsay, “Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future,” p. 27.

[6] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634 [Online]. Available: http://ieeexplore.ieee.org/document/8100117/. [Accessed: 09-Mar-2021]

[7] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne, and R. Rodrigo, “DeepCaps: Going Deeper With Capsule Networks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR), 2019, pp. 10717–10725, doi: 10.1109/CVPR.2019.01098.

[8] S. B. S. Bhamidi and M. El-Sharkawy, "3-Level Residual Capsule Network for Complex Datasets," p. 4.

[9] *Neural Network Intelligence (NNI)*. Microsoft [Online]. Available: https://github.com/microsoft/nni

[10] Z. Qiumei, T. Dan, and W. Fenghua, "Improved Convolutional Neural Network Based on Fast Exponentially Linear Unit Activation Function," *IEEE Access*, vol. 7, pp. 151359–151367, 2019, doi: 10.1109/ACCESS.2019.2948112.

[11] C. A. Stewart, V. Welch, B. Plale, G. Fox, M. Pierce, and T. Sterling, "Indiana University Pervasive Technology Institute," Sep. 2017, doi: 10.5967/K8G44NGB. [Online]. Available: https://scholarworks.iu.edu/dspace/handle/2022/21675. [Accessed: 11-Jan-2021]

[12] R. C. Calik and M. F. Demirci, "Cifar-10 Image Classification with Convolutional Neural Networks for Embedded Systems," in *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Aqaba, 2018, pp. 1–2, doi: 10.1109/AICCSA.2018.8612873 [Online]. Available: https://ieeexplore.ieee.org/document/8612873/. [Accessed: 11-Jan-2021]

Image Classification with CondenseNeXt for ARM-Based Computing Platforms

Priyank Kalgaonkar
Department of Electrical and Computer Engineering
Purdue School of Engineering and Technology
Indianapolis, Indiana 46202, USA.
pkalgaon@purdue.edu

Mohamed El-Sharkawy
Department of Electrical and Computer Engineering
Purdue School of Engineering and Technology
Indianapolis, Indiana 46202, USA.
melshark@purdue.edu

***Abstract*—In this paper, we demonstrate the implementation of our ultra-efficient deep convolutional neural network architecture: CondenseNeXt on NXP BlueBox, an autonomous driving development platform developed for self-driving vehicles. We show that CondenseNeXt is remarkably efficient in terms of FLOPs, designed for ARM-based embedded computing platforms with limited computational resources and can perform image classification without the need of a CUDA enabled GPU. CondenseNeXt utilizes the state-of-the-art depthwise separable convolution and model compression techniques to achieve a remarkable computational efficiency.**

Extensive analyses are conducted on CIFAR-10, CIFAR-100 and ImageNet datasets to verify the performance of CondenseNeXt Convolutional Neural Network (CNN) architecture. It achieves state-of-the-art image classification performance on three benchmark datasets including CIFAR-10 (4.79% top-1 error), CIFAR-100 (21.98% top-1 error) and ImageNet (7.91% single model, single crop top-5 error). CondenseNeXt achieves final trained model size improvement of 2.9+ MB and up to 59.98% reduction in forward FLOPs compared to CondenseNet and can perform image classification on ARM-Based computing platforms without needing a CUDA enabled GPU support, with outstanding efficiency.

***Index Terms*—CondenseNeXt, Convolutional Neural Network, Computer Vision, Image Classification, NXP BlueBox, ARM, Embedded Systems, PyTorch, CIFAR-10, CIFAR-100, ImageNet.**

I. Introduction

ARM processors are widely used in electronic devices such as smartphones and tablets as well as in embedded computing platforms such as the NXP BlueBox, Nvidia Jetson and Raspberry Pi for computer vision purposes. ARM is RISC (Reduced Instruction Set Computing) based architecture for computer processors which results in low costs, minimal power consumption, and lower heat generation compared to its competitor: CISC (Complex Instruction Set Computing) architecture based processors such as the Intel x86 processor family. As of 2021, over 180 billion ARM-based chips have been manufactured and shipped by Arm and its partners around the globe which makes it the most popular choice of Instruction Set Architecture (ISA) in the world [1].

The roots of ARM processors trace back to December 1981 when the first widely successful design, BBC Micro (British Broadcasting Corporation Microcomputer System), was introduced by Acorn Computers [2]. Due to the use of DRAM (Dynamic Random Access Memory) in its design, it outperformed nearly twice as that of Apple II, an 8-bit personal computer, which was world's first successfully mass-produced publicly available computer designed by Steve Wozniak, Steve Jobs and Rod Holt in June 1977 [3].

Fast forwarding to the 21st century, due to constant advances in computing and VLSI technology, ARM-based chips are found in nearly 60% of all mobile devices and computing platforms produced today. With processor performance doubling approximately every two years with a focus on parallel computing technologies such as multi-core processors, computer vision researchers can now implement sophisticated neural network algorithms to perform complex computations for OpenCV applications without a requiring a GPU support.

Convolutional Neural Networks (CNN), a class of Deep Neural Networks (DNN) first introduced by Alexey G. Ivakhnenko and V. G. Lapa in 1967 [4], have been gaining popularity in recent years as researchers focus on creating more advanced intelligent systems. CNNs are popularly used in machine (computer) vision applications such as image classification, image segmentation, object detection, etc. However, implementing a CNN on embedded systems with constrained computational resources for applications such as autonomous cars, robotics and unmanned aerial vehicle (UAV), commonly known as a drone, is a challenging task. In this paper, we present image classification performance results of CondenseNeXt CNN on NXP BlueBox, an ARM-based embedded computing platform for automotive applications.

II. Related Work

Following work has contributed to the research and implementation results presented within this paper:

CondenseNeXt: An ultra-efficient deep convolutional neural network for embedded systems, introduced by P. Kalgaonkar and M. El-Sharkawy in January 2021 [5] has been utilized to train and evaluate image classification performance on three benchmarking datasets: CIFAR-10, CIFAR-100 and ImageNet.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

III. NXP BlueBox 2.0

The BlueBox 2 family developed and manufactured by NXP Semiconductors N.V, a Dutch-American semiconductor manufacturer with headquarters in Eindhoven, Netherlands and Austin, United States of America, is a Automotive High Performance Compute (AHPC) platform that provides essential performance and reliability for engineers to develop sensor fusion, automated drive and motion planning applications along with functional safety, vision acceleration and automotive interfaces for self-driving (autonomous) vehicles.

NXP BlueBox Gen1 was first introduced in May 2016 at the 2016 NXP FTF Technology Forum held in Austin, Texas, USA. This opened avenues to a host of autonomous and sensor fusion applications. Shortly after, NXP introduced BlueBox Gen2 (BlueBox 2.0), a significant improvement over Gen1, incorporating three new processors: S32V234 ARM-based automotive computer vision processor, LS2084A high performance ARM-based compute processor and S32R274 ASIL-D RADAR microcontroller.

S32V234: The S32V234 automotive computer vision processor comprises of a quad core ARM Cortex-A53 CPU running at 1.0 GHz paired with a ARM Cortex-M4 functional safety core which utilizes the ARMv8-A 64-bit instruction set developed by ARM Holdings' Cambridge design centre. It has a 4MB internal SRAM in addition to a 32bit LPDDR3 memory controller for external memory support. It is an on-chip Image Signal Processor (ISP) designed to meet ASIL-B/C automotive safety standards and optimized for obtaining maximum performance per watt efficiency.

LS2084A: The LS2084A high performance compute processor comprises of an octa core ARM Cortex-A72 CPU running at 1.8 GHz which utilizes the ARMv8-A 64-bit instruction set developed by ARM Holdings' Austin design centre. It has two 72 bytes DDR4 RAMs running at up to 28.8GB/s memory bandwidth. The LS2 provides software compatibility with next generation LayerScape LX2 family and offers AEC Q100 Grade 3 reliability with 15 years product longevity.

S32R274: The S32R274 radar micro-controller comprises of a dual core Freescale PowerPC e200z7 32-bit CPU running at 240 MHz and a dual core Freescale PowerPC e200z4 32-bit CPU running at 120 MHz with an additional checker core. It has a 2 MB Flash and 1.5 MB SRAM for radar application storage, message buffering and radar data stream handling. The S32R processor is optimized for on-chip radar signal processing to maximize performance per watt efficiency. It has been designed by NXP to meet the ASIL-D automotive applications standards.

IV. RTMaps Remote Studio Software

Figure 1. NXP BlueBox 2.0 ARM-based Automotive High Performance Compute (AHPC) embedded development platform. It delivers necessary prerequisites to help develop high-performance computing systems, analyze driving environments, assess risk factors, and then direct the car's behavior. BlueBox 2.0 also supports OpenCV applications using an external camera for real-time image classification object detection and image segmentation.

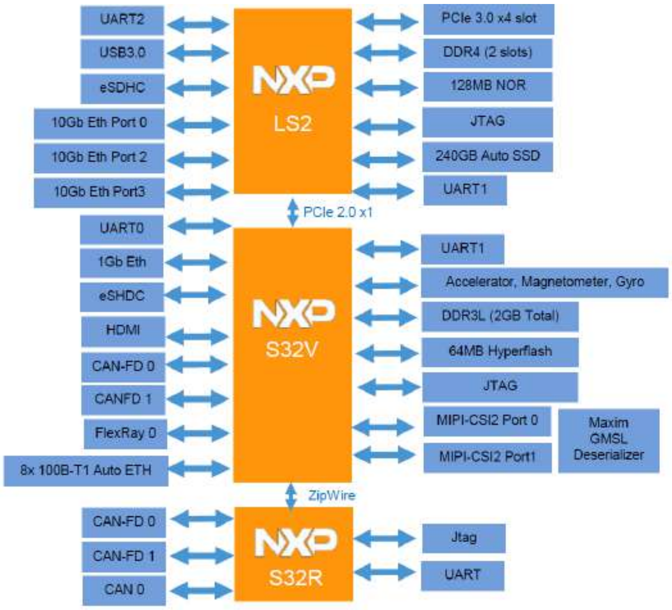


Figure 2. High-Level View of NXP Bluebox 2.0 Gen2 Architecture [6]. S32V processor utilizes two CAN-FD (Flexible Data Rate) with enhanced payload and data rate, PCIe, Ethernet, FlexRay, Zipwire, one SAR-ADCs, four SPI and one SD card connectivity. LS2 processor utilizes two DUART, four I2C, SPIO, GPIO and two USB 3.0 interfaces. S32R processor utilizes JTAG, UART, three FlexCAN and Zipwire to connect to a radar ASIC.

RTMaps (Real-Time Multisensor applications) developed by Intempora is a powerful GUI software that aids in development of applications for advanced driver assistance systems, autonomous driving and robotics. It helps in capturing, processing and viewing data from multiple sensors and offers a multi-modular development and run-time environment for ARM-based computing platforms such as the NXP BlueBox 2.0. This data can also be reviewed and play-backed at a later time for offline development and testing purposes.

RTMaps Remote Studio supports PyTorch, an open-source machine learning library based upon the Torch library, widely used for real-time computer vision (OpenCV) development. Algorithms for OpenCV can be developed using Python scripting language and by the means of block diagrams. It also facilitates the development of algorithms directly on to any supported embedded system without having to connect external user interfacing peripheral devices.

V. CondenseNeXt

CondenseNeXt is an ultra-efficient deep convolutional neural network architecture designed for embedded systems introduced by P. Kalgaonkar and M. El-Sharkawy in January 2021. CondenseNeXt refers to the *next* dimension of cardinality. In this section, we describe in detail the architecture of this neural network that has been utilized to train and evaluate image classification performance on three benchmarking datasets: CIFAR-10, CIFAR-100 and ImageNet.

A. Convolution Layers

One of the main goals of CondenseNeXt is to reduce the amount of computational resources required to train the network from scratch and for real-time inference on embedded systems with limited computational resources. Following state-of-the-art technique has been incorporated into the design of this CNN:

- Depthwise convolution layer: It acts like a filtering layer where convolution to a single input channel is applied separately instead of applying it to all input channels. Assume there is an input data of size $A \times A \times C$ and filters (kernels) K of size $F \times F \times 1$. If there are C number of channels in the input data, the output will be of size $B \times B \times C$. At this point, the spatial dimensions have shrunk. However, the depth C has remained constant and the cost of this operation will be $B^2 \times F^2 \times C$.
- Pointwise convolution layer: It acts like a combining layer where a linear combination is carried out for each of these layers. At this stage, a 1×1 convolution is applied to C number of channels in the input data. Thus, the size of filter for this operation will be $1 \times 1 \times C$ and size of the output will be $B \times B \times D$ for D such filters.

Assume a standard convolutional filter K of size $F \times F \times A \times B$ where A is the number of input channels and B is the number of output channels with an input feature map A of size $D_x \times D_x \times A$ that produces an output feature map Z of size $D_y \times D_y \times B$ can be mathematically represented as follows:

$$Z_{k,l,n} = \sum_{i,j,m} k_{i,j,m,n} \cdot A_{k+i-1,l+j-1,m} \tag{1}$$

In case of a depthwise separable convolution, (1) is factorized into two stages: the first stage applies a 3×3 depthwise convolution $\hat{K}$ with one filter for every input channel:

$$\hat{Z}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \cdot A_{k+i-1,l+j-1,m} \tag{2}$$

Consequently, in the second stage, a 1×1 pointwise convolution $\tilde{K}$ is applied to carry out linear combination and combine the outputs of depthwise convolution from previous stage as follows:

$$Z_{k,l,n} = \sum_{m} \tilde{K}_{m,n} \cdot \hat{Z}_{k-1,l-1,m} \tag{3}$$

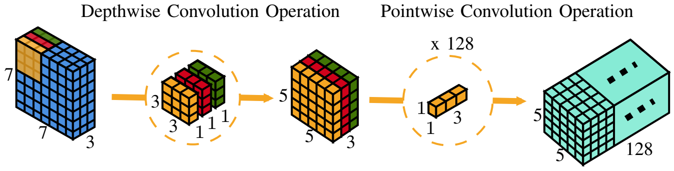

Depthwise Convolution Operation
Pointwise Convolution Operation
x 128
7
7
3
3
3
1
1
5
5
3
1
1
3
5
5
128

Figure 3. A 3D illustration of the overall process of depthwise separable convolution. An image is transformed 128 times whereas an image is transformed by depthwise separable convolution only once and then this transformed image is stretched to 128 channels which allows the neural network to process more data while consuming fewer FLOPs (Floating Point Operations).

This methodology splits a kernel into two discrete filters for filtering and combining stages as shown in Figure 3 above, which results in reduction of computational resources required to train the network from scratch as well for real-time inference.

A widely used model compression technique is also implemented into the design of this CNN where a further significant impact, both in computational efficiency at training time and on the final trained model size is seen.

B. Model Compression

A widely popular model compression technique called Group-wise Pruning is implemented to make CondenseNeXt neural network computationally more efficient by discarding redundant elements without influencing the overall performance of the network.

Group-wise Pruning: The purpose of group-wise pruning is to remove trivial filters for every group g during the training process which is based on the L_1-Normalization of $A^{g_{ij}}$ where for every group g, a is the input and z is the output. A pruning hyper-parameter p is established and set to 4 which allows the network to decide the number of filters to remove before the first stage of depthwise separable convolution. A class balanced focal loss function [7] is also added to assist and ease the effect of this pruning process.

Consider a group convolution comprised of G groups of size $F \times F \times C_A \times C_B$ where $C_A = \frac{A}{G}$ and $C_B = \frac{B}{G}$. The total number of trivial filters that will be pruned before the first stage of depthwise separable convolution is mathematically represented as follows:

$$G \cdot C_x = A \cdot C - p \cdot A \tag{4}$$

Cardinality: A new dimension to the network called *Cardinality* denoted by C is incorporated into the design of CondenseNeXt neural network in addition to the existing width and depth dimensions so that loss in accuracy during the pruning process is reduced. Experiments prove that increasing cardinality is a more efficacious way of accruing accuracy than going deeper or wider, especially when width and depth starts to provide diminishing returns [8].

C. Activation Function

In deep neural networks, activation functions determine the output of a neuron at particular input(s) by restricting the amplitude of the output. It aids in neural network's understanding and learning process of complex patterns of the input data. Furthermore, non-linear activation functions such as ReLU6 (Rectified Linear Units capped at 6) enable neural networks to perform complex computations using fewer neurons [9].

CondenseNeXt applies ReLU6 activation function in addition to Batch Normalization technique prior to each convolutional layer. In ReLU6, units are capped at 6 to promote an earlier learning of sparse features and to prevent a sudden blowup of positive gradients to infinity. ReLU6 activation function is defined mathematically as follows:

$$f(x) = min(max(0, x), 6) \tag{5}$$

VI. Cyberinfrastructure

A. Training Infrastructure

- Intel Xeon Gold 6126 12-core CPU with 128 GB RAM.
- NVIDIA Tesla V100 GPU.
- CUDA Toolkit 10.1.243.
- PyTorch version 1.1.0.
- Python version 3.7.9.

This cyberinfrastructure for training is provided and managed by the Research Technologies division at the Indiana University which supported our work in part by Shared University Research grants from IBM Inc. to Indiana University and Lilly Endowment Inc. through its support for the Indiana University Pervasive Technology Institute [10].

B. Testing Infrastructure

- NXP BlueBox 2.0 ARM-based autonomous embedded development platform.
- Intempora RTMaps Remote Studio version 4.8.0.
- CIFAR-10, CIFAR-100 and ImageNet Datasets.
- PyTorch version 1.1.0.
- Python version 3.7.9.

VII. Experiment and Results

Training results presented in this report are based on the evaluation of image classification performance of CondenseNeXt CNN on three benchmarking datasets: CIFAR-10, CIFAR-100 and ImageNet. CondenseNeXt was designed and developed in PyTorch framework and trained on NVIDIA's Tesla V100 GPU with standard data augmentation scheme [11], Nesterov Momentum Weight of 0.9, Stochastic Gradient Descent (SGD), cosine shape learning rate and dropout rate of 0.1 for all three datasets discussed in this section.

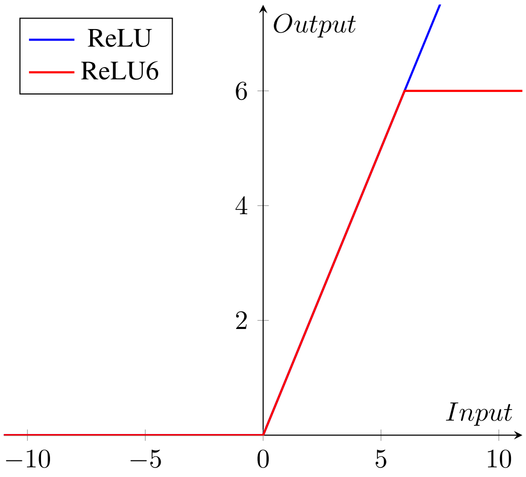


Figure 4. Difference between ReLU and ReLU6 activation functions.

A. CIFAR-10 Classification

CIFAR-10 dataset [9], [12] was first introduced by Alex Krizhevsky in [13]. It is one of the most widely used datasets for evaluating a CNN in the field of deep learning research. There are 60,000 RGB images of 10 different classes of size 32×32 pixels divided into two sets of 50,000 for training and 10,000 for testing.

CondenseNeXt was trained with a single crop of inputs on CIFAR-10 dataset for 200 epochs, batch size of 64 and features k of 8-16-32. Using RTMaps Remote Studio, an image classification script was developed using Python scripting language and evaluated on NXP BlueBox for single image classification analysis. Table I provides a comparison of performance between CondenseNet and CondenseNeXt CNN in terms of FLOPs, parameters, and Top-1 and Top-5 error rates. Figure 5 provides a screenshot of the RTMaps console.

B. CIFAR-100 Classification

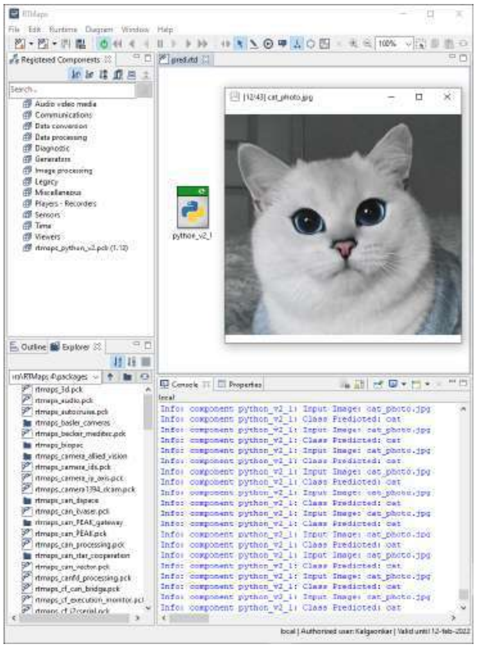
CIFAR-100 dataset was also first introduced by Alex Krizhevsky in [13] along side CIFAR-10 dataset. It is also one of the many popular choices of datasets in the field of deep learning research. Just like CIFAR-10 dataset, there are 60,000 RGB images in total. However, it has 100 different classes, where each class contains 600 images of size 32×32 pixels divided into two sets of 50,000 for training and 10,000 for testing. CIFAR-100 classes are mutually exclusive of CIFAR-10 classes. For example, CIFAR-100's baby, chimpanzee and rocket classes are not part of the CIFAR-10 classes.

CondenseNeXt was trained with a single crop of inputs on CIFAR-100 dataset for 600 epochs, batch size of 64 and features k of 8-16-32. Using RTMaps Remote Studio, an image classification script was developed using Python scripting language and evaluated on NXP BlueBox for single image classification analysis. Table I provides a comparison of performance between CondenseNet and CondenseNeXt CNN in terms of FLOPs, parameters, and Top-1 and Top-5 error rates. Figure 6 provides a screenshot of the RTMaps console.

Table I
COMPARISON OF PERFORMANCE

Dataset	CNN Architecture	FLOPs (in millions)	Parameters (in millions)	Top-1 % Error	Top-5 % Error
CIFAR-10	CondenseNet	65.81	0.52	5.31	0.24
	CondenseNeXt	**26.35**	**0.18**	**4.79**	**0.15**
CIFAR-100	CondenseNet	65.85	0.55	23.35	6.56
	CondenseNeXt	**26.38**	**0.22**	**21.98**	**6.29**
ImageNet	CondenseNet	529.36	4.81	26.2	8.30
	CondenseNeXt	**273.16**	**3.07**	**25.8**	**7.91**

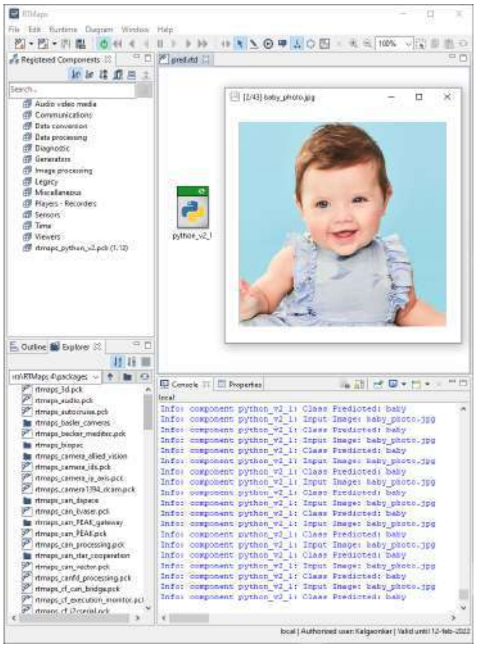
Table I provides a comparison between CondenseNet (the baseline architecture) vs. CondenseNeXt (our ultra-efficient deep neural network architecture) in terms of performance each utilizing the training setup and infrastructure as outlined in section 6 and 7 in this paper.



```
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
Info: component python_v2_1: Input Image: cat_photo.jpg
Info: component python_v2_1: Class Predicted: cat
```

[12/43] cat_photo.jpg
python_v2_1
Console
Properties
local | Authorised user: Kalgaonkar | Valid until 12-feb-2022

Figure 5. Evaluation of CondenseNeXt on CIFAR-10 dataset when deployed on NXP BlueBox 2.0 using RTMaps Remote Studio 4.8.0 for classifying an image of a cat and outputting the predicted class in RTMaps console.



```
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
Info: component python_v2_1: Input Image: baby_photo.jpg
Info: component python_v2_1: Class Predicted: baby
```

[2/43] baby_photo.jpg
python_v2_1
Console
Properties
local | Authorised user: Kalgaonkar | Valid until 12-feb-2022

Figure 6. Evaluation of CondenseNeXt on CIFAR-100 dataset when deployed on NXP BlueBox 2.0 using RTMaps Remote Studio 4.8.0 for classifying an image of a baby and outputting the predicted class in RTMaps console.

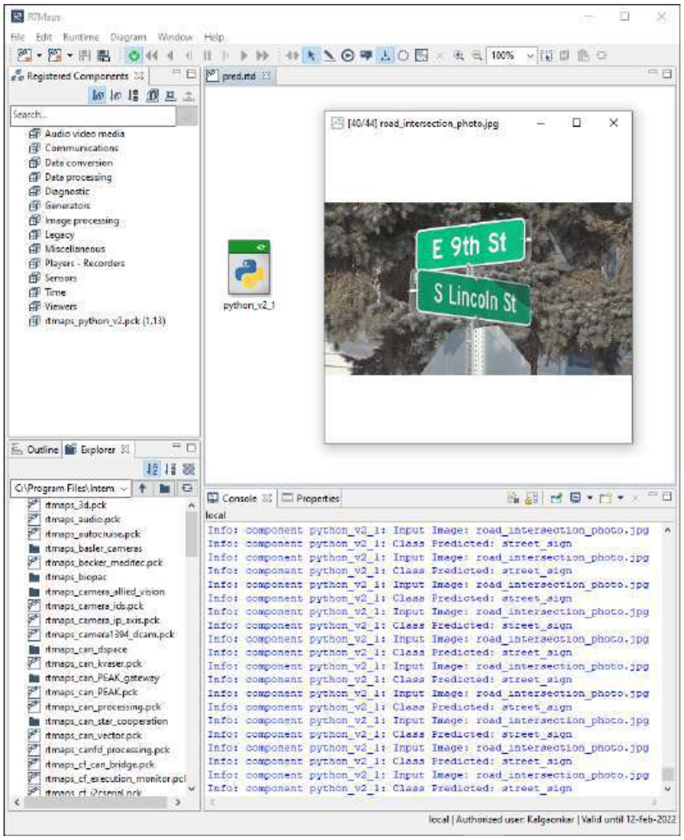
C. ImageNet Classification

ImageNet was introduced by an AI researcher Dr. Fei-Fei Li along with a team of researchers at a 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in Florida [14]. This dataset is built according to the WordNet hierarchy where each node in the hierarchy corresponds to over five hundred images. In total, there are over 14 million images in this dataset that have been hand-annotated and labelled by the team.

CondenseNeXt was trained with a single crop of inputs on the entire ImageNet dataset for 120 epochs with a Group Lasso rate of 0.00001, batch size of 256, features of 8-16-32-64-128 and four Nvidia V100 GPUs using Data Parallelism technique. An image classification script was developed in RTMaps Remote Studio and evaluated on NXP BlueBox for single image classification analysis. Table I provides a comparison of performance between CondenseNet and CondenseNeXt CNN in terms of FLOPs, parameters, and Top-1 and Top-5 error rates. Figure 7 provides a screenshot of the RTMaps console.

VIII. CONCLUSION

In this paper, we demonstrate the performance of CondenseNeXt CNN which is an ultra-efficient deep convolutional neural network architecture for ARM-based embedded computing platforms without CUDA enabled GPU(s). Extensive training from scratch and analysis have been conducted on three benchmarking datasets: CIFAR-10, CIFAR-100 and ImageNet. It achieves state-of-the-art image classification performance on CIFAR-10 dataset with a 4.79% Top-1 error rate, on CIFAR-100 dataset with a 21.98% Top-1 error rate and ImageNet dataset with a 7.91% single model and single crop Top-5 error rate. Our experiments on NXP's BlueBox further validate the effective use Depthwise Separable Convolutional layers and Model Compression techniques implemented to discard inconsequential elements and to reduce FLOPs without affecting overall performance of the neural network. In the future, we will explore different applications with CondenseNeXt such as image segmentation and object detection to better exploit different opportunities for OpenCV applications.



```
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
Info: component python_v2_1: Input Image: road_intersection_photo.jpg
Info: component python_v2_1: Class Predicted: street_sign
```

local | Authorized user: Kalgaonkar | Valid until 12-Feb-2022

Figure 7. Evaluation of CondenseNeXt on ImageNet dataset when deployed on NXP BlueBox 2.0 using RTMaps Remote Studio version 4.8.0 for classifying an image of a street sign and outputting the predicted class in the RTMaps console.

Acknowledgment

The authors would like to acknowledge the support of Indiana University Pervasive Technology Institute for providing supercomputing and storage resources that have contributed to all research results reported within this paper.

References

[1] A. Ltd, "The Arm ecosystem ships a record 6.7 billion Arm-based chips in a single quarter," Arm — The Architecture for the Digital World. https://www.arm.com/company/news/2021/02/arm-ecosystem-ships-record-6-billion-arm-based-chips-in-a-single-quarter (accessed Mar. 01, 2021).

[2] "History of ARM: from Acorn to Apple," *The Telegraph*, Jan. 06, 2011. https://www.telegraph.co.uk/finance/newsbysector/epic/arm/8243162/History-of-ARM-from-Acorn-to-Apple.html (accessed Feb. 25, 2021).

[3] "Total share: 30 years of personal computer market share figures — Ars Technica." https://arstechnica.com/features/2005/12/total-share/3/ (accessed Mar. 02, 2021).

[4] A. G. Ivakhnenko and V. G. Lapa, Cybernetics and Forecasting Techniques. American Elsevier Publishing Company, 1967.

[5] P. Kalgaonkar and M. El-Sharkawy, "CondenseNeXt: An Ultra-Efficient Deep Neural Network for Embedded Systems," in *2021 IEEE 11th Annual Conputing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, pp. 559–563.

[6] C. Cureton and M. Douglas, "Bluebox Deep Dive – NXP's AD Processing Platform," p. 28, Jun. 2019.

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," arXiv:1901.05555 [cs], Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.05555.

[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *arXiv:1611.05431 [cs]*, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1611.05431.

[9] A. Krizhevsky, "Convolutional Deep Belief Networks on CIFAR-10," p. 9.

[10] C. A. Stewart, V. Welch, B. Plale, G. Fox, M. Pierce, and T. Sterling, "Indiana University Pervasive Technology Institute," Sep. 2017, doi: 10.5967/K8G44NGB.

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[12] R. Doon, T. Rawat, and S. Gautam, "Cifar-10 Classification using Deep Convolutional Neural Network," Nov. 2018, pp. 1–5, doi: 10.1109/PUNECON.2018.8745428.

[13] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," p. 60.

[14] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

IEMTRONICS 2021 (International IOT, Electronics and Mechatronics Conference)

Deployment of Compressed MobileNet V3 on iMX RT 1060

Kavyashree Prasad S P
IOT Collaboratory IUPUI
Department of Electrical and Computer Engineering
Purdue School of Engineering and Technology
Indianapolis, USA
kshalini@purdue.edu

Mohamed El -Sharkawy
IOT Collaboratory IUPUI Department of Electrical and Computer Engineering
Purdue School of Engineering and Technology
Indianapolis, USA
melshark@iupui.edu

***Abstract*—Deep Neural Networks (DNN) are prominent in most applications today. From self-driving cars, sentiment analysis, surveillance systems, and robotics, they have been used extensively. Among DNNs, Convolutional Neural Networks (CNN) have achieved massive success in computer vision applications as the human visual system inspires their architecture. However, striving to achieve higher accuracies, CNN complexity, parameters, and layers were increased, which led to a drastic surge in their size, making their deployment challenging. Over the years, many researchers have proposed various techniques to alleviate this issue—one of them being Design Space Exploration (DSE) to minimize size and computation with little compromise to accuracy. MobileNet V3 is one such architecture designed to achieve good accuracy while being mindful of resources. It produces an accuracy of 88.93% on CIFAR-10 with a size of 15.3MB. This paper further reduces its size to 2.3MB while boosting its accuracy to 89.13% using DSE techniques. It is then deployed into NXP's i.MX RT1060 Advanced Driver Assistance System (ADAS) platform.**

Keywords—MobileNet V3, Convolution Neural Networks, Depthwise Pointwise Depthwise blocks, Compressed MobileNet V3, CIFAR-10, Design space exploration, TensorFlow, i.MX RT 1060.

I. Introduction

Convolutional neural networks have shown commendable performance in various computer vision tasks due to their compelling ability to use multiple feature description stages to grasp representations from images. They first came into the spotlight through the work of LeCun et al. 1989. Since the victory of AlexNet in the ImageNet challenge in 2012, they have gained high popularity [1]. The appealing factor of CNNs is their intelligence to extract spatial and temporal content from crude data. The design of CNN incorporates many convolutional layers, subsampling units, and nonlinear activation functions. Convolution operation guides extraction of valuable features from locally connected information points. Its output is then passed to nonlinear activations that produce diverse activations for various responses and encourage semantic contrasts in pictures. Subsampling is used in CNNs to make it invariant to the location of features. Subsequently, in this way, CNN learns pictures without the need for human involvement in feature extraction. The human visual cortex profoundly propels the building plan of CNNs. In the course of learning, CNN alters weights employing a backpropagation calculation. This ability to move towards the target is comparable to the brain's capacity to memorize based on responses. CNN's multi-layered structure helps gather low, mid, and high-level features, with high-level features being an aggregate of mid and low-level features. This dynamic learning ability of CNN emulates the neocortex in the human cerebrum and is responsible for its pervasiveness.

Over the years, there have been many advances in their architectures, activation functions, regularization, and parameter tuning. In a race to achieve higher accuracy, CNN model complexity and parameters have escalated [2] [3] [4] [5]. This has led to increased demand for resources needed for their storage and computation. Many small-sized architectures were proposed to ease this problem, such as MobileNet [6], SqueezeNet [7], ShuffleNet [8], and so on.

There are various advantages to using small architectures. Firstly, they are more suitable for embedded resource-constrained applications. Due to their size, computations can be performed in place for tasks such as image recognition, semantic segmentation, etc., rather than sending it to the cloud. It reduces latency and assures the privacy of data. In autonomous driving, companies make updates to the model and load it into customer vehicles from their servers [9]. Small-size CNNs make this update more convenient. Hence, to benefit from these advantages, many techniques were developed. Some of them include knowledge distillation [10] [11], pruning, and network quantization [12][13] , low rank and sparse decomposition [14], and developing new innovative architectures[8] [15][16].In knowledge distillation, a small model learns from a large model using a teacher-student approach. In pruning, weights that are insignificant to network performance are zeroed out based on a criterion [17] [18], and in network quantization, filter kernels and weights in fully connected layers are quantized. This quantization can be achieved by various methods such as k-means, Huffman coding, etc. Sparse decomposition and low-rank approximations achieve compression by reducing the parameter dimension of the network. This paper accomplishes a similar purpose by reducing MobileNet V3 small by making architectural modifications and changing the baseline model's activation functions.

Baseline architecture is demonstrated in section 2. Section 3 illustrates changes made to MobileNet V3 small to produce CMV3. Training setup is detailed in section 4. Section 5 has implementation details. Results and conclusion are mentioned in sections 6 and 7, respectively.

II. Prior Work

A. *Baseline Architecture*

MobileNet V3 is the latest variant of MobileNets. It was designed using a platform-aware network architecture search, and net adapt algorithm. MobileNet V3 small and MobileNet V3 large are two forms of this model developed to serve

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

different resource constraints [15]. MobileNet V3 large has lesser latency than MobileNet V2 while being 3.2% more accurate on the ImageNet dataset [19].

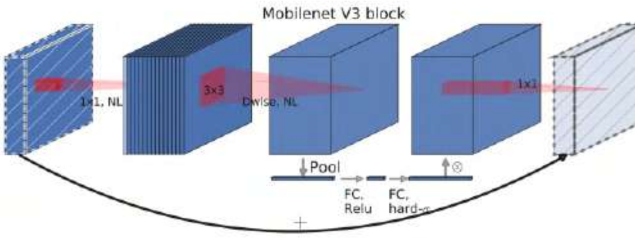


Fig. 1. MobileNet V3 Block [15].

MobileNet V3 encompasses the best practices from MobileNet V2 and Squeeze and excitation networks[20] .It is a combination of inverted residual bottlenecks and squeeze and excitation blocks. These SE blocks are added to improve networks' representational power by suppressing neurons that do not contribute to performance and enhancing those that do. The bottlenecks consist of an initial 1x1 pointwise expansion layer, a depthwise convolution layer (DWC) with a kernel of size 3x3 or 5x5, and a final 1x1 projection layer. The architecture of MobileNet V3 is shown in Fig.1. The model uses the H-swish activation function.

III. MODIFICATIONS

Modifications made to MobileNet V3 are described below. Table I summarizes CMV3 architecture.

A. *Convolution Layers*

CNN with more Depthwise convolutions than pointwise convolutions have shown better performance [21]. This fact has been exploited to make architectural changes by emphasizing spatial information rather than aggregating channel information. Depthwise Pointwise Depthwise (DPD) blocks, as shown in Fig.2, were used for the model.These blocks comprise a 3×3 Depthwise convolution, with a stride s that expands the number of channels and performs down-sampling. They also consist of a 1×1 pointwise convolution that collects information along channels, merges them, and finally, a 3×3 Depthwise separable convolution layer. All DPD blocks were followed by batch normalization and RELU.As more Depthwise convolutions are used than pointwise, good compression is achieved. The ratio of number of parameters in Pointwise to Depthwise convolutions is displayed below:

$$\frac{W \times H \times C \times mC}{W \times H \times k \times k \times mC} = \frac{C}{k^2} \tag{1}$$

(W × H) is the input dimension, C is the number of channels, m is the channel multiplier, and (k × k) is the filter size. Since the number of channels is much higher than the filter size, the ratio to be greater than 1.

B. *Mish Activation Function*

Mish possesses the self-regularizing capacity and lessens overfitting. It outruns other activation functions in performance. It also keeps negative gradients, has better generalization and eliminates saturation due to near-zero gradients [22]. The formula below can describe it:

$$f(x) \ = x \cdot \tanh(softplus(x)) \tag{2}$$

$$softplus(x) = \ln(1 + e^x) \tag{3}$$

The Mish activation function followed DPD blocks in the new architecture. It improved the accuracy from 88.14% to 89.13%. Fig.3 depicts the graph of Mish.

Compressed MobileNet V3 Architecture						
Input	***Operator***	***e***	***c***	***SE***	***NL***	***s***
32^2×3	Conv2d 3×3	-	16	-	HS	1
32^2×16	Bneck 3×3	48	32	✓	HS	1
32^2×32	DPD 3×3	88	40	-	MH	1
32^2×40	DPD 3×3	240	40	-	MH	1
32^2×40	Bneck 5×5	160	48	✓	HS	2
16^2×48	DPD 5×5	288	96	-	MH	1
16^2×96	DPD 5×5	592	128	-	MH	1
16^2×128	Conv2d 1×1	-	256	✓	HS	1
16^2×256	Pool 16×16	-	-	-	-	1
1^2×256	Conv2d 1×1	-	576	-	HS	1
1^2×576	Conv2d 1×1	-	k	-	-	1

TABLE I
WHERE E: EXPANSION FACTOR, C: NUMBER OF OUTPUT CHANNELS, SE: SQUEEZE AND EXCITE BLOCKS, NL: ACTIVATION, HS: H-SWISH, MH: MISH AND S: STRIDE

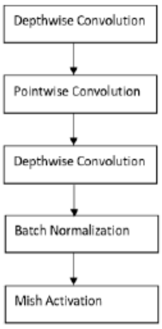


Fig. 2. DPD Blocks

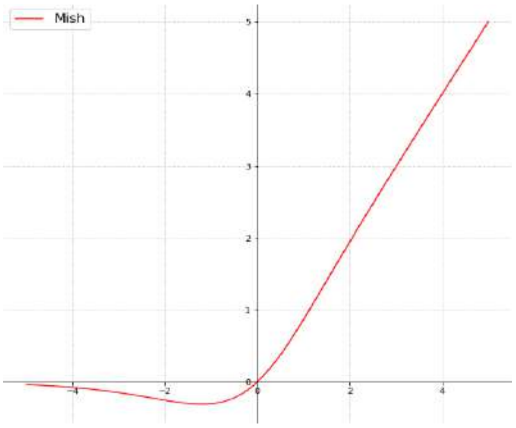


Fig. 3. Mish Activation Function [23].

C. *Expansion filters*

Mobilenet V3 uses expansion filters to extend to a high dimensional feature space to intensify non-linear transformation on channels [15]. This technique is used on CMV3 as well. Expansion filters in a few layers are increased. It boosted accuracy from 84.56% to 88.14%.

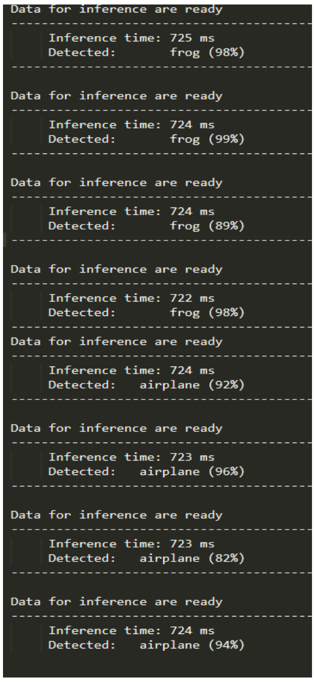
IV. TRAINING SETUP

The modified model was trained with Intel Xenon Gold 6126 processor with 32GB RAM and NVIDIA Tesla P100 GPU. An l2 weight decay of 1e-5 was used. A dropout of 0.8 and a cosine decay type scheduler were added. A width multiplier of 0.5 was selected to reduce overfitting and maintain a good trade-off between accuracy and size.

V. Deployment

NXP eIQ is a software platform comprising resources and tools to help machine learning deployment on NXP hardware. It has Neural Network (NN) compilers, libraries, inference engines, Hardware Abstraction Layers (HAL) to support TensorFlow lite (TFLite), ARM NN, glow, Cortex Microcontroller Software Interface Standard (CMSIS)-NN, and OpenCV [24].

The model used TFLite for deployment into iMX RT1060. It is available in both yocto and MCUXpresso environments. It is faster and consumes less memory than TensorFlow, making it suitable for use in low-resource devices. We used MCUXpresso IDE and built the SDK for iMX RT 1060 using eIQ middlewares. This middleware comes with a lot of demo examples. CIFAR-10 label image example was used. This example uses a DL model to classify images captured by the camera attached on board. CMV3 was then included in the header files, and many images were tested to decipher model accuracy and inference after deployment. Fig 4. shows the results observed on the console. Fig 5. shows the block diagram for eIQ inference procedure for TensorFlow Lite.



```
Data for inference are ready
--------------------------------------------
    Inference time: 725 ms
    Detected:        frog (98%)
--------------------------------------------

Data for inference are ready
--------------------------------------------
    Inference time: 724 ms
    Detected:        frog (99%)
--------------------------------------------

Data for inference are ready
--------------------------------------------
    Inference time: 724 ms
    Detected:        frog (89%)
--------------------------------------------

Data for inference are ready
--------------------------------------------
    Inference time: 722 ms
    Detected:        frog (98%)
--------------------------------------------
Data for inference are ready
--------------------------------------------
    Inference time: 724 ms
    Detected:    airplane (92%)
--------------------------------------------

Data for inference are ready
--------------------------------------------
    Inference time: 723 ms
    Detected:    airplane (96%)
--------------------------------------------

Data for inference are ready
--------------------------------------------
    Inference time: 723 ms
    Detected:    airplane (82%)
--------------------------------------------

Data for inference are ready
--------------------------------------------
    Inference time: 724 ms
    Detected:    airplane (94%)
```

Fig. 4. Classification on i.MX RT 1060 as displayed using semi hosting [24]

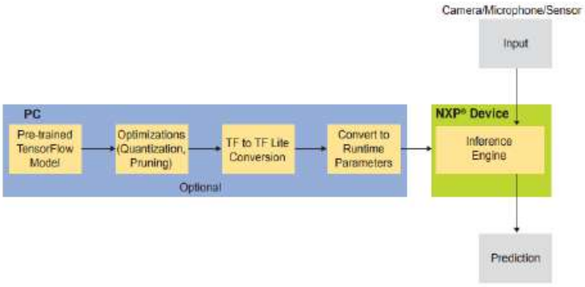


Fig. 5. eIQ inference procedure for TFLite models [24]

VI. Results

MobileNet V3 small was modified to give rise to CMV3 with no compromise to accuracy. The revised model has a size of 2.3 MB with an accuracy of 89.13%. Its parameter count has been reduced from 1,846,930 in baseline to 171,946 after compression. It was then deployed onto iMX RT 1060 for inference. It gave an average inference time of 720ms. A plot of proposed model accuracy vs the number of epochs using the Tensorboard visualization tool is shown in Fig. 6.

Various Scaling factors for CMV3.		
Width Multiplier	***Model Accuracy***	***Model size***
1.5	91.39%	10.5 MB
1.0	90.64%	5.2 MB
0.75	90.10%	3.6 MB
0.5	89.13%	2.3 MB
0.35	87.36	1.9 MB

TABLE II

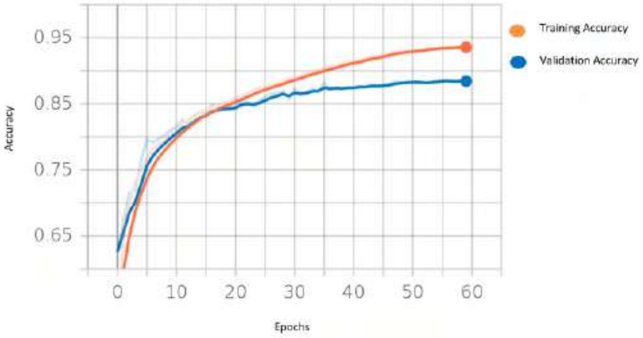


Fig. 6. Compressed MobileNet V3

VII. Conclusion

In this paper, using DPD blocks, mish activation function, and increase in expansion filters, an architecture that is 84.96% smaller in size and 0.2% more accurate than baseline is accomplished. It can be successfully used in various embedded vision platforms. Table II shows different width scaling factors that can be used with the model. Based on the application, a suitable configuration can be used to achieve optimal trade-off.

Acknowledgment

The authors would like to thank Lilly Endowment Inc for their support to Indiana University Pervasive Technology Institute.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, Dec. 2015, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1512.03385.

[3] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” *arXiv:1602.07261 [cs]*, Aug. 2016, Accessed: Mar.

09, 2021. [Online]. Available: http://arxiv.org/abs/1602.07261.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *arXiv:1512.00567 [cs]*, Dec. 2015, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1512.00567.

[5] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556 [cs]*, Apr. 2015, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *arXiv:1704.04861 [cs]*, Apr. 2017, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/1704.04861.

[7] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size,” *arXiv:1602.07360 [cs]*, Nov. 2016, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/1602.07360.

[8] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[9] F. Iandola and K. Keutzer, “Keynote’ESWEEK’2017:’Small’Neural’Nets’Are’Beautiful:’ Enabling’Embedded’Systems’with’Small’DeepBNeuralB Network’Architectures’,” *arXiv:1710.02759 [cs]*, p. 10, Sep. 2017.

[10] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, “Learning Efficient Object Detection Models with Knowledge Distillation,” p. 10.

[11] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” *arXiv:1503.02531 [cs, stat]*, Mar. 2015, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1503.02531.

[12] Y. Gong, L. Liu, M. Yang, and L. Bourdev, “Compressing Deep Convolutional Networks using Vector Quantization,” *arXiv:1412.6115 [cs]*, Dec. 2014, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1412.6115.

[13] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, “Quantized Convolutional Neural Networks for Mobile Devices,” *arXiv:1512.06473 [cs]*, May 2016, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1512.06473.

[14] X. Yu, T. Liu, X. Wang, and D. Tao, “On Compressing Deep Models by Low Rank and Sparse Decomposition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 67–76, doi: 10.1109/CVPR.2017.15.

[15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, “Searching for MobileNetV3,” *arXiv:1905.02244 [cs]*, Nov. 2019, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1905.02244.

[16] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, “MnasNet: Platform-Aware Neural Architecture Search for Mobile,” *arXiv:1807.11626 [cs]*, May 2019, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/1807.11626.

[17] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning Convolutional Neural Networks for Resource Efficient Inference,” *arXiv:1611.06440 [cs, stat]*, Jun. 2017, Accessed: Dec. 23, 2020. [Online]. Available: http://arxiv.org/abs/1611.06440.

[18] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, “Pruning Filters for Efficient ConvNets,” *arXiv:1608.08710 [cs]*, Mar. 2017, Accessed: Dec. 23, 2020. [Online]. Available: http://arxiv.org/abs/1608.08710.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *arXiv:1409.0575 [cs]*, Jan. 2015, Accessed: Mar. 09, 2021. [Online]. Available: http://arxiv.org/abs/1409.0575.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” *arXiv:1709.01507 [cs]*, May 2019, Accessed: Mar. 10, 2021. [Online]. Available: http://arxiv.org/abs/1709.01507.

[21] G. Li, M. Zhang, Q. Zhang, Z. Chen, W. Liu, J. Li, X. Shen, J. Li, Z. Zhu, and C. Yuen, “PSDNet and DPDNet: Efficient Channel Expansion,” p. 15.

[22] D. Misra, “Mish: A Self Regularized Non-Monotonic Activation Function,” *arXiv:1908.08681 [cs, stat]*, Aug. 2020, Accessed: Mar. 10, 2021. [Online]. Available: http://arxiv.org/abs/1908.08681.

[23] D. Misra, “Mish: A Self Regularized Non-Monotonic Neural Activation Function,” p. 13.

[24] NXP, “EIQ-FS.pdf.” Aug. 05, 2020, Accessed: Mar. 12, 2021. [Online]. Available: https://www.nxp.com/docs/en/fact-sheet/EIQ-FS.pdf.

Practical Limits and Challenges for Powering of Wireless Systems in Implantable and Lab on a Chip Biomedical Devices and Review of the Low Power Design Techniques

Bahram Ghafari
Department of Electrical and Electronics Engineering
The university of Melbourne
Parkville, Victoria 3010, Australia
e-mail: b.ghafari@pgrad.unimelb.edu.au

***Abstract*— There is a high demand for research into innovation and development of miniaturized electronic devices for biomedical applications such as lab on a chip device, implantable medical devices (IMD), and wireless biosensor systems (WBS), etc. These electronic systems must be wireless as wires penetrating through human skin increase the risk of infections as they act as conduits for viruses and bacteria and they also limit the flexibility of movement for patients. Ultra-low power transceivers are essential because wireless communication subsystems consume most of the power in wireless biomedical implants and implanted batteries are undesirable due to their limited lifespan and the risk of infection they pose. Also, a limited amount of power can be transferred through the wireless power link system. The design requirements for wireless communication and power source subsystems for wireless biomedical implants can be determined based on the specific application of the wireless biomedical implants. Practical limits and challenges in low power and low voltage design of wireless systems in lab-on-a-chip devices and implantable biomedical devices need to be considered for investigating various techniques for low power design with the advantages and the trade-offs of each design technique. This paper outlines the practical limits and challenges for powering wireless systems in implantable and lab-on-a-chip biomedical devices and reviews low power design techniques.**

Keywords*— body sensor networks, wireless, phase noise, ultra-low power, implantable medical sensors, medical Implant Communication Service (MICS) frequency band.

I. Introduction

Lab on a chip and implantable biomedical devices have become an increased research focus. These devices have various applications such as chronic monitoring implants, nervous system stimulating implants, brain-machine interface systems and temporary implants like intelligent pills and wireless capsule endoscopy (WCE) systems. Most of these devices have sensing, signal processing, wireless communication, and power source subsystems [1].

Wireless communication subsystems consume most of the power in wireless biomedical implants [3], and because of the limited available power in miniaturized energy sources and battery-less operating systems [2], designing a power efficient wireless communication and power source subsystem is very challenging.

The design requirements for wireless communication and power source subsystems for wireless biomedical implants can be determined based on the specific application of the wireless biomedical implants.

Chronic monitoring and nervous system stimulating implants need to operate for several years after insertion in the body through a surgical procedure. These implants transmit and receive data with different bit rates for different applications. For example, chronic monitoring implants have higher data transmission and so for the higher sample rate and resolution, they require higher bit rates [3].

Nervous system stimulating implants have higher bit rates for receiving data depending on the required stimulation cycle rate and the number of stimulating channels.

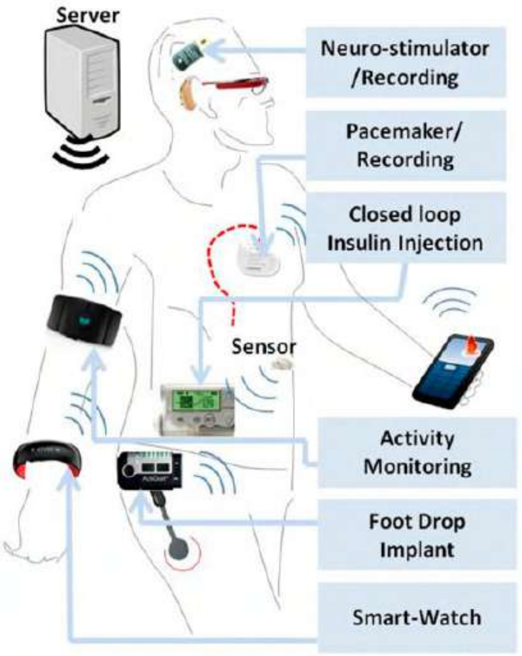


Figure 1. An overview of the Wireless Biosensor Systems (WBS) for monitoring and diagnostics of various diseases [34].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Brain-machine interfaces (BMI) have a high number of channels for monitoring and stimulating and require high bit rates for both transmitting and receiving data. Designing a power efficient wireless communication subsystem for these devices is very challenging.

Temporary implants like intelligent pills and wireless capsule endoscopy (WCE) systems have a limited operation lifetime inside the body, and there are more available options for designing the power source subsystem of these temporary implants.

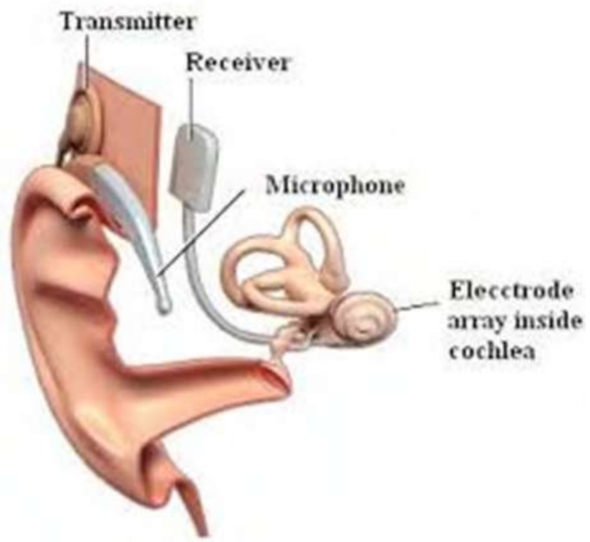


Figure 2. A system overview of the cochlear implant (bionic ear) [33].

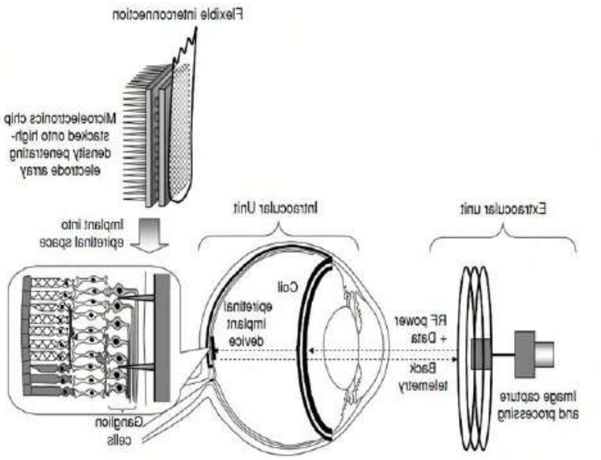

Flexible interconnection
Microelectronics chip stacked onto high-density penetrating electrode array
Implant into epiretinal space
Intraocular Unit
Extraocular unit
Coil
Epiretinal implant device
RF power + Data
Back telemetry
Image capture and processing
Ganglion cells

Figure 3. The system overview of the bionic eye device [35].

II. Available Power Sources for Wireless Biomedical Implants

Two important factors for power sources are the capability to deliver instantaneous power and the ability of integrated power delivery. For nervous system stimulating application, the power source has to have the capability for instantaneous power and integrated power delivery, but for chronic monitoring application, the ability of integrated power delivery is more critical.

Supercapacitors have a high capability of instantaneous power delivery, but their ability for integrated power delivery is limited. Alternatively, power harvester systems like solar, thermal and glucose fuel cells that gather energy from the implant surrounding area and convert it to electrical power have a higher ability for integrated power delivery, but reduced capability of instantaneous power delivery. Rechargeable batteries have a moderate capability for instantaneous and integrated power delivery and are suitable options for applications where the power source needs the capability for instantaneous and integrated power delivery and their operation lifetime is matched with that of a battery’s lifetime.

Battery-less operation systems are a suitable option for applications with an operational lifetime longer than that of a battery. For battery-less operation systems where the power source has to have the capability for instantaneous power and integrated power delivery, power harvester systems can be combined with supercapacitor systems through implementing extra circuitry for sleep and active mode control. However, available energy needs to be allocated to the extra circuitry for the sleep and active mode controls but having power-efficient design for this extra circuit is very challenging. For backup supplies of these systems, rechargeable batteries can be used.

A. *Rechargeable batteries*

Rechargeable batteries have moderate capability of instantaneous power delivery and integrated power delivery. Thin film batteries with mechanical flexibility are a cost-effective and attractive option for biomedical implants and lab-on-a-chip devices [4]. The challenge, however, of using rechargeable batteries in wireless lab-on-a-chip and biomedical implants is finding ways to decrease their recharge capacity over recharge cycles because currently expensive surgical procedures for battery replacement are required. Also, these batteries require special packaging to eliminate the risk of poisonous chemicals leaking. This naturally increases the device costs.

For power harvester systems, electromagnetic field power transfer systems and ultrasound power transfer systems, rechargeable batteries can be used as a backup supply.

B. *Electromagnetic field power transfer systems*

Wireless biomedical implants and lab-on-a-chip devices can be powered by an electromagnetic field generated by an external device. This system can be used in combination with a rechargeable battery as a backup [5].

One of the challenges for implementing electromagnetic field power transfer systems is designing an optimal external antenna size that fits with the implant size and still receives maximum power from the external electromagnetic field. Matching the external antenna size with the implant size creates some limitations in using compact external antennas and selecting an external electromagnetic field frequency. To minimize energy absorption in the tissue, electromagnetic field frequencies less than 30MHz are used in most commercial devices. However, for an efficient power transfer system in low frequency, the size of the implanted antenna has to be tens of centimetres, which is an unacceptable size for an implant [7][37]. In recent designs, to maximize power transfer efficiency and maintain a small sized implanted antenna, the implant size has been optimized to operate with a high electromagnetic field frequency in the range of 100MHz to 5GHz [6][36]. Human tissue absorbs some part of the transmitted power because of finite tissue conductivity and the maximum radiated power from the external electromagnetic field is limited.

One limiting factor of maximum radiated power from the external electromagnetic field is the operating life of the power source of the external electromagnetic field system. For example, if a battery is used as the power source of the external electromagnetic field system, the maximum radiated power is limited by the operating lifetime of the battery.

Another limiting factor for the maximum radiated power from the external electromagnetic field is IEEE standard for maximum allowable specific absorption rate (SAR) on chronic exposure [8]. This standard applies a limit on the maximum allowable radiated power, based on the carrier frequency and antenna size.

C. *Ultrasound power transfer systems*

The wireless biomedical implants and lab-on-a-chip devices can be powered by an external ultrasound transducer that transfers power through generating pressure waves and an implanted piezoelectric transducer inside the body absorbs and converts incoming pressure waves into electrical energy [7, 9]. A rechargeable battery can be used as a backup source in combination with ultrasound power transfer systems.

Designing a miniaturized ultrasound transducer and solving the mismatch issue between air acoustic impedance and tissue acoustic impedance is a challenge for using ultrasound power transfer systems as power sources for wireless biomedical implants.

D. *Power harvester systems*

Power harvester systems (ambient power) that gather energy from the implant’s surrounding area and convert it to electrical power can be a power source option for wireless biomedical implants [10, 11].

One power harvester system is solar cells which are the topic of significant commercial and academic research [12, 13].

Piezoelectric in some power harvester systems absorbs and converts incoming pressure waves into electrical energy. Electromagnetic and capacitive power harvesting systems absorb and convert electromagnetic and inductive waves into electrical energy [14].

One of the challenges of using these systems is the requirement for larger device volume size to generate more output power. Another challenge for designers in implementing these systems is the need to match excitation frequency with the size of these devices which is sometimes impracticable. Another challenge using these systems is the additional circuitry needed to rectify the generated AC output voltage which reduces their efficiency.

Power harvesting systems that use electrochemical [11] and glucose fuel cells [10] can produce power by implanting in specific locations of the human body with higher densities of energy. The challenge here is their large size and limited ability to integrate power delivery.

Power harvester systems that use thermoelectric generators can be used as power sources for implants under the skin as the temperature inside the human body does not vary significantly [15]. The amount of generated power in thermoelectric generators is related to the surface area of the device.

III. Wireless communication subsystem for wireless biomedical implants and lab-on-a-chip devices

Selection of correct system and architecture level design is critical for designing low power systems. One of the critical steps in system and architecture level design is selecting the modulation technique.

Modulation techniques such as on-off keying (OOK), amplitude shift keying (ASK), binary frequency shift keying (BFSK) and binary phase shift keying (BPSK) are the most popular digital modulation techniques for low power applications.

Design and implementing of OOK and ASK modulation techniques are more straightforward than BFSK and BPSK, but their reliability to the noise is less than BFSK and BPSK modulation techniques. BFSK modulation technique has better spectrum efficiency, and BER performance compares with OOK modulation technique [20].

Wireless biomedical implants and lab-on-a-chip devices can have either a remote power source or a local power source and can communicate via the back-scattering transmitter method or up-conversion transmitter method.

A. *Wireless biomedical implants with the local power source*

Wireless biomedical implants and lab-on-a-chip devices with local power and up-conversion transmitters are the most popular systems that can be powered by local power harvester systems and rechargeable batteries as backup. They are compatible with MICS technology and use broadband or narrow-band modulation techniques like BFSK, BPSK, MSK and ASK. These devices have good frequency selection capability, are reliable against noise and can be adopted by other systems easily [16].

The challenge for wireless lab-on-a-chip devices and biomedical implants with local power and up-conversion transmitters is the high-power consumption in transmit mode.

Ultra-wideband systems like on-off keying (OOK) or pulse position modulation (PPM) have simpler transmitter designs and lower power consumption in transmit mode, but require more complex receiver architecture and have higher power consumption in receive mode. There is a trade-off between the complexity of the transmitter and the receiver. Ultra-wideband systems are the preferred design for applications with more transmit cycles whereas narrowband systems are the preferred design for more receive cycles.

The back-scattering technique can be implemented in wireless biomedical implants and lab-on-a-chip devices with a local power source to decrease power consumption in transmit mode. The complexity of the architecture design in this technique is transferred to the receiver side and external carrier emitter device which has more power consumption for generating the carrier frequency. Implanted transmitter systems with no carrier generation and power amplification remain simple and power efficient.

B. *Wireless biomedical implants with the remote power source*

The back-scattering technique can be implemented in wireless biomedical implants and lab-on-a-chip devices with a remote power source to decrease power consumption in transmit mode. The remote power source for these devices can be electromagnetic field power transfer systems or ultrasound power transfer systems.

In the back-scattering technique, the external carrier emitter device generates the carrier signal. The back-scattering transmitter in the implant modulates the carrier signal which is picked up by the antenna through changing the impedance loading of the data. This modulated signal will be detected by the external receiver/reader which has more complex architecture and sensitivity, and higher power consumption. Implanted back-scattering transmitters remain simple and power efficient with no carrier generation and power amplification [17]. Single compact antennas for both wireless power transfer and data communication are the preferred option in recent back-scattering implant designs. The challenge for designers is to optimize the antenna size for both carrier frequency and wireless power transfer systems. Other issues with backscattering systems is their requirement for an external carrier signal emitter system, their limited capability for calibration and their performance changing with variations in environmental factors [18].

Wireless biomedical implants and lab-on-a-chip devices with a remote power source can be used in combination with the up-conversion transmitter [19]. Up-conversion transmitters use mixer circuits to modulate the baseband data signal and they require two separate (or multi-mode) antennae for data communication and the power transfer system. The issue using up-conversion transmitters in wireless biomedical implants is the higher noise bandwidth and power consumption, compared with backscattering transmitters.

IV. Low power design techniques

In order to significantly save power, consideration of a low power design for the system and architecture level are critical. Applying low-power methodology from system-level to device-level is essential for optimizing the power dissipation. Based on the system requirements and architecture proposal, the required process technology can be determined. The detailed circuits and device requirements are other critical factors. Also, the selection for types of architecture, circuit, and device can be limited by selected process technology. For analogue blocks such as amplifiers, oscillators, and filters, the fundamental minimum power consumption can be defined as [30]:

$$P_{min} = 8kTf(S/N) \qquad (1)$$

where f is the operating frequency, the S/N ratio (SNR) is the signal to noise power ratio, and k is the Boltzmann's constant.

Fundamental minimum power consumption is calculated based on the operation frequency and signal to noise power ratio (SNR), and only has an asymptotic limit realistic restrictions such as voltage swing, circuit topology, and nonlinearity are not considered in this calculation. The power consumption for typical analogue circuits is several orders of magnitude larger than the fundamental minimum power consumption.

For designing low power analogue circuits, finding the best trade-off among the basic analogue design parameters such as bandwidth, linearity, gain, noise, and accuracy is essential. The first step in the design process can be selecting the circuit topology, then the supply voltage based on specifications such as voltage swing and dynamic range. The selection of the operating region of transistors such as subthreshold operation is another important step.

CMOS technology has become the technology of choice for radio frequency (RF) and integrated circuit (IC) designers because of its low fabrication cost and the potential to integrate with accompanying digital circuits. Various low power design techniques for radio frequency (RF) and integrated circuits (IC) are available with modern CMOS technology, such as subthreshold operation of CMOS field-effect transistor (FET) devices, which are popular in low power designs.

A. *Low supply voltage and subthreshold operation*

Subthreshold or weak inversion (WI) operating regions for transistors are used in many biomedical electronic circuit implant applications as they are considered the most power efficient region. The transconductance is higher in a subthreshold device operation for a given bias current which is a desirable property for low power circuits design, but is let down by lower transit frequency (f_T). Therefore, the subthreshold operation has been associated with low frequency applications. In new MOS technology, with each generation of scaling, f_T increases by more than 75% in all regions of operation which provides the possibility of using subthreshold operation regions for low to mid-range frequency applications.

Most of the biomedical signals are in the low to mid-range frequency. Therefore, the subthreshold operation region is a practical option for these biomedical devices. The roadmap for CMOS, technology scaling report for submicron processes, shows that the device gate length and supply voltage decrease, thereby reducing power consumption and enabling a smaller chip size [21]. Total power consumption in electronic circuits is divided into dynamic power and static power. The dynamic power consumption relation to the supply voltage (V_{dd}), the load capacitance C_L, the clock frequency (f_{clk}), and α is as follows:

$$P_{Dynamic} = \alpha \,.\, C_L \,.\, (V_{dd})^2 \,.\, f_{Clk} \qquad (2)$$

Decreasing the supply voltage may reduce power consumption, but it creates a longer propagation delay, which is proportional to the square of the difference between the supply voltage and the threshold voltage as follows:

$$T_{pd} \propto V_{dd}/(V_{dd} - V_{TH})^2 \qquad (3)$$

Therefore, decreasing the supply voltage without decreasing the threshold voltage, increases the propagation delay and slows the circuit. The threshold voltage has to be scaled down to minimize the propagation delay. The static power consumption increases by reducing the threshold voltage because of leakage current. Static power has a direct relationship to leakage current and the supply voltage source as follows:

$$P_{Static} = I_{Leakage} \text{-} V_{dd} \qquad (4)$$

The total power is expressed as follows:

$$P_{Total} = P_{Dynamic} + P_{Static} \qquad (5)$$

Subthreshold or weak inversion operation is a suitable technique for low power design for low to mid-range frequency applications but the leakage current has to be minimised to decrease the static power specifically for devices with a long standby mode. There are several techniques for reducing the leakage current, such as variable-threshold CMOS (VTMOS), multi-threshold voltage CMOS (MTCMOS) [22], super-cut-off CMOS (SCCMOS), double-gate-dynamic-threshold CMOS (DGDTMOS), and multiple supply voltage and transistor stacks [23].

B. Important leakage current sources in MOS transistors

1) Gate-induced drain leakage current (I_{GIDL}): Gate-induced drain leakage current (I_{GIDL}) is caused by the gate-field influence on the depletion region of the drain. Increasing gate voltage makes the depletion layer thinner, causing overlap between the gate field and the depletion layer. This overlap makes band-to-band-tunnelling (BTBT) and increases the gate induced drain leakage current [23].
2) Gate leakage current (I_G): The thickness of CMOS oxide in recent technology is just a few nanometers, which allows electrons from the gate tunnel through the SiO2 layer to the substrate or in the reverse direction, thereby creating gate leakage current (I_G). Increasing the gate voltage makes the electric field stronger and helps the holes in the gate to overcome the potential barrier at the interface. Direct hot carrier injection is possible through three mechanisms: electron conduction band (ECB) tunnelling, electron valance band (EVB) tunnelling, and hole valance band (HVB) tunnelling. The gate leakage is produced more by electrons than holes because of the lower effective mass and barrier height (3.1 eV) in electrons holes have a higher barrier height (4.5 eV) [23]. Therefore, the hole valance band (HVB) tunnelling current is smaller, and the gate leakage current in the p-channel device is lower. The contribution of gate leakage to the static power is important because it increases by a factor of 4000 when changing technology from 90 nm to 50 nm, but the subthreshold current increases by a factor of 25 at the same time [24]. New gate material such as metal gate and high-k dielectric is being used in nanometers technology to keep the gate leakage under control [25].
3) Leakage current from the reverse-bias p-n junction (I_{rev}): Reverse-bias p-n junction leakage current (I_{rev}) is produced from the p-n junction diodes of the source and drain to the substrate in a reverse-biased situation. The generation of the electron-hole pair inside the depletion region and minority carrier diffusion/drift produce the reverse bias leakage current. The magnitude of the p-n junction leakage current is directly related to the doping concentration and the junction area when the high electric field is applied across the p-n junction. In this state, the BTBT current increases because of electrons from the p-side's valance band tunnel to the n-side's conduction band [23].
4) Channel punch-through leakage current: Two depletion regions of the source and drain in a CMOS device merge by applying the voltage to the drain in nanometre technology because of the very small channel length. Channel punch-through leakage current occurs when these two depletion regions merge. Most of the source carriers overcome the energy barrier in this condition and are either collected by the drain or enter the substrate.
5) Subthreshold leakage current (I_{sub}): As shown in (6), an equation for the drain current in the subthreshold region; the leakage current increases exponentially by scaling of the threshold voltage.
$$I_{DSub} = I_0 e^{\frac{V_{GS} - V_{TH} + \eta V_{DS}}{n \cdot V_T}} \left[1 - e^{\frac{-V_{DS}}{V_T}}\right] \;\; ; V_{GS} \leq V_{TH}$$
(6)
Where V_T is thermal voltage, n is subthreshold slop factor, V_{GS} is gate – source voltage, I_0 is reference static current, η is DIBL coefficient and V_{DS} is drain – source voltage.

V. THRESHOLD VOLTAGE MODIFICATION

Variations in the threshold voltage effect the leakage current leakage current reduces by increasing the threshold voltage. The width of the transistor has an effect on the threshold voltage and leakage current. Reducing the width of the transistor causes short channel effects (SCE) such as narrow width effect [26]. Local oxidation of silicon (LOCOS) or the shallow trench isolation is used in MOS transistors from isolating transistors from each other [26].

Because of the two-dimensional oxidation effect, the gate oxide can sometimes be thicker near the edges of the channel which increases the threshold voltage by enlarging the total depletion charge of the substrate.

Oxide thinning near the edges or fringing fields can sometimes reduce the thickness of the effective oxide near the edges of the device in shallow trench isolation (STI) devices. In this condition, reverse-narrow-width-effect occurs, and the threshold voltage decreases because a higher voltage is required to invert the channel [26].

A. Threshold voltage variation because of the body effect

The threshold voltage changes between the body and source of the transistor (V_{BS}). The relationship between the threshold voltage and V_{BS} is expressed in (7):

$$V_{TH} = V_{THo} - \eta V_{DS} - \gamma\left(\sqrt{2\varphi_B - V_{BS}} - \sqrt{2\varphi_B}\right) \qquad (7)$$

Where V_{TH0} is zero threshold voltage, V_{BS} is voltage between body and source, γ is body bias coefficient, η is DIBL coefficient and Φ_B is Fermi potential. The depletion region in the body increases by applying the voltage to the bulk of the transistor. In this state, the threshold voltage increases because of the reverse bias.

B. Threshold voltage variation because of drain-source voltage

Increasing the drain-source voltage decreases the threshold voltage which is called drain induced barrier lowering (DIBL). Depending on the channel length in SCE, the depletion regions around the drain and source junctions can be large, and these depletion regions increase with the drain-source voltage, causing channel inversion with lower gate-to-source voltage (V_{GS}). In fact, the threshold voltage is modulated by reducing the barrier potential. The DIBL effect has no effect on the subthreshold slope. Applying a high surface and channel doping or selecting shallow source and drain junction depth, can reduce the DIBL effect [27]. As expressed in (8), in minimum length devices, the DIBL effect decreases early voltage and the intrinsic voltage gain of the device.

$$A_v = g_m \,.\, r_{ds} = (g_m / I_D) \,.\, V_A \qquad (8)$$

Where A_V is intrinsic voltage gain, g_m is transconductance of the device, r_{ds} is output resistor, I_D is drain current and V_A is early voltage. The minimum length should be selected with caution where the gain of the circuit is important to the weak inversion region [32].

VI. CIRCUIT TECHNIQUES FOR REDUCING THE LEAKAGE CURRENT TO REDUCE THE STATIC POWER CONSUMPTION

One circuit technique for reducing leakage current is the dual-threshold CMOS (DTCMOS) method. The DTCMOS is an effective method to reduce static power consumption. The circuit in the DTCMOS method has low threshold voltage devices on the critical path and high threshold voltage devices off the critical path. The leakage current reduces because of the high threshold devices, but the delay of the critical path increases because of the high threshold devices [23]. In active mode, high threshold devices footer/or header from the power rails are able to enable regular circuit operation and turn off in standby mode to reduce the leakage current. There are several other techniques for reducing the leakage current such as variable-threshold CMOS (VTMOS), multi-threshold voltage CMOS (MTCMOS) [22], super-cut-off CMOS (SCCMOS), double-gate-dynamic-threshold CMOS (DGDTMOS), multiple supply voltage and transistor stacks [23].

VII. CIRCUIT TECHNIQUES FOR REDUCING THE LEAKAGE CURRENT TO REDUCE THE DYNAMIC POWER CONSUMPTION

Pass transistor logic (PTL) is one technique to reduce the leakage current to lower the dynamic power consumption [31]. The drain and source of the transistors in the pass transistor network are not connected to the ground and supply voltage, thus significantly decreasing the number of leakage paths. The pass transistor logic (PTL) is an interface between the drivers which generate the signal and the receiver circuits which recover the voltage swing and signal and therefore the leakage current is confined to the transistors for the drivers and receiver circuits. Increasing the delay, the effective channel length and number of sneak paths in the circuit, which allow the leakage current to flow, are the main drawbacks for pass transistor logic (PTL). The sense amplifier-based pass-transistor logic (SAPTL) method is offered to overcome the limitations of the PTL method. The SAPTL circuit consists of the stack, which is the pass transistor network and computes the logic; the node driver, which injects the signal to the stack section; and the sense amplifier, which recovers the voltage swing and performance. Considering the bidirectional operation of the pass transistors, an inverted pass transistor tree is utilized as the stack to mitigate the sneak path limitations of the pass transistor logic (PTL). The drain and source of the transistors in the stack are not connected to the supply rail and have delay paths. The stack has two pseudo-differential outputs, where a signal is present in one of the two stack outputs at the same time. The sneak path limitations are mitigated in the stack because the input signal can only pass from the root to the output of the stack. Therefore, reducing V_{TH} to near zero is possible. The reduction of the threshold voltage reduces the resistance and the propagation delay without increasing the leakage current [31]. The first transistor in the chain needs voltage drop to maintain the drive current flow, and therefore the maximum voltage that can appear at the output of the stack could be ($V_{dd} - V_{TH}$). The time it takes for the input signal to reach the stack output is dependent on the number of transistors in the series in the stack loop from the root node to the output of the stack. A sense amplifier (SA) which consists of the pre-amplifier and cross-coupled latch is added to the output of the stack to recover the voltage degradation of the signal, to improve the performance of the SAPTL, and to provide sufficient buffering for driving a reasonable load capacitance. There is a trade-off between the power consumption and sensitivity of the sense amplifier in the SAPTL.

VIII. MOS Transistor Model for Different Operation Region

The world's first industry standard model for a MOS transistor, introduced in 1997, is the UC-Berkeley short-channel insulated-gate field-effect transistor (IGFET) model (BSIM) which has become more complicated to increase accuracy as the technology geometry shrinks. Selecting the correct value of (V_{GS} – V_{TH}) and the transistor length L is essential for IC design and the design parameters are related to the threshold voltage-based model. The threshold-based model causes inconsistency because there are no unified equations covering the whole region of operation and dividing the CMOS operation region into pieces when an allocated set of equations is required. For low power (micro-power) analogue circuit design, the Enz-Krummenacher-Vittoz (EKV) model was derived to substitute the threshold voltage based-model by using the surface potential-based model or the charge-based model. The advanced compact current-based (ACM) model is based on the EKV model but avoids the use of equations for the weak and strong inversion models [28]. The availability and acceptability of the ACM model are less than the BSIM model [29]. The drain current (I_D) is expressed as the difference between the forward and the reverse current in the ACM model [28].

$$I_D = I_S\,(i_f - i_r) \qquad (9)$$

The forward normalized current (i_f) depends only on the source-gate voltage whereas the reverse normalized current ($\acute{\imath}_r$) depends only on the drain-gate voltage. I_S is the normalized coefficient or specific current [28]. The value of $\acute{\imath}_f$ and $\acute{\imath}_r$ specify different modes of transistor operation. The reverse current can be ignored in the saturation region and the forward current is roughly equal to the drain current. If $\acute{\imath}_f$ <1 and $\acute{\imath}_r$ <1, then the whole channel is only weakly inverted and the transistor is in the weak inversion region [28]. The normalized coefficient or specific current (I_S) is defined as:

$$I_S = 1/2 \,.\, n \,.\, \mu \,.\, C_{ox} \,.\, V_T{}^2 \,.\, W/L \qquad (10)$$

Where W is the width and L is the length of the transistor, V_T is the thermal voltage, C_{OX} is oxide capacitance, μ is mobility and n is the slope factor. The normalized reverse and forward current are defined as:

$$i_{f(r)} = I_{F(R)} / I_S \qquad (11)$$

The most accurate parameters for the ACM model can be obtained from the foundry, but another less accurate option is by extracting from the BSIM model.

IX. Parameter Extraction for Normalized Coefficient (I_S) and the G_M/I_D Design Methodology

The normalized coefficient or the specific current (I_S) can be extracted through a method explained in [23] but a more accurate method is by calculating transconductance efficiency to determine the normalized coefficient or the specific current (I_S). Transconductance efficiency is the ratio of the transconductance (g_m) to the drain current (I_D) [29]. The transconductance (g_m) is defined as δI_D / δV_{GS} and is equal to:

$$g_m = (2 \,.\, I_D) / (V_{GS} - V_{TH}) \rightarrow g_m / I_D = 2 / (V_{GS} - V_{TH}) \qquad (12)$$

The (V_{GS} – V_{TH}) and g_m/I_D are related. The transistor operating region can be determined by the transconductance value. The level of inversion can be determined by the normalized drain current value. The level of inversion is moderate when the normalized drain current value is equal to one. The intersection of the weak and strong inversion is the centre of moderate inversion. Based on the normalized drain current ($\acute{\imath}_S$) value, the transistor is in strong inversion when $\acute{\imath}_S$ >10, the transistor is in weak inversion when $\acute{\imath}_S < 0.1$, and the transistor is in moderate inversion when $0.1 < \acute{\imath}_S < 10$.

In order to obtain the same inversion coefficient, PFETs must have a larger aspect ratio than NFETs because of lower mobility and ($\acute{\imath}_S$) in PFETs. For most of the IC design, gm is the key design parameter and the gm/ID design methodology is a well-known and effective method of designing analogue circuits. The gm/ID ratio is plotted against the normalized inversion coefficient, which defines the operation region of the transistor. The first step for the gm/ID design methodology procedure is setting the target gm, then figuring out the required DC bias current in the targeted region and determining the transistor aspect ratio (W/L) according to the corresponding IC and also determining the channel length and width.

X. Conclusion

The characteristics and design requirements of the wireless biomedical implants and lab-on-a-chip devices for different applications were reviewed. The alternative power sources for the wireless biomedical implants and lab-on-a-chip devices were considered, and their advantages and limitations were discussed. Different architecture methods for wireless systems in lab-on-a-chip devices and implantable biomedical devices with local and remote power sources were reviewed.

Practical limits and challenges in low power and low voltage design of wireless systems in lab-on-a-chip devices and implantable biomedical devices were reviewed, and various techniques for low power design were investigated, with the advantages and the trade-offs of each design technique identified.

Subthreshold or weak inversion operation was highlighted as a suitable technique for low power design for low to mid-range frequency applications, but the leakage current has to be reduced to decrease the static power specifically for devices with a long standby mode. There are several techniques for reducing the leakage current such as variable-threshold CMOS (VTMOS), multi-threshold voltage CMOS (MTCMOS) [22], super-cut-off CMOS (SCCMOS), double-gate-dynamic-threshold CMOS (DGDTMOS), multiple supply voltage and transistor stacks [23].

However, a combination of low power techniques can be used for designing different modules to achieve the best performance for individual sections. Investigating the whole system functionality before selecting a modular design for each section may be more efficient. For example, different biomedical implants and lab-on-a-chip devices require different levels of efficiency even when performing the same tasks such as monitoring, stimulation or wireless transmission. Therefore, considering the required level of efficiency for the whole system functionality is critical for designing low power biomedical implanted devices.

REFERENCES

[1] A. Virdis, et al.: 'Evaluation of off-the-shelf NFC Devices for Biomedical Applications', IEEE International Conference on RFID Technology and Applications (RFID-TA), Sept. 2019, pp. 387-392.

[2] S. Wang, et al.:' A 0.35-V 240-μW Fast-Lock and Low-Phase-Noise Frequency Synthesizer for Implantable Biomedical Applications', IEEE Transactions on Biomedical Circuits and Systems, vol. 13, Issue 6, pp. 1759-1770, 2019.

[3] S. Song, et al.:' A 769 μW Battery-Powered Single-Chip SoC With BLE for Multi-Modal Vital Sign Monitoring Health Patches'. IEEE Transactions on Biomedical Circuits and Systems, vol. 13, Issue 6, pp. 1506-1517, 2019.

[4] E. Prawiro, et al.:' A Wearable System That Detects Posture and Heart Rate: Designing an Integrated Device With Multiparameter Measurements for Better Health Care', IEEE Consumer Electronics Magazine, Vol. 8, Issue 2, pp. 78-83, 2019.

[5] B. Rao, et al.:' Joint Wireless Charging and Data Collection using Mobile Element for Rechargeable WSNs', International Conference on Computing, Power and Communication Technologies (GUCON), Sept. 2019, pp. 837-844.

[6] M. Mark: 'Powering mm-Size Wireless Implants for Brain-Machine Interfaces', 2011.

[7] S. K. Kelly, et al.: 'Optimal primary coil size for wireless power telemetry to medical implants', in Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2010 3rd International Symposium on, 2010, pp. 1-5.

[8] "IEEE Standard for Safety Levels with Respect to Human Exposure to Electric, Magnetic, and Electromagnetic Fields, 0 Hz to 300 GHz," in IEEE Std C95.1-2019 (Revision of IEEE Std C95.1-2005/ Incorporates IEEE Std C95.1-2019/Cor 1-2019) , vol., no., pp.1-312, 4 Oct. 2019.

[9] J. Liou, et al,:' Piezoelectric Micro-Vibration Effective Energy Harvesting System', IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), Oct. 2019, pp. 63-64.

[10] C. Wang et al., "Wearable Textile-Based Glucose Fuel Cell Using Mositure Management Fabrics for Improved & Long-Term Power Generation," 2019 20th International Conference on Solid-State Sensors, Actuators and Microsystems & Eurosensors XXXIII (TRANSDUCERS & EUROSENSORS XXXIII), Berlin, Germany, 2019, pp. 2543-2544.

[11] P. D. Artyukhov, I. Burmistrov and I. Artyukhov, "Electric Power Supply of Wireless Sensors by Thermo-Electrochemical Cells," 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), Varna, Bulgaria, 2019, pp. 1-5.

[12] T. Wu, J. -M. Redouté and M. R. Yuce, "Subcutaneous Solar Energy Harvesting for Self-Powered Wireless Implantable Sensor Systems," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 4657-4660.

[13] H. Sharma, A. Haque and Z. A. Jaffery, "An Efficient Solar Energy Harvesting System for Wireless Sensor Nodes," 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), Delhi, India, 2018, pp. 461-464.

[14] S. J. Oh et al., "A Solar/Thermoelectric/Triboelectric/Vibration/RF Hybrid Energy Harvesting based High Efficiency Wireless Power Receiver," 2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Genoa, Italy, 2019, pp. 911-914.

[15] D. Artyukhov, I. Burmistrov, I. Artyukhov and V. Alekseev, "Using Thermoelectrics for Power Supplying of Wireless Sensors Network," 2019 3rd School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), Innopolis, Russia, 2019, pp. 13-15.

[16] H. Rahmani and A. Babakhani, "A 434 MHz Dual-Mode Power Harvesting System with an On-chip Coil in 180 nm CMOS SOI for mm-Sized Implants," 2018 IEEE/MTT-S International Microwave Symposium - IMS, Philadelphia, PA, USA, 2018, pp. 1130-1133.

[17] Q. Chen, X. Zhang, J. Li and J. Zhou, "Priority-based Access Strategy for Multi-transmitter Multi-receiver Ambient Backscatter Communication System," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020, pp. 1-5.

[18] S.-S. Lee, et al.: 'A new TX leakage-suppression technique for an RFID receiver using a dead-zone amplifier', in Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International, 2013, pp. 92-93.

[19] J. Masuch, et al.: 'Co-Integration of an RF Energy Harvester Into a 2.4 GHz Transceiver', Solid-State Circuits, IEEE Journal of, vol. PP, pp. 1-10, 2013.

[20] H. Seo, Y. Moon, Y. Park, D. Kim, D. S. Kim, Y. Lee, K. Won, S. Kim, and P. Choi: 'A low power fully CMOS integrated RF transceiver IC for wireless sensor networks', IEEE Trans. on VLSI Systems, vol. 15, no. 2, pp. 227–231, Feb. 2007.

[21] International Technology Roadmap for Semiconductors. 2011, International Technology Roadmap for Semiconductors (ITRS).

[22] Kao, J.T. and A.P. Chandrakasan: 'Dual-threshold voltage techniques for low-power digital circuits', IEEE Journal of Solid-State Circuits, 2000. 35(7): p. 1009-1018.

[23] Helms, D., E. Schmidt, and W. Nebel: 'Leakage in CMOS circuits – An Introduction integrated circuit and system design. power and timing modeling, optimization and simulation', E. Macii, V. Paliouras, and O. Koufopavlou, Editors. 2004, Springer Berlin /Heidelberg. p. 17-35.

[24] Roy, K., S. Mukhopadhyay, and H. Mahmoodi-Meimand: 'Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits', Proceedings of the IEEE, 2003. 91(2): p. 305-327.

[25] Bohr, M.T., et al.: 'The High-k Solution', IEEE Spectrum, 2007. 44(10): p. 29-35.

[26] Taur, Y. and T.H. Ning: 'Fundamentals of Modern VLSI Devices', 2009: Cambridge University Press.

[27] Schuster, C.: 'Leakage aware digital design optimization for minimal total power consumption in nanometer CMOS technologies', 2007: University of Neuchâtel.

[28] Cunha, A.I.A., M.C. Schneider, and C. Galup-Montoro: 'An MOS transistor model for analog circuit design', IEEE Journal of Solid-State Circuits, 1998. 33(10): p. 1510-1519.

[29] Terry, S.C., et al.: ‘Comparison of a BSIM3V3 and EKV MOSFET model for a 0.5 μm CMOS process and implications for analog circuit design’, IEEE Transactions on Nuclear Science, 2003. 50(4): p. 915-920.

[30] M. A. Sanduleanu and E. A. van Tuijl: ‘Power Trade-offs and Low-power in Analog CMOS ICs’, Kulwer Academic Publishers, Boston, 2002.

[31] R. S. Shelar and S. S. Sapatnekar: ‘BDD decomposition for delay oriented pass transistor logic synthesis’, IEEE Transactions on Very Large Scale Integration Systems, vol. 13, pp. 957-970, 2005.

[32] Binkley, D.M.: ‘Tradeoffs and optimization in analog CMOS design’, 2008: John Wiley and Sons Ltd.

[33] Blake S. Wilson, Michael F. Dorman, “Cochlear implants: Current designs and future possibilities,” Journal of Rehabilitation Research & Development (JRRD), vol. 45, no. 5, pp. 695-730, 2008.

[34] E. Ghafar-Zadeh, “Wireless integrated biosensors for point-of-care diagnostic applications,” J. Sensors, vol. 15, issue 2, pp. 3236-3261, Feb. 2015.

[35] D. C. Ng, S. Bai, J. Yang, N. Tran, and E. Skafidas, “Wireless technologies for closed loop retinal prostheses,” J. Neural Engineering, vol. 6, no. 6, pp. 1–10, Oct. 2009.

[36] P. Soontornpipit, "Design and Delevopment of a Dual-band PIFA Antenna for Brain Interface Applications," 2019 7th International Electrical Engineering Congress (iEECON), Hua Hin, Thailand, 2019, pp. 1-4.

[37] L. Huang, A. Murray and B. W. Flynn, "Optimal Design of a 3-Coil Wireless Power Transfer System for Deep Micro-Implants," in IEEE Access, vol. 8, pp. 193183-193201, 2020.

IT Infrastructure Agile Adoption for SD-WAN Project Implementation in Pharmaceutical Industry: Case Study of an Indonesian Company

Anita Nur Fitriani
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
anita.nur02@ui.ac.id

Teguh Raharjo
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
teguhr2000@gmail.com

Bob Hardian
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
hardian@cs.ui.ac.id

Adi Prasetyo
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
adip12@ui.ac.id

***Abstract*— Agile project management has been widely applied, especially in software development projects, but there are still very few implementations in IT Infrastructure projects. This study aims to improve knowledge about the implementation of agile project management on IT Infrastructure. This paper uses case studies and Design Science Research Methodology on a pharmaceutical company that transforms the IT network infrastructure for branch offices from conventional WAN topology to SD-WAN topology, managed by agile project management Lean and Kanban. As a contribution, this study proposes a project management model for IT Infrastructure agile adoption inspired by Kanban and Lean agile project management frameworks for implementing network transformation.**

Keywords*— *Agile Project Management, Design Science Research, IT Infrastructure, Kanban, Lean, Network Transformation, SD-WAN

I. Introduction

In a growing digital economy, the increasing and changing demands of internal consumers for infrastructure services, technological developments, and high technology will need to face formidable challenges as digitization changes the industrial landscape [1]. Information Technology (IT) Infrastructure has an essential role in business resources for competitive advantage because it is enabled cross-functional initiatives and processes for the business. Previous studies discussed in a dynamic and ever-changing environment characterized by complexity and uncertainty, new infrastructure services developed on top of existing infrastructure services [2].

This paper uses a case study on a pharmaceutical company among 100 branch offices in Indonesia. The entire network for branch offices managed by the Corporate IT (CIT) team. The IT team's challenge is to manage the complex network, and there is a network Service Level Agreement (SLA) target to be fulfilled every month as a part of the commitment to the company's services. The SLA Network Availability report in 2019 shows the average SLA has not achieved 98.17% from the SLA target of 98.80% see Fig 1. The impact of this condition is decreasing branch productivity. High network branch utilization in 2020 because of pandemic conditions prompted companies to implement a new standard office. From the network utilization report in 2020, 29% of branch offices reach the maximum throughput. This condition has made the branch performance decrease since delays in the branch administration and operational processes can affect the branch office's sales activities and productivity. Another challenge for the IT team is the enormous investment costs for the procurement of network equipment, and the impact is network operating costs increase every year. The pandemic conditions in 2020 require significant changes to IT infrastructure to improve immediately. The operational processes can run according to business needs, as discussed with IT Infrastructure Head, especially for network transformation. The research question is about how to transform IT Infrastructure agile.

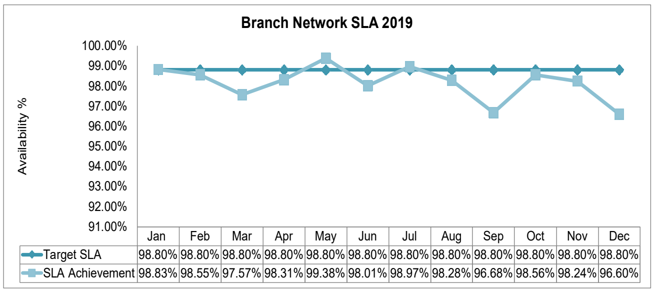


Fig. 1. SLA Achievement 2019

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

This study aims to learn adaptations of agile project management in IT infrastructure transformation, limitations on the Software-Defined Wide Area Network (SD-WAN) implementation project on the research object. This research's novelty is the implementation of agile using Kanban- Lean, which has been implemented in the company and custom to transform IT Infrastructure—using the research science design methodology (DSRM), through six stages, problem identification, the objective definition for solutions, design, and development, demonstration, evaluation, and communication [3]. Three expert judgments help the author formulated an agile adoption model, Lean and Kanban. As an academic contribution, this study proposes a project management model for IT Infrastructure inspired by Kanban and Lean agile project management frameworks for implementing network transformation SD-WAN. Simultaneously, the organization can apply agile methods to project management to be more efficient, effective and answer business needs or requirements more quickly for IT Infrastructure transformation.

II. LITERATURE STUDY

A. Agile IT Infrastructure

IT Infrastructure is the arrangement of IT stocks' technical components and services necessary for its IT environment's existence, operation, and management [1]. Three companies took a unique approach to agile transformation based on principles tailored to business needs in the previous research. Agile transformation helps organizations modernize their IT infrastructure while increasing performance significantly. The processes by which companies provide infrastructure services have grown to be more complex and labor-intensive as the company grows, so it can take months to bring new products and features to the market [1]. Previous studies have shown that agile IT Infrastructure transformation practices at team level IT combine close dialogue with stakeholders and an agile method. The scalable approach requires investment in automation, the key to a more agile delivery infrastructure. In this way, the end-to-end sending process can be better developed, and infrastructure needs should be better understood and implemented appropriately. From the above explanation, here is an example of Agile IT Infrastructure transformation inspiration from Scrum's agile framework, Kanban, and SAFe [2].

B. Agile Project Management

The agile project management approach is currently widely practiced in software development [4]. An agile project is better to adapt to changing business requirements. To increase visibility, customer collaboration, and limited-time iteration delivery provides an opportunity to re-evaluate priorities regularly [5]. Agile project management represents the management approach, and a productivity framework supports incremental progress on work priorities and continuous, even in the face of changes [6]. Agile project management is iterative, incremental, self-organizing, and emergent. The Agile Manifesto inspires the agile methodology, created by a group of software practitioners in response to this growth IT-related project failures [7]. Based on previous research, the authors conclude agile project management is a group of project management methodologies based on an agile mindset and principles.

C. IT Infrastructure

IT Infrastructure is a platform technology hardware and operating systems, network and communication technologies, data, and critical software applications [8]. IT Infrastructure collects various components can be individualized hardware entity or Infrastructure Group comprised of the server cluster infrastructure components[9]. IT Infrastructure is a technical and human device managed to provide the foundation for specific IT applications. IT technical Infrastructure includes computing platforms, hardware, operating systems, communication networks, data, and IT applications [10].

D. Kanban

Previous research explained Kanban in Japanese means signboard developed by Toyota in the 1940s to enhance its manufacturing process. Kanban is now a popular Agile methodology visualizing and controls work progress through the Kanban board. Kanban helps improve transparency, planning, and efficiency. Kanban boards can be physical or virtual and can vary in complexity [1]. Here are three principles of how the team can run the Kanban process [11]:

1. Visualize the workflow.

Divide the work into sections. Write the card's tasks and place them on the wall either physically or on a virtual computer system. Making work visible is believed as increased communication and collaboration.

2. Limit Work-in-Process (WIP)

Limit WIP focuses on WIP and designation, limiting total items in each workflow state. By limiting unfinished work in progress, the team can reduce the time it takes for an item to travel through a Kanban pipe. Limit WIP can avoid problems caused by task shifting and provide agility by activating new tasks prioritized effectively.

3. Focus on Flow

Using WIP constraints and growing team-driven policies, the team can smooth the workflow and complete the team job. Kanban works in an organization. From small to big companies and a personal level. This technique's advantage is the old practices do not have to be removed entirely but adapted to the new rules. Next to the principles, Kanban comes in six ways to implement this new working method in a team [12].

E. Lean

Lean is Kanban principles and practices based on a management philosophy, which Toyota also developed and concentrated on understanding things generate additional costs and human resources and elements in the production [11]. Lean is an effort to eliminate waste. The principle relies on improvements in administration, strategies, production [13]. Using Lean and Kanban at research object in all manufacturing processes, and Lean is also an excellent method there is no waste in all available resources during the project. This study, the agile management project approach by Lean-Kanban.

F. SD-WAN

SD-WAN has also understood the application of SDN (Software Defined Network) in WAN to integrate software-defined network technology and WAN. Besides the development of SD-WAN, various technology architectures emerge endlessly [14].

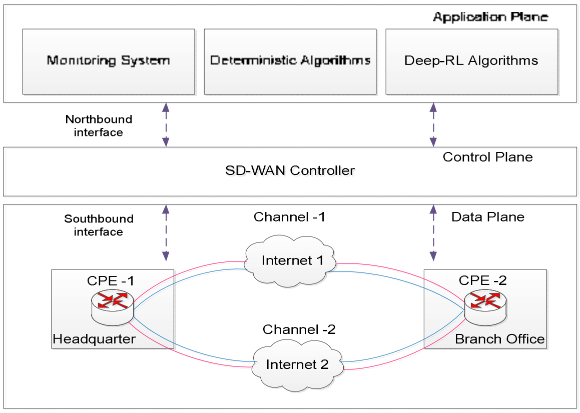


Fig. 2. SD-WAN Network Architecture [15].

SD-WAN relies on a Software Defined Network. It simplifies network control, network management and enabling innovation through network programmability. SD-WAN separates physical forwarding elements from the data field, from the network control logic, called control-plane is implemented in a logically centralized controller [15]. This control logic is what differentiates SD-WAN from VPN. SD-WAN is possible to provide good QoS. Control of the network from centrally, which continually monitors the network conditions and, consequently, adjusts the connection options [16]. SD-WAN has been implemented only on the edge of the overlay network, inside the Residing Customer Premises Edge (CPE) device at branches, head office, and the company's main office.

G. Related Research About IT Infrastructure Transformation

The related study performed agile techniques in three other infrastructure-related projects. A unique infrastructure point of view indicates that the term agile infrastructure consists of several layers. We need to address each layer [17]. The Agile methodology, adapted and applied for the Information Technology (IT) team, has also been discussed in previous studies. The research methodology practices for managing change as well as running managerial practices smoothly. The method development procedure uses a design science research approach, including problem identification, designing the proposed methodology, and evaluating case studies to show a strong relationship between the application of agile methods and IT staff members' performance [18]. IT technology models the development of infrastructure operations and the scope of agile management. It aims to analyze changes and the contribution of operational processes within the agility scope in traditional methods. An agile transformation contributes to the process's corporate structure within the system infrastructure's operational scope [19].

H. Theoretical Framework

The theoretical framework is a framework formed from related theories, previous studies, and appropriate methods in a study [20]. This paper's theoretical framework uses the theory and previous study agile project management, IT Infrastructure agile, and transformation. Designed to build agile adoption model solutions in case studies See Fig.3.

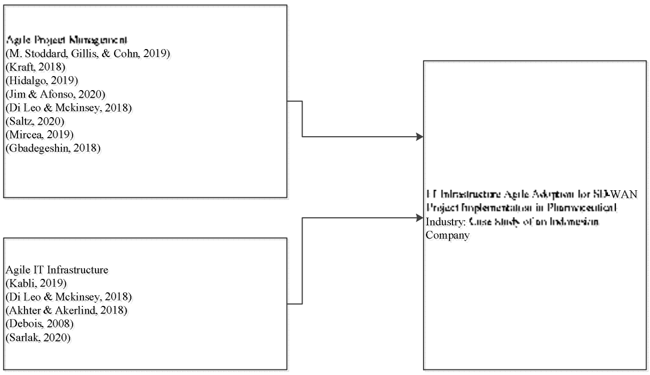

Agile Project Management
(M. Stoddard, Gillis, & Cohn, 2019)
(Kraft, 2018)
(Hidalgo, 2019)
(Jim & Afonso, 2020)
(Di Leo & Mckinsey, 2018)
(Saltz, 2020)
(Mircea, 2019)
(Gbadegeshin, 2018)
Agile IT Infrastructure
(Kabli, 2019)
(Di Leo & Mckinsey, 2018)
(Akhter & Akerlind, 2018)
(Debois, 2008)
(Sarlak, 2020)
IT Infrastructure Agile Adoption for SD-WAN Project Implementation in Pharmaceutical Industry: Case Study of an Indonesian Company

Fig. 3. Theoretical Framework

III. RESEARCH METHODOLOGY

Based on research questions and the theoretical framework formulated, this research is a case study by design science research (DSR). Design science research combines the necessary principles, practices, and procedures [21]. Through the following six stages of DSR [3], (1) Identification of problems and motivation based on the business needs and operational issue, (2) defining the objective of the solution, determining the goals of the projects, and measurable organizational value (MOV), provide the business case thru three option solutions, (3) design and development project planning, design project management model agile Lean and Kanban, design proof of concept (POC) for SD-WAN topology, (4) implementation the project execution stage SD-WAN using Kanban and Lean in Indonesian pharmaceutical company, (5) Evaluation is carried out during the SD-WAN project stage, on this project using two evaluation, weekly for IT project team, and monthly for all stakeholder review, conduct with three expert judgments from the internal and external company (6) this communication during the SD project -WAN to all IT project member, stakeholder, and publish the journal model adoption in international conference IEEE.

IV. RESULT AND DISCUSSION

This section contains the result from the DSR methodology and a discussion of the research results.

A. Identify Problem and Motivate

Conducted a discussion on current conditions attended by the IT Project Team, Stakeholder, and Branch Subsidiary IT. The result is the organization's needs. The result of this stage is a business case agreed upon by all parties involved in the project. The option agreed on in this project is to improve the branch office network services SD-WAN.

B. Define Objectives of a solution

The measurable organizational value (MOV) is determined to be achieved after understanding the business needs. The IT Project Team collects approximate options from the solutions that can help as solutions.

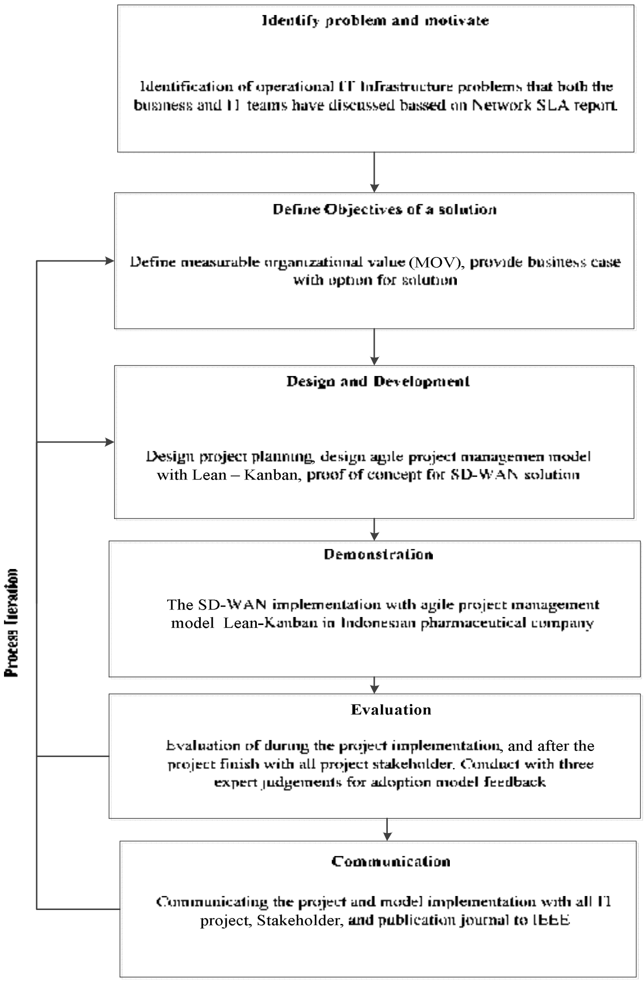

Identify problem and motivate
Identification of operational IT Infrastructure problems that both the business and IT teams have discussed bassed on Network SLA report.
Define Objectives of a solution
Define measurable organizational value (MOV), provide business case with option for solution
Design and Development
Design project planning, design agile project managemen model with Lean – Kanban, proof of concept for SD-WAN solution
Demonstration
The SD-WAN implementation with agile project management model Lean-Kanban in Indonesian pharmaceutical company
Evaluation
Evaluation of during the project implementation, and after the project finish with all project stakeholder. Conduct with three expert judgements for adoption model feedback
Communication
Communicating the project and model implementation with all IT project, Stakeholder, and publication journal to IEEE
Process Iteration

Fig. 4. Design Science Research Methodology

TABLE I. MOV PROJECT

Potential Area	Desire Objectives	Change Factor	Time
Financial	the efficiency of network operating costs	10%	Two years
	investment cost efficiency of network equipment	100%	Two years
Operational	improve network service performance	98%	Two years
Customer	increase the productivity of branch offices	Meet SLA	Two years

C. Design and Development

Design and development project plan, define the project implementation methods, SD-WAN implementation concept. In this stage, a selected SD-WAN model developed, and a proof of concept (POC) with several internet service providers to choose the best suits your needs. The results of this stage are the project implementation model, project plan, and POC.

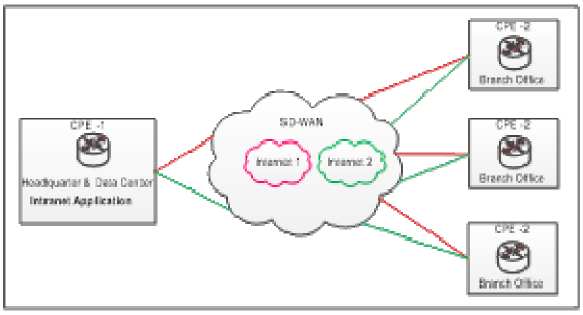


Fig. 5. Proposed Design Topology SD-WAN

D. Demonstration

The stage is the implementation of the SD-WAN project using the Lean-Kanban method. Kanban board to monitoring

all processes and project tasks until SD-WAN project complete. In this stage, the model for adopting Kanban and Lean for the SD-WAN project was implement in the Indonesian pharmaceutical company. The demonstration stage does inspect and adapts from agile to speed up feedback from all project members and stakeholders. During the implementation, Kanban helps to monitor the project task to eliminate the stopper.

E. Evaluation

The SD-WAN implementation project uses the agile Lean-Kanban method, so the evaluation process at the end of the project, but the evaluation in each period is by the IT project team every week. Moreover, at the end of each month, an evaluation is by the project team, stakeholders, and IT SBU. This evaluation aims for all teams to have the same perception and update of the ongoing project. The kanban board helps to evaluate the process that is inhibiting and must be accelerated more quickly. This evaluation also further accelerates decision-making if there is a need for changes related to ongoing projects.

To ensure the method used is correct, the researcher conducts expert judgment. It is from an internal company IT head Infrastructure and two people from external companies experts for agile and Kanban. Expert judgment with semi-structural interviews and presentation of the agile adoption model designed for the SD-WAN project. Three experts gave different input. According to business objectives, the first internal expert provided lean input as the project's main principle. The second expert provides process input to Kanban and must be detailed, and there is a "Done" in each process to make it easier to see the stopper project task. The third expert provides input to add an expedite in Kanban as a unique path if there is an urgent request, such as a VIP request. The final adoption model sees Figure 6.

F. Communication

Communication to ensure the project is going well. This stage generates inspection and adoption reports for improvements during the project. This stage's result is an update report for the project until it is done and accepted by all project teams and stakeholders. The author communicated the SD-WAN model implementation to all the project members and stakeholders and published the journal adoption model to the international conference IEEE.

G. Discussion

This study produces an agile adoption model for IT infrastructure projects, as shown in Figure 6. This agile adoption model suitable for current business needs. It was created a business case and project plan by the IT project team. The adoption of Lean in this project embodies one of the business goals with network operating costs' efficiency. Project implementation is carried out by monitoring tasks using Kanban, consisting of a task list, project processes, and a WIP 3 process limit. Each process has a done column to facilitate the complete task in the project stage.

Moreover, the completed column is for the project complete in all phases. This model has an expedite thru the provisions of WIP + 1. Expedite is used for exceptional cases if urgent conditions must be processed immediately without considering the WIP is currently running. This model evaluation and communication in each project process to ensure changes can be adapted quickly.

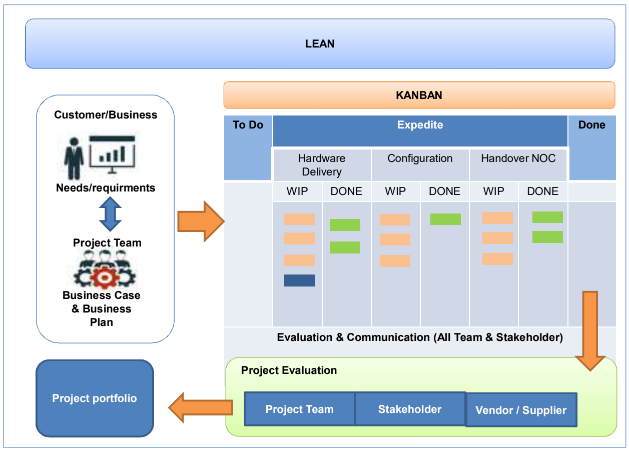


Fig. 6. Agile Adoption Model Lean-Kanban.

The project evaluation of the whole team and then a project portfolio is compiled and archived for the following project implementation.

The IT infrastructure transformation results in the branch office's SD-WAN implementation project see Figure 5. From the results of the project team's report on the performance of the SD-WAN implementation project, it has a good value, reaching 100% for the following project components, namely (1) project scope, (2) delivery increments, (3) and human resources. (4) While costs reach 98%, (5) the project implementation schedule gets 96% due to delays in completing four cities those experience problems in setting up network infrastructure from ISPs.

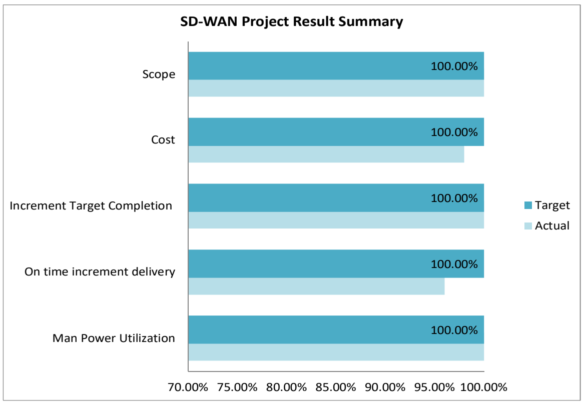


Fig. 7. Project Performance Result

The advantages of adopting an agile project management model lean and Kanban are as follows:

1. Lean by efficiency network operational, and investment cost practice SD-WAN solution.
2. Kanban board helps Project task monitoring, so all project teams and stakeholders can give feedback sooner.
3. It is facilitating project implementation has been accepted by users and stakeholders.
4. It minimizes waste for all resources in the project, costs, time, and human resources.
5. Assist in the implementation of projects more very flexible and need various adjustments.

Based on the results and discussions in this study, the agile adoption model in IT Infrastructure projects, especially the SD-WAN implementation, has a very effective and efficient impact on the organization. The IT project team's report shows with Kanban and Lean that completing the project for phase 1 in 2020 has reached on-time delivery 100%. All project tasks are well monitored through the Kanban board to minimize the risk of project delays. With WIP restrictions on each process, the project flow and tasks are controlled and carried out more quickly. This project had a cost-efficiency impact on network operation cost, which has decreased by 11.34% of total network expense.

Changes to the network topology with SD-WAN were carried out very well and smoothly in phase 1 of 2020 implementation due to the project risk mitigation arrange details. The changes made using the first SD-WAN installation scenario moved the old network to the new network during work breaks, making this activity not disturbing branch operations. Furthermore, with careful planning through a pilot project at the beginning of the proof of concept with the vendor, the SD-WAN implementation process runs smoothly.

V. CONCLUSIONS AND FUTURE WORK

A. Conclusions

This research aims to implement IT infrastructure agile project management using the Lean-Kanban approach. Lean helps achieve cost efficiency for network investment and operation. Kanban helps ensure the entire team is consistent, making it easier for all teams to monitor the project and understand its goals. Kanban can also be a solution IT

Infrastructure team to maintain flow during the transformation process as they have to ensure the IT Infrastructure project runs smoothly and existing operations are not interrupted. Based on this study's implications, IT Infrastructure at pharmaceutical companies, through complexity, can implement an agile project management approach for IT infrastructure transformation projects to accelerate the delivery process to internal businesses and increase company competitiveness due to accelerated changes according to the market and business needs.

B. Limitation and Future Work

This research's limitation is the specific adoption of Agile Lean and Kanban project management in IT infrastructure projects (SD-WAN) using the DSR method. Future research can further develop IT Infrastructure project applications, other agile project management and develop more complex iterative DSR models. This case study is specifically in organizations pharmaceutical companies in Indonesia. In future research, the next adoption model can use IT Infrastructure projects in other industrial fields suitable using Kanban and Lean.

REFERENCES

[1] V. Di Leo and P. Mckinsey, "Transforming IT infrastructure organizations using agile," pp. 1–9, 2018.

[2] T. Akhter and T. Akerlind, "Agile IT Infrastructure Transformation: A Case Study of a Nordic Incumbent Telco," 2018.

[3] R. Almeida, J. M. Teixeira, M. Mira, and P. Faroleiro, "A conceptual model for enterprise risk management management," vol. 32, no. 5, pp. 843–868, 2019, doi: 10.1108/JEIM-05-2018-0097.

[4] M. M. Stoddard, B. Gillis, and P. Cohn, "Agile Project Management in Libraries : Creating Collaborative, Resilient, Responsive Organizations Agile Project Management in Libraries : Creating Collaborative, Resilient, Responsive," *J. Libr. Adm.*, vol. 59, no. 5, pp. 492–511, 2019, DOI: 10.1080/01930826.2019.1616971.

[5] B. C. Kraft, "Management on Government Finance Projects," pp. 12–19, 2018.

[6] E. S. Hidalgo, "Adapting the Scrum framework for agile project management in science : a case study of a distributed research initiative," *Heliyon*, no. December 2018, p. e01447, 2019, DOI: 10.1016/j.heliyon.2019.e01447.

[7] V. Jim and P. Afonso, "Using Agile Project Management in the Design and Implementation of Activity-Based Costing Systems," pp. 1–24, 2020.

[8] A. Ashrafi, "How to market orientation contributes to innovation and market performance : the roles of business analytics and flexible IT infrastructure," no. April 2018, DOI: 10.1108/JBIM-05-2017-0109.

[9] R. Govindaraju, R. Akbar, and K. Suryadi, "IT Infrastructure Transformation and its Impact on IT Capabilities in the Cloud Computing Context," vol. 10, no. 2, pp. 395–405, 2018, DOI: 10.15676/ijeei.2018.10.2.14.

[10] J. Benitez and G. Ray, "IMPACT OF INFORMATION TECHNOLOGY INFRASTRUCTURE FLEXIBILITY ON MERGERS AND ACQUISITIONS 1," vol. 42, no. 1, pp. 25–43, 2018.

[11] J. Saltz, "Exploring Which Agile Principles Students Internalize When Using a Kanban Process Methodology," vol. 31, no. 1, pp. 51–61, 2020.

[12] E. Mircea, "Project Management using Agile Frameworks," vol. 19, no. 1, pp. 34–45, 2019.

[13] S. Gbadegeshin, "Lean Commercialization: A New Framework for Commercializing High Technologies," *Technol. Innov. Manag. Rev.*, vol. 8, no. 9, pp. 50–63, 2018, DOI: 10.22215/timereview/1186.

[14] G. Ming, L. Jin, J. Hai, and Z. Huiying, "A Design of SD-WAN Oriented Wide Area Network Access," pp. 174–177, 2020, DOI: 10.1109/CCNS50731.2020.00046.

[15] S. Troia, F. Sapienza, and L. Var, "On Deep Reinforcement Learning for Traffic Engineering in SD-WAN," vol. 8716, no. c, pp. 1–15, 2020, DOI: 10.1109/JSAC.2020.3041385.

[16] Y. Chen, Q. Wu, W. Zhang, and Q. Liu, "SD-WAN source route based on protocol-oblivious forwarding," *ACM Int. Conf. Proceeding Ser.*, pp. 69–73, 2018, DOI: 10.1145/3290480.3290486.

[17] P. Debois, "Agile infrastructure and operations: How infra-gile are you?," *Proc. - Agil. 2008 Conf.*, pp. 202–207, 2008, doi: 10.1109/Agile.2008.42.

[18] M. R. K. Mohammad R. Kabli, "Improve the Information Technology Infrastructure at Saudi Airlines (SAUDIA) by Adapting and Implementing Agile Project Management Methodology," *J. King Abdulaziz Univ. Eng. Sci.*, vol. 30, no. 2, pp. 25–34, 2019, DOI: 10.4197/eng.30-2.3.

[19] B. Sarlak, "Agile Methodology for Project/Process Management IT System Infrastructure," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, 2020, DOI: 10.1109/ICCCNT49239.2020.9225593.

[20] J. Braithwaite *et al.*, "Built to last ? The sustainability of healthcare system improvements, programs and interventions : a systematic integrative review," pp. 1–11, 2020, DOI: 10.1136/BMJ open-2019-036453.

[21] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, and P. Taylor, "Linked references are available on JSTOR for this article : A Design Science Research Methodology for Information Systems Research," vol. 24, no. 3, pp. 45–77, 2008, DOI: 10.2753/MIS0742-1222240302.

Dips Detection Techniques Discussion

Noureddine Brakta
Department of Eelectrical Engineering
LREEI, Faculty of Hydrocarbons
& Chemestry
University of Boumerdes
Boumerdes, Algeria
n.brakta@univ-boumerdes.dz

Omar Bendjeghaba
Department of Eelectrical Engineering
LREEI, Faculty of Hydrocarbons
& Chemestry
University of Boumerdes
Boumerdes, Algeria
bendjeghaba@univ-boumerdes.dz

Mohamed Yazid Zidani
Department of Electrical Engineering
Faculty of Technology
University of Mostefa Ben Boulaïd
Batna 2
Batna, Algeria
zidanikarim212@yahoo.fr

***Abstract*— In today's industrial world, renewable energy represents 16% of the total produced and consumed energy. Wind energy is gaining popularity and imposing bigger footprint in the natural sources of electricity and plays a key building block for economic recovery from the impact of COVID-19 according to the Global wind Energy Council.**

Yet, because its presence in the energy resource share is below 5%, since most of the wind farms use Doubly Fed Induction Generator (DFIG) it will suffer from grids faults such as symmetric dips. The scary part of the dips is that it will increase the currents at the back-to-back converters and even when dealing with them oscillation might destroy the turbine and generator assembly. In this paper we will review techniques used to detect the dips and override the heavy increase of the current using a crowbar, we will concentrate only on the detection and compare the resulting parameters while keeping the controllers and the value of the crowbar resistance unchanged

Keywords—DFIG, PI Controller, Voltage Dips, Crowbar, Grid Connection, Energy Conversion, Vector Control, Distributed Generation, Wind Turbines

I. Introduction

Distributed generation (DG) is the only answer to the growth of population, and the relocation of noisy industrial zones outside cities. This increased dramatically the challenges of producers of electricity in terms of cost of production, environment cost (usually emission), and reliability. The structure of the grid offers by itself a solution by itself as with DGs we can isolate regions to increase local reliability of the systems. Reliability of the system related to the adoption of wind energy is seen from two different perspective: does it offer enough energy with less cost and during the time, and weather it is immune against grid faults specially voltage dips.

In this paper we will talk about the effect of voltage dips on the Wind Turbines (WT) build around Doubly Fed Induction Generator, which is the most used machine in such field. [1,2].

Sudden drops of voltages in the grid are described as Voltage Dips, they are usually caused by faults occurring in the grid [1], there are two types of voltage dips symmetric and asymmetric voltage dips. We will be only talking about symmetric dips.

Due to the impact on the various connected DGs, several studies have been conducted to detect the dip and to limit its effect on the DFIG connected to the grid. Different scholars from diversity disciplines worked together to solve this issue mainly limit the effect on DFIG's which direct maintenance and down time cost cut. Thus, we will find people form electrical engineering working with artificial intelligence experts to predict behavior of the machine during dips and deploy necessary counter measure.

A. Stanisavljevic, et al in their work reference [4] described clearly the threat towards the DGs degenerated from voltage sags (dips), and then presented an overview of all the techniques used when the paper was published. They mainly worked on the detection of the fault by the grid administration. Once the fault is detected different generation unites are informed accordingly. These techniques are either based on analysis of harmonics, RMS value, or hybrid methods. Then they have established a Key Performance Index (KPI) to define the effectiveness of each method. Since we will not work on the grid side rather, we will deal with DFIG itself we will consider on the simplest technique base on RMS value of the nominal voltage and because it is the most used [5]

The early a voltage dips are detected the best it will be and multiple stages are always better, for large grids the dip might be seen by the grid operator (or control room) but before the DGs are notified the fault is fixed or disappeared as the cause disappeared, this justifies the need to have a at least a detection process at the DG level (DFIG in our case), this was elaborated in reference [6] a cleaver technique was used to coordinate different individual converts of each DFIG in the plant in order to regulate and voltage and respond rapidly to voltage dips.

Furthermore, work described in reference [7] finds a solution using combination of power electronics and apply crowbars in different circuits depending on the classification of the fault. The classification of the fault is achieved using "Adaptive Neuro-Fuzzy Inference System."

Other research found the solution in eliminating or reducing the fault magnitude by introducing dynamic voltage restorer (DVR) in attempt to enhance the fault ride-through of a DFIG based WT. the control is also based on "novel hybrid genetic algorithm optimized Elman neural network" [8], the cured system will help then to cure the grid.

Even though all the doors have been knocked, there would each time be some of the power electronics that needs to be called for help. To this end earlier work presented in reference [10] is revised thoroughly, we will use simple logic parameters and keep the crowbar resistance value fixed while we will be playing on the triggering parameters of the switching circuit to protect the IGBTs of the converters, while meeting the grid code of operation.

The introduction is followed by section II describing fault detection at the wind turbine level, then we describe the whole electrical configuration of the DFIG with different controllers and crowbar connection, followed by simulation and results we describe thoroughly the different control signals. Finally,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

we wrap up the findings by a description of the results versus the targeted objectives and future work.

II. Fault Detection

To achieve voltage Dips counter measure and save our DFIG from the voltage dip consequence we will need the deployment of a crowbar and a rectifier. During the voltage dip both rotor current and the dc-link voltage will increase and hence it will damage the rotor side converter components. Once the control switch CB is closed the current will be diverted towards the crowbar resistance meanwhile the RSC will have to be disabled so the current will not flow through.

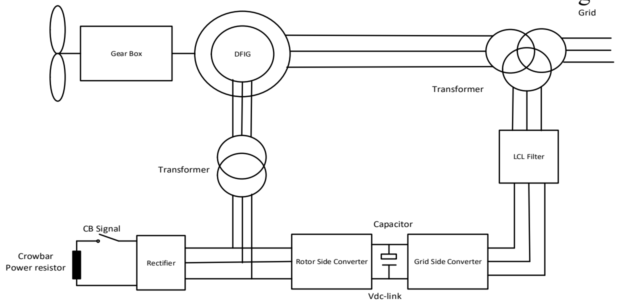


Fig. 1. DFIG with crowbar, RSC, GSC and power rectifier

We need also to detect the fault to decide when to apply a supportive reactive power from both Grid side converter and rotor side converter. We choose a simple method based on threshold RMS value of the nominal value of V_s we will consider anything below 90% as a voltage dip [5] this is considered as adequate choice as we will be dealing only with symmetric voltage dips. To increase the freedom of the fault detection and to simulate the possible delays of the other systems we can insert a delay after the fault is detected and another delay after the fault is cleared.

III. Electrical configuration of a DFIM based wind generation

We will consider a DFIG connected directly to the grid while the rotor is fed from rotor side converter, the last is also fed by grid side converter generating necessary dc-link.

This justified by its simplicity and its ability to deal with generated power as explained deeply on reference [1] [2]

We will be using PI controllers we will not go through the justification of such technique as it is well described in almost all previous work and we will not add anything extra to it. Yet we will have to find what are the limits of the crowbar, and what are limits of the dc-link and the rotor current to set our simulation to work within accepted operation range.

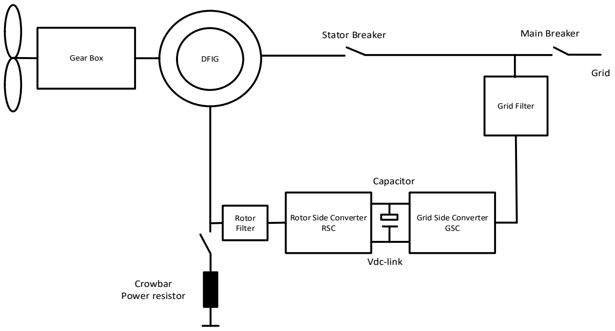


Fig. 2. DFIG Electrical configuration [2]

During a voltage dip the GSC will keep controlling the current, while the stator disturbance will be carried out to the rotor causing an overcurrent and overvoltage observed at the dc-link capacitor, this is due to the RSC currents control lose. This means that the rotor converter now cannot control the generator and has been restricted [2]

The aim of the crowbar is to short circuit the rotor at the right time with right duration to achieve a full protection to the system. Its minimum and maximum values are well developed in reference [10]

$$R_{cbmin} = \frac{\omega_r}{I_{safe}} \sqrt{\left(\frac{V_s}{\omega_1}\right)^2 - (L_\sigma I_{safe})^2} \qquad (1)$$

Where ω_1 , ω_r are synchronous and rotor angular frequencies, V_s is the stator voltage, R_{cb} crowbar resistor, I_{safe} is the upper extreme current at the rotor converter that will not damage it , finally L_σ is the sum of the leakage inductance of rotor and stator.

$$R_{cbmax} = \frac{V_{dc\_safe}\omega_r L_\sigma}{\sqrt{3(V_s\omega_r/\omega_1)^2 - V_{dc\_safe}^2}} \qquad (1)$$

$V_{dc_safe}$is the highest safe DC-link

IV. Simulation and Results

In this work we have adopted all justifications described in the literature to give renaissance to simple logic described in flow chart figure 3.

Case1:

We will use hysteresis compare described in paper of Y. Ling and X. Cai.

We built logical blocks to generate the different control signals as follows:

- CB: is the control signal for the crowbar.
- CE: is the control signal for the RSC enable or disable
- ED: is to enable the I_{dr} component to generate the supportive reactive power from the RCS as recommended by most of early references and literature.
- DI: is set the reference torque and the Pref (found by the MPPT) to zero to avoid saturating the PI feedback loop.
- EQ: is the control for the supportive reactive power on the GSC, it is mainly “true” during all the time on which the fault is detected.

The flow chart on figure 3 explains all connections with different control signals and how each one is active, again this just an interpretation of the logical sequence presented in most of the papers and books referenced here.

Case 2:

In the second experiment we have only 3 main control signals the remaining 2 are just a replay of the other two signals:

- CB: is the control signal for the crowbar.

- CE: is the control signal for the RSC enable or disable
- ED: is to enable the I_{dr} component to generate the supportive reactive power from the RCS as per earlier work and it is set to be identical to CB signal
- DI: is set the reference torque and the Pref (found by the MPPT) to zero to avoid saturating the PI feedback loop it equals the CE signal
- EQ: is the control for the supportive reactive power on the GSC, it is mainly "true" during all the time on which the fault is detected.

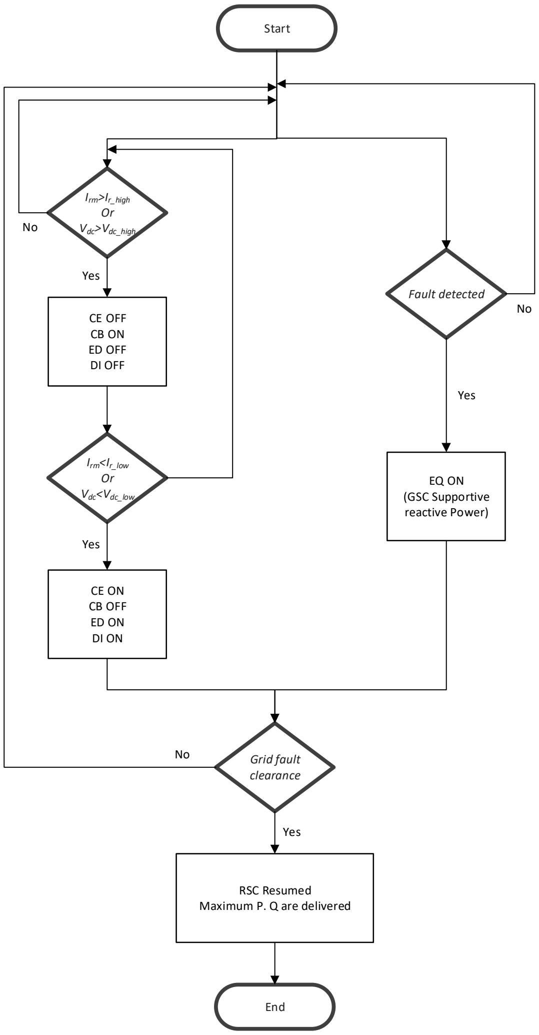

Start
$I_{rm}>I_{r\_high}$ Or $V_{dc}>V_{dc\_high}$
No
Yes
CE OFF
CB ON
ED OFF
DI OFF
$I_{rm}<I_{r\_low}$ Or $V_{dc}<V_{dc\_low}$
Yes
CE ON
CB OFF
ED ON
DI ON
Fault detected
No
Yes
EQ ON
(GSC Supportive reactive Power)
Grid fault clearance
No
Yes
RSC Resumed
Maximum P. Q are delivered
End

Fig. 3. Flow chart of propoused logical controls-case 1 during voltage dips

The flow chart presented on figure 4 depicts this rig up and explains the interconnection between the different signals.

The resistance of the crowbar is set to a value within the range $[R_{cbmin}, R_{cbmax}]$ calculated after setting the

$$I_{safe} = I_{th\_high} = 1.6\,pu$$

$$I_{th\_low} = 1 \quad pu$$

$$V_{dc\_safe} = V_{dc\_high} = V_{dc-no} \ \ +100 \text{ Volts}$$

$$V_{dc\_low} = V_{dc-nom}$$

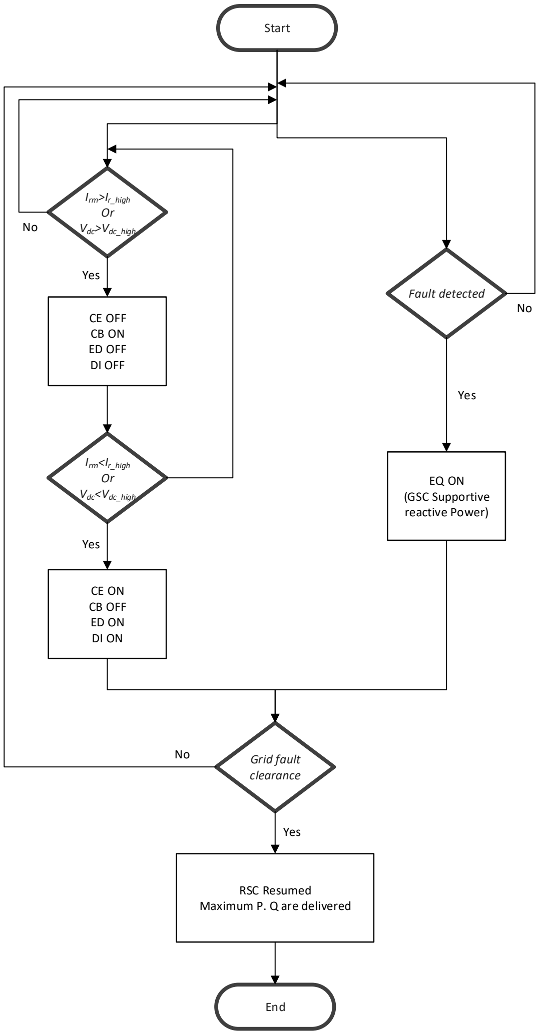

Start
$I_{rm}>I_{r\_high}$ Or $V_{dc}>V_{dc\_high}$
No
Yes
CE OFF
CB ON
ED OFF
DI OFF
$I_{rm}<I_{r\_high}$ Or $V_{dc}<V_{dc\_high}$
Yes
CE ON
CB OFF
ED ON
DI ON
Fault detected
No
Yes
EQ ON
(GSC Supportive reactive Power)
Grid fault clearance
No
Yes
RSC Resumed
Maximum P. Q are delivered
End

Fig. 4. Flow chart of propoused logical controls-case2 during voltage dips

Figure 5 shows the current flowing from the RSC, it shows that both techniques limit the current

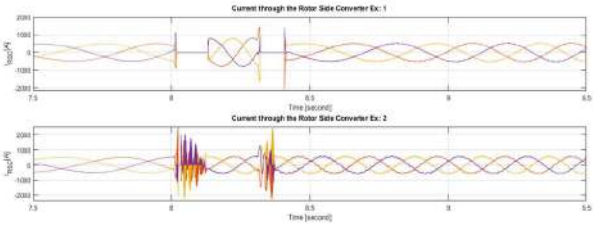


Fig. 5. Current flowing from or to the rotor convert side expirment 2 is done without fault detection

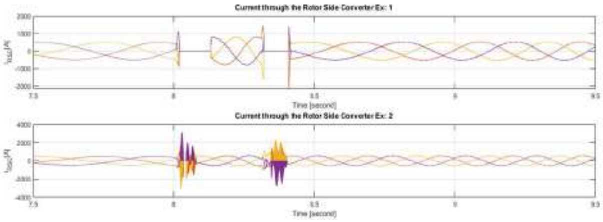


Fig. 6. Current flowing from or to the rotor convert side expirment 2 is done with fault detection

Figure 7 shows the Simulink® model created on the simulation software MATLAB® where the different values of the machine are shown in Table 1.

The different measured parameters on the machine for both cases are presented in figure 8. 1st method shows better control in terms of currents, Torque and represent less oscillation and better timing for stability of the machine after the fault, yet the power inserted to the grid remains higher than the one generated by the comparison technique on case 2

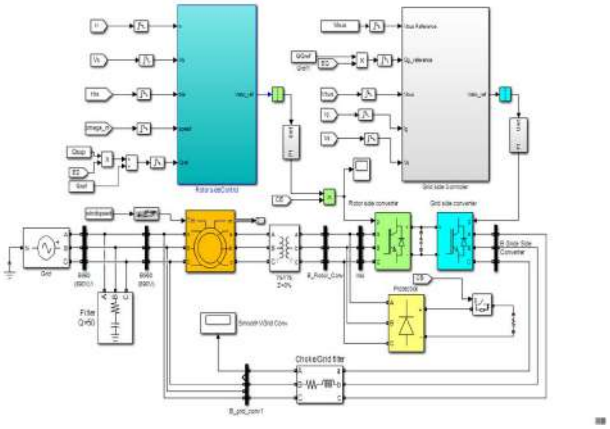
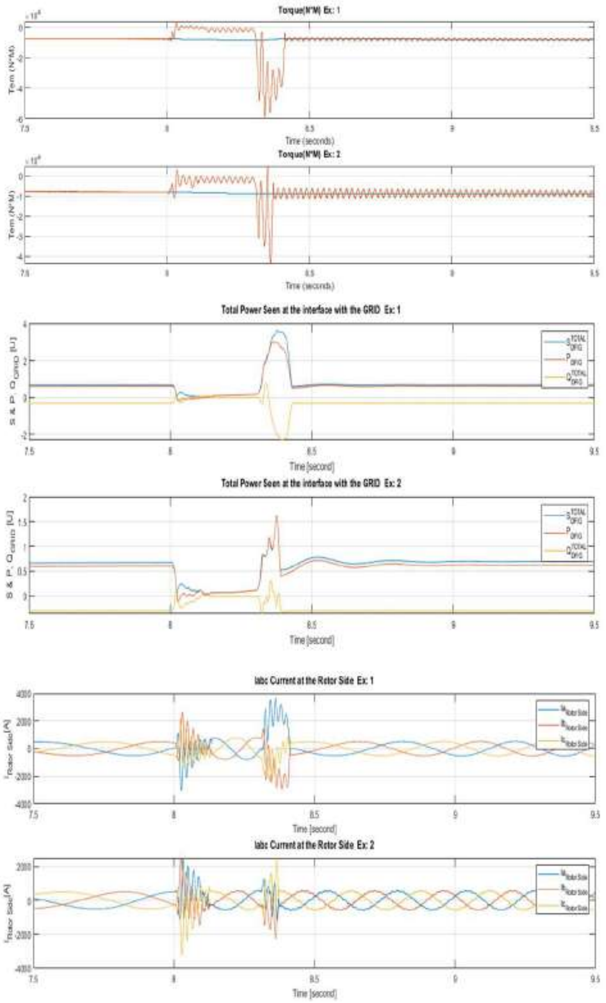


Fig. 7. Simulation model on MATLAB®



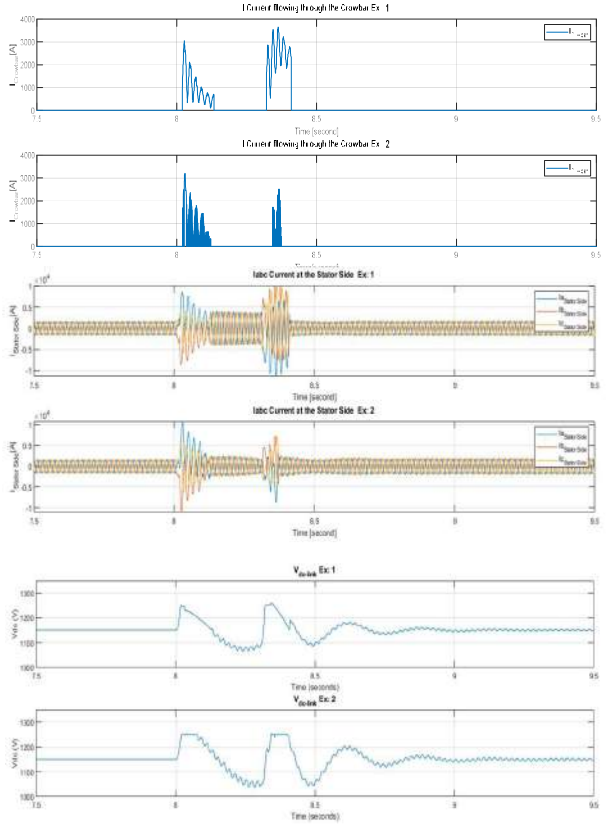


Fig. 8. Results of the simulation

V. Conclusion

The voltage dips are major issue that might kill the electricity production based on DFIG and wind turbines. Scientists tried to solve the issue using techniques that are far from power electronics or at least ignore the effect of power electronics in the whole system design, in our work we tried to polish legacy power electronics and bring it back using only simple logical methods and introducing different control signals we counted 5 in our proposed revision of crowbar technique. Essentially the Dc-link is better controlled in case 2, and might happen to go little bit beyond the upper limit in the first method or case 1, yet the first method is better in terms it offers the possibility to invest in simple logic as it offers lot of parameters that can reduce the power or increase it depending on the KPI required by the grid operator, it opens also the door for applying optimization algorithms to enhance it according to the requirements or constraints of the operator. Each of the graphs shared in this work can be shaped by this parameters and fair resolution between all the requirements is always possible.

Another side finding occurred when generating a reactive power from the GSC in the first experiment shows less current flowing in the inverter compared to the one flowing when running case 2 with the same requirement that is providing supportive reactive power from the GSC while the RSC is off (disabled and restricted)

We believe that the rig up of case 2 is much easier to implement but we doubt that such high frequency switching will not cause issue in real live converters. Also, we simulated the action of IBGTs but we did not simulate the heat dissipation.

The rig up of the whole system is much closer to the real-world system as we are not using average model to replace the PWM and the converters instead we are using a model of IBGTs and the closest possible model of PWM.

TABLE I. MACHINE PARAMETERS

Parameter	***Value***
Rated stator voltage Vs	690 V
Rated stator power Ps	2 MW
Rated stator current Is	1760 A
Stator frequency f	50 Hz
Rated torque Tem	12732 N.M
Rated rotor voltage	2070 V
Leakage inductance (stator and rotor)	0.087e-3 H
Magnetizing / Mutual inductance	2.5e-3 H
Stator inductance	Ls = Lm + Lsi
Rotor inductance	Lr = Lm + Lsi
Rotor resistance referred to stator Rr	2.9e-3 ohm
DC de bus voltage referred to stator Vdc	1150 V

REFERENCES

[1] D. H. Abu-Rub, M. Malinowski, and K. Al-Haddad, *Power electronics for renewable energy systems, transportation and industrial applications*. Hoboken, NJ: Wiley-Blackwell, 2014.

[2] G. Abad, J. Lopez, M. Rodriguez, L. Marroyo, and G. Iwanski, *Doubly fed induction machine: Modeling and control for wind energy generation*. New York, NY: Wiley-IEEE Press, 2011.

[3] “Green Recovery Data & analysis,” *Gwec.net*, 08-Dec-2020. [Online]. Available: https://gwec.net/green-recovery-data-analysis/. [Accessed: 10-Nov-2020]

[4] A. M. Stanisavljevic, V. A. Katic, B. P. Dumnic, and B. P. Popadic, “Overview of voltage dips detection analysis methods,” in 2017 *International Symposium on Power Electronics (Ee)*, 2017.

[5] M. H. Bollen, *Understanding Power Quality Problems*. IEEE, 1999.

[6] J. Kim, E. Muljadi, J.-W. Park, and Y. C. Kang, “Flexible IQ–V scheme of a DFIG for rapid voltage regulation of a wind power plant,” *IEEE Trans. Ind. Electron*., vol. 64, no. 11, pp. 8832–8842, 2017.

[7] O. Noureldeen and I. Hamdan, “A novel controllable crowbar based on fault type protection technique for DFIG wind energy conversion system using adaptive neuro-fuzzy inference system,” *Prot. control mod. power syst*., vol. 3, no. 1, 2018.

[8] Sitharthan, Sundarabalan, Devabalaji, S. K. Nataraj, and Karthikeyan, “Improved fault ride through capability of DFIG-wind turbines using customized dynamic voltage restorer,” *Sustain. Cities Soc*., vol. 39, pp. 114–125, 2018.

[9] Y. Ling and X. Cai, “Rotor current dynamics of doubly fed induction generators during grid voltage dip and rise,” *Int. j. electr. power energy syst*., vol. 44, no. 1, pp. 17–24, 2013.

[10] Y. Ren and W. Zhang, “A novel control strategy of an active crowbar for DFIG-based wind turbine during grid faults,” in 2011 IEEE International Electric Machines & Drives Conference (IEMDC), 2011..

[11] G. Pannell, D. J. Atkinson, and B. Zahawi, “Minimum-threshold crowbar for a fault-ride-through grid-code-compliant DFIG wind turbine,” *IEEE trans. energy convers.*, vol. 25, no. 3, pp. 750–759, 2010

[12] “Documentation Home,” *Mathworks.com*. [Online]. Available: https://www.mathworks.com/help/. [Accessed: 15-Sep-2020].

[13] *Matlab* (2013b), *Mathworks.com*.

Smart Cooling System for Milk Transportation in Rural Areas

Tankiso A. Komako[1], Ashleigh K. Townsend[1], Immanuel N. Jiya[2] and Rupert Gouws[1]
[1]*School of Electrical, Electronic and Computer Engineering, North West University*
Potchefstroom 2520, South Africa
[2]*Department of Engineering and Science, University of Agder,*
Grimstad 4879, Norway
ashleighktownsend2@gmail.com

***Abstract*—In the dairy industry, road milk tankers transport milk from one location to another. The milk inside the tanker needs to be kept between 3-5 °C to ensure that the quality of milk is always preserved. The tanker needs to be kept running at all time with the case of sufficient energy being continuously supplied to the cooling unit. The source of energy normally used for this application is a typical generator which needs fuel to operate. This is expensive and is not environmentally friendly. To address this problem, generator as a source of energy needs to be replaced by solar energy to lower the costs associated with cooling of the tanker. In this research a small scale solar powered intelligent cooling system was developed. This system was designed to make use of thermoelectric cooler as a viable cooling unit and operating it intelligently with a programmable logic controller. The system was designed in a way that it can also display the volume of milk inside the tanker and the system power consumption.**

Keywords—PLC, thermoelectric cooler, milk tanker, milk cooling, TECM

I. Introduction

Road milk tankers, transport raw milk from dairy farms to dairy processing factories where the milk is processed and utilized to produce dairy products like cheese, yoghurt etc. [1]. To ensure that the quality of the milk from milked cows is maintained during transportation, the temperature of the milk is kept between 3-5 °C [2]. This implies that the milk tank, during transportation, needs to be kept running by ensuring that sufficient charge is provided to the cooling unit of the tank. Road milk tank trucks make use of fuel-based energy sources like DC generator to cool down and control the temperature of the milk during transportation. The fuel-based energy machine is usually kept always running to provide sufficient charge to the cooling unit. The process to maintain the temperature of the milk is costly due to the fuel being used as a source of energy [3].

Rapid increase in price for fuel can have devastating effects on the freight management companies, particularly, road milk trucks. The fuel price increase can also decrease the profit that dairy processing plants make [4]. The use of renewable energy sources could help reduce the high fuel costs associated with cooling of the milk tank. The petrol cost associated with the cooling of the milk tank is high and this causes milk trucking/logistics companies not to make maximum profit due to petrol expenses. The fluctuation of petrol price constantly has evolving effects on the milk price rates [5]. The petrol as a source of energy is not environmentally free, and it contributes to global warming. Environmentally friendly and renewable energy powered intelligent milk tanks need to be developed in this case to lower the fuel costs and fuel related global warming effects, associated with cooling of the milk tank [6].

Heat pumps are very efficient for heating and cooling systems, they can significantly reduce the energy cost. They keep energy cost as low as possible which is the most important thing for any industry. The performance of the heat pump can also be described by COP. This method of cooling uses the refrigeration evaporator pipes or specially designed flat pillow-shaped plates that are attached to the bottom outer side of the tank containing milk. The heat of the milk is transferred through the walls of the tank to the refrigerant. The refrigerant then absorbs the heat of the milk inside the tank. The milk cooling process takes place inside the tank and normally an agitator is used to continuously mix the milk to prevent it from mixing at the bottom of the tank. Ice water cooling is another method used for the cooling of liquids in bulk tanks. The small compressors are normally used to build up a reserve of 'coolth' in the form of ice over a long period of time. The ice is normally formed on the copper pipes through which the refrigerant passes and the ice is surrounded by the water close to freezing point [7]. The water generally surrounds the outside of the tank while agitator is used to mix the milk inside the tank. The advantage of this method is that the ice can be made anytime and not only during the time when content must be cooled. This method is not as effective as a direct expansion method because the heat must be transferred several times (from ice, to water, to the milk). The other disadvantage of this method is if the ice has melted, it must be made all over resulting in high electricity usage.

In this research a smart cooling system is proposed based on the thermoelectric cooling module (TECM). A similar approach has been proposed in [8]. TECM, also known as Peltier cooler, because it is based on the Peltier effect is a semiconductor-based electronic component that functions by removing the heat from one side to the other when a low DC power source (12 V) is applied. One side of the module is cooled while the opposite side is heated [9]. The working principle of the TECM is shown in [10]. The heat flow out of the module has a significant impact on the overall system performance. The larger temperature difference that the module must work against, the smaller the quantity of the heat removed for the cold side. The performance of TECM and any other cooling device can be compared by determining their coefficient of performance (COP) [11]. The advantage that a TECM has over a compressor is that they have no moving parts and do not require the use of chlorofluorocarbons. This makes them

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

perfect for application since they are environmentally friendly [12].

II. Proposed Design

Fig. 1 shows the overview and operational analysis of the proposed system which describes the actions that the end-user will have to perform in order to operate and monitor the system.

The end user will at first, must switch the system on and check the alarm indications - whether all the components of the system are connected and functional. If the alarm system indicates certain components of the system are disconnected or malfunctioning, the end user will have to replace or connect them to the system and monitor the alarm system again to verify that they all connected. When the alarm system indicates positive results, the end user will have to operate the controller with the aim to check/ monitor the temperature and the level of the milk. The end user will also monitor the power being consumed by the system. After the process of operating the system, the end-user will then have to deliver the cooled milk inside the tanker to respective dairy processing plants. The control unit will have the buttons and screen through which the end-user will operate to view the status (level and temperature) of the milk inside the tanker.

The TECM from laird technology/TE technology Inc., usually comes with heat sink and fan already assembled, depending on the application for which they are needed. The cooling capacity of the thermoelectric cooling system is dependent on the operating temperature of the content desired for cooling, quantity of thermoelectric modules used, the type of thermoelectric module used and the applied power. The Marlow guide [13], will be used in this case to select the applicable TECM which will be compatible for the research. Firstly, the heat load capacity of the milk needs to be determined to select the applicable TECM which will be able to provide cooling capabilities. As such, the selected TECM must be able to handle the temperature difference of 34 °C. The cooling capacity of the milk, in this case, is determined to select the right TECM for the application. Equation (1) can be used to estimate the heat load/capacity of liquids [14].

$$Q_C = mC\Delta T\Delta t = \rho VC\Delta T \qquad (1)$$

Where, m is cooled substance mass (kg), C is cooled substance specific heat capacity (J/(kgK)), ΔT is cooling temperature difference (K), ρ is cooled substance density (kg/L), V is Volume of the cooled substance (dm^3) and Δt is cooling time (s). Although our research was based on milk cooling, water was used as a viable substance to cool.

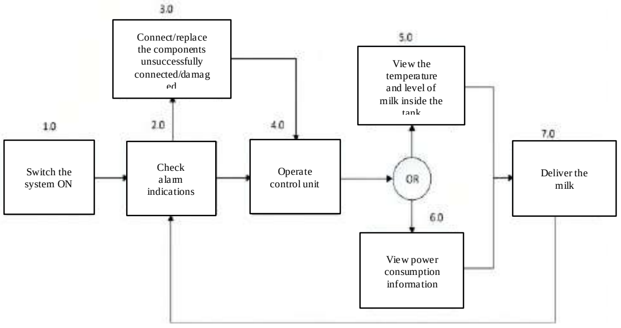

3.0
Connect/replace the components unsuccessfully connected/damag ed
5.0
View the temperature and level of milk inside the tank
1.0
Switch the system ON
2.0
Check alarm indications
4.0
Operate control unit
OR
6.0
View power consumption information
7.0
Deliver the milk

Fig. 1. Operational flow diagram of proposed system.

The density of the water was 1 kg/L, its heat capacity was 4.2 KJ/(kg*K). The heat load capacity of 50 L of water is: 661.1 W. Considering the safety factor of 20%, cooling capacity of 50 L of water/milk is: 1.2 *661.1 = 793.3 W. The TECMs chosen, must be able to pump this amount of energy to cool 50 L of milk to 4 °C within the estimated time of 3 hours.

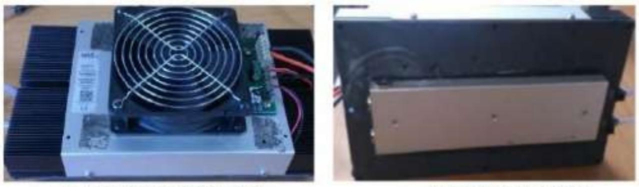
Table I, shows the heat load for different levels of water. It is evident that the heat load doubles whenever the quantity of the milk doubles. The heat loads calculated using Equation (1) is only an estimate and something to work on regarding choosing the right TECM for the research. The liquid cooled TECM (Laird Technologies LA-115-24-02-0710) that was used is shown in Fig. 2.

This module has the voltage rating of 24 V, current rating of 5.8 A and power rating of 139 W. This Single Assembly, from our heat load capacity calculation of the water/milk, would be able to cool 8 L of milk/water to a temperature of 4 °C within the period of 3 hours. Approximately 7 of these TECM assemblies will be able to cool 50 L of water/milk. These approximations are only theoretical, and they do not account for the shape of the tanker and all other factors which might affect cooling process of the tanker.

Simulation was done later in SolidWorks® to estimate the cooling capacity that is needed for 50 L of milk. Simulation results provided more accurate approximations because in this case other factors which can influence the cooling process were considered. The eTape liquid level sensor from the Milone Technologies Inc was selected for this research. This sensor output was used by the controller to calculate/determine the actual capacity of milk inside the tanker in L. The controller also used eTape results to determine the temperature sensors that must be activated and considered to calculate and ensure that accurate temperature results of the milk are displayed. eTape was placed in a vertical position inside the tanker and was in contact with the liquid. The eTape used the radius/diameter measurement of the tanker which was then used altogether with the tanker dimensions to determine and display the liters of the liquid inside the tanker.

Table I. Heat Loads for Different Quantities of Water

Liters (L)	Q_c heat load (W)
1.0	15.86
3.0	47.80
5.0	79.33
10	159.67
20.0	317.33
30.0	476.00
40.0	634.67
50.0	793.33



a) Heat disposal side of Available TECM

b) Liquid cooling assembly side

Fig. 2.Utilized liquid cooled TECM.

A thermocouple was selected as the temperature sensor to be used for the research. Three thermocouples were placed at different levels inside the tanker and their output was connected to the controller. The controller was programmed to only display the average temperature sensed by the thermocouples in contact with the liquid. One thermocouple was used to sense the temperature surrounding the tanker. Four thermocouples were needed for this for the research. The thermocouple sensor had the requirement of being able to measure the minimum temperature of -5 °C and maximum of 40 °C, as the temperature ranges of the liquid sensed were within this range. The controller also used the imbalance of 10% (within the sensors placed inside the tanker) to activate the mixer to ensure that temperature of the milk was distributed and accurately displayed. Type K thermocouple sensor from RS-pro was used. This industrial manufactured sensor has a temperature range of -50 to +400 °C [54], which was viable for the research.

The alarm system for the research consisted of light emitting diodes (LEDs) which indicated the status of the system components such as photovoltaic (PV) (1st Power Supply), the battery (2nd power supply) and the state of the mixing unit. The green LED was used for indicating the connection of the PV panel to the system, yellow LED was used for indicating the presence of the battery in the system and lastly the blue LED and red LEDs were used for indicating the status of the mixing unit. The blue LED was activated when the mixing unit was in operation and the red LED was used to indicate the off state of the mixing unit. The operator (driver) was able to view the status of the components and able to make informed decisions thereafter. The alarm system also indicated whether the control system was supplied through the PV or the battery and this was indicated by means of green and yellow LEDs, respectively.

This controller was used to implement the temperature control unit and energy management unit of the system. The PLC read the information, gathered through the sensors discussed, and took control steps to meet the specifications of the research, such as maintaining temperature of the liquid between 3-4 °C. In the temperature control unit, the PLC used the temperature sensors in contact with the fluid and displayed their average sensed results. The PLC required level measurements from eTape to decide which temperature sensor to consider. In the energy management unit, the PLC decided on the energy source that needed to be used depending on weather conditions. The PV was the primary source while the battery was the secondary source. The battery was utilized when there was insufficient energy from the PV due to weather conditions. The PV charged the battery and supplied the system when there was sufficient energy during sunny days. The PLC also managed the alarm system of the research by sending the switching instructions to the LEDs depending on the input information it received from the system's components.

Fig. 3 shows the suggested control flow diagram that was implemented. In this flow diagram, the system first determines the amount of milk that is inside the tanker. The system uses the minimum level (L_1) as the level at which the cooling occurs, as to avoid wasting energy during the time when the tanker was empty. TECM started to be activated when the water level was above the minimum level.

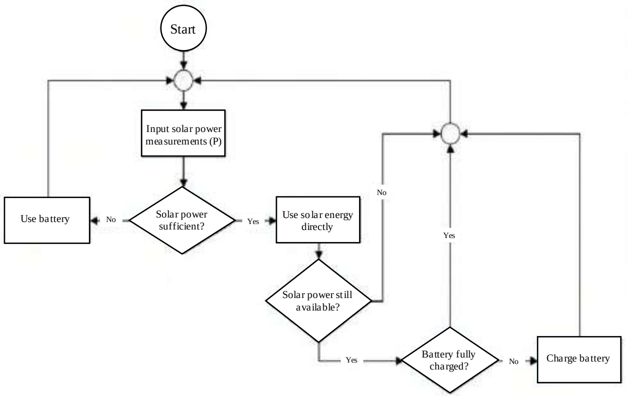


Fig. 3. Implemented control flow diagram.

When the measured liquid level was between the minimum level and the median level (L_2), only the first thermocouple (T_1) measurement was considered. T_1 and T_2 were considered when the water level was between L_2 and L_3 (maximum level).

T_3 together with T_1 and T_2 were considered when the liquid level was above L_3. The average temperature (T_{avg}) sensed by the thermocouples, was calculated together with their temperature deviation (T_{dev}). When the average temperature of the liquid was greater than 4 °C, the cooling units continued to be in activation state until the average temperature was less than 4 °C. The mixer was activated when the temperature deviation was greater than 10% and deactivated when the temperature deviation was less than 10%.

The energy management control flow (shown in Fig. 4) was implemented on the PLC controller. Table II summarizes the abbreviations used in Fig. 4.

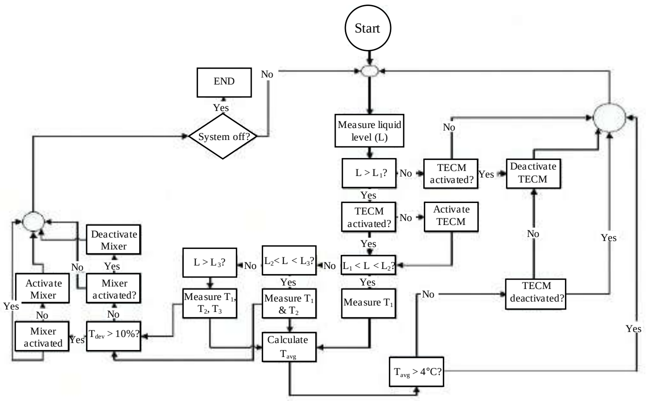


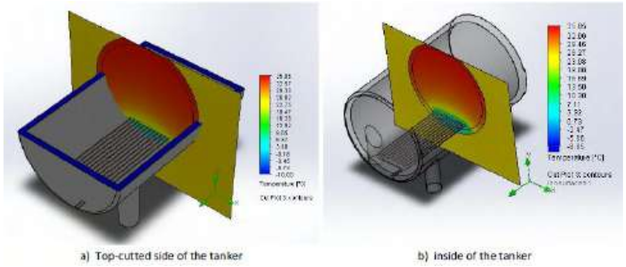
Fig. 4. Energy management control flow diagram.

TABLE II. SUMMARY OF ABBREVIATIONS USED IN FIG.4

Abbreviation	Description
L	Measured liquid level
L_1	Minimum liquid level of the tanker
L_2	Medium liquid level of the tanker
L_3	Upper Liquid level of the tanker
T_1	Temperature sensor placed at minimum liquid level
T_2	Temperature sensor placed at medium liquid level sensor
T_3	Temperature sensor placed at the upper liquid level
TECM	Thermoelectric cooling unit
Mixer	Agitator
T_{avg} & T_{dev}	Average temperature and temperature deviation

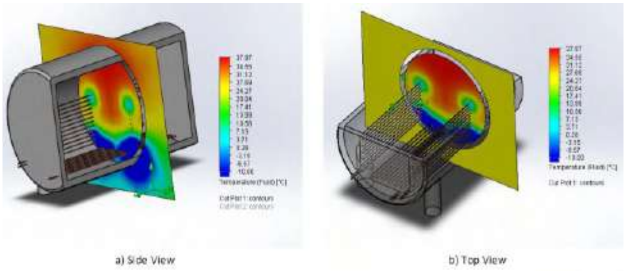
In the flow diagram, it is seen that the energy of the PV was first determined/measured and decided if it's enough to power the system. This required both the voltage sensor and current sensor be implemented to determine and monitor the power from the PV. During sunny days, the energy of the PV is likely to be enough to power the system and simultaneously charge the battery. During the night or bad weather conditions, the energy from PV is likely insufficient to power the system directly, hence a battery was needed as a backup source. The power measurement from the PV allowed the controller to determine which power source to use between the PV and the battery. The voltage sensor (voltage divider) was also used to sense the voltage capacity of the battery to determine or monitor its energy capability. This sensor was connected to the analog input of the PLC to allow the implementation of flow diagram shown in Fig. 3. Since there three sensors were used to implement the energy management unit, three analog inputs of the PLC were used for this purpose.

For current sensing measurements, the shunt resistor, together with the current sense amplifier, was used to take the current measurements from the PV. The power of the PV was then determined from the current and voltage measurement of the PV. This power calculation was then used to determine if the PV is able to supply sufficient power or not. Fig. 5-Fig. 7 shows the temperature analysis of the tanker.


a) Top-cutted side of the tanker

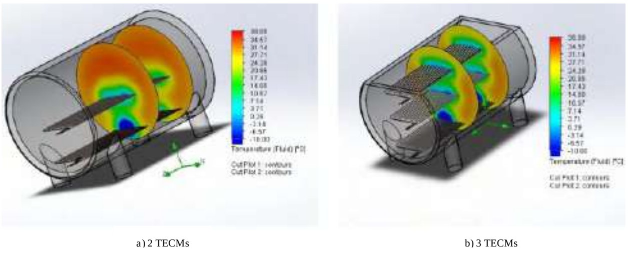
b) Inside of the tanker

Fig. 5. Temperature analysis with one TECM placed at lowest level in tanker.


a) Side View

b) Top View

Fig. 6. Temperature analysis with three TECMs placed along sides of the tanker.


a) 2 TECMs

b) 3 TECMs

Fig. 7. Temperature analysis with TECMs placed parallel to bottom of tanker.

The TECM was operated outside the tanker with copper coils placed inside, for direct contact with the milk/water. The purpose thereof, was to see how the temperature was distributed within the milk upon the placement of the copper coil to the relevant place inside the tanker. Only one TECM was available for the research and the SolidWorks® simulation was done to investigate the effect of this module inside the tanker. The minimum temperature of the provided cooling module was -10 °C. Ethanol was chosen as the circulating liquid which will flowed through the heat exchanger of the TECM. In the simulation, the temperature of the ethanol entering the copper coil was set at -10 °C. Fig. 5 shows the temperature results of our analysis. It was considered that the coils were placed at the minimum level of the tanker to allow cooling when the liquid level is at that point.

The surrounding temperature of the tanker was set at 25 °C since this was the temperature regarded as room temperature and the water was used as a liquid to be cooled with an initial temperature of 38 °C. In Fig. 5, the copper coil only managed to cool the liquid near it and as such, the agitator was needed to mix the liquid and ensure that the temperature of the liquid was well distributed. This setup of the tanker required that the agitator be placed in the parallel position to the tanker to allow the movement of the cooled liquid to mix with the hot liquid as seen in Fig. 5(a).

The thermocouples for this scenario, had to be placed at different levels of the tanker. One thermocouple was placed at the bottom level of tanker, the second thermocouple was placed in the middle of the tanker and the third thermocouple in the top level of the tanker. The fourth thermocouple sensor was placed outside of the tanker to read the outside temperature. This setup in Fig. 5, was implemented as there was only one available TECM. The tests on the tanker were conducted to evaluate the cooling capability of this cooling module.

Fig. 6 shows the thermal analysis of the tanker when there is multiple liquid cooled thermoelectric cooling modules. The position placement of three TECMs was investigated in this regard, to determine how the liquid was cooled in these scenarios. In Fig. 6, the tanker was connected to three TECMs whose coils are placed perpendicular to each other. This set up of coil placements is not proper since the temperature at the top is not cooled. The liquid at the bottom of the tank will also freeze due to high concentration of cooling happening at the bottom. Hence the choice of placing the cooling coils in this manner is not proper. In Fig. 7, the three TECM copper coils are placed in a parallel in accordance with the levels of the tank. The cooling in this case happens much better that in Fig. 6. The temperature of the water in Fig. 7 is mostly seen to be around 3-7 °C. This proves the setup in Fig. 7 is better than the setup in Fig. 6.

III. Results

The subsystems discussed above were all integrated to realize the intelligent solar powered TECM road milk tanker shown in Fig. 8. This figure shows the setup of the research where 8.91 L of water was inside the tanker as indicate on the LCD screen. The LCD screen also displays power consumption. The PLC recorded the temperature to be 7℃ inside the tanker and the surrounding temperature was 18 °C.

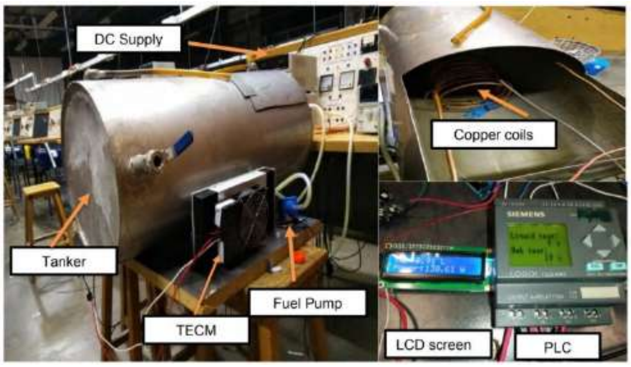


Fig. 8. Integrated intelligent solar powered TECM road milk tanker.

Two tests were performed in order to evaluate the system performance and determine the efficiency of the system. The results were used in order to draw a conclusion on how, on the big scale, the system performed. The ability of TECM as a cooling unit was evaluated and tested as to predict how many TECMs were needed to ensure that the system becomes more efficient in terms of cooling. The ability of the tank to preserve and maintain the cooled temperature will be discussed. The critical tests that were conducted were the cooling ability and the temperature control.

A. *Test 1: Cooling ability*

This test determined the ability of the tanker to cool the milk. Water was used as a viable liquid that was cooled and tested even though the objective of this research was to design the tanker which will be able to cool the temperature of the milk from 38°C to 4 °C within a time period of 3 hours to ensure that the quality of the milk is preserved. 7 L and 21 L of water were externally heated up to a temperature of 38 °C. 7 L of this hot water was filled inside the tanker first and the test was performed. 21 L of hot water was filled secondly, and the cooling test was conducted. The temperature measurements of the critical sensors were recorded and noted while the system was being operative for 3 hours for both cases of 7 L and 21 L. Fig. 9 shows the cooling test results for both the cases. The system managed to cool down 21 L of water from a temperature of 33 °C to a temperature of 19°C within the period of 3 hours. On the other hand, the system only managed to cool down the temperature of 7 L of water from the temperature of 30 °C to the final temperature of 13 °C.

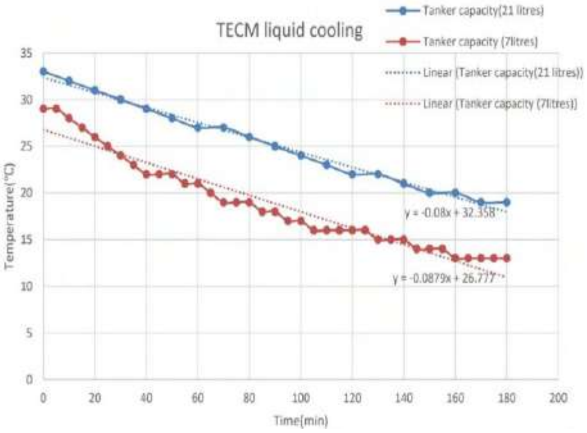


Fig. 9. Liquid cooling test results for 7 L and 21 L, repsectively.

The cooling test results from Fig. 9 indicate linearity hence the trend line was drawn. With these results, we were able to estimate after how long the system will be able to cool the liquid to the temperature of 4°C. These results shows us that the cooling unit (TECM) had insufficient cooling capacity for this tanker. The TECM that was used in this research has a cooling capacity of 113 W which is insufficient to ensure that the temperature of the liquid is cooled and maintained up to the temperature of 4°C within the period of 3 hours. The heat load for specific amounts of water were calculated and estimated in Table I. From this table, this cooling unit was to cool 7 L content to 4 °C within 3 hours. Because of the surrounding factors which dissipate heat, like the electric fuel pump, this was not realized. From Fig. 9, it is safe to conclude that two of these TECMs should be able to cool down 7 L of water from initial temperature of 38 °C to the final temperature of 4 °C within 3 hours.

B. *Test 2: Temperature control*

The objective of this test was to determine the ability of the tanker to maintain the temperature of the liquid between 3 and 5 °C. The ice cubes were filled inside the tanker so that the temperature of water can initially be below 4 °C. Ice cubes were added to the 7 L of water that was tested for cooling. The water capacity after the ice cubes were filled changed to an average water capacity of 9 L. The critical sensor, which in this case is sensor number one, initially registered a temperature of 2 °C as can be seen in Fig. 10.

The system was then off for a period of 40 minutes. The system turned on when the temperature of 4 °C was exceeded. This is indicated by the profile of current measurements plotted in Fig. 10. The temperature of water continued to rise even when the system cooling unit was activated. The temperature of 9 L of water continued to rise until it settled at 7 °C after 120 minutes. The temperature of the liquid continued to be between 6 and 7 °C for the remaining period of 1 hour. The behavior of current when this happened is shown in Fig. 10. During the time when the system was off, the temperature of liquid inside the tanker changed from 2 °C to 4 °C within the space of 40 min.

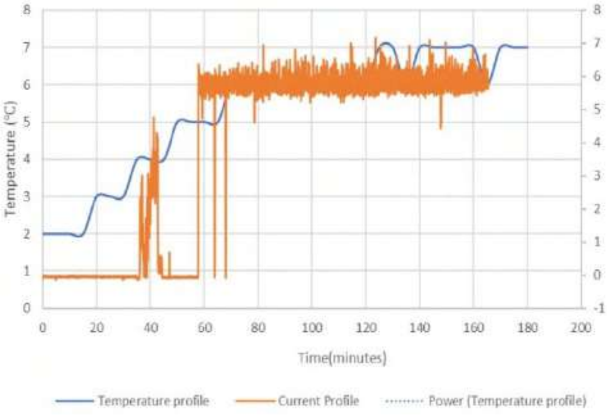


Fig. 10. Results for temperature control test.

This result proves that the tanker was not efficient enough to hold and maintain the temperature of the liquid for an extended time. This is not surprising since the tanker did not have insulation between the outside part of the tanker interacting with the environment and the inside of the tanker which interacts with the liquid. This means that insulations needs to be applied to this tanker. The tanker can store 50 L of milk, as such this required a cooling unit with minimum cooling capacity of approximately 800W according to Table I. Taking into account environmental factors, it would be safe to use a cooling unit with cooling capacity of 1200 W. Using TECM as a cooling unit, 12 will be required. For an actual tank of 8000 L, the cooling unit used shall require a cooling capacity of 192 kW. This will require 1920 TECMs.

IV. Conclusion

This research was based on designing and implementing a small scale intelligent TECM road milk tanker. The proposal was drafted and a brief background about the research was presented. The problem stated that an intelligent solar powered road milk tanker needed to be developed to lower the cost associated with cooling of the milk. Research was compiled on available solutions which can be used in order to implement the research. All the critical subsystems like liquid measuring, power monitoring and temperature control of the research were integrated and found to be working accordingly. Two types of tests were performed to determine the cooling ability of the tanker. It was found that the cooling of the tanker was not efficient as the temperature of 4 ℃ could not be reached by the minimum capacity of the tanker in a period of 3 hours. It was decided that the cooling capacity of the TECM is not sufficient to realize the objectives of cooling the liquid to temperature below 4 ℃. The temperature control of the tanker was also not efficient due to the fact that the temperature of 9 liters of water was not controlled at the required set point for the period of 1 hour. It was suggested cooling units with sufficient cooling capacity be used instead. Comment was made regarding how many of these TECMs were needed and what can be done to improve the tanker.

References

[1] M. Frazer, K. Som, B. Anthony, B. T. Tek, and M. G. Sonnet, *Technical and investment guidelines for milk cooling centres*. Rome, Italy: FAO, 2016.

[2] T. Abebe and T. Markos, *Hygienic milk processing: clean environment, clean utensils | FAO*. International Center for Agriculture Research in the Dry Areas (ICARDA), 2009.

[3] S. Monteleone, M. Sampaio, and R. F. Maia, “A novel deployment of smart Cold Chain system using 2G-RFID-Sys temperature monitoring in medicine Cold Chain based on Internet of Things,” in *2017 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, Sep. 2017, pp. 205–210, doi: 10.1109/SOLI.2017.8120995.

[4] W. Jacobs, M. R. Hodkiewicz, and T. Bräunl, “A Cost-Benefit Analysis of Electric Loaders to Reduce Diesel Emissions in Underground Hard Rock Mines,” *IEEE Trans. Ind. Appl.*, vol. 51, no. 3, pp. 2565–2573, 2015, doi: 10.1109/TIA.2014.2372046.

[5] W. A. Bisschoff, I. N. Jiya, and R. Gouws, “Novel Intelligent Energy Management System for Residential PV Systems in Non-feed-in Tariff Countries,” *Int. J. Appl. Eng. Res. ISSN 0973-4562*, vol. 13, no. 10, pp. 8457–8466, 2018.

[6] B. K. Bose, “Global Warming: Energy, Environmental Pollution, and the Impact of Power Electronics,” *IEEE Ind. Electron. Mag.*, vol. 4, no. 1, pp. 6–17, Mar. 2010, doi: 10.1109/MIE.2010.935860.

[7] I. Saritas and M. Okay, “Design of portable medical cooler with artificial intelligent control,” in *International Conference on challenges in IT, Engineering and Technology (ICCIET’2014)*, 2014, pp. 39–44, doi: 10.15242/IIE.E0714028.

[8] R. Foster *et al.*, “Direct Drive Photovoltaic Milk Chilling Experience in Kenya,” in *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, Jun. 2017, pp. 2014–2018, doi: 10.1109/PVSC.2017.8366541.

[9] N. Zabihi and R. Gouws, “Verifying the cooling capacity and power consumption of thermoelectric cooling holders for vaccine storage,” in *2015 International Conference on the Domestic Use of Energy (DUE)*, Mar. 2015, pp. 115–119, doi: 10.1109/DUE.2015.7102970.

[10] Laird-Technologies, “LA PowerCool Series, LA-115-24-02 Thermoelectric Assembly,” 2010.

[11] H. Eilers, “Thermoelectric (TECM) Cooling Holder,” North West University, Potchefstroom, 2012.

[12] W. Salah, S. Taib, and A. Al-Mofleh, “Development of TEC System for Commercial Cooling Applications,” *Mod. Appl. Sci.*, vol. 3, no. 4, p. p203, Mar. 2009, doi: 10.5539/mas.v3n4p203.

[13] H. Buitendach, I. N. Jiya, and R. Gouws, “Solar powered peltier cooling storage for vaccines in rural areas,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 17, no. 1, pp. 36–46, 2020, doi: 10.11591/ijeecs.v17.i1.pp36-46.

[14] J. Tsado, M. K. Mahmood, A. G. Raji, A. U. Usman, and I. N. Jiya, “Solar Powered DC Refrigerator with a Monitoring and Control System,” in *2018 IEEE PES/IAS PowerAfrica*, Jun. 2018, pp. 591–594, doi: 10.1109/PowerAfrica.2018.8521158.

Automatic Diagnosis of Venous Thromboembolism Risk based on Machine Learning

Autcharaporn Sukperm
Department of Computer Science
Faculty of Science
Srinakharinwirot University
Bangkok, Thailand
autcharaporn.sukperm@g.swu.ac.th

Polapat Rojnuckarin
Department of Medicine
Faculty of Medicine
Chulalongkorn University and King Chulalongkorn Memmorial Hospital
Bangkok, Thailand
rojnuckarinp@gmail.com

Benjaporn Akkawat
Department of Medicine
Faculty of Medicine
Chulalongkorn University and King Chulalongkorn Memmorial Hospital
Bangkok, Thailand
smedbam@gmail.com

Vera Sa-ing
Department of Computer Science
Faculty of Science
Srinakharinwirot University
Bangkok, Thailand
vera@g.swu.ac.th

***Abstract*— Venous thromboembolism (VTE) is an important disease to increase the number of patients because of lacking awareness in Thailand to block the blood flow in the vein. In addition, an effective assessment model of VTE risk is the most important for medical doctors to diagnose. So, this paper represents an automatic diagnosis model by using effective machine learning to predict the important risk factors of VTE from collecting patient data of the medical ward at King Chulalongkorn Memorial Hospital. This research prepares the 83,850 raw data and investigates the missing values for transforming the data to ready import into each model and then separates the adjusted data for training and testing in the ratio of 70:30. The experimental results were compared to the effectiveness of three machine learning algorithms that consist of the decision tree, logistic regression, and neural network. From the experimental result of the decision tree, this model represents the best assessment model with an accuracy of 96.6% by adjusting the balance data with the class weight method for assisting diagnose the medical doctor.**

***Keywords*— venous thromboembolism, machine learning, automatic diagnosis**

I. Introduction

Venous thromboembolism (VTE) is an important problem to increase the number of patients because of lacking awareness in Thailand [1]. The common complication of VTE occurs during and after hospitalization for acute medical illness or surgery that 5–10% of patients die in hospital. Moreover, VTE is a long-term treatment of the post-thrombotic syndrome. So, these complications contribute to patient disease and the cost of management. VTE is the main cause of a blood clot that slowly blocks the blood flow in a vein. Deep Vein Thrombosis (DVT) occurs in a deep vein in the legs because a blood clot to the pulmonary artery can cause immediate death [2]. However, an assessment model of VTE risk only have been developed and validated in Caucasians that reports the assessment model of VTE risk for Asians by using a machine learning algorithm is lacking [3,4].

This research investigates the research of Agharezaei L. et al. [5] to propose the prediction of the risk level of pulmonary embolism in patients by using the mean of artificial neural networks. There are two types of artificial neural networks consists of Feed-Forward Back Propagation and Elman Back Propagation that were compared in this study and analysis data by using MATLAB software. The result of this study proposed the optimized artificial neural network model that represent an accuracy and risk level index of 93.23 percent. In addition, Fei Y. et al. [6,7] constructed and validated the artificial neural networks (ANN) for predicting portosplenomesenteric venous thrombosis (PSMVT) by comparing the predictive ability of ANN and logistic regression. This research represented the result of ANN that was more accurate than logistic regression in predicting the occurrence of PSMVT following acute pancreatitis. Moreover, Qatawneh et al. [8] presented a clinical decision support system that classifies the risk of venous thromboembolism using ANN. The system uses Multilayer Perceptron (MLP) feed forward neural network and uses Resilient Backpropagation algorithm (Rprop) for training. The developed system classifies the risk of VTE into five risk levels ranging from low to high. The results of the experiment show that the accuracy of the system is 81%. Furthormore, Ferroni Z.et al. [9,10] represented to use a machine learning (ML) and random optimization (OR) techniques developed for predicting the risk of venous thromboembolism to devise a web interface for VTE risk stratification in chemotherapy-treated cancer patients. This research compared the effectiveness of both techniques with the current technique which is the khorana score (KS). The result of this research shown the ML-RO to provide more performance than khorana score (KS). It is therefore suitable and useful to design a web service. In addition, Liu S. et al. [11] proposed a ML to conduct an effective assessment of peripherally inserted central venous catheter (PICC) related thrombosis. It uses five models: Seely, Seely-RF, Seely-LASSO-RF, RF and LASSO-RF to compare the model performance. The result shown machine learning model provides better performance to help prevent disease decision and effectively reduce the rate of thromboembolism in cancer patients with PICC catheters. Moreover, Nafee T. et al. [12] evaluated the performance of super model (sML) and reduced model (rML) compared by using International Medical Prevention Registry on Venous Thromboembolism (IMPROVE) score for predicting occurrence venous thromboembolism. The result shown sML method has c-statistic result highest for predicted venous thromboembolism.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

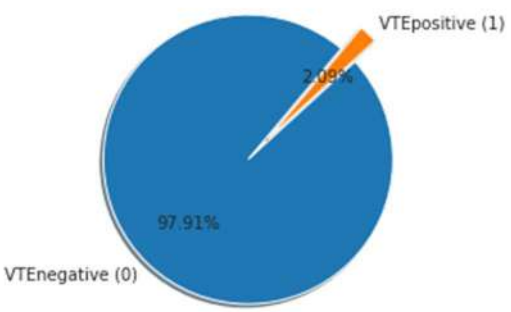


Fig. 1. The percentage of VTE positive and VTE negative was derived

Therefore, this research will find an effective machine learning to predict the risk of VTE from collecting patient data of the medical ward at King Chulalongkorn Memorial Hospital to assist the medical doctors that make more accurate diagnosis and treatment decisions. The remaining of this paper is organized as follows : Section 2 presents the methodology followed to design the prediction model of machine learning. Section 4 presents the experimental result. Finally, Section 5 provides the conclusion.

II. METHODOLOGY

From the data of King Chulalongkorn Memorial Hospital, this research was used for comparing the performance of each algorithm as following in these steps. First, this research imported raw data by using the Python program and then using the Exploratory Data Analysis (EDA) for analyzing the relationship between the variables or risk factors for thromboembolism. Second, this research prepared the raw data and investigated the missing values for transforming the data to ready import into the model for further analysis. Third, this research separated the adjusted data for training and testing in the ratio of 70:30. Finally, this research created a predicting model based on machine learning by using 3 techniques including Decision tree, Logistic regression and Neural network technique for comparing and finding the best performance of prediction model.

A. Gathering Data

This research is based on the information that was collected by the doctors who researched to determine risk factor for VTE. The data were collected from patients who were admitted to internal medicine wards of Chulalongkorn hospital in 2009 and approved by the Ethical committee of the faculty of medicine, Chulalongkorn University. The data consisted of a total 1290 rows that were the number of patient's survey and 65 columns that were the risk factors of VTE. Analyze the correlation of data all risk factors for venous thromboembolism as shown in Fig.2 and Fig.3. This figure represented the relationship of each risk factor for venous thromboembolism that was related each columns variable has very little correlation with the class label because all the data collected are discrete variables.

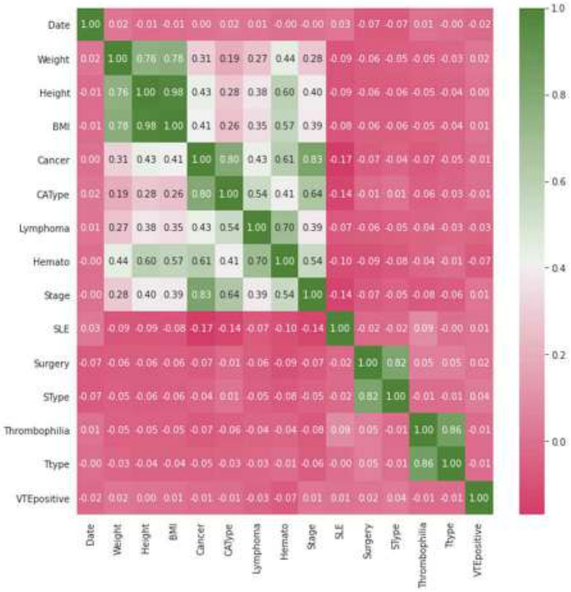


Fig. 2. The relationship between each risk factor for venous thromboembolism

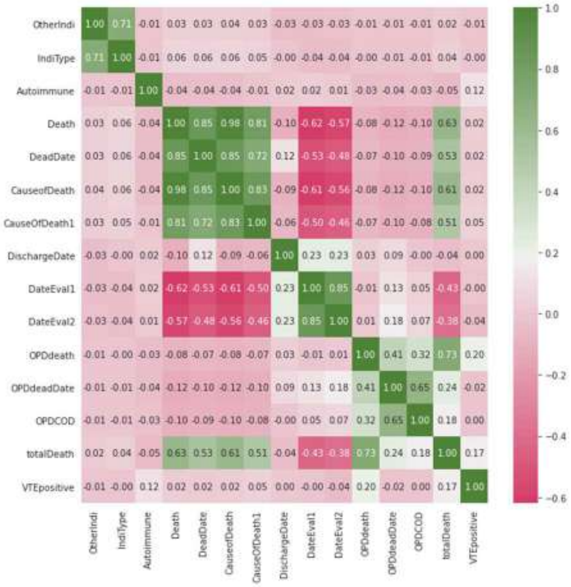


Fig .3. The relationship between each risk factor for venous thromboembolism

B. Filtering Data

This research filtered the gathering data by using a professional suggestion and a filtering algorithm, as following:

1) Filtering by a computer principle algorithm: This research filtered the data to be analyzed for risk factors for VTE as shown in Fig. 1. This figure represented the patients having venous thromboembolism (VTE positive) 2.09% and patients without venous thromboembolism (VTE negative) 97.91%. Then, we had classified it as a class label with the following: 0 is a patient without venous thromboembolism (VTE negative) and 1 is a patient with venous thromboembolism (VTE positive).

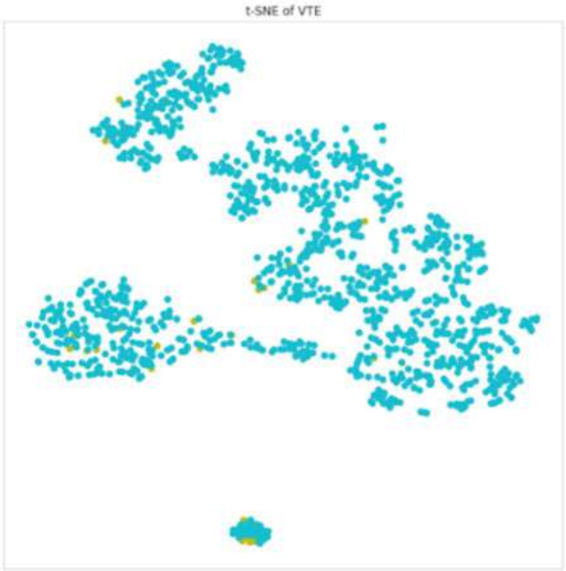


Fig. 4. The relationship between patients with venous thromboembolism and all risk factors.

Then we looked at the relationship between patients with venous thromboembolism and all risk factors represented by t-distributed Stochastic Neighbor Embedding : t-SNE method shown in Fig.4.

C. Preparing Data

1) Missing Values: The data had many missing values for considering to delete some of the columns that were not affected or analysis and manage the remaining missing values with fill missing values principle using "fill with zero" method.

2) Transform data: As a result of the exploration data, there some columns were an object type to be analyzed data must be change to numbers. Therefore the data type was transformed data with a label encoder method

3) Split data: This research splited the data for training and testing with a ratio 70:30.

4) Balance data: From the Fig. 1, it represented an imbalance data problem was more in patients VTE negative than VTE positive. It would be sensitive to a majority class data because bias on the data that would not be able to detect data in a minority class. Therefore, we managed an imbalance data problem with 5 methods include Oversampling, Undersampling, Synthetic Minority Oversampling TEchnique : SMOTE, Class weight and Ensemble sampling.

D. Model Building

This paper is proposed an automatic diagnosis model for risk of symptomatic VTE by analyzing the important risk factors of VTE data 1290 rows and 65 columns. From review the related works, we selected the 3 effective models and tunning parameters of each model to evaluate and compare the performance of the best model as follows:

1) Decision Tree model : Decision tree model would to try to divide the data into classes as clearly. Modeling principle was to choose the variables can separate the answer classes.

2) Logistic Regression model : Logistic Regression model was an extended version of linear regression used to solve binary classification problems for predicted 2 classes such as 0 with 1. The sigmoid function used to normalize the value between 0 to 1. The result from the sigmoid function was the probability that classes can be divided.

3) Neural Network model : The Neural Network model simulated the function of the human brain which consists of consists of input layer, hidden layer and output layer. neurons. In a neural network, it was divided into 3 layers

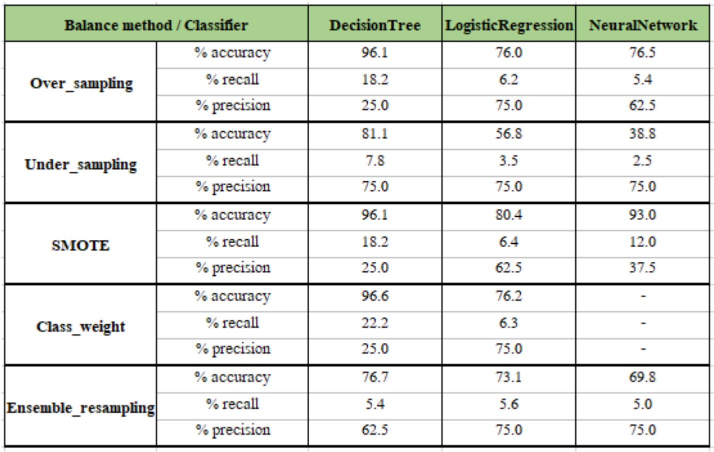
E. Evaluate Model

Each model evaluation had the different targets. This research based on classification to use a confusion matrix table for the evaluation model which was a table used to evaluate prediction results compared to the actual. The indicated accuracy was the recall and precision of models by representing to graph by using the receiver operating characteristic curve (ROC) and area under the curve (AUC).

III. EXPERIMENT RESULTS

This section presented the experimental results of each automatic diagnosis model for risk of symptomatic venous thromboembolism. After preprocessing, we divided the VTE data with ratio training data per testing data 70:30. After that, we balanced the VTE data with 5 methods include Oversampling, Undersampling, Synthetic Minority Oversampling TEchnique : SMOTE, Class weight and Ensemble sampling. Finally, we apply to create the model results as presented in Table I.

TABLE I. THE EXPERIMENTAL RESULTS OF EACH MODEL



Balance method / Classifier		DecisionTree	LogisticRegression	NeuralNetwork
Over_sampling	% accuracy	96.1	76.0	76.5
	% recall	18.2	6.2	5.4
	% precision	25.0	75.0	62.5
Under_sampling	% accuracy	81.1	56.8	38.8
	% recall	7.8	3.5	2.5
	% precision	75.0	75.0	75.0
SMOTE	% accuracy	96.1	80.4	93.0
	% recall	18.2	6.4	12.0
	% precision	25.0	62.5	37.5
Class_weight	% accuracy	96.6	76.2	-
	% recall	22.2	6.3	-
	% precision	25.0	75.0	-
Ensemble_resampling	% accuracy	76.7	73.1	69.8
	% recall	5.4	5.6	5.0
	% precision	62.5	75.0	75.0

IV. Conclusion

In this paper, we proposed to select an effective of the automatic diagnosis model for risk of symptomatic venous thromboembolism based on machine learning. This model will assist to diagnosis of professionals and medical doctor for treating patients more easily. The proposed system compared the 3 effective techniques of machine learning for evaluating multifactorial venous thromboembolism. Developing the system passed through several steps starting from collecting, preparing and stratifying the data set and choosing the most accurate system model. Decision tree model represented the best performance model with an accuracy of 96.6 percentage by adjusting the balance data with the class weight method.

Acknowledgements

We would like to thank King Chulalongkorn Memorial Hospital for providing effective data and good diagnostic information in this research. In addition, our research was funded by the master of science in data science from the department of computer science and the Graduate School of at Srinakharinwirot University of Thailand to provide the scholarship for publishing this article.

References

[1] Rojnuckarin, P. et al. “Risk factors for symptomatic venous thromboembolism in Thai hospitalised medical patients.” Thrombosis and haemostasis vol. 106, no. 6, 2011, 1103-1108.

[2] Cohen, A. T. et al. “Venous thromboembolism risk and prophylaxis in the acute hospital care setting (ENDORSE study): a multinational cross-sectional study.” Lancet (London, England) vol. 371, no. 9610, 2008, pp. 387-394.

[3] Motwani M., et al. “Advanced cardiovascular magnetic resonance myocardial perfusion imaging: high-spatial resolution versus 3-dimensional whole-heart coverage,” Circ Cardiovasc Imaging. vol. 6, no. 2, 2013, pp. 339–348.

[4] Laohakiat S., Sa-ing V., “An incremental density-based clustering framework using fuzzy local clustering,” Information Sciences, vol. 547, 2021, pp. 404-426.

[5] Agharezaei, L., Agharezaei, Z., Nemati, A., Bahaadinbeigy, K., Keynia, F., Baneshi, M., Agharezaei, a. “The Prediction of the Risk Level of Pulmonary Embolism and Deep Vein Thrombosis through Artificial Neural Network,” Acta Informatica Medica, vol. 24, no. 5, 2016, pp. 354-359.

[6] Fei, Y., Hu, J., Gao, K., Tu, J., Li, W. Q., & Wang, W., “Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models,” J Crit Care, vol. 39, 2017, pp. 115-123.

[7] Fei, Y., Hu, J., Li, W. Q., Wang, W., & Zong, G. Q., “Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis,” J Thromb Haemost, vol. 15, no. 3, pp. 439-445.

[8] Qatawneh, Z., Alshraideh, M., Almasri, N., Tahat, L., & Awidi, A. , “Clinical decision support system for venous thromboembolism risk classification,” Applied Computing and Informatics, vol. 15, no. 1, 2019, pp. 12-18.

[9] Ferroni, P., Zanzotto, F. M., Scarpato, N., Riondino, S., Guadagni, F., & Roselli, M., “Validation of a Machine Learning Approach for Venous Thromboembolism Risk Prediction in Oncology,” Dis Markers, 2017, pp. 1-7.

[10] Ferroni, P., Zanzotto, F. M., Scarpato, N., Riondino, S., Nanni, U., Roselli, M., & Guadagni, F., “Risk Assessment for Venous Thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patients,” Med Decis Making, vol. 37, no. 2, 2017, pp. 234-242.

[11] Liu, S., Zhang, F., Xie, L., Wang, Y., Xiang, Q., Yue, Z., Yu, C., “Machine learning approaches for risk assessment of peripherally inserted Central catheter-related vein thrombosis in hospitalized patients with cancer,” Int J Med Inform, vol. 129, 2019, pp. 175-183.

[12] Nafee, T., Gibson, C. M., Travis, R., Yee, M. K., Kerneis, M., Chi, G., Goldhaber, S. Z., “Machine learning to predict venous thrombosis in acutely ill medical patients,” Res Pract Thromb Haemost, vol. 4, no. 2, 2020, pp. 230-237.

E-Commerce Case Collection

William Haggerty
Information Technology Department
Georgia Southern University
Statesboro, GA USA
whaggerty@georgiasouthern.edu

***Abstract*—The purpose of this paper is to provide descriptions of E-Commerce case situations and draw insightful conclusions from them. The document will be in a format with Ten E-Commerce Rules that are numbered here from 1 - 10 within the cases they represent. They can be applied in retrospect for discussion and in the future of new E-Commerce opportunities to help identify risk and give insight into ways to avoid it. This is a subset of available cases that will be released in a larger more detailed document for use in an academic environment for teaching about E-Commerce. Most of the SMB companies and participants have been hidden to protect those whose revelation could prove harmful.**

Keywords—E-Commerce; Risk; Shopping Cart; Payment Processing; Web Sites.

I. Introduction

My internet experience since 1998 includes a long list of E-Commerce opportunities and situations where my services were employed to make SMB on-line shopping work. In some cases, it was merely to see if it could work. From those experiences comes a list of Ten E-Commerce Rules that are time tested and true from the real-world case summaries included here. Admittedly, they are based on the internet prior to Web 2.0 but the resulting Rules are still applicable even today.

II. CASES

1. A commercial cleaning products company that was selling 55-gallon drums and 22 and 32 oz bottles of cleaning products using ShopSite [2] shopping cart installed and subsequent conversion to OpenCart [3].

In early 2003 my four-year customer, a local wholesale distributor of his and I had a meeting to discuss installing a shopping cart on their website. As the meeting began, it was clear that the distributor had an idea to sell 55-gallon drums of cleaning solutions in an online shopping cart. “Those large drums won’t fit in a shopping cart” I said, indicating that the idea was not practical at the time. Most shopping carts then were B2C with B2B only just beginning and not likely to penetrate institutional cleaning purchase agents anytime soon. The distributor insisted, however, and the owner acquiesced yet cleverly agreed to a one-year test. So, I built the shopping cart with descriptions of twenty products with product “image not available” superimposed over images of 55- gallon drums.

Image #1 – 55-Gallon Drum [1]

After one year, only one 55-gallon drum was sold. As agreed, I was called in to discuss taking the shopping cart down. Before leaving, I suggested that since the cart was operational, and the product descriptions were already there, all I had to do was add images of their 22 oz and 32 oz containers with current pricing and he could sell them in the shopping cart. The owner agreed to a new test for that concept with me managing the cart and sending the orders to his facility for processing.

Ironically, the shopping cart soon took off and within approximately 90 days, the owner asked me to build him another shopping cart with the disclaimer that it was being offered for customers that had no retail location within their local area.

~E-Commerce rule #1. Try anything with a time sensitive test. Be willing to change course as needed and if it takes off, take immediate action.

2. An independent drop ship distributor selling cleaning products in 22 and 32 oz bottles to consumers using ShopSite shopping cart.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Below is an image reflecting the sales from the shopping cart I continued after such a robust beginning. Of course, with the revenue I received from product sales commissions, it was agreed that I would complete the web site maintenance for them at no additional charge.

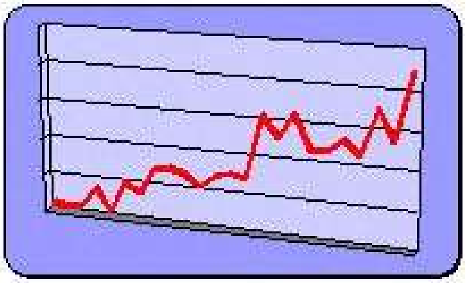


Image #2 - 3-2009 (3 years of growth) [1]

That arrangement seemed to work fine until the government dictated that the Material Safety Data Sheet (MSDS) was to be changed to Safety Data Sheet (SDS). All links between the Technical Data Sheet (TDS) for the product were linked to the product number followed by …msds. Consequently, all the MSDS file names and TDS links had to be manually changed to …sds on two websites for over 20 products to meet the compliance deadline.

Fortunately, this presented us with a good time to update the SDS's as pdf's that I could put 'inline' in the pages. Then we no longer needed to update those SDS pages in HTML detail.

~E-Commerce rule #2. When the government makes a compliance decree, waste no time in responding no matter how cumbersome it appears to be. Try to use the opportunity to your advantage to accomplish something on the 'nice to do' list simultaneously.

Finally, as their investment in IT grew, they eventually ported their website over to different servers. With that migration came the shopping cart as well. But, their financial officer wanted a less expensive alternative to the monthly charge for ShopSite. That is how both carts had been installed initially.

Because of my experience of migrating another customer from MivaMerchant [4] for the same reason, I quickly converted them to OpenCart [3], which was a proven and reliable object-oriented public domain shopping cart still in operation today. My ShopSite [2] version continued until it was mutually ended in 2019, almost 15 years after initially started as a test to see if it could work. It most certainly did!

~E-Commerce Rule #3 – There is a time for everything and if the situation dictates a change, cut your losses, make the change, and move on.

3. A spa company selling chemicals, filters and accessories on-line using MivaMerchant [4], OS Commerce [5] and OpenCart [3] shopping cart software products at different times.

I met the female half of the husband-and-wife team when she opened a new location next to one of my customers that I used to visit often. We became fast friends and when her website developer left for Texas, she turned to me to provide the local support she was looking to have.

~E-Commerce Rule #4 – Keep your technical support nearby. Even with today's collaboration, it is best they be near you.

One challenge was to learn the functions of MivaMerchant, their shopping cart product. Of course, my first objective was to get a handle on their website, but it was not long before I tackled MivaMerchant as well. They became conscious of the fees for that service and rather than suggest moving to the more familiar ShopSite with fees of its own, I looked at the Object-Oriented offerings my hosting company made available at no additional charge.

We looked at a couple of them such as OS Commerce [5], OpenCart [3] and ZenCart [6] to name a few. We landed on OS Commerce as it offered a sort/list capability that was of interest to the owner.

After a minor technical glitch with OS Commerce, we switched to OpenCart and OpenCart continued to run just fine.

Image #3 – Original Flash Template 2010 [7]

4. BBD – One of the previous customer's friends was interested in also selling Hot Tub Accessories online via a shopping cart. After completing modification of a suitable template, I built the cart and loaded up the products. But it never seemed to get off the ground. To stop the bleeding, we pulled the plug not long after it had begun.

Image #4 Online Accessory Template [8]

~E-Commerce Rule #5 – Having a retail establishment in a large market to fuel online sales of additives, supplies and accessories does not necessarily translate well to a fully online entity somewhere else without that connection.

Unfortunately, it was not allowed to be viewed as a test, and with an expectation of success, it was sad to end. If we had applied E-Commerce Rule #1 it would have worked better for all involved.

5. A wedding location listed online images for both on and off-site wedding and event planning services.

One of my earlier customers was an event facility. They performed event and wedding planning and on-site wedding ceremonies including dinners and receptions. They did very well and as I offered directory listing for customers, I even featured them on my wedding directory pages with the wedding picture of one of their staff members,

Image #5 Wedding Directory [9]

as well as an image from one of their wedding receptions.

On his site we had created images from letters of reference and added those to his site. (E-Commerce Rule #10). Also, we designed an on-line form that had drop down lists for party size and estimated budget amount in ranges.

Eventually, the owner wanted to attract some of the customers they were not able to reach because of their location. He decided to update his site with images of his site location and images of other sites in Greater Atlanta that he was able to use for events.

With the acquisition of another domain name, we created a mirror image web site and linked the images on the left of the home page to the new one.

Including directions for each location, the on-line form generated increased targeted results.

Image #6 Pictures of event locations [1]

We updated that on-line form with a location dropdown list and it worked just fine. Business was booming. Regrettably, the owner met a premature death from an accident in the Pacific Ocean not long after that.

~E-Commerce Rule #6 – Do not be limited by what your situation dictates. Expand by offering what is available per diem and limit acquisition expenses.

6. A delicatessen specializing in sandwich rolls uploaded their menu planning to sell the idea of creating a franchise.

My regional prospecting one day brought me to a very good delicatessen for lunch. I asked for the owner and upon meeting her we began to talk about a website for the restaurant and how it might be designed.

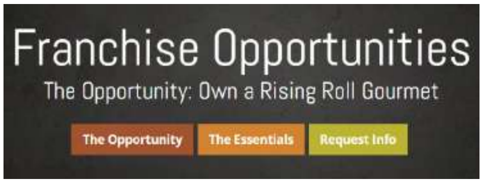
It seemed she had almost been waiting for me as when she began to describe how she wanted her menu displayed she knew exactly what she wanted. She also knew exactly how she wanted it to look. I had done a few menus in the past and it appeared that this would be a similar experience. That is, until I had given her a link to review what I had put together.

The minute detail she had me go through seemed to be beyond the usual menu display, but I figured that was her style.

Consequently, I eagerly followed her instructions. What we ended up with was quite a menu indeed. And even I was impressed with the final version.

Unfortunately, not long after we parted ways as she had sold her restaurant to a major franchising company complete with their own professional web site design staff. They were in position to take over the site and handle the marketing package for any additional franchisees.

It turns out, she was using the design of her menu to attract suitors to establish the franchise tag all along. And although I did not get paid any more than I did for the other menus in the past, somehow there was a sense of pride that my design work had done the trick for her. She was always a most gracious customer and although tedious, my visits with her were always a pleasure. Not to mention the occasional exceptional free lunch.



Franchise Opportunities

The Opportunity: Own a Rising Roll Gourmet

Image #7 The 'Franchise Opportunity' (Rising Roll Gourmet Café' web site) [10]

I must admit, after going to their new website to research this paper, it appears that the menus have lost something in the translation.

~E-Commerce Rule #7 – Always put your best foot forward online when looking to market your company. Suitors usually do not take a second look.

7. A vacation property manager that was renting from an inventory of over 40 properties got a Flash web site, property rotating image galleries, and videos.

At a local business networking event, I met the owner of a company that was renting property to vacationers in condos at the beach in Florida. We negotiated for a web site and I produced an exciting Flash site (2010) complete with music, property rotating image galleries and videos that replaced them as the videos became available.

The ordering was conducted through email and phone and things were going along fine. Then, a service combined a calendar and scheduling automated service that far outstripped my capabilities and we amicably parted ways.

Not long after that, I found out the Tenant associations in the condo buildings they rented out of changed their rules to restrict access for short term rental tenants. Their business dried up overnight.

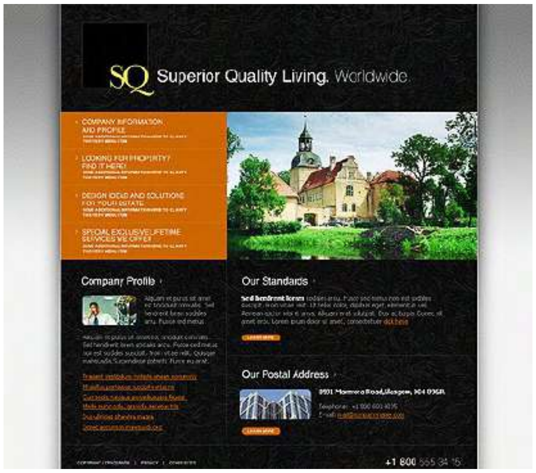
~E-Commerce Rule #8 – The assumptions underlying your business operations can change with the wink of an eye. Stay aware about outside related influences.

8. A real estate group with three owner principals and multiple agents listed their properties online.

Through another friend I was referred to three women that had started a real estate company and had been burned by a web site developer. They were left with virtually nothing to show for a substantial investment they had made.

I went to work right away and with the help of a talented graphics artist family member of one of the three, and got the site running quickly. They had been left wounded by the previous web developer and did not have much left for my services, but my interest in quality was well respected, nonetheless.

The problem was that after they finally realized the problem they were in, got together with me and got me going, the real estate market of 2009 started to nosedive.



SQ Superior Quality Living. Worldwide
COMPANY INFORMATION AND PROFILE
LOOKING FOR PROPERTY? FIND IT HERE!
DESIGN IDEAS AND SOLUTIONS FOR YOUR ESTATE
SPECIAL EXCLUSIVE LIFETIME SERVICES WE OFFER
Company Profile
Our Standards
Our Postal Address
+1 800 555 34 15

Image #8 Ended before it really began [11]

It was not long before the real estate group of three had disbanded with each person going their own separate way.

~E-Commerce rule # 9 – Make certain that you do not spend all your money in one place. Hold back a good percentage until you see actual working documents (with sources) that can be transferred to someone else if need be later. Remain cautious of web site prototypes.

Unfortunately, they also were victims of E-Commerce Rule #8.

9. Three different food caterers who listed their menus online with party and event platter images. 1st caterer used a block of images representing menu items.

2nd caterer web site image is below:

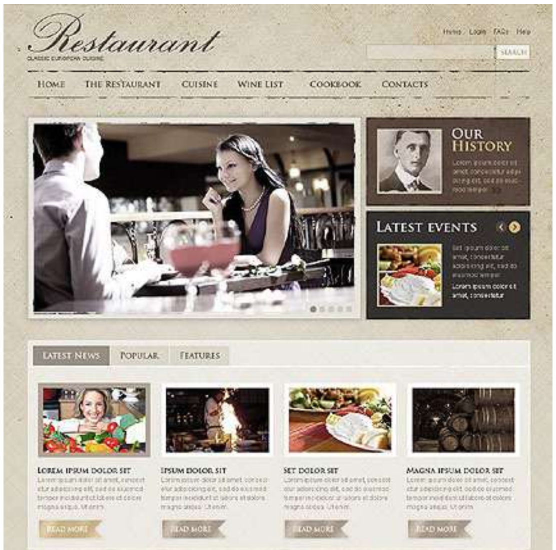


Image #9 Home Page Template [12]

All three companies were eager to display images of their most delectable offerings (1st and 3rd not shown). But what was missing was a customer reference tied to each one of those images. The first and third company had never thought to document those originally and the second one was too new in the business to have enough to be worthwhile.

In all cases, it was a challenge to convince them that images alone were not enough to attract new customers. The cost associated with setting up that capability was beyond their budget expectations.

~E-Commerce Rule #10 – Before beginning a web site, be certain to document images with customer reactions to your services. On the site, allow customers to post references with images and if possible, capture their reactions on video as part of your standard operating procedures at your events.

Today, of course, customer testimonials in the form of videos are catching on, but even in the past, the most effective images were those that were directly representing actual customer testimonials.

III. TEN E-COMMERCE RULES

#1. Try anything with a time sensitive test. Be willing to change course as needed and if it takes off, take immediate action.

#2. When the government makes a compliance decree, waste no time in responding no matter how cumbersome it appears to be. Try to use the opportunity to your advantage to accomplish something on the ‘nice to do’ list simultaneously.

#3. There is a time for everything and if the situation dictates a change, cut your losses, make the change and move on.

#4. Keep your technical support nearby. Even with today’s collaboration, it is best they be near you.

#5. Having a retail establishment in a large market to fuel online sales of additives, supplies and accessories does not necessarily translate well to a fully online entity somewhere else.

#6. Do not be limited by what your situation dictates. Expand by offering what is available per diem if possible and limit acquisition expenses.

#7. Always put your best foot forward online when looking to market your company. Suitors usually do not take a second look.

#8. The assumptions underlying your business operations can change with the wink of an eye. Stay aware about outside related influences.

# 9. Make certain that you do not spend all your money in one place. Hold back a good percentage until you see actual working documents (with sources) that can be transferred to someone else if need be later. Remain cautious of web site prototypes.

#10. Before beginning a web site, be certain to document images with customer reactions to your services. On the site, allow customers to post references with images and if possible, capture their reactions on video as part of your standard operating procedures at your events.

IV. CONCLUSION

I have close to twenty additional case studies and in the future, my plan is to expand this document to edit a small book of SMB E-Commerce web cases for use in the academic classroom environment.

V. References

#1 Images used with permission.

#2 ShopSite E-Commerce Product
https://www.shopsite.com/

#3 OpenCart E-Commerce Product
https://www.opencart.com/

#4 MivaMerchant E-Commerce Product
https://www.miva.com

#5 Os Commerce E-Commerce Product
https://www.oscommerce.com/

#6 Zen Cart E-Commerce Product
https://www.zen-cart.com/

#7 Expired Template # 8467 from
https://www.templatemonster.com/

#8 Expired Template # 17362 from
https://www.templatemonster.com/

#9 Image from http://haggo.com/weddings.htm
- used with permission.

#10 Rising Roll Gourmet Café – website:
https://risingroll.com/franchise-opportunities/

#11 Expired Template # 13311 from
https://www.templatemonster.com/

#12 Expired Template # 33580 from
https://www.templatemonster.com/

COVID-19 Zero-Interaction School Attendance System

Emily Sawall
Computer Science
University of Wisconsin-Green Bay
Green Bay, Wisconsin, USA
Sawallemily@gmail.com

Amber Honnef
Computer Science
University of Wisconsin-Green Bay
Green Bay, Wisconsin, USA
Amberhonnef@gmail.com

Mohamed Mohamed
Computer Science
University of Wisconsin-Green Bay
Green Bay, Wisconsin, USA
Moham22@uwgb.edu

Ali Abdullah S. AlQahtani
Computer Science
University of Wisconsin-Green Bay
Green Bay, Wisconsin, USA
AlQahtani.AASA@gmail.com

Thamraa Alshayeb
Physics
University of Wisconsin-Green Bay
Green Bay, Wisconsin, USA
Alshayeb.T@gmail.com

***Abstract*—For decades, school attendance has an important role in determining student success and ensuring that students regularly come to classes. Traditionally, a sheet of paper and a pencil or pen is used to document school attendance at universities. In most cases, the professor would hand out a piece of paper to a student. This student would mark their presence on the paper sheet and pass it to the person sitting next to them. The piece of paper would then travel throughout the classroom until the final student received the paper. Finally, the piece of paper would be given back to the professor or the teacher's assistant, so they may take note of which student is present and which student is absent. In the age of the COVID-19 pandemic, this method of taking attendance poses a great risk to possible disease transmission. Everyone within the room would be touching the same piece of paper, and sometimes students may even share the same pen or pencil with their classmates. Everyone in the classroom, including the professor, would be at risk if one student were infected with the COVID-19 virus. This is attributed to the attendance method. This paper proposes a zero-effort, zero-interaction attendance method based on two aspects that are currently present in a typical classroom: access points and a student's Wi-Fi enabled devices (e.g., smartphones). Our research found this method to be a no-hassle attendance-taking method that pinpoints where a student is within the classroom, without the class suffering from any disruption in learning.**

***Index Terms*—access point, attendance system, RSSI, Wi-Fi, trilateration system, zero-effort, zero-interaction.**

I. Introduction

USER authentication is critical in confirming a user's identity and their presence [1]. This same idea is used in classroom attendance. Traditional methods of taking attendance can be inaccurate. Students can mark their absent classmates as present, so they do not miss out on points. Besides, if attendance is taken at the beginning of class, students could leave right after attendance is counted. Having the professor take time out of their lesson in order to search through the attendance or seating chart for each student takes away valuable lesson time from the professor and the students. The traditional pen-and-paper methods of attendance take effort and cost the professor and class time.

COVID-19 pandemic has also shown how obsolete the traditional methods of conducting attendance have become—traditional pen-and-paper methods of attendance risk the transmission of possible disease particles. The passing of paper and the possibility of sharing writing utensils make it challenging to keep others safe from the virus. As schools are preparing to conduct more in-person classes, there is a growing need for a safer and more hygienically friendly way of conducting attendance. This can be achieved utilizing technology (i.e., access points broadcast signals) already available in most modern classrooms and schools.

The proposed scheme aims to present a zero-effort and zero-interaction method of classroom attendance. The presented attendance system utilizes students' smartphones and access points located within the school. Most college students within the twenty-first-century own Wi-Fi enabled devices that connect to the outside world through the Internet. The principal benefit of broadcast signals that the Wi-Fi-enabled devices can receive and read them from different radio frequency technologies within range without being connected to the broadcaster. The broadcast signals are utilized to take a student's attendance and submit it to his/her professor with requiring zero-effort and zero-interaction to conduct attendance within the confines of a worldwide pandemic safely.

In the next section, existing research regarding attempts to modify school attendance systems is discussed technologically, followed by the proposed scheme's model and the experiments that have been conducted.

The research has been funded by:
(1) WiSys Spark Grant, Code WISYS-AAI7814, www.wisys.org/grants/spark.
(2) University of Wisconsin-Green Bay, College of Science, Engineering, and Technology.

II. Literary Review

Several methodologies and solutions have been proposed to mitigate the issue of classroom attendance and monitoring. Each methodology and solution are unique, solving the problems in their own unique way.

In 2019 Koppikar, et al., showcase their solution in their work titled, ”IoT Based Smart Attendance Monitoring System using RFID” [2]. The authors propose an RFID-based attendance monitoring system using student or employee-based identification cards. Students and faculty would need to register themselves into the RFID attendance monitoring program before attending their class. Registration would occur through a graphical user interface. On the day of class, the user would need to place their RFID tag near the RFID reader. The information is then sent to an Arduino module to process the information.

In 2020 Liu, et al., proposed using R-FCN technology and video surveillance to monitor student attendance in their work titled, ”Intelligent Counting System for Classroom Numbers Based on Video Surveillance” [3]. R-FCN is used in this scenario to detect targets within a given space. The work also features different formulas for calculating the position of a student. The layered feature maps ensure targets will not be lost within the system. This solution needs stable classroom lighting in order to remain effective.

Bai, et al., proposed in 2020 a facial recognition solution in their work titled ”Design of Attendance System Based on Face Recognition and Android Platform” [4]. This system requires the use of the Adaboost algorithm and includes the calculation of errors within the dataset. The system needs to detect a face, run it through the algorithms, and then use facial recognition algorithms to determine if the system detected the correct person. This would require users to have pictures taken of themselves while showing different facial expressions, and if the user wears glasses, they will need to have pictures taken with or without glasses.

In 2019 Zhi-heng and Yong-Zhen proposed using facial recognition technology with a camera placed within the classroom to monitor attendance, in their work ”Design and Implementation of Classroom Attendance System Based on Video Face Recognition” [5]. The video feed from the classroom camera is separated into still frames and relies on pictures with stable lighting and clear facial features in order to accurately verify the students. Like the previous work, this work also uses the Adaboost cascade classifier to detect the faces quickly. This proposal also discusses the advantages of using machine learning to train the software. However, as mentioned, this proposal relies heavily on clear lighting and positioning of the camera.

Chennattu, et al., in 2019, proposed a fingerprint identification system to monitor classroom attendance in their work titled, ”Portable Biometric Attendance” [6]. On the user's end, the system would use an LCD and button interface. The teacher and the students would need to scan their fingerprints, where the teacher would be required to authenticate first before the students. The teacher would terminate the attendance cycle by scanning their fingerprint again. The data is then sent to an AWS database. After the data is assorted into the correct tables, it is then shown to all students.

As of today, there are a numerous number of attendances extracting techniques. In 2019 a paper was published that was based on a frequency distribution algorithm with passive RFID tags to take attendance [7]. In 2020, another paper utilized beacon frames to prove the location of users [8]. In addition, a proposed system was published in 2020 that uses a single image to tack attendance [9].

In this section, numerous attendance extracting techniques have been discussed. The following section presents the proposed scheme from different aspects.

III. System Model

The proposed scheme aims to utilize the access points broadcast signals to pinpoint students' location by applying the trilateriation approach. If the smartphone is found to be within a certain distance and within the classroom confines, the student is confirmed to be attending the class. However, if the student's cellular device is not found or is not determined to be within the classroom's confines, the student is found not to be in attendance of the class. In the proposed scheme, students' devices measure the Received Signal Strength Indicator (RSSI) reading utilizing the broadcast signals, then transmit them to a centralized application. The centralized application then utilizes the collocated RSSI reading to determine a student's attendance status. A completed list of attendees vs. non-attendees will then be sent to the professor, requiring zero-interaction and zero-effort from neither students nor the professor.

There are benefits to this method compared to other proposed school attendance methods. Wi-Fi-enabled devices do not necessarily need to be connected to the school's access points in order to receive the broadcast signals. Also, the transmitted signals are limited to a physical area that limits students from spoofing their locations or impersonating other students, which means the student needs to be physically within the classroom to be counted as present.

A. Architecture

The proposed scheme architecture consists of:

1) Participating devices: Wi-Fi-enabled devices, such as a smartphone installed on it a relevant application.
2) Wireless access points are located within the school: Access points utilize to determine a student's position within the room by automatically scanning the student's environment and sending the measured RSSI readings off to the remote server.
3) A remote server collocates the RSSI reading received from a student's Wi-Fi-enabled device and determines which students are present and which students are absent.

4) An application located on the server which totals the calculations and sends the attendance list to the instructor.

The Proposed school attendance system is zero-effort and zero-interaction as the data collection and verification occurs without the users' involvement, whether the user is the professor conducting attendance or the student who is present in class.

B. Operation

- Client-side interaction:
  The student will need to register their device prior to the beginning of the first day of class or the first day of implementation. Then, the student enters the classroom and finds a seat within the room. While this is occurring, the professor enters the classroom and begins to prepare for the lesson. As students begin to enter the classroom, their wireless devices are started to receive the broadcast signals that transmitted by schools' routers. As the class continues, the messages continue, sending the results back to the server. When a class is over, the attendance list is sent to the professor, that shows which students were present in comparison to the students who were absent.
- Server-side interaction:
  This section shows the authentication procedure between the student devices, the access points, and the server. The steps and an overview can be found in Figure 1
  1) A student enters a classroom.
  2) The server instructs the student's device to scan.
  3) The student's device sends the needed information (i.e., RSSI values) back to the server.
  4) The RSSI values will be used with equation (1) to calculate the distance from the student's device to the access points.

$$PL_{log} = PL_0 + 10\gamma \; log_{10} \frac{d}{d_0} \tag{1}$$

  where PL_{log} is the transmitted power minus the power which is received. PL_0 is the path loss in dB, γ is the path loss exponent, and d_0 is the reference distance (1m).
  5) The results from step 4 will be plugged in equation (2) to determine the student's position (i.e., the student's x and y coordinates).

$$\begin{aligned} (x - x_1)^2 + (y - y_1)^2 &= d_1^2 \\ (x - x_2)^2 + (y - y_2)^2 &= d_2^2 \\ (x - x_3)^2 + (y - y_3)^2 &= d_3^2 \end{aligned} \tag{2}$$

  6) The system uses the student's position to determine whether the student is present or absent. These results get sent to the instructor at the end of the lesson.

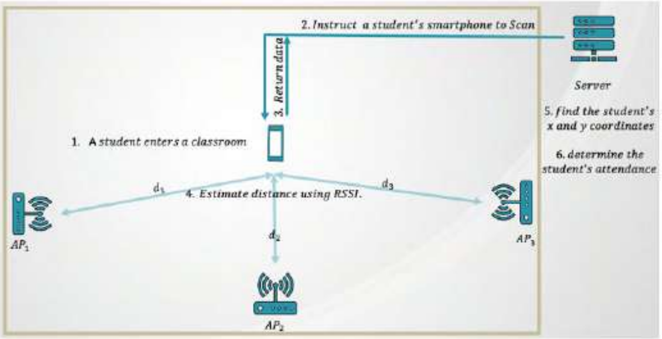

2. Instruct a student's smartphone to Scan
3. Return data
Server
5. find the student's x and y coordinates
6. determine the student's attendance
1. A student enters a classroom
d_1
4. Estimate distance using RSSI.
d_3
d_2
AP_1
AP_3
AP_2

Fig. 1. School Attendance System Overview

IV. Experiment

This section describes the setup for our experiment and the multiple experiments which were conducted. A smart device was placed throughout the 18ft by 9ft room. The access points were placed in a triangular position within the room. The positions of the access points were measured for our experiment. The smart device was also located in specified, physically mapped locations, so we may scan the access points along with the corresponding RSSI values. Our data was then collected from the physically mapped access points and calculated for authenticity.

A. Location Accuracy

Three experiments were conducted to test the location accuracy of the devices using RSSI values. In each experiment, the access points scanned for the devices 30 times from 7 different points. These values were then calculated, giving us the first values, the maximum values, and the average values. Results are shown in Tables I, II, and III below.

TABLE I
Experiment One (First)

Actual points	*Calculated points*	*Difference*
(2 , 17)	(4.2 , 14.3)	1.1 m
(3 , 2)	(1.3 , 7.7)	1.8 m
(4.5 , 9)	(4.3 , 9.3)	0.1 m
(5 , 14)	(4.1 , 12.3)	0.6 m
(6 , 2)	(10.7 , -1)	1.7 m
(7 , 2)	(-0.1 , 7.4)	2.7 m
(8 , 17)	(6.4 , 9)	2.2 m

TABLE II
Experiment Two (Mix)

Actual points	*Calculated points*	*Difference*
(2 , 17)	(4.2 , 14.3)	1.1 m
(3 , 2)	(1.5 , 7.3)	1.7 m
(4.5 , 9)	(3.3 , 10)	0.5 m
(5 , 14)	(3.8 , 12.5)	0.6 m
(6 , 2)	(5.3 , 4)	0.6 m
(7 , 2)	(1.3 , 6.6)	2.3 m
(8 , 17)	(4.6 , 9.9)	2.1 m

TABLE III
EXPERIMENT THREE (AVERAGE)

Actual points	***Calculated points***	***Difference***
(2 , 17)	(5.2 , 15.4)	1.1 m
(3 , 2)	(1.4 , 7.5)	1.7 m
(4.5 , 9)	(3.9 , 9.9)	0.3 m
(5 , 14)	(4 , 12.6)	0.5 m
(6 , 2)	(7 , 2.4)	0.3 m
(7 , 2)	(3.2 , 4.5)	1.4 m
(8 , 17)	(10.5 , 15)	0.8m

B. Success Rate

In this experiment, the proposed scheme's success rate was assessed by applying every RSSI value to calculate the location of the device using equation (1) and equation (2). The results are showing in Table IV. The proposed attendance system provided a 95.24% success rate that was calculated using equation (3).

TABLE IV
SUCCESS RATE

	Actual	
N=210	***Positive***	***Negative***
True	TP = 108	FP = 4
False	FN = 6	TN = 92

$$\frac{True\ Positive + True\ Negative}{N, Total\ number\ of\ a\ dataset} \times 100 \tag{3}$$

V. CONCLUSION

As shown from the experiments, the proposed attendance system is an effective, zero-effort and zero-interaction method for conducting attendance in a post-pandemic society. From the results of our experiment, we believe our proposed attendance system is a sound method due to its success rate of 95.24%. The experiments were performed during different times of the day, and we believe this system will encourage instructors to take attendance in a safe and fool-proof way. In the age of the pandemic, implementing an effective and socially distant attendance method is integral for schools across the world.

REFERENCES

[1] AlQahtani, A., Alamleh, H. and Gourd, J. 0EISUA: Zero Effort Indoor Secure User Authentication. IEEE Access, 8, pp.79069-79078, 2020.

[2] U. Koppikar, S. Hiremath, A. Shiralkar, A. Rajoor and V. Baligar, ”IoT based Smart Attendance Monitoring System using RFID”, 2019 1st International Conference on Advances in Information Technology (ICAIT), pp. pp. 193-197, 2019.

[3] M. Liu, X. Zhang and Y. Han, ”Intelligent Counting System for Classroom Numbers Based on Video Surveillance”, 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2020.

[4] X. Bai, F. Jiang, T. Shi and Y. Wu, ”Design of Attendance System Based on Face Recognition and Android Platform”, 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2020.

[5] Z. Lin and Y. Li, ”Design and Implementation of Classroom Attendance System Based on Video Face Recognition”, 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2019.

[6] S. Chennattu, A. Kelkar, A. Anthony and S. Nagdeote, ”Portable Biometric Attendance System Using IOT”, 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019.

[7] Q. Miao, F. Xiao, H. Huang, L. Sun and R. Wang, ”Smart attendance system based on frequency distribution algorithm with passive RFID tags”, Tsinghua Science and Technology, vol. 25, no. 2, pp. 217-226, 2019.

[8] A. Alqahtani, H. Alamleh and J. Gourd, ”BF2FA: Beacon Frame Two-factor Authentication”, 2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat), 2020.

[9] M. Ali, H. Usman Zahoor, A. Ali and M. Ali Qureshi, ”Smart Multiple Attendance System through Single Image”, 2020 IEEE 23rd International Multitopic Conference (INMIC), 2020.

Machine Learning Approach For Clustering Of Countries To Identify The Best Strategies To Combat Covid-19

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Venkatasubramanian J
Student – AIML
Great Learning
Bengaluru, India
srikanth1995jvs@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Bengaluru, India
anwesh@greatlearning.in

Prashant Kumar
Student – AIML
Great Learning
Bengaluru, India
kr12.prashant@gmail.com

Vigneswar S
Student – AIML
Great Learning
Bengaluru, India
vigneswarsundaramurthy@gmail.com

Krishna Chaitanya Thangeda
Student – AIML
Great Learning
Bengaluru, India
tkchaitanya05@gmail.com

Abhaya Thakur
Student – AIML
Great Learning
Bengaluru, India
abhaya.thakur@gmail.com

***Abstract*— The purpose of this study is to identify how different government measures impacted the level of Covid-19 influence on countries of similar nature. Demographic, economic, health, and weather conditions were considered to identify countries that are inherently similar in nature. This grouping along with Covid-19 epidemiology data was used to cluster countries over a period of time after Covid-19 struck. We identified those countries which changed clusters over a time period and were influenced differently by the impact of Covid-19. We then looked at the government measures through the stringency index of containment measures and observed a relation in how different stringency measures impacted the countries differently even though they belonged to the same original group. We also observed that countries that eased restrictions quickly after containment measures had to go back to the earlier stringent measures. Gradual ease of containment measure was more efficient in tackling Covid-19. The inherent grouping of countries done in our study can be used in the future as well to deploy similar measures when faced with Covid-19 like pandemic situation. The strategies adopted on average by countries within each inherent cluster can become the base for handling Covid-19 or any such pandemic in the future. The significance of the work resides in the fact that the strategies would not be aligned to economic conditions of a nation (developed versus developing) or a single factor like healthcare facilities but based on a varied list of inherent factors using machine learning methods.**

Keywords — Covid-19, k-means, clustering, unsupervised learning, machine learning

I. Introduction

After Covid-19 struck the world in early December 2019 and until the World Health Organization (WHO) declared the novel coronavirus outbreak as a public health emergency of global concern on 30th January 2020 [1], none of the countries were prepared on how to deal with such an emergency. There were varying degrees of strategies and responses adopted by different countries with varying degrees of effectiveness [2], [3], [4]. Some countries went as far as complete lockdown [5], [14] and other countries went for herd immunity (Sweden [16], [17]). Some countries tried for herd immunity, but due to increased infection rates and the associated burden on the health infrastructure of the nations they had to change the strategy to lock down and restrictions (United Kingdom [5]). There is no clear solution that people know might work for any particular country at any stage of the pandemic impact.

The purpose of this study is to help the countries identify what strategies they can adopt at what stage of a pandemic and what might work best for that country based on their socio-economic and other relevant parameters. The premise is that what might work best for one country may not be ideal for other countries at any given stage of the pandemic. Covid-19 was one such pandemic which had a global impact where the majority of the countries were unprepared to handle such a crisis. This research using Machine Learning techniques on historical data of Covid-19 is an attempt to get meaningful suggestions for policy makers and governments to be better prepared for any such future risks based on the countries classification type at that point in time.

There are already some studies done using machine learning techniques to predict the Covid-19 cases and its impact. One such study tried to identify the covariates that are important to predict the total confirmed cases and whether there are clusters of countries possibly associated with these covariates [6]. Data considered for this study was the global Covid-19 data for 133 countries from the worldometer website [18]. Their study indicated 4 major clusters of countries that contributed towards the global number of confirmed cases. This study only considered the Covid-19 related data for the clustering purpose and doesn't consider what type of country it was before the pandemic struck. There is an opportunity here to analyze why two countries which might be similar by demographic, economic, health and weather parameters before the pandemic could end up in

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

different clusters when affected by the pandemic. The difference could be because of the response action of the government agencies, which is what we aim to analyse as part of this study.

Another study applied supervised machine learning (ML) approaches to identify the key factors affecting COVID-19 infected and death counts from a list of features including GDP, gender, ethnicity, health care, homeless, lockdown type, population density, airport activity, and age groups [5]. They concluded that population density, testing numbers and airport traffic emerge as the most discriminatory factors, followed by higher age groups (above 40 and specifically 60+). We wish to consider these factors as part of our analysis for clustering purposes.

Other studies on impact of weather [14], age [12], comorbidities [11], [12], vaccinations [13], economic and social factors [5] on the Covid-19 cases have also been carried out. The aim of those studies was to see to what extent each of the individual attributes are correlated to the Covid-19 cases. But to the best of our knowledge there is a lack of study that looks at demographic, economic, health and weather parameters holistically together with the infection rate to identify clusters of countries.

In this paper we tried unsupervised machine learning techniques like k-means clustering and agglomerative clustering to cluster a set of 142 countries. We took a 2-stage process where we first tried to identify which countries are similar in nature (fall into the same group) based on demographic, economic, health and weather conditions. This would give us an indication of the countries that are similarly placed with respect to the potential influence of pandemic on them. We then considered this information as a feature along with the global covid-19 data to see how the clustering of the countries within each original group changes over a period of time. This is a unique approach taken by the authors as the inherent nature of a country is considered as one of parameters in clustering along with Covid-19 data. Further based on the strategies adopted by different countries within the clusters we will aim to identify what strategy or combination of state policies might work best for the cluster to deal with Covid-19 impact. Our hypothesis is that most of the variation within the same inherent group could be explained by a difference in the government response and mitigation strategies adopted by various countries. This analysis can give rich insights on what response strategy is best for which inherent country group at any particular stage during the fight against pandemic like Covid-19.

II. Dataset and Methods

A. *Dataset used for the Analysis*

We wanted to process several open-access datasets to collect parameters like the weather of the nation, general health index of people, medical facilities, economic conditions, poverty levels, population density, demographic details etc. to create an integrated dataset of overall factors that might impact COVID cases in the countries. One of the repositories that compiles this information is [19] and available here: *https://github.com/GoogleCloudPlatform/covid-19-open-data*.

We have used the GoogleCloudPlatfrom Covid-19 open data from 1st January 2020 till 11th of December 2020 for the purpose of this study. The main data set is this repository is daily level time series data that comprises demographic, economic, epidemiology, geography, health, hospitalizations, government response and weather data. We considered the data at country level and identified data attributes under 3 categories:

1. Demographic, economic, health and weather data (pre-Covid data)
2. Covid-19 epidemiology data
3. Data related to government measures and responses

We have identified 142 countries for our analysis from the main data set. These countries have data captured for all the important attributes which are required for analysis.

B. *Methodology*

The data from the first category mentioned in the above section was used for clustering exercise to identify the inherent clusters that each of the 142 countries falls into. Since the data points considered in the first category don’t change by time series on a daily or monthly level, we have considered the data points for a date before the impact of pandemic. We selected a data view on the 15th of March 2020 for the filtered 142 countries to carry out the clustering exercise. The mentioned data set has so many dimensions and few are similar to each other. In order to improve efficiency in forming clusters, PCA is performed. principal components are extracted with a goal of at least 90% explained variance. K-means clustering (Fig. 1) and agglomerative clustering (Fig. 2) was done on the data set and we found 4 prominent clusters emerging out of the analysis.

We used the k-means clustering method with 4 clusters to identify the inherent groups of the 142 countries. Table 1 captures the grouping of the countries from this clustering exercise.

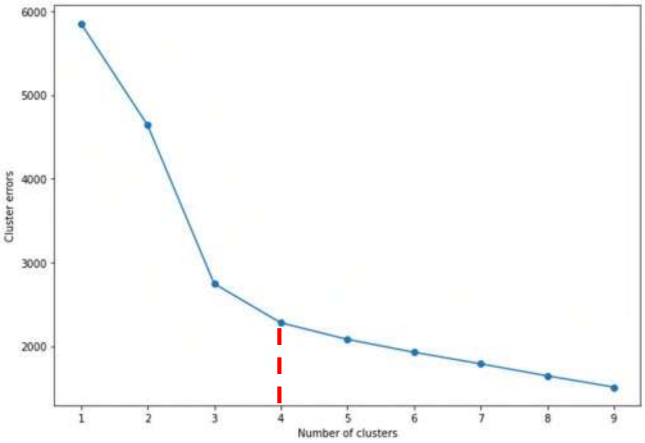


Fig. 1. Elbow plot of k-means clustering to identify inherent country clusters. There are 4 major clusters observed from the plot.

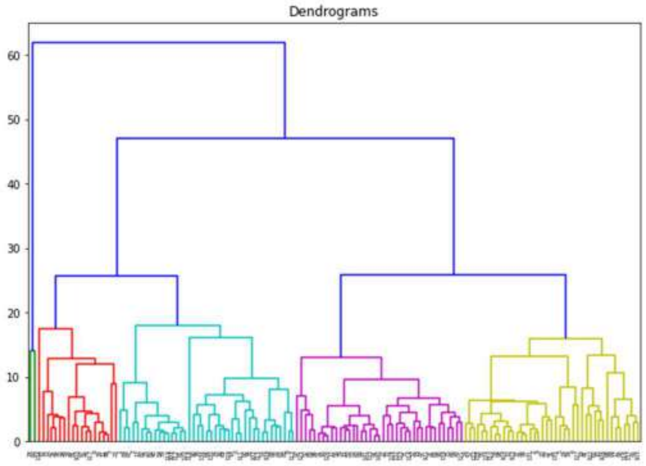


Fig. 2. Four major inherent country clusters observed from Dendrogram distances from agglomerative clustering exercise.

We then used the epidemiology data of Covid-19 to generate a Covid-19 indexed monthly time series data. The day when each of the country reaches 10 Covid-19 cases is indexed as the start of the 0th time period. We considered a moving time period of 30 days to get to the monthly indexed time series data. The epidemiology data was normalized by per million population or as a percentage of Covid-19 cases to get all countries' data to a comparable scale.

TABLE I. INHERENT CLUSTERS WITHOUT COVID-19 DATA

Group 0	Afghanistan, Angola, Bangladesh, Burkina Faso, Burundi, Benin, Bhutan, Botswana, Ethiopia, Fiji, Gabon, Ghana, Gambia, Guinea, Equatorial Guinea, Guinea-Bissau, Guyana, Haiti, Indonesia, Kenya, Comoros, Laos, Liberia, Mali, Myanmar, Malawi, Mozambique, Namibia, Niger, Nigeria, Nepal, Papua New Guinea, Philippines, Pakistan, Rwanda, Senegal, São Tomé and Príncipe, Chad, Togo, East Timor, Tanzania, Uganda, Vanuatu, South Africa, Zambia, Zimbabwe
Group 1	Armenia, Austria, Australia, Bosnia and Herzegovina, Belgium, Bulgaria, Belarus, Canada, Switzerland, Cyprus, Czech Republic, Germany, Denmark, Estonia, Spain, Finland, France, United Kingdom, Greece, Croatia, Hungary, Israel, Iceland, Japan, South Korea, Lithuania, Luxembourg, Latvia, Moldova, Mongolia, Malta, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Sweden, Slovenia, Slovakia, United States of America
Group 2	China, India
Group 3	United Arab Emirates, Antigua and Barbuda, Argentina, Barbados, Bahrain, Brunei, Bolivia, Brazil, Bahamas, Belize, Chile, Colombia, Costa Rica, Cuba, Cape Verde, Dominican Republic, Algeria, Ecuador, Egypt, Grenada, Guatemala, Honduras, Iraq, Iran, Jamaica, Jordan, Kuwait, Lebanon, Saint Lucia, Sri Lanka, Morocco, Mauritius, Maldives, Mexico, Malaysia, Nicaragua, Oman, Panama, Peru, Paraguay, Qatar, Saudi Arabia, Solomon Islands, Singapore, Suriname, El Salvador, Thailand, Tunisia, Turkey, Trinidad and Tobago, Uruguay, Vietnam, Samoa

The inherent group number from the earlier mentioned clustering exercise is also considered as one of the attributes for identifying the cluster of the countries post Covid-19 impact. We then performed the k-means clustering to identify 7 clusters (Fig. 3) into which the countries can fall into as the impact of Covid-19 progressed over the indexed time period.

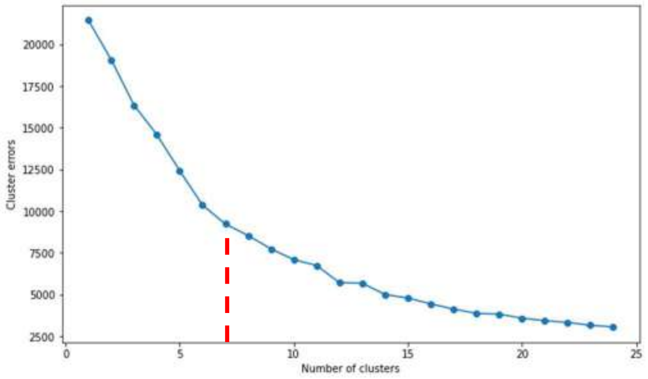


Fig. 3. Elbow plot of k-means clustering of post Covid-19 impact on countries shows 7 major clusters.

We then looked at each of the inherent groups individually to see how the countries which stacked up together in the same cluster pre-Covid moved into different clusters based on the extent of Covid-19 impact. The difference and movement of different countries into different clusters from the same inherent clusters can be attributable to the responses and measures taken by the local governments which we evaluated through the stringency index parameter provided by Oxford Covid-19 government response tracker [20]. The stringency index captures all the containment and closure policy indications along with the public information campaign indicator.

III. RESULTS AND OBSERVATIONS

A. *Results from Group Level Clusters*

When we observed the countries from Group 0 from Table 1, for the differences in impact of Covid-19 we didn't notice much difference. All those countries continued to fall in the same cluster post Covid-19 impact which also didn't change over a period of time. Same goes for the Group 2.

There were some differences noticed between a few countries from Group 1 as depicted in Fig. 4. The differences were pronounced after time period 6 and later. Till period 6 mostly all countries continued to fall under the same cluster barring a couple of exceptions. For the purpose of further analysis in this paper we are considering the time period 6 as a representative view of the differences between the countries. This also provides a good time period for evaluating the different measures adopted by various countries as the governments would have got enough time of 6 months to act upon the pandemic threat.

From Fig. 4 we could notice that while most countries had similar impact there were a few countries which ended up being different. Belgium, Denmark, France, Iceland, Israel, Malta and the United States are countries which fell into different clusters compared to the rest of the nations. Among them Belgium and France seem to have experienced unique scenarios because of Covid-19.

Similarly, for Group 3 there were a few countries which had different impact than the rest of the countries in the same group. Argentina and Brazil had similar experiences while Bahrain, Jordan, Kuwait, Maldives, Panama, Qatar and Trinidad and Tobago had different experiences in dealing with Covid-19 as shown in Fig. 5.

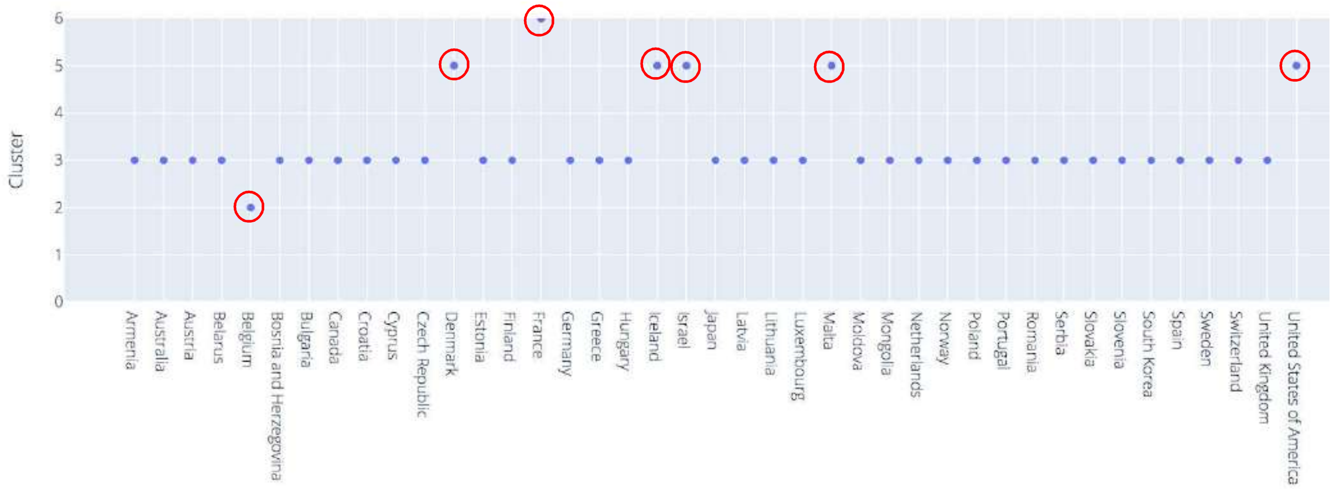


Fig. 4. Clustered view of countries in Group 1 at time period 6 (roughly 6 months) after the start of Covid-19 shows countries that show different behavior compared to others (Belgium, Denmark, France, Iceland, Israel, Malta and the USA).

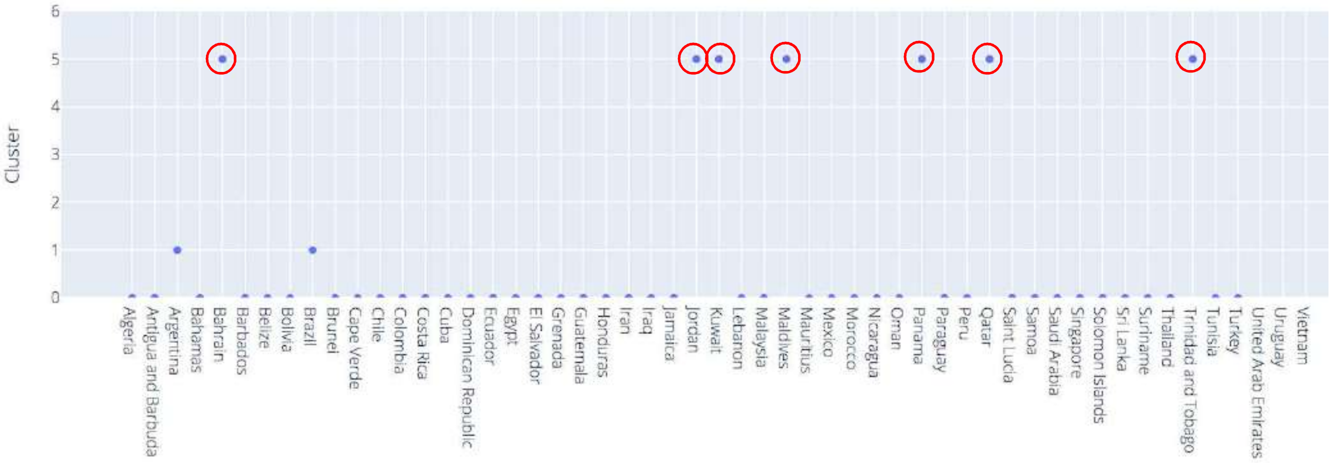


Fig. 5. Clustered view of countries in Group 3 at time period 7 (roughly 7 months) after the start of Covid-19 shows countries that exhibit different behavior than the others in the group (Bahrain, Jordan, Kuwait, Maldives, Panama, Qatar, Trinidad and Tobago)

B. Analysis of Government Measures

To analyze the government measures we looked at the stringency index of countries within each inherent group and the observations for Group 1 are captured in Fig. 6, Fig. 7 and Fig. 8. Note that as a representative of Group 1 we considered Germany as reference which happens to fall in the same cluster as the rest of the typical countries in Group 1.

Measures taken by the government generally played an important role to curb Covid. As mentioned earlier, countries from the group 1 like Belgium, Denmark, Malta, Israel, Iceland and France not only had a difference in Covid spread but also in the measures taken by the government in comparison to other countries. Countries like Denmark and Iceland Israel did not have a complete lockdown in the month of march to June unlike other countries which can clearly be observed in Fig. 7. Iceland never had a complete lockdown through 2020. Though these countries had restrictions in a limited scale, the lockdown and strict curfew were imposed widely later in Q3 2020.

Belgium had early government measures related to containment compared to other countries as observed from Fig. 6. This resulted in fewer cases in Belgium and because the government continued with the containment measures, it stayed comparatively different from the rest of the countries in the group.

On the other hand France imposed a lockdown in march and relaxed the lockdown at the end of April. Similarly Israel had varying government measures related to containment and lockdown. As France and Israel had a rise in cases, the countries went into lockdown again and had curfew during later parts of 2020.

It can be observed from Fig. 9 that both Argentina and Brazil had stricter government measures as compared to the average of the countries in the inherent Group 3. While no specific pattern could be observed for the other countries (Bahrain, Jordan, Kuwait, Maldives, Panama, Qatar and Trinidad and Tobago) that differed from average of Group 3, it can be observed from Fig. 10 that these countries tried to alter the containment policies a bit too quickly resulting in going back to the earlier stricter measures.

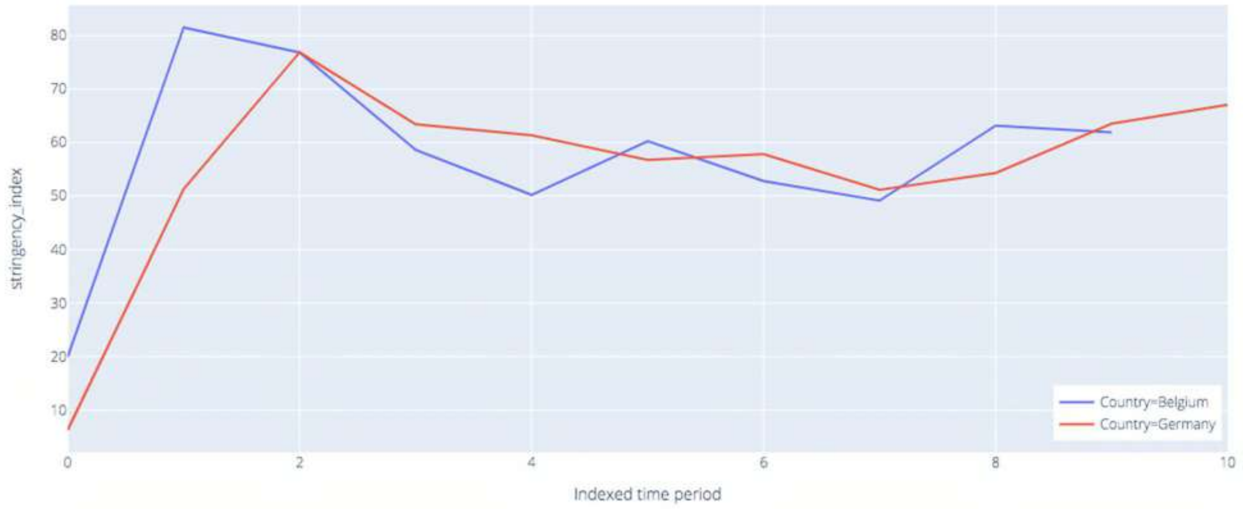


Fig. 6. Stringency index comparison between Germany & Belgium over a time period of 10 months from start of Covid-19. Notice the early government measures adopted by Belgium.

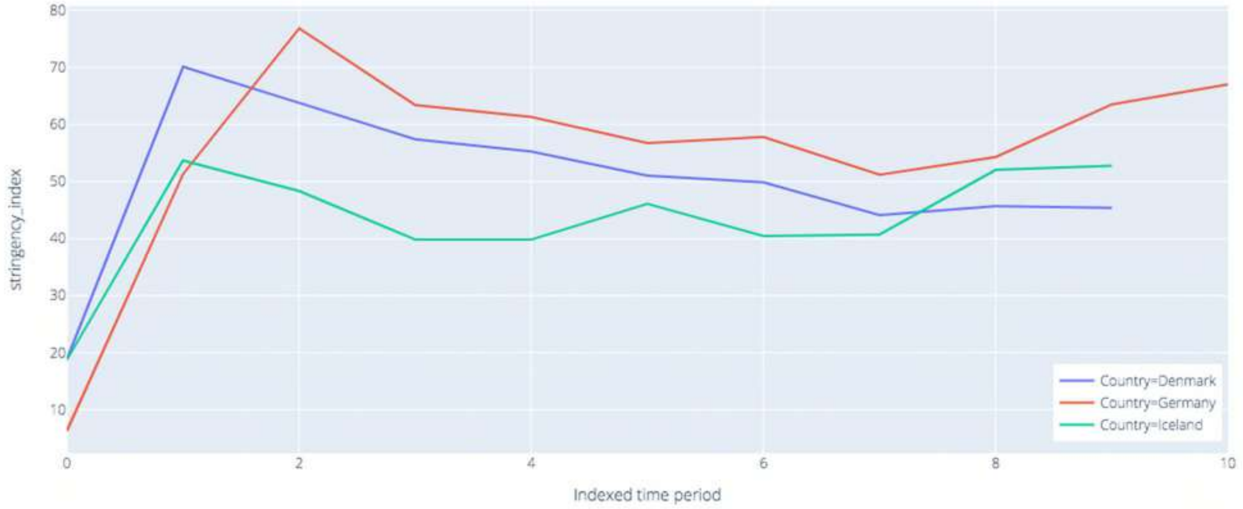


Fig. 7. Stringency index comparison between Germany, Denmark and Iceland. Notice Iceland had little government restrictions while Germany was more proactive to take appropriate containtment measures.

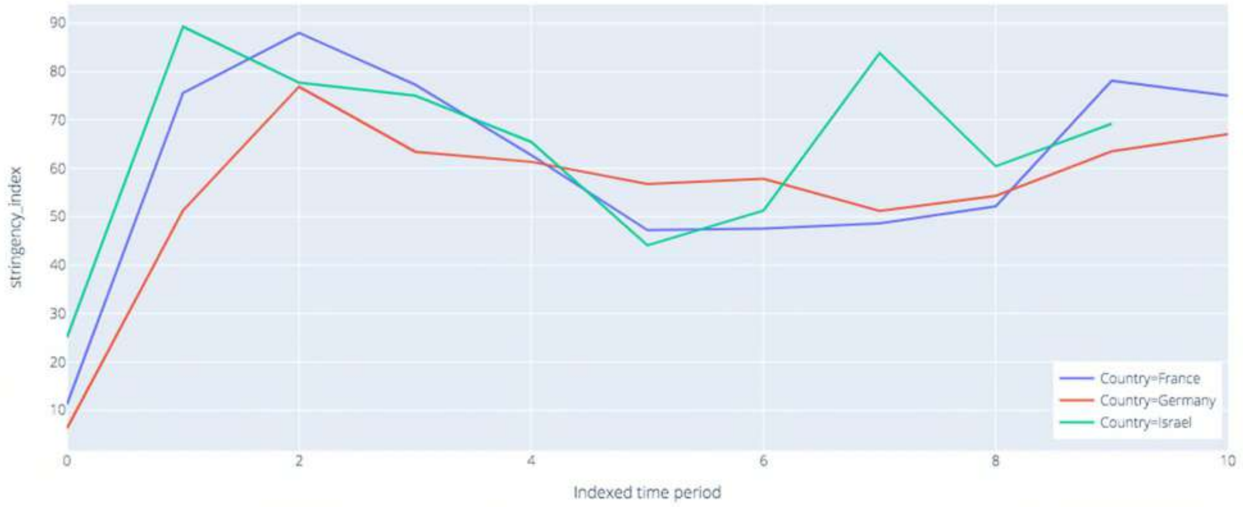


Fig. 8. Stringency index comparison between Germany, France and Israel. Notice how France and Israel had to go back to stricter restrictions because of faster easing of government measures.

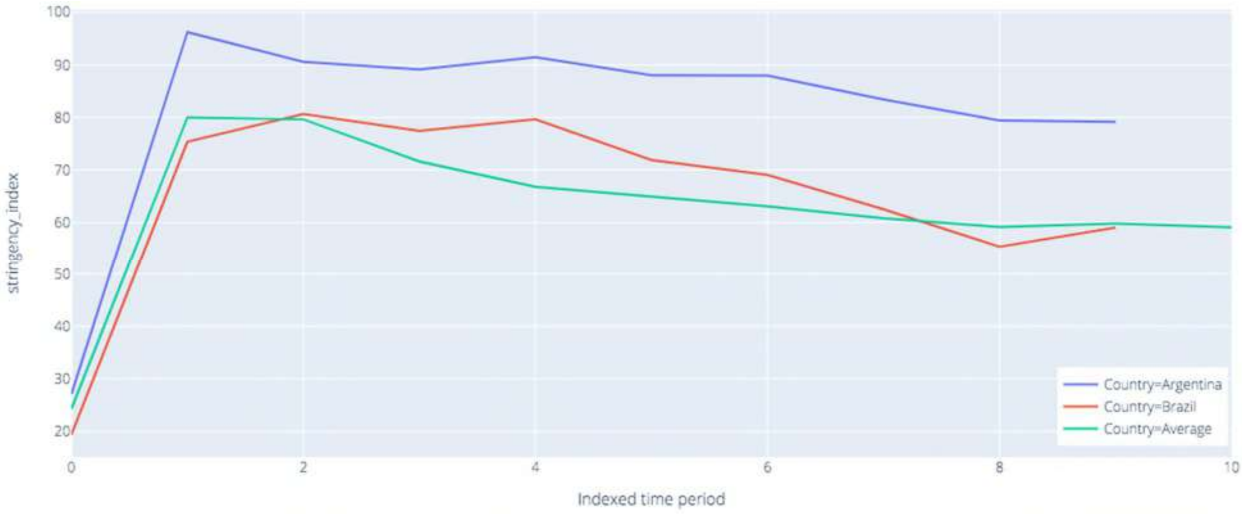


Fig. 9. Stringency index comparison between Average of Group 3 countries, Argentina and Brazil. Argentina had more stricter measures which were eased out in a very gradual manner which seemed to help fight Covid-19 effectively.

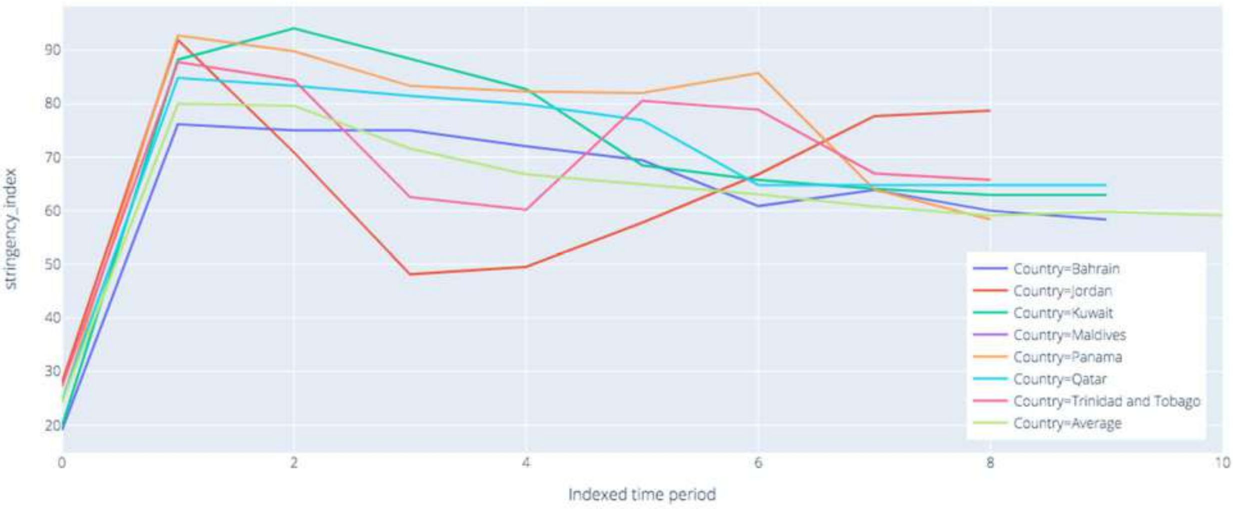


Fig. 10. Stringency index comparison between average of Group 3 countries and other deviating countries. Gradual easing of restrictions seem to be ideal.

IV. Discussions and conclusions

This study looked at identifying inherent groups of countries from a perspective of potential impact to pandemic like Covid-19. These countries within each group can continue to be looked at as a cluster for policy decisions making in future and potential impact from Covid-19 like scenario in future. The impact of Covid-19 on some of the countries within each inherent group was different most possibly due to the government containment measures and policies adopted locally. While it is obvious that countries which took a stringent containment policy stood out from the rest and also the other way around as well, one observation that came out from our study is that countries which eased out the restrictions within a month or two had to suffer adverse consequences and had to go back to the earlier restrictions. Examples being France, Israel (Fig. 8), Jordan and Trinidad and Tobago (Fig. 10). The observation infers the countries to start with stricter containment measures quickly and ease out restrictions slowly and gradually rather than sudden eases. Argentina (Fig. 9) is a good example of how the restrictions could be eased out in a controlled way, so the countries won't have to go back to stricter measures to avoid a spike in cases.

Oxford Covid-19 government response tracker [20] captures various containment, health system, economic policies taken by 180 countries. A combination of various parameters are used to calculate four major indices - overall government response index, containment and health index, stringency index and economic support index. This study only evaluated the effect of the stringency index (which majorly captures the containment and closure policies) on the differences of Covid-19 impact on similarly grouped countries. We didn't look at the combination of all the indices or permutations and combinations of the indices on what impact it would have on the way countries were impacted by Covid-19. That could be a good next step to this exercise to understand which indices could have a major impact on countries falling into different clusters and which indices might have a high correlation with the Covid-19 cases.

Further analysis could also be conducted at the individual indicators which contributed significantly towards the stringency index like closing public transport, restrictions on public gathering etc. This will help to understand what measures works better for a specific cluster.

References

[1] Timeline: WHO's COVID-19 response. *(https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#)*

[2] Krishnakumar, B., & Rana, S. (2020). COVID 19 in INDIA: Strategies to combat from combination threat of life and livelihood. Wei Mian Yu Gan Ran Za Zhi [Journal of Microbiology, Immunology, and Infection], 53(3), 389–391.

[3] Huang, N. E., Qiao, F., & Tung, K.-K. (2020). A data-driven model for predicting the course of COVID-19 epidemic with applications for China, Korea, Italy, Germany, Spain, UK and USA (p. 2020.03.28.20046177). doi:10.1101/2020.03.28.20046177

[4] Peak, C. M., Childs, L. M., Grad, Y. H., & Buckee, C. O. (2017). Comparing nonpharmaceutical interventions for containing emerging epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(15), 4023–4028.

[5] Roy, S., & Ghosh, P. (2020). Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. *PloS One*, *15*(10), e0241165.

[6] Rahaman Khan, M. H., & Hossain, A. (2020). Countries are Clustered but Number of Tests is not Vital to Predict Global COVID-19 Confirmed Cases: A Machine Learning Approach (p. 2020.04.24.20078238). doi:10.1101/2020.04.24.20078238

[7] Khrapov, P., & Loginova, A. (2020). Mathematical modelling of the dynamics of the Coronavirus COVID-19 epidemic development in China. *International Journal of Open Information Technologies*, *8*(4), 13–16.

[8] Tang, B., Bragazzi, N. L., Li, Q., Tang, S., Xiao, Y., & Wu, J. (2020). An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). *Infectious Disease Modelling*, *5*, 248–255.

[9] Gompertz, B. (1825). XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. *Philosophical Transactions of the Royal Society of London*, *115*(0), 513–583.

[10] Khalifa, S. A. M., Mohamed, B. S., Elashal, M. H., Du, M., Guo, Z., Zhao, C., … El-Seedi, H. R. (2020). Comprehensive overview on multiple strategies fighting COVID-19. *International Journal of Environmental Research and Public Health*, *17*(16), 5813.

[11] Singh, A. K., & Misra, A. (2020). Impact of COVID-19 and comorbidities on health and economics: Focus on developing countries and India. *Diabetes & Metabolic Syndrome*, *14*(6), 1625–1630.

[12] Sanyaolu, A., Okorie, C., Marinkovic, A., Patidar, R., Younis, K., Desai, P., … Altaf, M. (2020). Comorbidity and its Impact on Patients with COVID-19. *SN Comprehensive Clinical Medicine*, *2*(8), 1–8.

[13] Patella, V., Delfino, G., Bruzzese, D., Giuliano, A., & Sanduzzi, A. (2020). The bacillus Calmette-Guérin vaccination allows the innate immune system to provide protection from severe COVID-19 infection. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(41), 25205–25206.

[14] Jeyanthi, V., & Department of Biotechnology, SRM Arts and Science College,, Kattankulathur, Chengalpattu, 603203, Tamil Nadu, India. (2020). COVID-19 outbreak: An overview and India's perspectives on the management of infection. *Indian Journal of Science and Technology*, *13*(36), 3716–3724.

[15] Malki, Z., Atlam, E.-S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons, and Fractals*, *138*(110137), 110137.

[16] Orlowski, E. J. W., & Goldsmith, D. J. A. (2020). Four months into the COVID-19 pandemic, Sweden's prized herd immunity is nowhere in sight. *Journal of the Royal Society of Medicine*, *113*(8), 292–298.

[17] McNaughton, C. D. (2020). Herd immunity: Knowns, unknowns, challenges, and strategies. *American Journal of Health Promotion: AJHP*, *34*(6), 692–694.

[18] Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus Pandemic (COVID-19). *Our World in Data*. Retrieved from https://ourworldindata.org/coronavirus

[19] GoogleCloudPlatform. (n.d.). GoogleCloudPlatform/covid-19-open-data. Retrieved February 21, 2021, from Goo.gle website: https://goo.gle/covid-19-open-data

[20] Coronavirus Government Response Tracker. (n.d.). Retrieved February 21, 2021, from Ox.ac.uk website: https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker

[21] S. Chikode, N. Hindlekar, P. Padhye, N. Darapaneni, and A. R. Paduri, "COVID-19: Prediction of Confirmed cases, active cases and health infrastructure requirements for India," International Journal of Future Generation Communication and Networking, vol. 13, no. 4, pp. 2479–2488–2479–2488, 2020.

Food Image Recognition and Calorie Prediction

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Subhav Kataria
Student – AIML
Great Learning
Delhi, India
kataria217@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Delhi, India
anwesh@greatlearning.in

Vishal Singh
Student – AIML
Great Learning
Delhi, India
vishalsngh1992@gmail.com

Nayana Bansal
Student – AIML
Great Learning
Delhi, India
naybansa09@gmail.com

Yasharth Singh Tarkar
Student – AIML
Great Learning
Delhi, India
yasharthsinghrajput@gmail.com

Abhijeet Kharade
Mentor – AIML
Great Learning
Delhi, India
Kharadeabhi@gmail.com

***Abstract* -- With the increasing number of health issues reported due to obesity and overeating, people have become cautious about their diet intake to prevent themselves from the diseases such as hypertension, diabetes, and other heart related problem which are caused due to obesity. As per the data shared by WHO, at least 2.8 million people are dying each year because of being overweight or obese. The important part of any healthy diet plan is its calories intake. Hence, we propose a deep learning-based technique to calculate the calories of the food items present in the image captured by the user. We used a layer-based approach to predict calorie in the food item which include Image Acquisition, Food item classification, Surface area detection and calorie prediction.**

Keywords – Mask R-CNN, ROI (Region of Interest), IOU (Intersection Over Union), Resnet 50

I. INTRODUCTION

It is very important in today's time that people should be aware of what they are consuming and what will be its impact on the body. So, a system that can help individuals to maintain their calories intake is very important. Most of the world's population live in countries where overweight and obesity kills more people than any other health disease. The problem here is not about having enough food, it is about the people not knowing what is in their diet. If people could estimate their calorie intake during a day, they can easily decide on the number of calories they want to consume. However, managing calorie intake is a very cumbersome task which involves the people to manually keep a track of food item they have consumed throughout the day and they must determine the calories they have consumed. This process is not only manual but also inaccurate as the calorie estimation not only depends on what you are eating but also depends on how much are having.

With the advancement in the field of image processing techniques, the image recognition models are in demand. Researchers are aggressively deploying image recognition model for various uses such as self-driving cars, cancer detection, video frame analysis etc. Researchers have also shown keen interest in predicting the calories present in the food item with the help of the image. Researchers have used various machine learning and deep learning techniques to perform the task of calories estimation with the help of supplied images.

Manal Chokr and Shady Elbassuoni (2017) [8] proposed a solution in which they used the supervised learning technique to perform the single food classification and its calorie prediction. They used the model which takes an image of the food item as input and provides the calorie as output. They developed a Mathwork Image Process tool which was used to extract features from the image. Extracted features were then compressed and fed to a classifier and regressor to identify the food type and determine the size of the food item respectively. Finally, the output of both classifier and regressor is fed to another regressor which provides the calories as the output.

Parisa Pouladzadeh1, Abdulsalam Yassine1, Shervin Shirmohammadi (2015) [9] developed a deep learning-based model which takes a food image clicked from the mobile camera which is capable of estimating calorie for the mixed portion of the food item as well. The dataset used contained 3000 images clicked under different condition with different camera model and then the clicked image is given as input. They used color segmentation, k-mean clustering, and texture segmentation tools. They employed Cloud SVM and deep neural network to increase the performance of the image identification model. For calorie prediction they used the reference object approach wherein they mandated the presence of the thumb in the image so that their model can use the thumb present in the mage as reference for the size estimation of food item present in the image which helped in calorie estimation.

In earlier researches, the researchers have collected 101 different food images [10] such as Chicken Wings, Tacos,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Bread pudding etc. The dataset has 101 food categories with 101000 images where most of the images contains mixed food items. They used 750 images for model training purpose and 250 images for testing purpose. They used model based on Random Forests to mine discriminative visual components and efficient classification of the images. However, they used the model for the image classification only

Nowadays, there are many android and web-applications available which allows users to track the food intake to help them manage their calorie. However, these applications rely on the users to manually select the food items they had and the size and dimension of the food items they ate to track record of the diet and to further estimated the calorie consumption. This process is not accurate as it is almost impossible for the user to input the accurate food size they ate.

Hence, we propose a solution to this problem. The proposed model can predict the calories user consumed with the help of the image of the food portion user had. The model basically makes use of layered approach wherein each layer performs different tasks. The first layer takes the image as the input from the user and prepares it to be fed to the second layer. The second layer employs the deep learning method known as Mask R-CNN which performs the food identification task along with the bounding box and mask generation. The third layer calculated the surface area covered by the food items which in turn is fed to the last layer. The fourth layer predicts the calories consumed by the user based on the surface area covered by the food item present in the supplied image.

The model provides satisfactory results in estimating the calories intake by the user with the help of image. This model uses the dataset which contains 638 images of food item which belongs to 6 different categories (Apple, Oranges, Omelet, Pizza toast, Banana and Idli). These images are collected from internet, hand-picked from the available dataset such as Food 101[4], UNIMIB2016[5] and UEC FOOD-100[6]. The images are preprocessed as per the model requirement and used to train our model.

In this paper, the model employs Mask R-CNN deep learning technique for the mask and bounding box determination for the image containing different food items which helps the model to identify the food item, estimate the surface area occupied by the food item and to determine the calorie associated with it with the help of defined mathematical formula. The model provides the satisfactory results in determining the calorie in the food portion present in the image.

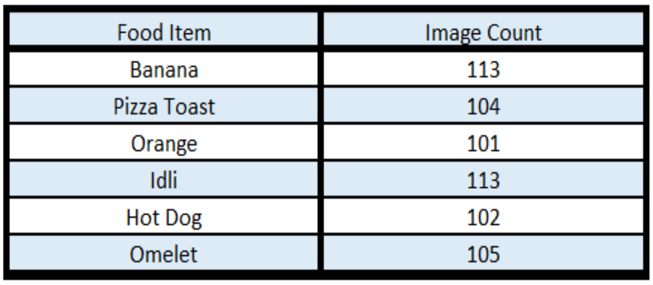
II. MATERIALS AND METHOS

A. Overview

Our model takes Image of the food items as input and give calories of the food item as output. In this there are number of middle process which are done to achieved to this. Firstly, food item whose calorie need to be predicted is identified in the captured image. After the food item identification its size and volume are determined and at last food calorie is estimated to 128*128 pixels before they are used in the model. Mask R-CNN Algorithm is used for image recognition and calorie prediction is done using approximate proportion approach format.

B. Dataset

Our dataset has six different food items. Dataset is custom generated by selecting food item image from internet and by selecting images from the existing dataset such as Food 101. Food Item are Banana, Pizza Toast, Orange, Idle, Hot Dog, Omelet.



Food Item	Image Count
Banana	113
Pizza Toast	104
Orange	101
Idli	113
Hot Dog	102
Omelet	105

Fig.1.: Dataset details

As images are gathered from different sources it is important to scale them, so they are scaled to 128*128 pixels before they are used in model Also images are manually annotated using Pixel Annotation tool and masks are created for model training purpose.

C. Food-Item Identification

We use instance segmentation to create a pixel wise mask of object in the image. This technique gives us more granular understanding of the food data in the image. Here we use Mask R-CNN for image segmentation which used ROI and IOU to generate bounding boxes, provide labels and mask.

As in figure 2. Bounding box is created across the identified food item and label omelet is assigned to the food item.

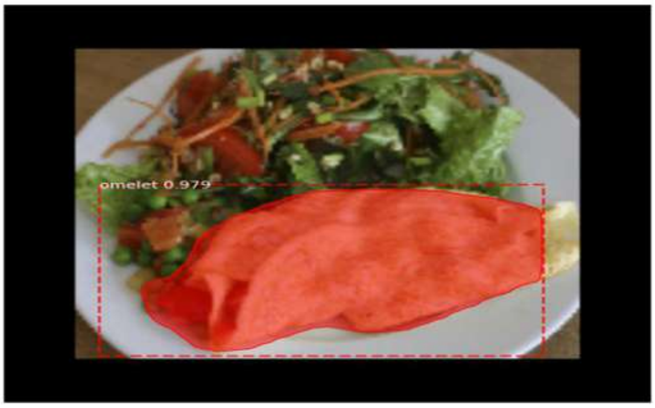


Fig.2: Mask output of omelet

Mask R-CNN is a deep neural network aimed to solve instance segmentation in machine learning. It separates different objects in an image or a video and gives out the objects bounding boxes, classes and masks. There are two stages in mask R-CNN first is to generate a proposal about the reason where there might be an object based on the input image second is to predict the class of the object, refines the boundary boxes and generate mask in pixel level. Both stages are connected to the backbone structure which helps in feature extraction. Here we are using Resnet 101 as backbone. Mask R-CNN Model is initialized with preloaded weights.

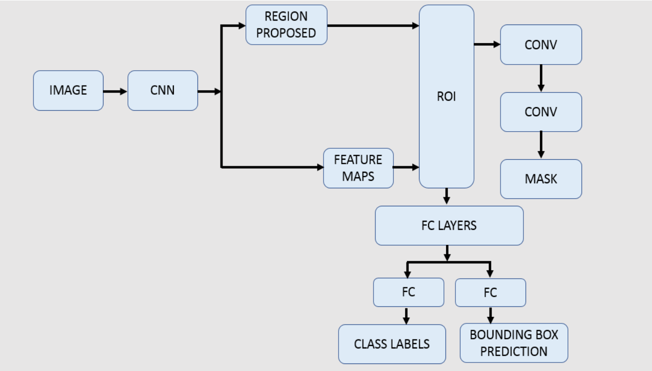


Fig.3: Mask R-CNN Model

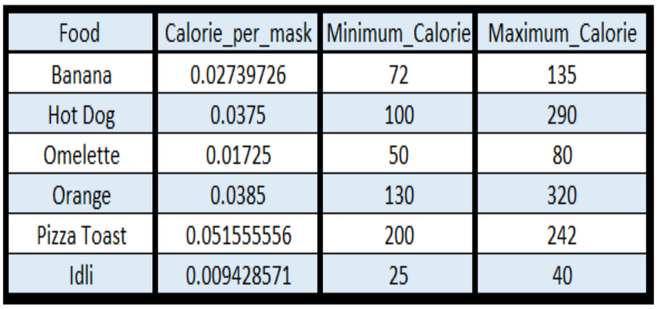
To train a Mask R-CNN based image recognition model, we need many images to train deep learning models. Since the dataset we collected is not large enough to train the model, we used transfer learning approach from matterport repository [7].

With the help of **transfer learning**, instead of training a model from scratch, we started with a weights file that is been trained on the COCO dataset. COCO dataset contains lot of images (~120K), so the trained model weights have already learned a lot of the features common in natural images, which really helps.

D. Food Calorie Prediction

As the same food can be taken at different depths to generate different picture sizes, we need a method to calculate calorie or estimate the size of the food in a real-world scenario. After we get the desired food items detected along with their masks, we need the real object sizes, which is not possible through a pin-hole camera image alone. So, we take a referencing approach that references the food-objects to the size of the pre-known object to extract the actual size of the food contained in that specific image.

For food calorie prediction, method used is approximation of proportions. In this approach calorie per mask of the food class is taken as reference for predicting calorie of the input image. A spread sheet has been prepared containing "Class", "Calorie Per Mask", "Minimum Calories", "Maximum calorie".



Food	Calorie_per_mask	Minimum_Calorie	Maximum_Calorie
Banana	0.02739726	72	135
Hot Dog	0.0375	100	290
Omelette	0.01725	50	80
Orange	0.0385	130	320
Pizza Toast	0.051555556	200	242
Idli	0.009428571	25	40

Fig.4: Calorie details used in predicting food calorie

In Proportion Approximation Approach each image has been scaled to the size of (128*128) and segmented area of the food class is calculated and further multiplied with "Calorie per Mask ".

III. EXPERIMENTAL RESULTS

In this section we present the experimental result of all the sections like Food identification which generate bounding boxes, provide labels and mask. Dataset is divided in train and test (Validate) set in the ration of 80:20.

A. Dataset Inspection

Our dataset has six food items named banana, Pizza Toast, omelet, orange and idle.

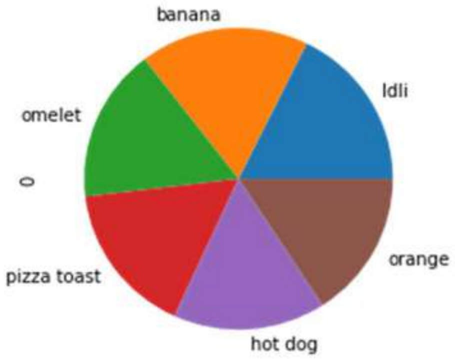


Fig.5: Food classes in dataset

Images are of different lengths, so we converted them to 128 * 128 and loaded them. Similarly, mask images had also converted them to 128 * 128. Initial dataset was class imbalanced, so we added more images and did data augmentation.

During Visual dataset inspection it was observed classes were initially imbalanced, which were balanced by adding new images. Also, Image size is changed to 128*128. Minimum class image is 101 and maximum class image is 113.

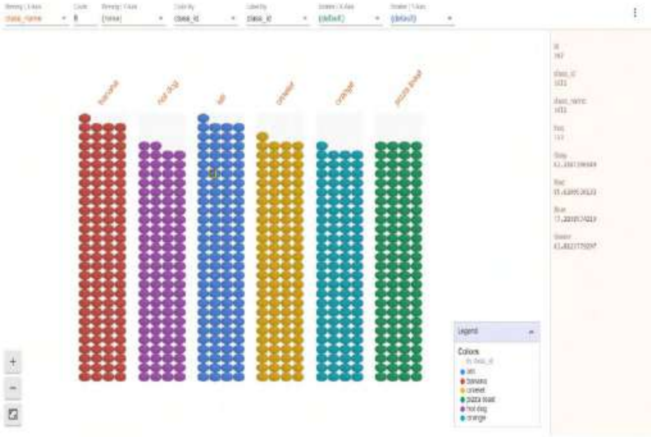


Fig.6: Initial Visual representation of dataset

B. Food Item Identification

Our goal is to identify the type of food item after extracting the features from the image. Here Mask R-CNN is used to generate bounding box, assign labels and mask to the image.

Analyzing images and mask after preprocessing, converting to 128 * 128 dimensions.

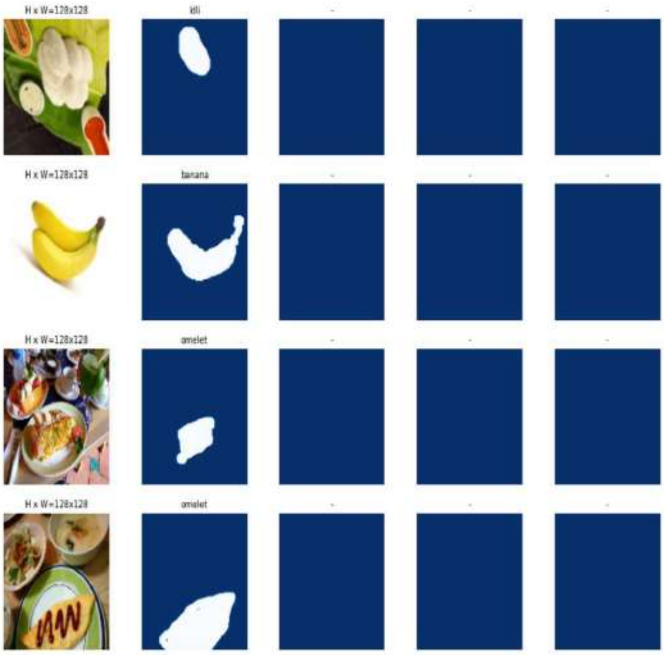


Fig.7: Analyzing images and mask after preprocessing

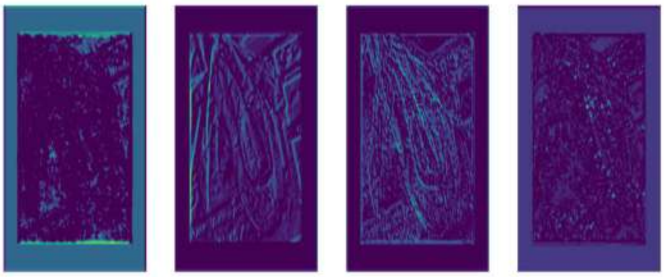


Fig.8: Analyzing activation of hidden layers

C. *Calorie Prediction*

A referencing approach that references the food-objects to the size of the pre-known object to extract the actual size of the food contained in that specific image. Refer figure 4

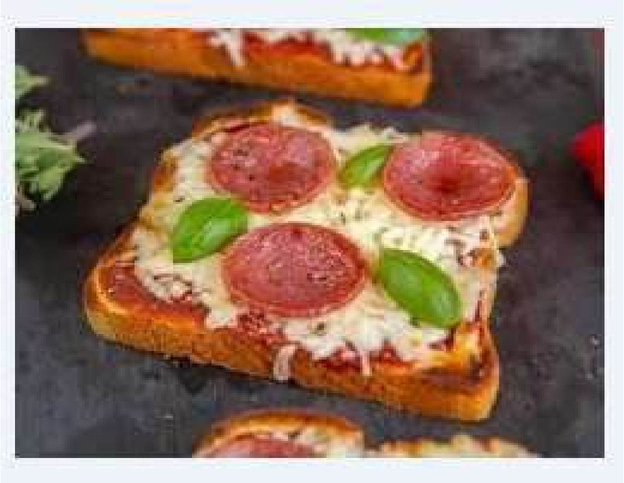


Fig.9: Actual Image- Image before Mask R-CNN and calorie prediction

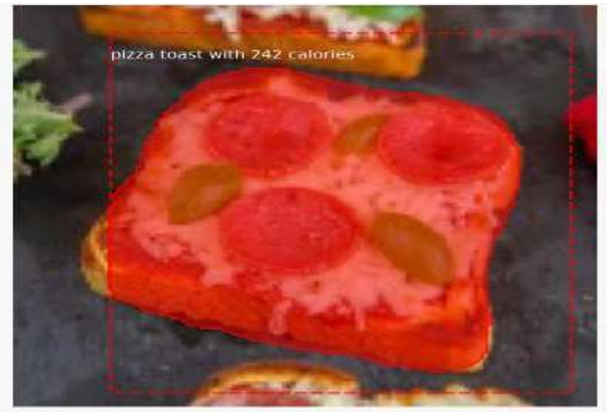


Fig.10: shows 242 calories for pizza toast - Image after Mask R-CNN and calorie prediction

D. *Visualization Of Model Accuracy & Loss*

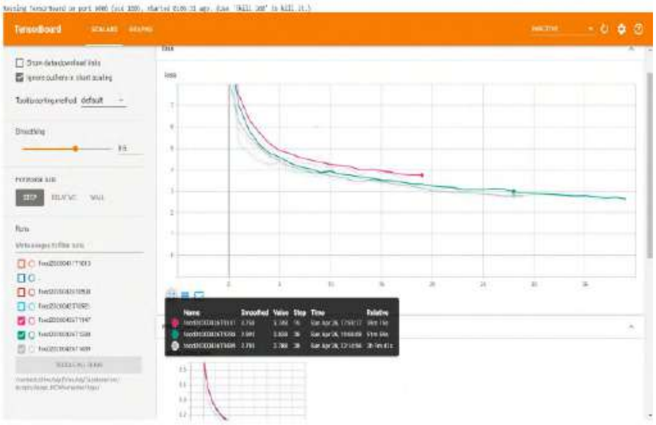


Fig.11: Overall loss while training the Train dataset

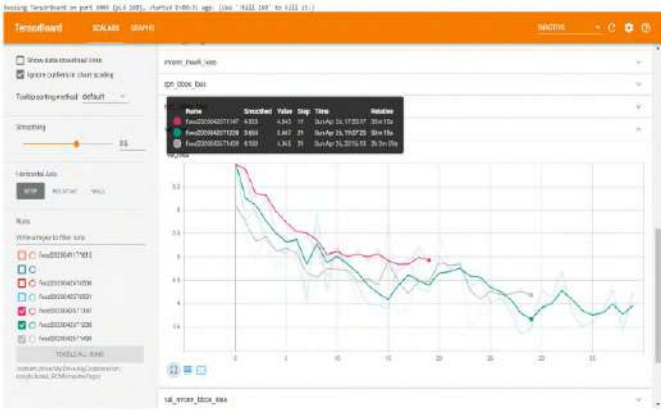


Fig.12: Overall loss while Validating/testing the Test dataset

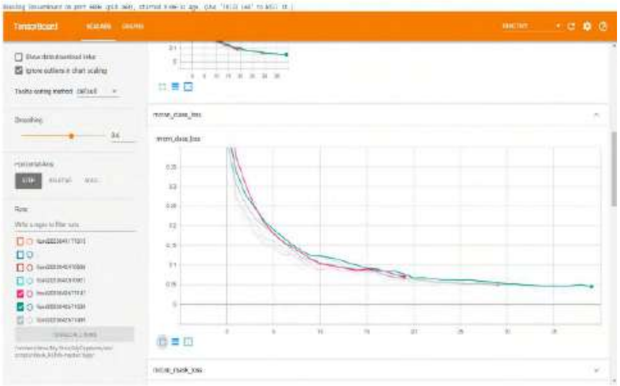


Fig.13: Classification loss while training the Train dataset

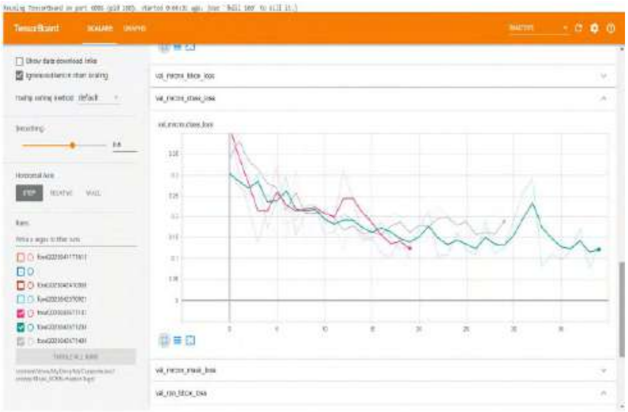


Fig.14: Classification loss while Validating/testing the Test dataset

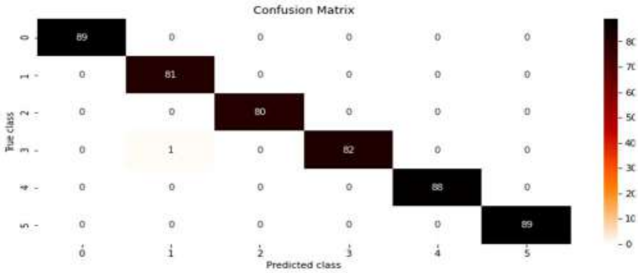


Fig.15: Confusion matrix while Training the Train data (0=BG ,1=Orange, 2=hot dog, 3= omelet ,4=banana, 5=pizza toast , 6=idli)

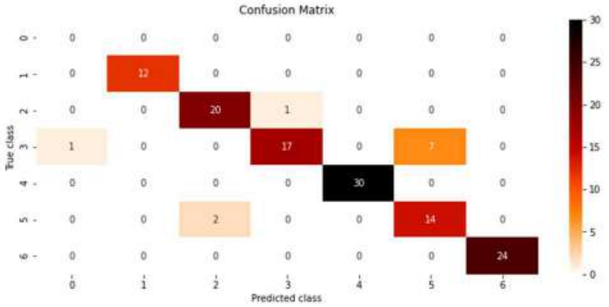


Fig.16: Confusion matrix while Validating/testing the Test data (0=Background ,1=Orange, 2=hot dog, 3= omelet ,4=banana, 5=pizza toast , 6=idli)

IV. CONCLUSION

In this paper, we have used the deep learning-based model to predict the total calories of the food item present in the image. To develop this solution, we used Mask R-CNN technique to create mask and bounding boxes. This in turn helped the model to calculate the surface area occupied by the different food items in the image which further facilitated the model with the ability to satisfactorily predict the calories associated with each food item. Calorie are estimated with the help of mathematical formulas which compares the proportion of the image occupied by each food items and determines the calories associated with it.

In future work, we plan to extend the scope of model by increasing the ability of the model to identify a greater number of food items instead of 6 food items, which we used in our current dataset. The dataset we used contain 638 images of food item with 6 different categories, which will be extended in our next model. Finally, we would like to calculate the calories of the food item based on their volume with the help of 3D images. As the technology is advancing, it would be interesting to work on the model development which can handle 3D image as input and predict the calories of the food item with better results.

V. REFERENCES

[1] World Health Orgamisation https://www.who.int/features/factfiles/obesity/en/

[2] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

[3] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[4] Bossard L., Guillaumin M., Van Gool L. (2014) Food-101 – Mining Discriminative Components with Random Forests. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8694. Springer, Cham

[5] Ciocca, Gianluigi & Napoletano, Paolo & Schettini, Raimondo. (2016). Food Recognition: A New Dataset, Experiments, and Results. IEEE Journal of Biomedical and Health Informatics. PP. 1-1. 10.1109/JBHI.2016.2636441.

[6] Ge, ZongYuan & Mccool, Chris & Sanderson, Conrad & Corke, Peter. (2015). Modelling local deep convolutional neural network features to improve fine-grained image classification. 4112-4116. 10.1109/ICIP.2015.7351579.

[7] Cao, Yuanzhouhan, et al. "Exploiting Depth From Single Monocular Images for Object Detection and Semantic Segmentation." IEEE Transactions on Image Processing, vol. 26, no. 2, Feb. 2017, pp. 836–46. DOI.org (Crossref), doi:10.1109/TIP.2016.2621673.

[8] Chokr, Manal and Shady Elbassuoni. "Calories Prediction from Food Images." AAAI (2017).

[9] Pouladzadeh, Parisa & Yassine, Abdulsalam & Shirmohammadi, Shervin. (2015). FooDD: Food Detection Dataset for Calorie Measurement Using Food Images. Lecture Notes in Computer Science. 9281. 10.1007/978-3-319-23222-5_54

[10] Food recognition: a new dataset, experiments and results (Gianluigi Ciocca, Paolo Napoletano, Raimondo Schettini) In IEEE Journal of Biomedical and Health Informatics, volume 21, number 3, pp. 588-598, IEEE, 2017

[11] He Y, Xu C, Khanna N, Boushey CJ, Delp EJ. FOOD IMAGE ANALYSIS: SEGMENTATION, IDENTIFICATION AND WEIGHT ESTIMATION. Proc (IEEE Int Conf Multimed Expo). 2013;2013:10.1109/ICME.2013.6607548. doi:10.1109/ICME.2013.6607548

[12] Estrada, F.J. & Jepson, A.D. Int J Comput Vis (2009) 85: 167. https://doi.org/10.1007/s11263-009-0251-z

[13] Liang, Yanchao & Li, Jianhua. (2017). Computer vision-based food calorie estimation: dataset, method, and experiment. https://arxiv.org/pdf/1705.07632.pdf.

[14] P. Pouladzadeh, S. Shirmohammadi and R. Almaghrabi, "Measuring Calorie and Nutrition from Food Image", *IEEE Transactions on Instrumentation & Measurement*, vol. 63, no. 8, pp. 1947-1956, August2014.

[15] *World Health Statistics 2012*, 2012, [online] Available: http://www.who.int/gho/publications/world_health_statistics/2012/en/index.html.

[16] *Obesity Study*, October 2011, [online] Available: http://www.who.int/mediacentre/factsheets/fs311/en/index.html.

Improved encryption scheme based on the automorphism group of the Ree function field

Gennady Khalimov
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
hennadii.khalimov@nure.ua

Yevgeniy Kotukh
Sumy State University
Sumy, Ukraine
yevgenkotukh@gmail.com

Svitlana Khalimova
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
svitlana.khalimova@nure.ua

Abstract. *The article describes an improved implementation of the encryption scheme based on the automorphism group of Ree function field. Our proposal is to use the Ree function field automorphism group to encrypt the whole group with key bindings. We extended the logarithmic signature to the entire group and got changed the encryption algorithm to protect against a sequential key recovery attack.*

Keywords: *MST cryptosystem, logarithmic signature, random cover, automorphism group, Ree function field.*

I. Introduction

The advent of quantum computers capable of solving arbitrarily complex problems casts doubt on the existence of modern cryptography. Classical public key cryptographic protocols exploiting the idea of the complexity of solving the problem of factoring large numbers. This idea become insecure with implementation of quantum algorithms. The implementation of cryptosystems resistant to quantum cryptanalysis becomes relevant. The idea of constructing public-key cryptosystems on the basis of the intractable word problem was proposed in 90th. The basis is the use of permutation groups. One possible solution is to exploit the idea of Wagner and Magyarik, which proposed the concept of a cryptosystem based on the unsolvable problem of words in groups and semigroups [1]. Birget et al. [2] projected this idea into a public key cryptosystem with finite generated groups. The type of cryptosystems based on group factorization has developed rapidly [3-7].

In 1986, Magliveras [5] proposed a symmetric cryptosystem based on factorization in finite permutation groups called logarithmic signature (LS). In 2009, Lempken et al. [4] developed a new public key cryptographic system - MST3, based on random covers and Suzuki 2-group. In 2008 Magliveras et al. [8] presented a comprehensive analysis of the MST3 cryptosystem and stated that the transitive logarithmic signature is not suitable for the MST3 cryptosystem. In 2010, Svaba et al. [6] analyzed all published references to attacks on MST cryptography and built a more secure eMST3 cryptosystem by adding a secret homomorphic coverage. In 2018, T. van Trung [9] proposed a general method for constructing strong aperiodic logarithmic signatures for abelian p-groups and contributed to the practical application of MST cryptosystems. Since the 2000s, several dozen group cryptosystems schemes have been proposed. Further development of MST3 cryptosystems is proposed in [11–13].

In this paper the implementation on multiparameter groups are considered for the first time. The main idea lies in the plane of solving the problem of optimizing overhead costs by reducing the large size of keys and increasing the efficiency of the encryption (decryption) algorithm. It was shown that on groups of large order it is possible to construct cryptosystems with calculations of the LS outside the center of the group over finite fields of small dimension.

The implementation of an encryption algorithm with LSs based on the automorphism group of the Ree function field [13] allows increasing the size of the cipher text and reducing the requirements for the size of LS. At the same time, despite the very large order of the group, the secrecy of this implementation turned out to be unrelated to the order of the automorphism group of the Ree function field and is determined by the dimension of the finite representation field. This article discusses brute force attacks on key recovery for a cryptosystem with of the Ree group of automorphisms and presents a new encryption algorithm with associated keys.

II. Encryption secrecy based on the group of automorphisms of the Ree functional field

The concept of constructing the MST cryptosystem is based on application of non-commutative group algebra to cover the logarithmic signature. LS as a mapping is given by the following definition 14].

Definition 1 (cover (logarithmic signature) mappings). Let $\alpha = [A_1, \dots, A_s]$ be a cover (logarithmic signature) of type $(r_1, r_2, \dots, r_s)$ for G with $A_i = [a_{i,1}, a_{i,2}, \dots, a_{i,r_i}]$, where $m = \prod_{i=1}^{s} r_i$. Let $m_1 = 1$ and $m_i = \prod_{j=1}^{i-1} r_j$ for $i = 2, \dots, s$. Let τ denote the canonical bijection

$$\tau : \mathbb{Z}_{r_1} \times \mathbb{Z}_{r_2} \times \ldots \times \mathbb{Z}_{r_s} \rightarrow \mathbb{Z}_m ,$$

$$\tau(j_1, j_2, \dots, j_s) = \sum_{i=1}^{s} j_i \cdot m_i .$$

Then the surjective (bijection) mapping $\alpha' : \mathbb{Z}_m \rightarrow G$ induced by is

$$\alpha'(x) = a_{1j_1} \cdot a_{2j_2} \cdots a_{sj_s}$$

where $(j_1, j_2, ..., j_s) = \tau^{-1}(x)$.

More generally, if $\alpha = [A_1, ..., A_s]$ is a LS (cover) for, then each element $g \in G \in$ can be expressed uniquely (at least one way) as a product of the form

$$g = a_1 \cdot a_2 \cdots a_s,$$

for $a_i \in A_i$.

Let G is ultimate nonabelian group with nontrivial center $\mathbf{Z}$, such that G does not decompose over $\mathbf{Z}$. Suppose that $\mathbf{Z}$ is quite large, such that the search is over $\mathbf{Z}$ is computational impracticable.

The cryptographic hypothesis, which is the basis for the cryptosystem, is that if $\alpha = [\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_s] := (a_{i,j})$ – accidental cover for a "large" matrices S at G, then search for the layout $g = a_{1j_1} a_{2j_2} \cdots a_{sj_s}$ for any element $g \in G$ relatively α is, in general, not a solvable problem.

One of the concepts of secrecy is based on use of aperiodic logarithmic signatures. Constructions of aperiodic logarithmic signatures are widely presented [9,16]. Research and estimates carried out in [9] are quite optimistic. There are also questions for MST cryptanalysis related to group algebra. We omit the subtle issues of basic attack analysis here, although the details are essential.

The MST implementation cryptosystem on the group of automorphisms of the Ree function field exploits the idea that good implementation and secrecy characteristics can be obtained on a large-order multivariable group.

The group of automorphisms of the Ree function field is defined over a finite field F_q, $q = 3^{2s+1}$, where $s \in N \setminus \{0\}$ and $q_0 = 3^s$ [10].

The Ree function field F over K is defined as $S = K(x, y, z)$ where

$$y^q - y = x^{q_0}(x^q - x),$$
$$z^q - z = x^{2q_0}(x^q - x).$$

It has $N = q^3 + 1$ rational places and genus $g = 3q_0(q-1)(q+q_0+1)/2$.

The automorphism group A of F is the Ree group $\mathrm{Ree}(q)$ and has order $|\mathrm{Ree}(q)| = q^3(q^3+1)(q-1)$.

Let P_∞ denote the unique pole x in F. Let

$$A(P_\infty) := \{\sigma \in A | \sigma(P_\infty) = P_\infty\} \subset A.$$

$A(P_\infty)$ consists of all automorphisms $\psi_{a,b,c,d}$ with $a, b, c, d \in K, a \neq 0$, which are defined as

$$\psi_{a,b,c,d} := \begin{cases} x \mapsto ax + b \\ y \mapsto a^{q_0+1} y + ab^{q_0} x + c \\ z \mapsto a^{2q_0+1} z - a^{q_0+1} b^{q_0} y + ab^{2q_0} x + d \end{cases}$$

We have $|A(P_\infty)| = q^3(q-1)$, and the subgroup $A(P_\infty)$ is a maximal subgroup of A.

The each element of $A(P_\infty)$ can be expressed uniquely

$$A(P_\infty) = \{S(a,b,c,d) | a \in F_q^* := F_q \setminus \{0\}, c, b, d \in F_q\},$$

where $S(a,b,c,d) = [a,b,c,d]$ and group operation is defined as

$$S(a_1, b_1, c_1, d_1) \cdot S(a_2, b_2, c_2, d_2) =$$
$$S(\ a_1 a_2, a_2 b_1 + b_2, a_2^{q_0+1} c_1 + a_2 b_1 b_2^{q_0} + c_2, a_2 b_1 b_2^{2q_0} - a_2^{q_0+1} b_2^{q_0} c_1$$
$$+ a_2^{2q_0+1} d_1 + d_2\ )$$

The identity is the 4-triple $[1,0,0,0]$ and the inverse of $S(a,b,c,d)$ is

$$S(a,b,c,d)^{-1} = S(\ a^{-1}, -a^{-1}b, (a^{-1}b)^{q_0+1} - a^{-(q_0+1)}c,$$
$$-(a^{-1}b)^{2q_0+1} - a^{-(2q_0+1)} b^{q_0} c - a^{-(2q_0+1)} d\ ).$$

The encryption scheme based on the automorphism group of the Ree function field was proposed in [13] The correctness of the proposed algorithm was also verified by practical assessment.

For the purposes of cryptoanalysis of encryption scheme let's consider key generation and encryption steps as follows.

Input: a large group on the field F_q, $q = 3q_0{}^2$, $q_0 = 3^s$

$$A(P_\infty) = \{S(a,b,c,d) | a \in F_q^* := F_q \setminus \{0\}, c, b, d \in F_q\}.$$

Choose a tame logarithmic signatures $\beta_{(k)} = [B_{1(k)}, ..., B_{s(k)}] = (b_{ij})_{(k)}$, $(b_{ij})_{(k)} \in A(P_\infty)$ of type $(r_{1(k)}, ..., r_{s(k)})$, $i = \overline{1, s(k)}$, $j = \overline{1, r_{i(k)}}$, $b_{ij(k)} \in F_q$, $k = \overline{1,3}$. Group element $(b_{ij})_{(k)}$ has a value in only one coordinates b, c or d, respectively.

Select a random covers $\alpha_{(k)} = [A_{1(k)}, ..., A_{s(k)}] = (a_{ij})_{(k)} = S(a_{ij(k)_1}, a_{ij(k)_2}, a_{ij(k)_3}, a_{ij(k)_4})$ of the same type as $\beta_{(k)}$, where $a_{ij} \in A(P_\infty)$, $a_{ij(k)_1}, a_{ij(k)_2}, a_{ij(k)_3}, a_{ij(k)_4} \in F_q \setminus \{0\}$, $k = \overline{1,3}$.

Choose $t_{0(k)}, t_{1(k)}, ..., t_{s(k)} \in A(P_\infty) \setminus Z$,

$t_{i(k)} = S(t_{i(k)_1}, t_{i(k)_2}, t_{i(k)_3}, t_{i(k)_4})$, $t_{i(\mathrm{k})_j} \in F^\times$, $i = \overline{0, s(\mathrm{k})}$, $j = \overline{1,4}$, $k = \overline{1,3}$. Let's $t_{\mathrm{s}(1)} = t_{0(2)}$, $t_{\mathrm{s}(2)} = t_{0(3)}$.

Construct a homomorphism f_k, $k = \overline{1,3}$ defined by

$$f_1(S(a_1, a_2, a_3, a_4)) = S(1, a_1, a_2, a_3),$$
$$f_2(S(a_1, a_2, a_3, a_4)) = S(1, 0, a_2, a_3),$$
$$f_3(S(a_1, a_2, a_3, a_4)) = S(1, 0, 0, a_3).$$

Let's do the following calculations

$$\gamma_{(k)} = [h_{1(k)}, ..., h_{s(k)}] = t_{(i-1)(k)}^{-1} f_k((a_{ij})_{(k)})(b_{ij})_{(k)} t_{i(k)},$$

where $k = \overline{1,3}$, $i = \overline{1, s(k)}$, $j = \overline{1, r_{i(k)}}$,

$$f_1((a_{ij})_{(1)})(b_{ij})_{(1)} =$$
$$S(1, a_{ij(1)_1} + b_{ij(1)}, a_{ij(1)_2} + a_{ij(1)_1} b_{ij(1)}^{q_0}, a_{ij(1)_1} b_{ij(1)}^{2q_0} + a_{ij(1)_3}),$$
$$f_2((a_{ij})_{(2)})(b_{ij})_{(2)} = S(1, 0, a_{ij(2)_2} + b_{ij(2)}, a_{ij(2)_3}),$$
$$f_3((a_{ij})_{(3)})(b_{ij})_{(3)} = S(1, 0, 0, a_{ij(3)_3} + b_{ij(3)}).$$

An output public key $[f_k,(\alpha_k,\gamma_k)]$, and a private key $\left[\beta_{(k)},\left(t_{0(k)},\dots,t_{s(k)}\right)\right]$, $k=\overline{1,3}$.

Encryption. Let's have message $m\in A\left(P_\infty\right)$, $m=S\left(m_1,m_2,m_3,m_4\right)$, $m_1\in F_q\setminus\{0\}$, $m_2,m_3,m_4\in F_q$, the public key $[f_k,(\alpha_k,\gamma_k)]$, $k=\overline{1,3}$, $R=R_1,R_2,R_3\in Z_{|F_q|}$

Compute

$$y_1=\alpha'(R)\cdot m=\alpha_1{}'(R_1)\cdot\alpha_2{}'(R_2)\cdot\alpha_3{}'(R_3)\cdot m,$$

$$y_2=\gamma'(R)=\gamma_1{}'(R_1)\cdot\gamma_2{}'(R_2)\cdot\gamma_3{}'(R_3)$$

$$=S(\ *,a_{(1)_1}(R_1)+\beta_{(1)}(R_1)+*,a_{(2)_2}(R_2)+\beta_{(2)}(R_2)+*,$$

$$a_{(3)_3}(R_3)+\beta_{(2)}(R_3)+*\ ).$$

Here, the $(*)$ components are determined by cross-calculations in the group operation of the product of $t_{0(k)},\dots,t_{s(k)}$ and the product of $a_{(k)_j}(R_k)+\beta_{(k)}(R_k)$.

Compute

$$y_3=f_1\left(\alpha_1{}'(R_1)\right)=S\left(1,a_{(1)_1}(R_1),*,*\right),$$

$$y_4=f_2\left(\alpha_2{}'(R_2)\right)=S\left(1,0,a_{(2)_2}(R_2),a_{(2)_3}(R_2)\right),$$

$$y_5=f_3\left(\alpha_3{}'(R_3)\right)=S\left(1,0,0,a_{(3)_3}(R_3)\right).$$

Output (y_1,y_2,y_3,y_4,y_5).

This encryption scheme has a serious flaw. In the proposed implementation of the algorithm, we have R_1 and R_2 as encryption keys. They are not bound and allow for a sequential key recovery attack. The keys can be restored on the basis of calculating the $\alpha_k{}'(R_k{}')$ for each $k=\overline{1,3}$ and comparing it with the y_3,y_4,y_5 according to the values of the corresponding coordinates.

$$y_3=f_1\left(\alpha_1{}'(R_1)\right)=S\left(1,a_{(1)_1}(R_1),*,*\right),$$

$$y_4=f_2\left(\alpha_2{}'(R_2)\right)=S\left(1,0,a_{(2)_2}(R_2),a_{(2)_3}(R_2)\right),$$

$$y_5=f_3\left(\alpha_3{}'(R_3)\right)=S\left(1,0,0,a_{(3)_3}(R_3)\right).$$

The complexity of key recovery attack of $R=(R_1,R_2,R_3)$ equals to $3q$.

III. Improved Encryption Based on the Automorphism Group of the Ree Function Field

In the new implementation of the cryptosystem, we changed the encryption algorithm in a way to bind the keys of the logarithmic signatures and protect against a sequential recovery attack. Our suggestion is to use the group of automorphisms of the Ree function field for the encryption on full group $A\left(P_\infty\right)=\{S(a,b,c,d)\}$ with the bound keys $R=(R_1,R_2,R_3)$. In such case, the brute force attack complexity with equal to q^3.

Description of the Scheme. Key Generation.

Input: a large group on the field F_q, $q=3q_0{}^2$, $q_0=3^s$

$$A\left(P_\infty\right)=\left\{S(a,b,c,d)\middle|a\in F_q^*:=F_q\setminus\{0\},c,b,d\in F_q\right\}.$$

Choose a tame logarithmic signatures $\beta_{(k)}=\left[B_{1(k)},\dots,B_{s(k)}\right]=\left(b_{ij}\right)_{(k)}$, $\left(b_{ij}\right)_{(k)}\in A\left(P_\infty\right)$ of type $\left(r_{1(k)},\dots,r_{s(k)}\right)$, $i=\overline{1,s(k)}$, $j=\overline{1,r_{i(k)}}$, $b_{ij(k)}\in F_q$, $k=\overline{1,3}$. Group element $\left(b_{ij}\right)_{(k)}$ has a value in only one coordinates b, c or d, respectively.

For example $\left(b_{ij}\right)_{(1)}=S\left(1,b_{ij(k)_a},0,0\right)$.

Select a random covers $\alpha_{(k)}=\left[A_{1(k)},\dots,A_{s(k)}\right]=\left(a_{ij}\right)_{(k)}=S\left(a_{ij(k)_a},a_{ij(k)_b},a_{ij(k)_c},a_{ij(k)_d}\right)$ of the same types as $\beta_{(k)}$, where $a_{ij}\in A\left(P_\infty\right)$, $a_{ij(k)}\in F_q\setminus\{0\}$, $i=\overline{0,s(\mathrm{k})}$, $j=\overline{1,r_{i(k)}}$, $k=\overline{1,3}$.

Choose $t_{i(k)}=S\left(t_{i(k)_a},t_{i(k)_b},t_{i(k)_c},t_{i(k)_d}\right)$ $t_{\mathrm{i}(k)}\in A(P_\infty)\setminus Z$, $t_{i(k)_a},t_{i(k)_b},t_{i(k)_c},t_{i(k)_d}\in F_q/\{0\}$, $i=\overline{0,s(\mathrm{k})}$, $k=\overline{1,3}$.

Let's $t_{\mathrm{s}(k-1)}=t_{0(k)}$, $k=\overline{1,3}$.

Construct a homomorphisms defined by

$$f_1\left(S(a,b,c,d)\right)=S(1,b,c,d),$$

$$f_2\left(S(a,b,c,d)\right)=S(1,0,c,d),$$

$$f_3\left(S(a,b,c,d)\right)=S(1,0,0,d).$$

Let's do the following calculations

$$\gamma_{(\mathrm{k})}=\left[h_{1(\mathrm{k})},\dots,h_{s(\mathrm{k})}\right]=t_{(i-1)(\mathrm{k})}^{-1}f_k\left(\left(a_{ij}\right)_{(\mathrm{k})}\right)\left(b_{ij}\right)_{(\mathrm{k})}t_{i(\mathrm{k})},$$

$i=\overline{1,s(\mathrm{k})}$, $j=\overline{1,r_{i(\mathrm{k})}}$, $k=\overline{1,3}$

and

$$f_1\left(\left(a_{ij}\right)_{(1)}\right)\left(b_{ij}\right)_{(1)}=$$

$$S\left(1,a_{ij(1)_b}+b_{ij(1)_b},a_{ij(1)_c}+a_{ij(1)_b}b_{ij(1)_b}^{q_0},a_{ij(1)_b}b_{ij(1)_b}^{2q_0}+a_{ij(1)_d}\right),$$

$$f_2\left(\left(a_{ij}\right)_{(2)}\right)\left(b_{ij}\right)_{(2)}=S\left(1,0,a_{ij(2)_c}+b_{ij(2)_c},a_{ij(2)_d}\right),$$

$$f_3\left(\left(a_{ij}\right)_{(3)}\right)\left(b_{ij}\right)_{(3)}=S\left(1,0,0,a_{ij(3)_d}+b_{ij(3)_d}\right).$$

An output public key $[f_k,(\alpha_k,\gamma_k)]$, and a private key $\left[\beta_{(k)},\left(t_{0(k)},\dots,t_{s(k)}\right)\right]$, $k=\overline{1,3}$.

Encryption

Input: a message $m\in A\left(P_\infty\right)$, $m=S\left(m_1,m_2,m_3,m_4\right)$, $m_1\in F_q\setminus\{0\}$, $m_2,m_3,m_4\in F_q$ and the public key $\left[f_k,\overset{\square}{f}_1,\overset{\square}{f}_2,(\alpha_k,\gamma_k)\right]$, $k=\overline{1,3}$.

Output: a ciphertext $\left(y_1,y_2,y_3\right)$ of the message m.

Choose a random $R=(R_1,R_2,R_3)$, $R_k\in Z_{|Z|}$, $k=\overline{1,3}$.

Let's set the encryption key through the mapping $R'=\pi(R_1,R_2,R_3)=(R_1{}',R_2{}',R_3{}')$.

Compute

$$y_1=\alpha'(R')\cdot m=\alpha_1{}'(R_1{}')\cdot\alpha_2{}'(R_2{}')\cdot\alpha_3{}'(R_3{}')\cdot m.$$

Compute component y_2.

$$\gamma(R)=\gamma_1'(R_1)\cdot\gamma_2'(R_2)\cdot\gamma_3'(R_3)=$$
$$S(\ 1,\sum_{i=1,j=R_{i(1)}}^{s(1)}\left(a_{ij(1)_b}+\beta_{ij(1)_b}\right)+*,\sum_{i=1,j=R_{i(2)}}^{s(2)}\left(a_{ij(2)_c}+\beta_{ij(2)_c}\right)+*,$$
$$\sum_{i=1,j=R_{i(3)}}^{s(3)}\left(a_{ij(3)_c}+\beta_{ij(3)_c}\right)+*\ ).$$

Here, the $(*)$ components are determined by cross-calculations in the group operation of the product of $t_{0(k)},\ldots,t_{s(k)}$ and the product of $a_{(k)}(R_k)+\beta_{(k)}(R_k)$.

$$y_2=\gamma(R)\cdot\overset{\square}{f}_2\left(\alpha_3'(R_3)\right)\cdot\overset{\square}{f}_1\left(\alpha_3'(R_3)\right)\cdot\overset{\square}{f}_1\left(\alpha_2'(R_2)\right),$$

where

$$\overset{\square}{f}_1\left(\alpha_k'(R_k)\right)=\prod_{i=1,j=R_{i(k)}}^{s(k)}S\left(1,a_{ij(k)_b},0,0\right),\ k=2,3$$

$$\overset{\square}{f}_2\left(\alpha_k'(R_k)\right)=\prod_{i=1,j=R_{i(k)}}^{s(k)}S\left(1,0,a_{ij(k)_c},0\right),\ k=3$$

and

$$y_2=S\left(1,\sum_{k=1}^{3}\sum_{i=1,j=R_{i(1)}}^{s(1)}a_{ij(k)_b}+\sum_{i=1,j=R_{i(1)}}^{s(1)}\beta_{ij(1)_b}+*,\right.$$
$$\sum_{i=1,j=R_{i(2)}}^{s(2)}\left(a_{ij(2)_c}+\beta_{ij(2)_c}\right)+\sum_{i=1,j=R_{i(3)}}^{s(3)}a_{ij(3)_c}+*,$$
$$\left.\sum_{i=1,j=R_{i(3)}}^{s(3)}\left(a_{ij(3)_d}+\beta_{ij(3)_d}\right)+*\right).$$

Compute component y_3.

$$\lambda(R)=f_1\left(\alpha_1'(R_1)\right)\cdot f_2\left(\alpha_2'(R_2)\right)\cdot f_2\left(\alpha_3'(R_3)\right),$$
$$y_3=\lambda(R)\overset{\square}{f}_1\left(\alpha_3'(R_3)\right)\cdot\overset{\square}{f}_1\left(\alpha_2'(R_2)\right),$$

where

$$f_1\left(\alpha_k'(R_k)\right)=\prod_{i=1,j=R_{i(k)}}^{s(k)}S\left(1,a_{ij(k)_b},a_{ij(k)_c},a_{ij(k)_d}\right),\ k=1$$

$$f_2\left(\alpha_k'(R_k)\right)=\prod_{i=1,j=R_{i(k)}}^{s(k)}S\left(1,0,a_{ij(k)_c},a_{ij(k)_d}\right),\ k=2,3$$

and

$$y_3=S\left(1,\sum_{k=1}^{3}\sum_{i=1,j=R_{i(k)}}^{s(k)}a_{ij(k)_b},\right.$$
$$\left.\sum_{k=2}^{3}\sum_{i=1,j=R_{i(k)}}^{s(k)}a_{ij(k)_c}+*,\sum_{k=2}^{3}\sum_{i=1,j=R_{i(k)}}^{s(k)}a_{ij(k)_d}+*\right).$$

Output (y_1,y_2,y_3).

Decryption

Input: a ciphertext (y_1,y_2,y_3) and private key $\left[\beta_{(k)},\left(t_{0(k)},\ldots,t_{s(k)}\right)\right]$, $k=\overline{1,3}$.

To decrypt a message m, we need to restore random numbers $R=(R_1,R_2,R_3)$.

Compute

$$D(R_1,R_2,R_3)=t_{0(1)}y_2y_3^{-1}t_{s(3)}^{-1}=S\left(1,\sum_{i=1,j=R_{i(1)}}^{s(1)}\beta_{ij(1)_a},*,*\right).$$

Restore R_1 with $\beta_{(1)}(R_1)=\sum_{i=1,j=R_{i(1)}}^{s(1)}\beta_{ij(1)_b}$ using $\beta_{(1)}(R_1)^{-1}$, because β_1 is simple. For further calculation, it is necessary to remove the component $\gamma_1'(R_1)$ from y_2 and $\alpha_1'(R_1)$ from y_3.

Compute

$$y_2^{(1)}=\gamma_1'(R_1)^{-1}y_2=S\left(1,\sum_{k=2}^{3}\sum_{i=1,j=R_{i(\mathrm{k})}}^{s(\mathrm{k})}a_{ij(k)_b}+*,\right.$$
$$\left.\sum_{i=1,j=R_{i(2)}}^{s(2)}\left(a_{ij(2)_c}+\beta_{ij(2)_c}\right)+*,\sum_{i=1,j=R_{i(3)}}^{s(3)}\left(a_{ij(3)_d}+\beta_{ij(3)_d}\right)+*\right)$$

and

$$y_3^{(1)}=f_1\left(\alpha_1'(R_1)\right)^{-1}y_3=S(1,\sum_{k=2}^{3}\sum_{i=1,j=R_{i(k)}}^{s(k)}a_{ij(k)_a},$$
$$\left.\sum_{k=2}^{3}\sum_{i=1,j=R_{i(k)}}^{s(k)}a_{ij(k)_b}+*,\sum_{k=2}^{3}\sum_{i=1,j=R_{i(k)}}^{s(k)}a_{ij(k)_c}+*\right).$$

Repeat the calculations for $D(R_2,R_3)$

$$D(R_2,R_3)=t_{0(2)}y_2^{(1)}\left(y_3^{(1)}\right)^{-1}t_{s(3)}^{-1}=$$
$$\left.S\left(1,0,\sum_{i=1,j=R_{i(2)}}^{s(2)}\beta_{ij(2)_c},\sum_{i=1,j=R_{i(3)}}^{s(3)}\left(a_{ij(3)_d}+\beta_{ij(3)_d}\right)+*\right)\right).$$

Restore R_2 with $\beta_{(2)}(R_2)=\sum_{i=1,j=R_{i(1)}}^{s(2)}\beta_{ij(1)_c}$ using $\beta_{(2)}(R_2)^{-1}$, because β_2 is simple. Remove the component $\gamma_2'(R_2)$ from $y_2^{(1)}$ and $f_1\left(\alpha_2'(R_2)\right)$ from $y_3^{(1)}$.

$$y_2^{(2)}=\gamma_2'(R_2)^{-1}y_2^{(1)}=S\left(1,\sum_{k=2}^{3}\sum_{i=1,j=R_{i(\mathrm{k})}}^{s(\mathrm{k})}a_{ij(k)_b}+*,\right.$$
$$\left.\sum_{i=1,j=R_{i(3)}}^{s(3)}a_{ij(3)_c}+*,\sum_{i=1,j=R_{i(3)}}^{s(3)}\left(a_{ij(3)_d}+\beta_{ij(3)_d}\right)+*\right)$$

and

$$y_3^{(2)}=f_2\left(\alpha_2'(R_2)\right)^{-1}y_3^{(1)}=S\left(1,\sum_{k=2}^{3}\sum_{i=1,j=R_{i(\mathrm{k})}}^{s(\mathrm{k})}a_{ij(k)_b},\right.$$
$$\left.\sum_{i=1,j=R_{i(3)}}^{s(3)}a_{ij(3)_c}+*,\sum_{i=1,j=R_{i(3)}}^{s(3)}\left(a_{ij(3)_d}+\beta_{ij(3)_d}\right)+*\right).$$

Compute

$$D(R_3)=t_{0(3)}y_2^{(2)}\left(y_3^{(2)}\right)^{-1}t_{s(3)}^{-1},$$

$$D(R_3)=t_{0(3)}y_2^{(2)}\left(y_3^{(2)}\right)^{-1}t_{s(3)}^{-1}==S\left(1,0,0,\sum_{i=1,j=R_{i(3)}}^{s(3)}\beta_{ij(3)_d}\right).$$

Restore R_3 with $\beta_{(3)}(R_3)$ using $\beta_{(3)}(R_3)^{-1}$.

We obtain $R'=\pi(R_1,R_2,R_3)=(R_1',R_2',R_3')$ and recovery the message $m=\alpha'(R_1',R_2',R_3')^{-1}\cdot y_1$.

Example

We will show the correctness of the obtained expressions in the following simple example.

Fix the subgroup $A(P_\infty)=\{S(a,b,c,d)\}$ for the group of automorphisms of the Ree function field over F_q, $q=3^5$, $q_0=3^2$, $g(\mathrm{x})=x^5+2x+1$.

First stage is to generate a tame logarithmic signature with the dimension of corresponding selected type $(r_{1(k)},...,r_{s(k)})$ and finite field F_q. The construction of arrays of logarithmic signatures is presented in [15].

For our example, we use the construction of simple logarithmic signatures without analyzing the details of their secrecy. Let's $\beta_{(k)}$, $k=\overline{1,3}$ have the types of $(3^2,3^2,3)$, $(3,3^2,3^2)$, $(3^2,3,3^2)$. They are represented as a strings and elements of the group over the field F_q.

$\beta_k=\left[B_{1(k)},B_{2(k)},B_{3(k)}\right]=\left(b_{ij}\right)_{(k)}$, $\left(b_{ij}\right)_{(k)}\in U(q)$, $k=\overline{1,3}$					
$B_{1(1)}$	$\left(b_{ij}\right)_{(1)}$	**$B_{1(2)}$**	$\left(b_{ij}\right)_{(2)}$	**$B_{1(3)}$**	$\left(b_{ij}\right)_{(3)}$
00 00 0	1,0,0, 0	**0** 00 00	1,0,0,0	**00** 0 00	1,0,0,0
10 00 0	$1,\alpha^{0},0,\ 0$	**1** 00 00	$1,0,\alpha^{0},0$	**10** 0 00	$1,0,0,\alpha^{0}$
20 00 0	$1,\alpha^{121},0,0$	**2** 00 00	$1,0,\alpha^{121},0$	**20** 0 00	$1,0,0,\alpha^{121}$
01 00 0	$1,\alpha^{1},0,0$	**$B_{2(2)}$**		**01** 0 00	$1,0,0,\alpha^{1}$
11 00 0	$1,\alpha^{69},0,0$	0 **00** 00	1,0,0,0	**11** 0 00	$1,0,0,\alpha^{69}$
21 00 0	$1,\alpha^{5},0,0$	2 **10** 00	$1,0,\alpha^{5},0$	**21** 0 00	$1,0,0,\alpha^{5}$
02 00 0	$1,\alpha^{122},0,0$	2 **20** 00	$1,0,\alpha^{190},0$	**02** 0 00	$1,0,0,\alpha^{122}$
12 00 0	$1,\alpha^{126},0,0$	1 **01** 00	$1,0,\alpha^{46},0$	**12** 0 00	$1,0,0,\alpha^{126}$
22 00 0	$1,\alpha^{190},0,0$	0 **11** 00	$1,0,\alpha^{70},0$	**22** 0 00	$1,0,0,\alpha^{190}$
$B_{2(1)}$		2 **21** 00	$1,0,\alpha^{222},0$	**$B_{2(3)}$**	
21 **00** 0	$1,\alpha^{5},0,\ 0$	1 **02** 00	$1,0,\alpha^{195},0$	00 **0** 00	1,0,0,0
12 **10** 0	$1,\alpha^{138},0,0$	2 **12** 00	$1,0,\alpha^{17},0$	10 **1** 00	$1,0,0,\alpha^{46}$
02 **20** 0	$1,\alpha^{191},0,0$	2 **22** 00	$1,0,\alpha^{131},0$	11 **2** 00	$1,0,0,\alpha^{101}$
12 **01** 0	$1,\alpha^{198},0,0$	**$B_{3(2)}$**		**$B_{3(3)}$**	
01 **11** 0	$1,\alpha^{11},0,0$	2 12 **00**	$1,0,\alpha^{17},0$	12 1 **00**	$1,0,0,\alpha^{138}$
20 **21** 0	$1,\alpha^{36},0,0$	1 21 **10**	$1,0,\alpha^{30},0$	02 0 **10**	$1,0,0,\alpha^{75}$
20 **02** 0	$1,\alpha^{86},0,0$	1 02 **20**	$1,0,\alpha^{109},0$	12 0 **20**	$1,0,0,\alpha^{170}$
11 **12** 0	$1,\alpha^{39},0,\ 0$	2 22 **01**	$1,0,\alpha^{105},0$	10 0 **01**	$1,0,0,\alpha^{189}$
11 **22** 0	$1,\alpha^{22},0,\ 0$	0 10 **11**	$1,0,\alpha^{228},0$	11 0 **11**	$1,0,0,\alpha^{34}$
$B_{3(1)}$		1 02 **21**	$1,0,\alpha^{154},0$	21 0 **21**	$1,0,0,\alpha^{212}$
01 12 **0**	$1,\alpha^{102},0,0$	1 01 **02**	$1,0,\alpha^{206},0$	00 2 **02**	$1,0,0,\alpha^{169}$
02 20 **1**	$1,\alpha^{150},0,0$	2 00 **12**	$1,0,\alpha^{220},0$	22 0 **12**	$1,0,0,\alpha^{180}$
22 20 **2**	$1,\alpha^{21},0,\ 0$	1 02 **22**	$1,0,\alpha^{239},0$	12 1 **22**	$1,0,0,\alpha^{97}$

Construct random covers α_k, for the same type as $\beta_{(k)}$

$$\alpha_{(k)}=\left[A_{1(k)},...,A_{s(k)}\right]=\left(a_{ij}\right)_{(k)}=S\left(a_{ij(k)_a},a_{ij(k)_b},a_{ij(k)_c},a_{ij(k)_d}\right)$$

where $a_{ij}\in A\left(P_\infty\right)$, $a_{ij(k)_a},a_{ij(k)_b},a_{ij(k)_c}\in F_q\setminus\{0\}$, $i=\overline{1,s}$, $j=\overline{1,r_{i(k)}}$, $k=\overline{1,3}$.

In the field representation α_k has the following form

$\alpha_k=\left[A_{1(k)},...,A_{s(k)}\right]=S\left(a_{ij(k)_a},a_{ij(k)_b},a_{ij(k)_c},a_{ij(k)_d}\right)$		
$k=1$	$k=2$	$k=3$
$A_{1(1)}$	**$A_{1(2)}$**	**$A_{1(3)}$**
$\alpha^{240},\alpha^{174},\alpha^{226},\alpha^{127}$	$\alpha^{223},\alpha^{101},\alpha^{7},\alpha^{221}$	$\alpha^{15},\alpha^{24},\alpha^{202},\alpha^{35}$
$\alpha^{138},\alpha^{119},\alpha^{129},\alpha^{201}$	$\alpha^{216},\alpha^{155},\alpha^{60},\alpha^{32}$	$\alpha^{93},\alpha^{97},\alpha^{87},\alpha^{51}$
$\alpha^{7},\alpha^{2},\alpha^{51},\alpha^{28}$	$\alpha^{17},\alpha^{185},\alpha^{157},\alpha^{28}$	$\alpha^{190},\alpha^{68},\alpha^{201},\alpha^{50}$
$\alpha^{170},\alpha^{155},\alpha^{165},\alpha^{221}$	**$A_{2(2)}$**	$\alpha^{193},\alpha^{190},\alpha^{13},\alpha^{68}$
$\alpha^{198},\alpha^{146},\alpha^{14},\alpha^{107}$	$\alpha^{112},\alpha^{143},\alpha^{1},\alpha^{134}$	$\alpha^{205},\alpha^{131},\alpha^{176},\alpha^{196}$
$\alpha^{2},\alpha^{154},\alpha^{152},\alpha^{102}$	$\alpha^{83},\alpha^{191},\alpha^{167},\alpha^{141}$	$\alpha^{60},\alpha^{177},\alpha^{21},\alpha^{168}$
$\alpha^{95},\alpha^{155},\alpha^{4},\alpha^{119}$	$\alpha^{214},\alpha^{28},\alpha^{66},\alpha^{1}$	$\alpha^{209},\alpha^{125},\alpha^{139},\alpha^{67}$
$\alpha^{63},\alpha^{35},\alpha^{96},\alpha^{25}$	$\alpha^{127},\alpha^{103},\alpha^{14},\alpha^{98}$	$\alpha^{148},\alpha^{64},\alpha^{102},0$
$\alpha^{120},\alpha^{100},\alpha^{81},\alpha^{45}$	$\alpha^{222},\alpha^{225},\alpha^{117},\alpha^{20}$	$\alpha^{223},\alpha^{170},\alpha^{149},\alpha^{88}$
$A_{2(1)}$	$\alpha^{113},\alpha^{14},\alpha^{98},\alpha^{26}$	**$A_{2(3)}$**
$\alpha^{46},\alpha^{70},\alpha^{68},\alpha^{172}$	$\alpha^{58},\alpha^{18},\alpha^{54},\alpha^{197}$	$\alpha^{51},\alpha^{124},\alpha^{209},\alpha^{73}$
$\alpha^{231},\alpha^{17},\alpha^{115},\alpha^{46}$	$\alpha^{131},\alpha^{171},\alpha^{212},\alpha^{190}$	$\alpha^{42},\alpha^{162},\alpha^{61},\alpha^{223}$
$\alpha^{232},\alpha^{52},\alpha^{168},\alpha^{15}1$	$\alpha^{73},\alpha^{36},\alpha^{40},\alpha^{126}$	$\alpha^{108},\alpha^{214},\alpha^{147},\alpha^{163}$
$\alpha^{8},\alpha^{157},\alpha^{8},\alpha^{239}$	**$A_{3(2)}$**	**$A_{3(3)}$**
$\alpha^{15},\alpha^{48},\alpha^{119},\alpha^{44}$	$\alpha^{184},\alpha^{83},\alpha^{17},\alpha^{110}$	$\alpha^{100},\alpha^{110},\alpha^{102},\alpha^{136}$
$\alpha^{1},\alpha^{109},\alpha^{147},\alpha^{210}$	$\alpha^{174},\alpha^{192},\alpha^{69},\alpha^{241}$	$\alpha^{163},\alpha^{216},\alpha^{174},\alpha^{168}$
$\alpha^{72},\alpha^{32},\alpha^{135},\alpha^{26}$	$\alpha^{87},\alpha^{174},\alpha^{179},\alpha^{18}$	$\alpha^{190},\alpha^{64},\alpha^{197},\alpha^{129}$
$\alpha^{148},\alpha^{127},\alpha^{167},\alpha^{188}$	$\alpha^{129},\alpha^{21},\alpha^{217},\alpha^{172}$	$\alpha^{188},\alpha^{137},\alpha^{32},\alpha^{193}$
$\alpha^{94},\alpha^{151},\alpha^{204},\alpha^{107}$	$\alpha^{178},\alpha^{232},\alpha^{16},\alpha^{162}$	$\alpha^{130},\alpha^{203},\alpha^{123},\alpha^{228}$
$A_{3(1)}$	$\alpha^{241},\alpha^{205},\alpha^{0},\alpha^{126}$	$\alpha^{86},\alpha^{58},\alpha^{86},\alpha^{191}$
$\alpha^{149},\alpha^{8},\alpha^{125},\alpha^{43}$	$\alpha^{155},\alpha^{179},\alpha^{39},\alpha^{201}$	$\alpha^{81},\alpha^{208},\alpha^{137},\alpha^{74}$
$\alpha^{169},\alpha^{161},\alpha^{182},\alpha^{180}$	$\alpha^{132},\alpha^{92},\alpha^{95},\alpha^{193}$	$\alpha^{4},\alpha^{98},\alpha^{100},\alpha^{127}$
$\alpha^{208},\alpha^{163},\alpha^{164},\alpha^{222}$	$\alpha^{216},\alpha^{180},\alpha^{9},\alpha^{92}$	$\alpha^{49},\alpha^{62},\alpha^{115},\alpha^{228}$

Choose random $t_{0(k)},t_{1(k)},...,t_{s(k)}\in A(P_\infty)\setminus Z$, $s_{(k)}=3$, $k=\overline{1,3}$ and $t_{3(1)}=t_{0(2)}$, $t_{3(2)}=t_{0(3)}$.

$t_{0(k)},t_{1(k)},...,t_{s(k)}\in A(P_\infty)\setminus Z$, $s=3$, $k=\overline{1,3}$		
$k=1$	$k=2$	$k=3$
$\alpha^{0},\alpha^{92},\alpha^{67},\alpha1^{04}$ $\alpha^{0},\alpha^{227},\alpha^{109},\alpha^{12}$ $\alpha^{0},\alpha^{180},\alpha^{204},\alpha^{22}$ $\alpha^{0},\alpha^{14},\alpha^{165},\alpha^{217}$	$\alpha^{0},\alpha^{14},\alpha^{165},\alpha^{217}$ $\alpha^{0},\alpha^{67},\alpha^{106},\alpha^{166}$ $\alpha^{0},\alpha^{76},\alpha^{221},\alpha^{150}$ $\alpha^{0},\alpha^{104},\alpha^{58},0$	$\alpha^{0},\alpha^{104},\alpha^{58},\ 0$ $\alpha^{0},\alpha^{89},\alpha^{84},\alpha^{169}$ $\alpha^{0},\alpha^{239},\alpha^{163},\alpha^{152}$ $\alpha^{0},\alpha^{72},\alpha^{88},\alpha^{91}$

The next step is to calculate the arrays γ_1, γ_2 and γ_3. By the condition of the example, we obtain

$$\gamma_{(k)}=\left[h_{1(k)},...,h_{s(k)}\right]=\left(h_{ij}\right)_{(k)}=t^{-1}_{(i-1)(k)}f_k\left(\left(a_{ij}\right)_{(k)}\right)\left(b_{ij}\right)_{(k)}t_{i(k)},$$

$k=\overline{1,3}$, $i=\overline{1,s(k)}$, $j=\overline{1,r_{i(k)}}$.

$\gamma_k=S(h_{ij(k)_a},h_{ij(k)_b},h_{ij(k)_c},h_{ij(k)_d})$, $k=\overline{1,3}$		
$h_{1(1)}$	**$h_{1(2)}$**	**$h_{1(3)}$**
$\alpha^{0},\alpha^{126},\alpha^{157},\alpha^{11}$	$\alpha^{0},\alpha^{192},\alpha^{99},\alpha^{198}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{9}$
$\alpha^{0},\alpha^{192},\alpha^{214},\alpha^{105}$	$\alpha^{0},\alpha^{192},\alpha^{79},\alpha^{208}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{12}$
$\alpha^{0},\alpha^{220},\alpha^{95},\alpha^{150}$	$\alpha^{0},\alpha^{192},\alpha^{92},\alpha^{125}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{9}$
$\alpha^{0},\alpha^{85},\alpha^{91},\alpha^{218}$	**$h_{2(2)}$**	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{167}$
$\alpha^{0},\alpha^{208},\alpha^{146},\alpha^{210}$	$\alpha^{0},\alpha^{112},\alpha^{146},\alpha^{132}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{176}$
$\alpha^{0},\alpha^{70},\alpha^{81},\alpha^{137}$	$\alpha^{0},\alpha^{112},\alpha^{193},\alpha^{16}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{203}$
$\alpha^{0},\alpha^{232},\alpha^{227},\alpha^{122}$	$\alpha^{0},\alpha^{112},\alpha^{1},\alpha^{70}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{162}$
$\alpha^{0},\alpha^{47},\alpha^{36},\alpha^{62}$	$\alpha^{0},\alpha^{112},\alpha^{93},\alpha^{218}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{181}$
$\alpha^{0},\alpha^{126},\alpha^{157},\alpha^{11}$	$\alpha^{0},\alpha^{112},\alpha^{69},\alpha^{73}$	$\alpha^{0},\alpha^{175},\alpha^{155},\alpha^{48}$
$h_{2(1)}$	$\alpha^{0},\alpha^{112},\alpha^{229},\alpha^{64}$	**$h_{2(3)}$**
$\alpha^{0},\alpha^{137},\alpha^{145},\alpha^{171}$	$\alpha^{0},\alpha^{112},\alpha^{227},\alpha^{187}$	$\alpha^{0},\alpha^{194},\alpha^{29},\alpha^{133}$
$\alpha^{0},\alpha^{108},\alpha^{83},\alpha^{95}$	$\alpha^{0},\alpha^{112},\alpha^{47},\alpha^{19}$	$\alpha^{0},\alpha^{194},\alpha^{29},\alpha^{232}$
$\alpha^{0},\alpha^{109},\alpha^{11},\alpha^{174}$	$\alpha^{0},\alpha^{112},\alpha^{159},\alpha^{14}$	$\alpha^{0},\alpha^{194},\alpha^{29},\alpha^{139}$
$\alpha^{0},\alpha^{137},\alpha^{230},\alpha^{115}$	**$h_{3(2)}$**	**$h_{3(3)}$**
$\alpha^{0},\alpha^{227},\alpha^{107},\alpha^{29}$	$\alpha^{0},\alpha^{30},\alpha^{126},\alpha^{154}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{171}$
$\alpha^{0},\alpha^{194},\alpha^{7},\alpha^{227}$	$\alpha^{0},\alpha^{30},\alpha^{15},\alpha^{12}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{164}$
$\alpha^{0},\alpha^{15},\alpha^{155},\alpha^{29}$	$\alpha^{0},\alpha^{30},\alpha^{66},\alpha^{223}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{117}$
$\alpha^{0},\alpha^{162},\alpha^{224},\alpha^{174}$	$\alpha^{0},\alpha^{30},\alpha^{53},\alpha^{110}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{60}$
$\alpha^{0},\alpha^{147},\alpha^{37},\alpha^{123}$	$\alpha^{0},\alpha^{30},\alpha^{19},\alpha^{186}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{229}$
$h_{3(1)}$	$\alpha^{0},\alpha^{30},\alpha^{219},\alpha^{188}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{122}$
$\alpha^{0},\alpha^{191},\alpha^{239},\alpha^{34}$	$\alpha^{0},\alpha^{30},\alpha^{81},\alpha^{131}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{217}$
$\alpha^{0},\alpha^{144},\alpha^{48},\alpha^{120}$	$\alpha^{0},\alpha^{30},\alpha^{78},\alpha^{196}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{37}$
$\alpha^{0},\alpha^{11},\alpha^{176},\alpha^{137}$	$\alpha^{0},\alpha^{30},\alpha^{62},\alpha^{238}$	$\alpha^{0},\alpha^{146},\alpha^{66},\alpha^{175}$

For example, let $R_1 = \left(R_{1(1)}, R_{2(1)}, R_{3(1)}\right) = (2,3,0) = 29$.

$\gamma_1(29) = h_{1(1)}(2)\, h_{2(1)}(3)\, h_{3(1)}(0) = S\left(\alpha^0, \alpha^{66}, \alpha^{62}, \alpha^{29}\right)$.

For $R_2 = \left(R_{1(2)}, R_{2(2)}, R_{3(2)}\right) = (1,1,1) = 31$. Compute γ_2

$\gamma_2(31) = h_{1(2)}(1)\, h_{2(2)}(1)\, h_{3(2)}(1) = S\left(\alpha^0, \alpha^{160}, \alpha^{112}, \alpha^{186}\right)$.

For $R_2 = \left(R_{1(2)}, R_{2(2)}, R_{3(2)}\right) = (3,1,1) = 39$. Compute γ_3

$\gamma_3(39) = h_{1(3)}(3)\, h_{2(3)}(1)\, h_{3(3)}(1) = S\left(\alpha^0, \alpha^{5}, \alpha^{72}, \alpha^{215}\right)$.

Encryption

Input: a message $m \in A(P_\infty)$, $m = S(m_1, m_2, m_3, m_4)$, $m_1 \in F_q \setminus \{0\}$, $m_2, m_3, m_4 \in F_q$ and the public key $[f_k, (\alpha_k, \gamma_k)]$, $k = \overline{1,3}$.

Let $m = \left(a^0, a^1, a^2, a^3\right) = S\left(a^0, a^1, a^2, a^3\right)$.

Choose a random $R = (R_1, R_2, R_3) = (29, 31, 39)$. Lets define a mapping for the encryption keys such as

$R' = \pi(R_1, R_2, R_3) = (R_3, R_2, R_1)$

Compute cipher text

$$y_1 = \alpha'(R)\cdot m = \alpha_1'(R_3)\cdot\alpha_2'(R_2)\cdot\alpha_3'(R_1)\cdot m =$$
$$S\left(\alpha^1, \alpha^{71}, \alpha^{159}, \alpha^{131}\right) S\left(a^0, a^1, a^2, a^3\right) = S\left(\alpha^1, \alpha^{11}, \alpha^{187}, \alpha^{216}\right)$$

Calculate

$$\gamma(R) = \gamma_1'(R_1)\cdot\gamma_2'(R_2)\cdot\gamma_3'(R_3) = S\left(\alpha^0, \alpha^{140}, \alpha^{184}, \alpha^{155}\right)$$

and second component of the cipher text

$$y_2 = \gamma(R)\cdot \square f_2\left(\alpha_3'(R_3)\right)\cdot \square f_1\left(\alpha_3'(R_3)\right)\cdot \square f_1\left(\alpha_2'(R_2)\right) =$$
$$S\left(\alpha^0, \alpha^{21}, \alpha^{191}, \alpha^{65}\right).$$

Compute component y_3.

$$\lambda(R) = \alpha_1'(R_1)\cdot f_1\left(\alpha_2'(R_2)\right)\cdot f_1\left(\alpha_3'(R_3)\right) =$$
$$S\left(\alpha^0, \alpha^{25}, \alpha^{142}, \alpha^{53}\right),$$

$$y_3 = \lambda(R)\square f_1\left(\alpha_3'(R_3)\right)\cdot \square f_1\left(\alpha_2'(R_2)\right) = S\left(\alpha^0, \alpha^{107}, \alpha^{60}, \alpha^2\right).$$

We obtained output $y_1 = \left(\alpha^1, \alpha^{11}, \alpha^{187}, \alpha^{216}\right)$, $y_2 = \left(\alpha^0, \alpha^{21}, \alpha^{191}, \alpha^{65}\right)$, $y_3 = \left(\alpha^0, \alpha^{107}, \alpha^{60}, \alpha^2\right)$.

Decryption

Input: a ciphertext (y_1, y_2, y_3) and private key $\left[\beta_k, \left(t_{0(k)}, \ldots, t_{s(k)}\right)\right]$, $k = \overline{1,3}$.

Output: the message $m \in A(P_\infty)$ corresponding to ciphertext (y_1, y_2, y_3).

To decrypt a message m, we need to restore random numbers $R = (R_1, R_2, R_3)$.

Compute

$$D(R_1, R_2, R_3) = t_{0(1)} y_2 y_3^{-1} t_{s(3)}^{-1} = S\left(\alpha^0, \alpha^2, \alpha^{148}, \alpha^{213}\right).$$

We get $\beta_1(R_1) = \alpha^2 = (00100)$.

Perform inverse calculations $\beta_3(R_3)^{-1}$.

<table>
<tr><td>00|10|<b>0</b></td><td>R_1=(*,*,0)</td></tr>
<tr><td>01|12|<b>0</b></td><td>row 0 from $B_{3(1)}$</td></tr>
<tr><td>00|10|<b>0</b>−01|12|<b>0</b>=02|<b>01</b>|0</td><td>R_1= (*,3,0)</td></tr>
<tr><td>12|<b>01</b>|0</td><td>row 3 from $B_{2(1)}$</td></tr>
<tr><td>02|<b>01</b>|0−12|<b>01</b>|0=<b>20</b>|00|0</td><td></td></tr>
</table>

We get $\beta_1(R_1)^{-1} = (2,3,0) = 29$

For further calculations, it is necessary to remove the components of the components $\alpha_1'(R_1)$ and $\gamma_1'(R_1)$ from ciphertext (y_2, y_3).

Compute

$$y_2^{(1)} = \gamma_1'(R_1)^{-1} y_2 = S\left(\alpha^0, \alpha^{37}, \alpha^{225}, \alpha^{62}\right)$$

and

$$y_3^{(1)} = \alpha_1'(R_1)^{-1} y_3 = S\left(\alpha^0, \alpha^{65}, \alpha^{129}, \alpha^{115}\right).$$

Compute

$$D(R_2, R_3) = t_{0(2)} y_2^{(1)} \left(y_3^{(1)}\right)^{-1} t_{s(3)}^{-1} = S\left(\alpha^0, 0, \alpha^{227}, \alpha^{175}\right).$$

We get $\beta_2(R_2) = \alpha^{227} = (10110)$. Restore R_2 with $\beta_2(R_2) = \alpha^{227} = (10110)$. We use the same calculations as in the example for $\beta_1(R_1)^{-1}$, and we get

<table>
<tr><td>1|01|<b>10</b></td><td>R_2=(*,*,1)</td></tr>
<tr><td>1|21|<b>10</b></td><td>row 1 from $B_{3(2)}$</td></tr>
<tr><td>1|01|<b>10</b>−1|21|<b>10</b>=0|10|00</td><td>R_2= (*,1,1)</td></tr>
<tr><td>2|<b>10</b>|00</td><td>row 1 from $B_{2(2)}$</td></tr>
<tr><td>0|<b>10</b>|00−2|<b>10</b>|00=1|00|00</td><td>R_2=(1,1,1)</td></tr>
</table>

$$\beta_2(R')^{-1} = 1|01|10 = \left(R_{1(2)}, R_{2(2)}, R_{3(2)}\right) = (1,1,1) = 31.$$

Perform inverse calculations $\beta_2(R_2)^{-1}$. Select bit groups in vector $\beta(R)$ according to type $\left(r_{1(2)}, \ldots, r_{s(2)}\right) = \left(3, 3^2, 3^2\right)$. We use the same calculations as in the example for $\beta_1(R_1)^{-1}$, and we get $\beta_2(R')^{-1} = (1,1,1) = 31$

Remove the component $\gamma_2'(R_2)$ from $y_2^{(1)}$ and $f_1\left(\alpha_2'(R_2)\right)$ from $y_3^{(1)}$

$$y_2^{(2)} = \gamma_2'(R_2)^{-1} y_2^{(1)} = S\left(\alpha^0, \alpha^{83}, \alpha^{211}, \alpha^{61}\right)$$

$$y_3^{(2)} = f_1\left(\alpha_2'(R_2)\right)^{-1} y_3^{(1)} = S\left(\alpha^0, \alpha^{65}, \alpha^{207}, \alpha^{160}\right).$$

Compute

$$D(R_3) = t_{0(3)} y_2^{(2)} \left(y_3^{(2)}\right)^{-1} t_{s(3)}^{-1} = S\left(\alpha^0, 0, 0, \alpha^{227}\right).$$

We get $\beta_3(R_3) = \alpha^{227} = (10110)$. Perform inverse calculations $\beta_3(R_3)^{-1}$. Select bit groups in vector $\beta(R)$ according to type $\left(r_{1(3)}, \ldots, r_{s(3)}\right) = \left(3^2, 3, 3^2\right)$. We get

<table>
<tr><td>10|1|<b>10</b></td><td>R_3=(*,*,1)</td></tr>
<tr><td>02|0|<b>10</b></td><td>row 1 from $B_{3(3)}$</td></tr>
<tr><td>10|1|<b>10</b>−02|0|<b>10</b>=11|1|00</td><td>R_3= (*,1,1)</td></tr>
<tr><td>10|<b>1</b>|00</td><td>row 1 from $B_{2(3)}$</td></tr>
<tr><td>11|<b>1</b>|00−10|<b>1</b>|00=01|0|00</td><td>R_3=(3,1,1)</td></tr>
</table>

$$\beta_3(R')^{-1} = 1|01|10 = \left(R_{1(3)}, R_{2(3)}, R_{3(3)}\right) = (3,1,1) = 39.$$

Receive a message $m = \alpha'(R)^{-1} y_1 = S\left(a^0, a^1, a^2, a^3\right)$.

Security Analysis

Consider a brute force attack of key recovery. There are three possible schemes for such an attack.

Brute force attack on cipher text. By selecting $R = (R_1, R_2, R_3)$ try to decipher the text $y_1' = \alpha'(R')\cdot m = \alpha_1'(R_1')\cdot\alpha_2'(R_2')\cdot\alpha_3'(R_3')\cdot m$. The covers $\alpha_{(k)} = \left(a_{ij}\right)_{(k)} = S\left(a_{ij(k)_a}, a_{ij(k)_b}, a_{ij(k)_c}, a_{ij(k)_d}\right)$ are selected randomly and the value is determined by multiplication in a group with no coordinate constraints. The resulting vector $\alpha'(R')$ depends on all components $\alpha_1'(R_1'), \alpha_2'(R_2'), \alpha_3'(R_3')$. Enumeration of key values $R = (R_1, R_2, R_3)$ has an estimation of complexity. For a practical attack, the message m is also unknown and has uncertainty to choose from q^3 . This makes a brute-force attack on a key infeasible. If we take an attack model with a known text, then the attack complexity still remains the same and equal to q^3 .

Brute force attack on the cyphertext y_2 . Select $R = (R_1, R_2, R_3)$ to match y_2 . The vector y_2 has a following definition over the components $\alpha_i'(R_i)$

$$y_2 = S\left(1, \sum_{k=1}^{3} \sum_{i=1, j=R_{i(1)}}^{s(1)} a_{ij(k)_b} + \sum_{i=1, j=R_{i(1)}}^{s(1)} \beta_{ij(1)_b} + *, \right.$$
$$\sum_{i=1, j=R_{i(2)}}^{s(2)} \left(a_{ij(2)_c} + \beta_{ij(2)_c}\right) + \sum_{i=1, j=R_{i(3)}}^{s(3)} a_{ij(3)_c} + *,$$
$$\left. \sum_{i=1, j=R_{i(3)}}^{s(3)} \left(a_{ij(3)_d} + \beta_{ij(3)_d}\right) + * \right)$$

The values of the coordinates y_2 are defined by calculations over the vectors $\alpha_1'(R_1), \alpha_2'(R_2), \alpha_3'(R_3)$. The keys R_1, R_2, R_3 are bound and changes in any of them leads to change y_2 . The brute force attack on key $R = (R_1, R_2, R_3)$ has a complexity equal to q^3 .

Brute force attack on the ciphertext y_3 . Select $R = (R_1, R_2, R_3)$ to match y_3 . The vector y_3 has a following definition over the components $\alpha_i'(R_i)$

$$y_3 = S\left(1, \sum_{k=1}^{3} \sum_{i=1, j=R_{i(k)}}^{s(k)} a_{ij(k)_b} , \right.$$
$$\left. \sum_{k=2}^{3} \sum_{i=1, j=R_{i(k)}}^{s(k)} a_{ij(k)_c} + *, \sum_{k=2}^{3} \sum_{i=1, j=R_{i(k)}}^{s(k)} a_{ij(k)_d} + * \right).$$

The values of the coordinates y_2 are defined by calculations over the vectors $\alpha_1'(R_1), \alpha_2'(R_2), \alpha_3'(R_3)$. The keys R_1, R_2, R_3 are bound and changes in any of them leads to change y_2 . The brute force attack on key $R = (R_1, R_2, R_3)$ has a complexity equal to q^3 .

Brute force attack on the $\left(t_{0(k)}, \ldots, t_{s(k)}\right)$. The brute force attack on $\left(t_{0(k)}, \ldots, t_{s(k)}\right)$ is a general for the MST cryptosystems and for the calculation in the field F_q over the group center $Z(G)$ has an optimistic complexity estimation equal to q . For the proposed algorithm all calculations are executed on whole group $|G| = q^3$ and is a such case the complexity of the brute force attack on $\left(t_{0(k)}, \ldots, t_{s(k)}\right)$ will be equal to q^3 .

Attack on the algorithm. The given estimation of such an attack will be valid for the MST cryptosystem implementation based on any non-commutative group and requires a separate analysis. This attack has many details that related to logarithmic signature vulnerabilities and possibly group operation.

IV. Conclusions

Our suggestion is to use the automorphism group of the Ree function field for full group $A(P_\infty)$ encryption with bound keys $R = (R_1, R_2, R_3)$ and brute force attack complexity q^3 . We have improved the encryption algorithm to bind logarithmic signature keys and protect against sequential recovery attacks.

References

[1] N.R. Wagner and M.R. Magyarik, A public-key cryptosystem based on the word problem", Proc. Advances in Cryptology – CRYPTO 1984, LNCS 196, Springer-Verlag (1985), 19–36.

[2] J. Birget, S. S. Magliveras, and M. Sramka, ``On public-key Cryptosystems based on combinatorial group theory," Tatra Mt. Math. Publ., vol. 33, pp. 137-148, Jan. 2006.

[3] A. Caranti and F. D. Volta, ``The round functions of cryptosystem PGM generate the symmetric group," Des. Codes Cryptogr., vol. 38, no. 1,pp. 147_155, 2006.

[4] W. Lempken, S.S. Magliveras, Tran van Trung and W. Wei, "A public key cryptosystem based on non-abelian finite groups", J. of Cryptology, 22 (2009), 62–74.

[5] S. S. Magliveras, "A cryptosystem from logarithmic signatures of finite groups," in Proceedings of the 29th Midwest Symposium on Circuits and Systems , pp. 972–975, Elsevier Publishing, Amsterdam, The Netherlands, 1986.

[6] P. Svaba and T. van Trung, "Public key cryptosystem MST3 cryptanalysis and realization", Journal of Mathematical Cryptology,vol.4,no.3,pp.271–315,2010.

[7] T. van Trung, ``New approaches to designing public key cryptosystems using one-way functions and trapdoors in finite groups," J. Cryptol., vol. 15, no. 4, pp. 285-297, 2002.

[8] Magliveras S S, Svaba P, van Trung T, et al. On the security of a realization of cryptosystem MST3. Tatra Mt Math Publ, 2008, 41: 1–13

[9] T. van Trung, ``Construction of strongly aperiodic logarithmic signatures," J. Math. Cryptol., vol. 12, no. 1, pp. 23-35, 2018.

[10]A.Bassa, L.Ma, Ch.Xing, S.L.Yeo Towards a characterization of subfields of the Deligne–Lusztig function fields, Journal of Combinatorial Theory, Series A Volume 120, Issue 7, September 2013, Pages 1351-1371

[11]G. Khalimov, Y. Kotukh, S.Khalimova "MST3 cryptosystem based on the automorphism group of the hermitian function field" // IEEE International Scientific-Practical Conference: Problems of Infocommunications Science and Technology, PIC S and T 2019 - Proceedings, 2019, pp. 865–868.

[12]G. Khalimov, Y. Kotukh, S.Khalimova "MST3 cryptosystem based on a generalized Suzuki 2 - Groups" // CEUR Workshop Proceedings, 2020, 2711, pp. 1–15.

[13]G. Khalimov, Y. Kotukh, S.Khalimova "Encryption scheme based on the automorphism group of the Ree function field" 2020 7th International Conference on Internet of Things: Systems, Management and Security, IOTSMS 2020, 2020, 9340192

[14]W. Lempken and T. van Trung, "On minimal logarithmic signatures of finite groups," Experimental Mathematics,vol.14, no. 3, pp. 257–269, 2005.

[15]P. Svaba, "Covers and logarithmic signatures of finite groups in cryptography", Dissertation, https://bit.ly/2Ws2D24

S. Szaby, Topics in Factorization of Abelian Groups, Birkhдuser, Basel, 2004.

A Systematic Literature Review on Malware Analysis

Fahad Mira
Fahad.Mira@beds.ac.uk

Department of Computer Science and Technology, University of Bedfordshire, Luton, UK

***Abstract*— Malware is a significant security danger on the Internet nowadays. Hostile to Virus organizations get a huge number of malwares tests each day. It is intended to harm PC frameworks without the information on the proprietor utilizing the framework and method headways are presenting enormous difficulties for scientists in both the scholarly world and the business. Malware tests are arranged and gathered for additional investigation. In this literature review, we did the manual research on the publications from the year 2014 to 2020. We selected about 27 articles out of 55 articles as primary studies and applied quality evaluation criteria and deducted research questions from them. The motivation behind this SLR is to inspect the accessible literary works on malware examination and to decide how exploration has developed and progressed regarding the amount, substance, and publication outlets. We also discussed the issues and challenges we are facing in malware analysis along with detection system requirements. Large numbers of the malicious programs are tremendous and confounded so it is difficult for researchers to fathom its subtleties. Scattering of malicious data beyond clients of the web and furthermore preparing them to effectively utilize against malicious items are critical to shielding clients from malicious attack. This review paper will give a comprehensive book index of techniques to help with battling malicious data.**

Keywords*— *Malware; Malware analytics; Malware code; Taxonomy; Signature-based; Anomaly-based; Malware system requirements.

I. Introduction

Malware is an overall term that incorporates infections, Trojans, Spywares, and other obtrusive code is far and wide today. Malware investigation is a multistep cycle giving knowledge into malware design and usefulness, encouraging the development of a cure.

The expression "malware" here is being utilized as the conventional name for the class of code that is pernicious, including infections, Trojans, worms, and spyware. Malware writers use generators, fuse libraries, and get code foremothers—there exists a hearty organization for trade, and some malware writers set aside an effort to peruse and comprehend earlier methodologies.

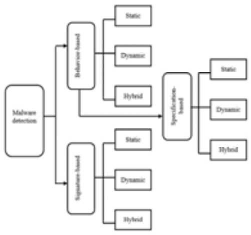
first-historically speaking PC infection (malware) Brain showed up in 1986 [1].

Malware is utilized to send spam messages, to perform web cheats, to take individual data like MasterCard data, and for some, other accursed assignments like Ransomware [2] and counterfeit antivirus programming [3].

Since 1988 [4, 5] the increment in the quantity of PC-based security penetrates affirms that noxious programming has arrived at practically unmanageable levels. Mulling over the degree of potential harm brought about by noxious programming, its discovery alone has caused huge issues for both the agents and the overall public. Recognition frameworks made by examiners are consistently put to broad use in identification works out. This paper is committed to investigating malignant programming location systems.

A. Malware analysis with respect to behaviour and signature-based:

Malware programming identification is divided into two parts: Signature-and Behavior-based advances and every innovation can be utilized with static, dynamic, or hybrid examination [5][16]. The specific technique for an oddity or mark put-together method is based on respect to how the innovation orders the data to distinguish noxious programming [6-9]. How noxious programming discovery is overseen is appeared in Fig. 1.



*F*ig.1 Flow-chart of malware detection

I. Behavior (Anomaly) based:

An anomaly-based put together discovery attracts with respect to its database to decide the presence of ordinary conduct to choose the firmness of malware in examination. Another type of inconsistency-based discovery is called specification-based recognition. This type of discovery examination occurs in 2 circumstances:

- In preparing and learning circumstance. While in the preparation circumstance, an indicator attempts to get familiar with a typical conduct. It is very conceivable that an indicator is learning the host's conduct or the PUI's or possibly the two joined. The principal advantage of anomaly-based identification is the capacity to identify ‘zero-day” interruptions.
- having location and observing circumstance

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

The 2 fundamental downsides of such system are:

- Immense bogus caution charges: It is apt of unreasonable bogus alert charges, which are characterized as 'typical' yields sorted as (bogus positive) and separated by the all-out figures of 'ordinary' conduct.
- Also getting trouble in guaranteeing what boundaries must realize in preparation circumstance.

II. Detection by signature-based

Signature based discovery uses technology-based character to observe malicious code and thus affirm a malignant idea of a program in examination. Putting aside, the signature-based discovery attempts by setting a criterion for utilizing a malicious code and accordingly uses it as the sort of perspective for distinguishing another noxious programming. In gathering every one of these models, the signature-based recognition creates an information base for itself. In an ideal framework, it is basic that the mark ought to perceive any program showing conduct fitting the mark's malignant data set. This data set contains all data required by the mark to recognize vindictive programming. This data set is counseled at whatever point here lies a possible issue with PUI.

The one of most fundamental issues of the signature-based recognition technique is the failure to perceive 'zero day' interruptions. A zero-day interruption means where there is none comparative mark lies in any information base to contrast and additionally, the accomplished individual is likely expected to plan a mark. Besides offering an approach to administrator mistakes it is a dreary cycle if the plan and establishment are not set up to work naturally. The way that certain malware can multiply the capacity to plan and introduce a more exact mark is amazingly basic. Engineers of such marks, which work on a programmed mode, could be found absent a lot of exertion, however altogether more energy should be placed into doing this. Be that as it may, all recognition systems could utilize one of three different philosophies.

TABLE 1 SUMMARY

Analysis tools	Purposes	Tools
Static	Use whatever number antivirus recognition motors as could reasonably be expected to help characterization.	“Virus Total” (2008)
	Search the body of the malware for strings.	‘Strings” (Microsoft, 2008)
Dynamic	Document respectability check to record gauge setup.	“Winalysis” (2008)
	Record observing. Discover which devices are opening, perusing and composing documents.	“Filemon” (2008)
	Vault observing. Screen vault exercises as they happen.	“Regmon” (Microsoft, 2008)
Hybrid	Dismantling, investigating	“IDA Pro” “OllyDbg” (Yuschuk, 2008)

B. Analysis Methodologies

The three principal malware investigation procedures are static, dynamic, and hybrid examination. Every examination strategy has its own favorable circumstances and weaknesses which have been talked about in this segment. Table 1 shows the outline of investigation devices.

1. Static Analysis

It is an examination of programming executed without truly performing a program [10]. Different procedures are carried out to play out this examination. While a few rely upon the characteristics of the twofold record, by removing “byte code courses” of action from a combined one, and isolating Op code progressions in the wake of destroying the twofold report, to eliminate the “control stream graphical figures” among the party record, and eliminating API calls, from the equal, and the like. Each addresses the rundown of capacities and anyone or many are utilized for malicious area.

2. Dynamic Analysis:

It is the analysis in which while executing the program, a programming is done [17]. An information segment can achieve by the one of the examinations that is API calls, structure calls, direction follows, dirty assessment, vault changes, memory makes, etc. A part of the malware acknowledgment procedure utilizing dynamic one that has been investigated formerly is according to the accompanying.

II. METHODOLOGY

In this section, we are going to apply the methodology for this systematic review that was proposed by ‘Kitchenham” [35]. This part presents the technique to achieve the objectives of literature review. The phases of our procedure include:

(1) Data planning
(2) Making of research questions
(3) Searching measures
(4) Addition and rejection measures

A. Data planning

In this phase, to achieve the goals of the current examination we recognized the important ways. In the given underlying process, it was guaranteed that the key and specialized plans were appropriately defined. This is guaranteed that different periods of the proposed philosophy were appropriately completed in a coordinated and standard way. This arranging stage shaped the reason for a fruitful usage of the proposed SLR strategy.

B. Forming research questions:

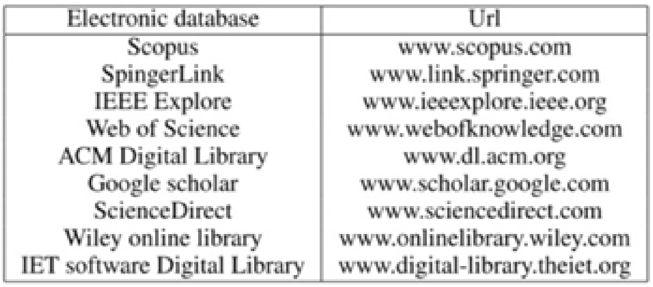
In the given section, we present the exploration addresses researched in the flow SLR study. The exploration addresses RQs examined in our investigation are:
RQ1. What number of yearly number of studies on malware examination have there been since 2010?
RQ2. What types of datasets are used?

RQ3. What is the size of datasets?
RQ4. What types of analysis methods are used in this research?
RQ5. Which are the requirements of malware programming detectors?
RQ6. What are the advantages and disadvantages of malware identification draws near?
RQ7. What are the issues and challenges faced in malware analysis?

C. Searching measures

In the following part, we illustrate the measures to distinguish the given articles for this examination. To remove pertinent SLR concentrates on malware investigation, various electronic information bases were thought of and gotten to. The rundown of information bases looked, and their comparing URL is introduced in Table 2.

TABLE 2 ELECTRONIC DATABASES



Electronic database	Url
Scopus	www.scopus.com
SpingerLink	www.link.springer.com
IEEE Explore	www.ieeexplore.ieee.org
Web of Science	www.webofknowledge.com
ACM Digital Library	www.dl.acm.org
Google scholar	www.scholar.google.com
ScienceDirect	www.sciencedirect.com
Wiley online library	www.onlinelibrary.wiley.com
IET software Digital Library	www.digital-library.theiet.org

We take out each article from the given electronic databases utilizing the customary searching measures from conferences and journals individually.

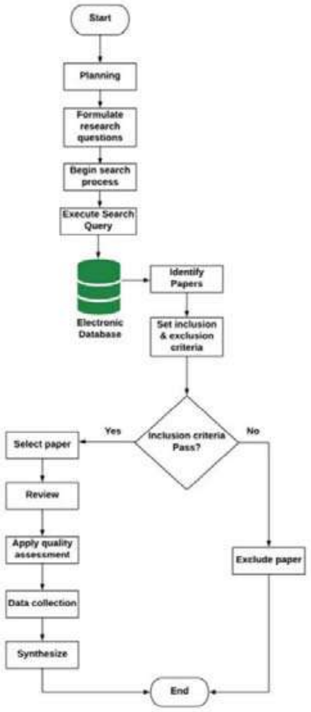


Fig.2. SLR process

The pursuit string was consequently adjusted to suit the necessities of every information base. We looked through every information base by titles, edited compositions, and catchphrases. The figure 2 depicts the rules of deliberate cycle.

D. Inclusion and rejection measures:

The investigations remembered for this SLR depended on specific measures that decided that if an examination matches the condition of inclusion, otherwise such an investigation will be precluded. Articles written in language other than English were excluded from the current examination on the grounds that such articles would be hard to peruse and comprehend.

	Table 5	**[Primary studies]**			
Ref. no.	**Author, year**	**Area of research**	**Institutions**	**Publications**	**Contributions**
[PS5]	B.Yu et al. 2018	A survey of malware behavior description and analysis	National university of defense technology	journal	This paper conducted a survey on malware behaviour description and analysis considering description, analysis and visualization methods.
[PS11]	Parmjit et al. 2014	Literature Analysis on Malware Detection	Chandigarh University	Journal	This paper gives the android architecture, various types of malware and literature analysis for security considerations in android smartphones.
[PS12	H.M Deylami et al. 2016	Taxonomy of malware detection techniques	Universiti Kebangsaan	Conference paper	This gives a comprehensive list of sources of strategies to help with battling malware.
[PS13]	Maigida et al. 2019	SLR and metadata analysis of ransomware attacks and detection mechanisms	Federal University of Technology	journal	This paper fills in as a benchmark for new analysts in proposing the ransomware discovery technique.
[PS14]	Ya Pan et al. 2020	An SLR of android malware detection using static analysis	Nanjing university	journal	A systematic literature review is performed to perform the clarification on Android malware detection by using malware detection method

III. Conclusion and Furture work

Malicious arracks lead to the typical danger for PC and correspondence frameworks to harm gadgets or take classified data. The reason for this SLR is to concoct a proficient procedure for malware location that consolidates the upsides of anomaly and signature based detection. This paper identified a definite order of malignant programming detection and evasion programs for specialists to 'nibble into'. Devoted significance was doled out to malignant programming necessities and acknowledgment given to the basics of malware recognition and avoidance strategies..

References

[1] Arief, B. & Bernard, D, “ Technical and human issues in computer-based systems security”, *University of Newcastle upon Tyne*, 2010.

[2] H. J. Highland, “A History of Computer Viruses -*The Famous ‘Trio’,” Computers & Security*, Vol. 60, No. 5, pp. 412-415, 1997.

[3] A. Gazet, “Comparative analysis of various ransomware virii”, *Journal in Computer Virology*, Vol. 6, No. 1, pp. 77-90, 2010.

[4] Barbara Guttman, Edward A. Roback, “An Introduction to Computer Security: The NIST Handbook”, *Computer Systems Laboratory, National Institute of Standards and Technology*, Gaithersburg, MD 20899-0001,1995.

[5] Howard F. Lipson, “Tracking and Tracing Cyber-Attacks: Technical Challenges and Global Policy Issues”, *PhD CERT ® Coordination Center, Networked Systems Survivability*
Program, 2002.

[6] Nwokedi Idika, Aditya P. Mathur, “A Survey of Malware Detection Techniques”, *Department of Computer Science Purdue University*, West Lafayette, IN 47907, 2007.

[7] Hao, S., Wang, W., Lu, H. and Ren, P. “AutoMal: automatic clustering and signature generation for malwares based on the network flow”, *Security Comm. Networks*, 2014.

[8] Muazzam Ahmed Siddiqui, “Data mining methods for malware detection”, PhD thesis, *College of Sciences, University of Central Florida*, Orlando, Florida, 2008.

[9] Xue, L., Sun, G., “Design and implementation of a malware detection system based on network behavior”, *Security Comm. Networks*, 2014.

[10] Threat Expert, “Threat Expert,” *Threat Expert*, [Online].*Available*: http://www.threatexpert.com/. [Accessed 20 01 2021].

[11] COMODO, “COMODO Automated Analysis System,”COMODO, [Online]. Available:http://camas.comodo.com/.[Accessed 19 01 2021].

[12] M. Egele, T. Scholte, E. Kirda, C. Kruegel, “A Survey on Automated Dynamic Malware Analysis Techniques and Tools”, *Journal ACM Computing Surveys*, Volume 44, Issue 2, Article No. 6, 2012.

[11] M. Christodorescu, S. Jha. “Static analysis of executables to detect malicious patterns”, *USENIX Security Symposium*, 2003.

[12] F. Leder, B. Steinbeck, P. Martini, “Classification and Detection of Metamorphic Malware using Value Set Analysis”, in 4th International Conference on Malicious and Unwanted Software (MALWARE), 2009.

[13] M. G. Schultz, E. Eskin, E. Zadok, S. J. Stolfo, “Data mining methods for detection of new malicious executables”, *In IEEE Symposium on Security and Privacy*, 2001.

[14] P.Deshpande, “Metamorphic Detection Using Function Call Graph Analysis”, *Master’s Projects*, Paper 336 http://scholarworks.sjsu.edu/etd projects/336,2013

[15] G. Shanmugam, R. Low, M. Stamp. “Simple Substitution Distance and Metamorphic Detection,” *Journal of Computer Virology and Hacking Techniques*, Volume 9, Issue 3, pp. 159–170, 2013.

Languages and Computing, Volume 23, Issue 3,pp. 154–162,2012.

[`16] Deylami, H.M., Muniyandi, R.C., Ardekani, I.T. and Sarrafzadeh, A., 2016, December. Taxonomy of malware detection techniques: A systematic literature review. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)* (pp. 629-636). IEEE.

Identifying Phasic dopamine releases using DarkNet-19 Convolutional Neural Network

Qasem Abu Al-Haija[1] *, Mahmoud Smadi[2], Osama M. Al-Bataineh [3]

[1] Department of Data Science & Artificial Intelligence, University of Petra, Amman 1196, Jordan
[2] Department of Electrical Engineering, The Hashemite University, Zarqa 13133, Jordan
[3] Department of Biomedical Engineering, The Hashemite University, Zarqa 13133, Jordan

[1*]qasem.abualhaija@uop.edu.jo, [2] smadi@hu.edu.jo, [3] omabio@hu.edu.jo

Abstract*—Understanding the role of neurotransmitter dopamine in brain function under normal or pathological states is one of the most active areas of research in neurosciences. Failures in dopamine neurotransmission affects tremendous amount of brain abilities including movement, mental, and motivation and reward systems of the brain. Ability to measure phasic release of dopamine in specific locations of the brain will lead to a powerful tool for the neuroscientists, However, the tremendous amount of image-formed data as produced from different locations of the brain makes the manual analysis of these data cumbersome. Luckily, image processing techniques will help in solving these problems effortlessly to ease and speed the analysis for neuro-physicians. In this paper, we propose a deep-learning based identification scheme to identify the release case of phasic dopamine by examining the dopamine analysis (DA) imaging attributes using a convolutional neural network (CNN). More precisely, the proposed scheme exploits the transfer learning based DarkNet-19 network to train and identify the phasic dopamine release-2019 (PDR19) dataset into two-classes; namely, "release images, "or "non-release images" The experimental outcomes demonstrated the distinction of our identification scheme, recording an identification accuracy of 99.1% with a cross entropy loss of 0.022 attained after 25 epochs each with 100 iterations (i.e., 2500 iterations) for the 2-class classifier. Besides, our identification scheme was assessed using many other assessment factors, such as the identification precision percentage (IPP), the identification sensitivity percentage (ISnP), the identification specificity percentage (ISpP), and the identification weighted average percentage (F1P). Consequently, the performance of the proposed scheme surpassed several existing dopamine identification schemes.

***Keyword* — DarkNet-19, Phasic dopamine release, Identification, Convolutional Neural Network, Deep learning, Image Classification.**

I. Introduction

Dopamine is a neurotransmitter produced in both central and peripheral nervous systems. Its receptors are commonly uttered in the body and function in both systems. The dopaminergic system shows important roles in neuromodulation of motor control, incentive, reward, mental function, maternal, and reproductive behaviors [1]. As part of the reward system, the brain releases dopamine once doing things that humans or animals crave to it, contributing to feeling of pleasure and fulfillment. It creates reward seeking loops, lifts mood, and helps regulate movement, learning and emotional responses [1]. Opiates, on the other hand, affect the brain by increasing the amount of dopamine available to act on D2 receptors on the nucleus accumbens of the brain thus acutely stimulates the reward system. Long term addiction involves the development of tolerance, i.e., the need for increased amount of drug to produce a "high" phenomenon [2].

Abnormal conditions affect the dopaminergic systems and leads to pathologies in movement (Parkinson's disease), psychiatric (schizophrenia) and reward / motivation related disorders (drug addiction) [7]. As an example, defects in mesocortical, long dopamine system which projects from the midbrain tegmentam to limbic cortex, was responsible for the development of some of the symptoms of schizophrenia disorder and D4 receptors in the brain had been reported to be increased six-folds in this condition. Researchers also approved that there was a steady loss of dopamine receptors in the basal ganglia of the brain with age while the loss was greater in men than in women [2]

Understanding the role of neurotransmitter dopamine in brain function under normal or pathological states is one of the most active areas of research in neurosciences. Fast dopamine release, in less than a second period depending on crave stimuli, causes phasic release of dopamine which increases the activation of specific neurons to start a learned action. Failures in dopamine neurotransmission affects tremendous amount of brain abilities including movement, mental, and motivation and reward systems of the brain [3]. Ability to measure phasic release of dopamine in specific locations of the brain will lead to a powerful tool for the neuroscientists. However, the tremendous amount of image-formed data as produced from different locations of the brain makes the manual analysis of these data cumbersome. Luckily, image processing techniques will help in solving these problems effortlessly to ease and speed the analysis for neuro-physicians.

In this paper, we make use of DarkNet-19 Convolutional neural network (CNN) to identify the dopamine analysis imaging (DA) using Phasic Dopamine Release-2019 (PDR-19) dataset to assist at delivering an early identifying of the failures in dopamine neurotransmission. The PDR-19 dataset [10] comprises a set of 6030 images distributed into two main classes; namely, "release images, "or "Non- release images". Considering the obtained dataset, after applying a number of data wrangling operations, we employ the DarkNet-19 Convolutional neural network with customized input and output layers, frozen network parameters using transfer learning technique, and fine-configuration for the network hyper-parameters. We show the supremacy of our identification scheme scoring an outstanding classification

978-1-7281-9615-2/20/$31.00 ©2020 IEEE

accuracy. More specifically, the primary contributions of this work can be itemized as follows:

- We provide an accurate and precise identification model utilizing the powerful DarkNet-19 CNN with pre-learned trainable parameters (i.e., weights and biases) to train and validate the phasic dopamine release dataset (PDR2019) using our commodity machine equipped with NVIDIA Quadro GPU for enhanced computing.
- We provide an inclusive simulation results and investigation of our identification scheme to extend the comprehension of proposed methodology and experimental check. Besides, we benchmark our results with the recent related work to demonstrate the gain of the achieved performance trajectories.

The subsequent sections of this paper are arranged as follows: Section II investigate the dataset of phasic dopamine releases employed in this study. Section III describes and discusses the phasic dopamine identification system using DarkNet-19, experimental setup and configurations. Section IV provides details about the model evaluation, results and discussion. Finally, Section V concludes the paper.

II. Dataset of Phasic Dopamine Releases

In our research, we utilized the latest dataset introduced by Matsushita et al. [7]. This dataset is extremely comprehensive, and it is the only public dataset of fast-scan cyclic voltammetry (FSCV) images that we are aware of. The authors in [4] used FSCV data to create images depicting plots of 20-second-long experimental recordings. These recordings were made at the Federal University of Parana's (UFPR) Laboratory of Central Nervous System in Curitiba, Brazil, and D. Robinson's Laboratory at the University of North Carolina (UNC) in Chapel Hill, United States of America. The experiment was carried out by the two laboratories on different animals and FSCV configurations, resulting in slightly different plots. The experiment was carried out using a total number of 30 recordings. In the same experiment, the UFPR lab used 29 male Swiss mice while the UNC lab used six male Sprague Dawley rats. Moreover, and for legalization purposes, all experiments were carried out in conjunction with the National Institutes of Health's Guide for the Treatment and Use of Laboratory Animals, with procedures approved by the University of North Carolina's Institutional Animal Care and Use Committee and the Federal University of Parana's Institutional Ethics Committee for Animal Experimentation. For more information, a detailed description of the dataset is available in [10].

The basic idea of Matsushita et al. dataset configuration is that for each image, columns from the beginning, middle, and end were chosen, and the values of those columns were subtracted from the values of the other columns, yielding three distinct images distinguished by their backgrounds. There are a total of 1005 induced dopamine release images and 1005 images without dopamine release for each background. To better view the transformations, FSCV analysis software was used to apply a regular false color palette. Except for the UNC laboratory's experiment "RIX2," which includes all the mixed readings collected from the other experiments, each image is classified by its experiment.

Matsushita et al. dataset was extensively studied by Patarnello et al. [6]. The authors in [6] branched Matsushita et al. approaches into two groups: Global methods and Patch methods. The global methods include original photos as well as zoning variations. The first zoning method was applied on the y-axis of the image from pixel 320 to pixel 520, resulting image size of 875x200 pixels. It has been observed that the phasic Dopamine release episodes frequently established in this zone, hence it was named “common release region of the dopamine”. The second zoning method was applied on the first 90 pixels of the y-axis and the common dopamine release region, causing an 875x290 pixels image size. This region was called “concatenated zones” since some releases include visual data that can be used to abstract features. The global methods are graphically described in Fig. 1.

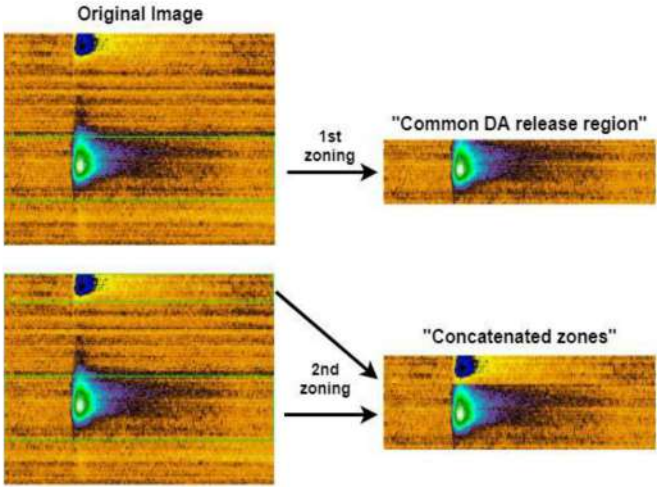


Fig. 1. Zoning variations and the original image [6].

Patch extraction methods, on the other hand, manually and automatically extract patches of size 200x200 from the common release area. The dataset includes a release peak position that used to remove manual patches, which is set as the image's core. Over the common release, a sliding window with a resolution of 200x200 pixels and a window slide of 135 pixels are used to extract automatic patches. The slide size is justified by the fact that it is a divider of the original image's width size; as a result, no pixels are left over, and six patches are removed per image. The authors in [6] recommended that size 290x290 patches be extracted from the concatenated zones in addition to the 200x200 patches. Although the method for applying manual patches hasn't changed, the fact that the image's width and height can easily be increased to 290 pixels, whereas the sliding process continues to use a 135-pixel slide. As a result, patches of size 200x290 are created by automated extraction, which are then padded to make them 290x290. Patches examples can be made clearer in Fig. 2.

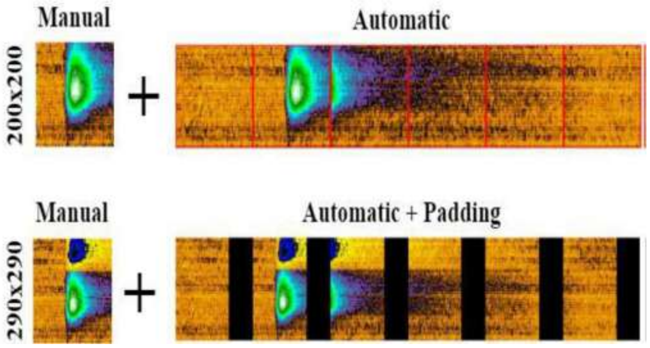


Fig. 2. Patches variations [6].

III. Identification Model Using DarkNet-19

To develop the proposed deep learning based identification system, we firstly collected the aforementioned recent dopamine analysis imaging introduced in 2019 [7]. Initially, the collected images were imported using Image-Data-Store component provided by the MATLAB's image processing objects. Thereafter, they undergone a set of preprocessing processes in order to be prepared for use by the deep learning model. Therefore, the preprocessing stage included (a) Image Resizing to unify the size of all images to $256 \times 256 \times 3$ all with JPG image extension, (b) Target Encoding to encode the two categories using binary as "0: release images" and "1: non-release images", (c) Shuffling operation to perform a stochastic re-distribution for images' dataset at every training epoch to ensure the unprejudiced distribution for the image categories, (d) Data Augmentation for dataset expansion using several image processing operations, as such, random image warping transformations, cropping transformations, color transformations, synthetic noise, synthetic blur and others [19], (e) Dataset Allocation to divide the dataset into training dataset and testing dataset using k-fold cross validation mechanism [20]. Afterward, the resulting preprocessed images can be fed to be trained and validated using the deep $DarkNet$ network, engaged in this research.

$DarkNet$ is a convolutional neural network, developed to enable the efficient training and validation of very deep neural network with more than 53+layers. Typically, $Darknet$ employs a number of global kernels and doubles the number of channels after every pooling layer [12]. The original version of $Darknet$ was written in C and CUDA, employing CPU and GPU computations, and its available as an open source framework [11]. $DarkNet$ is developed based on several preceding ideas such as $Network\ In\ Network$ [13], $Inception$ [14] and $Batch\ Normalization$ [15].

In this work, we have used the 19-convolutional layers' deep version of $DarkNet$, which is known as $DarkNet-19$. DarkNet-19 comprises exactly 19 convolutional layers and 5 max-pooling layers, involving only 3×3 and several 1×1 convolutional kernels to minimize the amount of trainable parameters [18]. We have employed the $DarkNet-19$ with the transfer learning techniques using a pretrained network that is trained on ImageNet dataset comprising a set of 1,000,000+ images [17], which provide the network with rich feature representations. While $DarkNet-19$ can be used to classify the images into 1000 classes, we have customized the output layer of $DarkNet-19$ to have two-classes to accommodate our identification outcomes. Also, since DarkNet-19 uses an input layer size of $256\ x\ 256$, we have resized all the images of the phasic dopamine releases dataset (PDR19) to be in $256\ x\ 256$ pixles to accommodate the input layer of the $DarkNet-19$. The simplified architecture of $Darknet-19$ neural network showing the main layers of the $DarkNet-19$ network is illustrated in Fig.3.

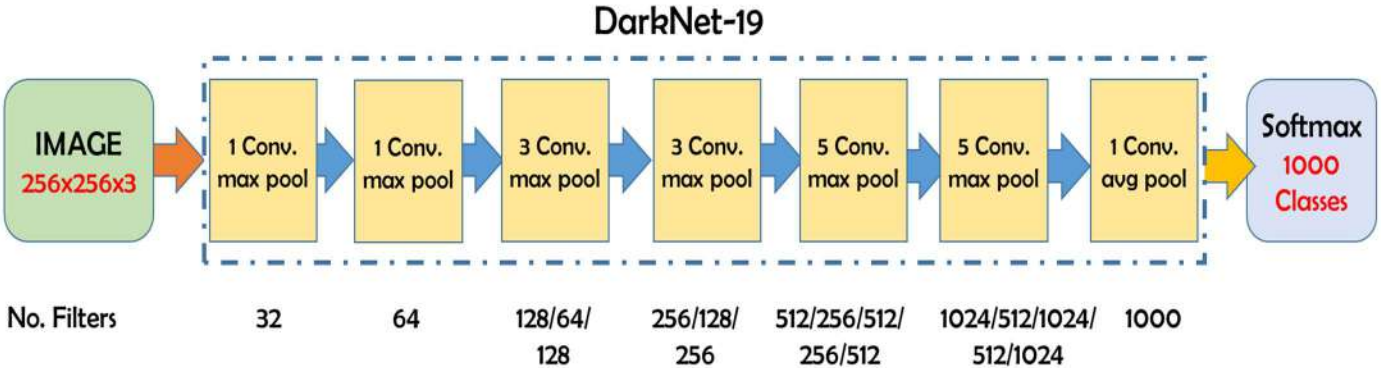


Fig. 3. Top-View Architecture of Darknet-19 neural network

Also, the prescribed identification scheme has been carried out using MATLAB computing system and its supplementary components including the image processing, the parallel computation and the deep learning toolboxes as well as the corresponding machine learning methods and approaches. Moreover, to train and test the performance of deep learning scheme, we have allocated 80% of the collected dopamine analysis imaging (i.e. PDR19) dataset for training process and 20% for the validation/testing process using random cross validation experiments. Besides, the rest of other system development are provided in Table 2, below.

TABLE. I. SUMMARY OF SYSTEM DEVELOPMENT CONFIGURATIONS

Development Items	**Description**
Training/Testing Model	Transfer Learning via DarkNet-19
Execution Environment	CPU + GPU/Auto
Input Image dimensions	256 x 256 x 3
Solver Approach	Stochastic Gradient Descent (SGD)
Loss Function Computation	Cross Entropy Loss Function
Binary-Classifier technique	Onevsall (one-vs-both classification)
Classification Learner	Linear learner algorithm
Validation Frequency	10-Fold Cross Validation
Max number of Epochs	25, with Shuffling at Every Epoch
No. of iterations per Epoch	100, accumulated 2500 iterations.
Mini Batch Size	32
Initial Learning Rate	0.01 with Momentum of 0.9
L2Regularization	0.0001

IV. EVALUATION AND DISCUSSION

A reliable auto-recognition deep-learning model to classify the weather condition images with high-level of classification accuracy, precision, and recall. To enhance the performance of feature extraction and learning, we have utilized the power of transfer learning technique with fine-tuning of the recognized deep ResNet-18 CNN pretrained on ImageNet dataset. The developed model uses the multi-class weather recognition dataset with 75% of the images used for training and 25% used for testing. Actually, the proposed work provides an inclusive framework model for multi-class image classification applications from input layer to the output layer. Finally, based on the comparison with other related research in the field, the obtained results outperform the results of existing automated classification models for weather conditions images.

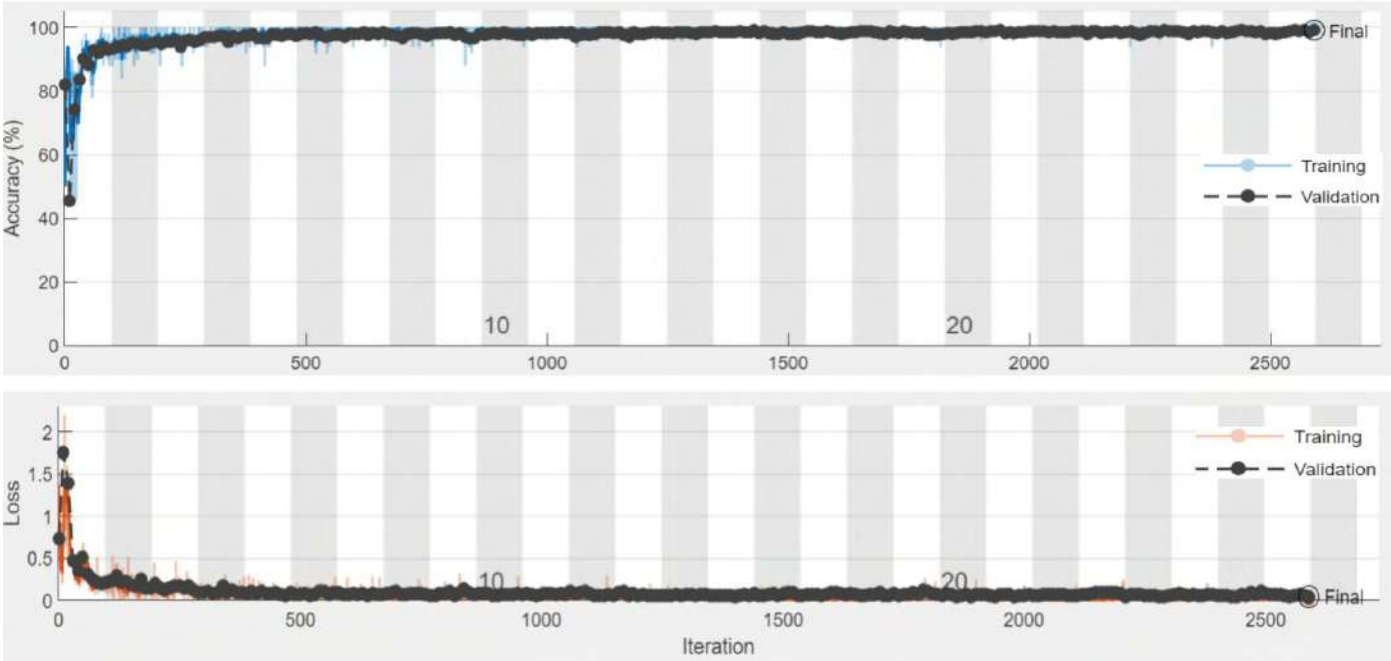


Fig. 4. Trajectories of {Identification Accuracy, Identification Loss} vs. Number of Epochs for training and testing datasets

The effectiveness of the PDR identification model can be measured using several performance assessment factors including the number of true positives samples (NTP), the number of true negatives samples (NTN), the number of false positives samples (NFP), the number of false negatives samples (NFN), the identification precision percentage (IPP), the identification sensitivity percentage (ISnP), the identification specificity percentage (ISpP), and the identification weighted average percentage (F1P). Hence, Table 1 shows summery of result obtained for the proposed PDR identification model.

TABLE II. SUMMARY OF EXPERIMENTAL RESULTS OBTAINED FOR THE AFORESAID EVALUATION METRICS

NTN	***NFP***	***NFN***	***NTP***	***IAC***	***IPP***	***ISnP***	***ISpP***	***F1P***	***IER***
605	6	5	590	99.10%	98.99%	99.16%	99.35%	99.08%	0.90%

Finally, to have additional perception into the benefits of our proposed solution, we have compared the performance measures of our identification system with the existing machine-learning-based dopamine release identification systems in terms of the identification accuracy factor. As a result, the benchmark evaluation revealed the preponderance of our proposed phasic dopamine release identification based DarkNet-19 model with greater identification accuracy using the PDR-19 dataset in comparison with other existing deep learning based models in the field by an improvement percentage (IM%) of ≈ (1.0% – 4.0%) and lower identification overhead.

TABLE III. COMPARISON WITH RELATED RESEARCH METHODS EMPLOYING SIMILAR DATA OF PHASIC DOPAMINE RELEASES.

Research Method		**Year**	**Classification Technique**	**Comments**	**Accuracy**
Russakovsky et. al.	[9]	2015	Inception v3 network + Roecker CNN	Hybrid, Complex design, Very high overhead	≈ 95.6%
Matsushita, et. al.	[7]	2019	Roecker CNN [5] + YOLOv3 Detectors	Hybrid, Larger dataset, with low harmonic mean	≈ 97.3%
L. Patarnello et. al.	[6]	2020	60-AlexNet CNN Networks + 3 YOLOv2 Detectors	Non-reasonable number of Combined CNNs, Complex design, Very high overhead	≈ 98.7 %
This Model		2021	DarkNet-19 Convolutional Neural Network	Employs Transfer learning from DarkNet-19, Low overhead	≈ 99.1 %

V. CONCLUSIONS

An automated intelligent identification model for dopamine analysis (DA) to classify the DA imaging for phasic dopamine releases has been developed and analyzed in this article. The model exploits of pretraind DrakNet-19 CNN with re-configuration of the simulation environment, the hyperparameters and the softmax layer (reduced to output 2-classes instead of 1000-classes). Also, the suggested model utilizes the images of dopamine analysis dataset, PDR2019, providing around 6,000 images, fairly balanced into two classes including “release images, “or “non-release images”. The images were distributed 80%: 20% for training and testing datasets respectively. Moreover, the identification model has been trained for 25 epochs each with 100 iterations scoring the best identification accuracy at 99.1% attained for the binary identification model. Furthermore, the identification model was validated through several other performance evaluation factors to achieve more perceptions of the model behavior identification precision percentage (IPP), the identification sensitivity percentage (ISnP), the identification specificity percentage (ISpP. In conclusion, the proposed identification model has beaten other available identification models for phasic dopamine releases in terms of identification accuracy and computation overhead.

REFERENCES

[1] M. Klein, D. Battagello, A. Cardoso, D. Hauser, “Dopamine: Functions, Signaling, and Association wih Neurological Diseases,” Cellular and molecular Neurobiolog, 2019.

[2] W. F. Ganong,” Review of Medical Physiology,” 19th Ediion, Appleon and lang Press. 1999.

[3] C. Gerfen, J. Sermeier, “Modulation of striatal projection systems by dopamine,” Annu. Rev. Neurosci. 2011. 34:441–66

[4] Q. A. Al-Haija, M. Smadi, S. Zein-Sabatto. Multi-Class Weather Classification Using ResNet-18 CNN for Autonomous IoT and CPS Applications. IEEE 7th Annual Conference on Computational Science & Computational Intelligence (CSCI'20), Las Vegas, USA. (2020).

[5] K.E. Koech, “Cross-Entropy Loss Function”, Medium: towards data science. (2020).

[6] L. Patarnello, M. Celin, L. Nanni. Phasic dopamine release identification using ensemble of AlexNet. arXiv:2006.02536. (2020).

[7] Matsushita, G., Sugi, A. H, Costa, Y., Gomez, A., Cunha, C., Oliveira, L. S., Automatic dopamine release identification using convolutional neural network, Computers in Biology and Medicine, 114, 2019

[8] Roecker, M. N., Costa, Y. M. G., Almeida, J. L. R., and Matsushita, G. H. G. (2018). Automatic vehicle type classification with convolutional neural networks. In 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP), pages 1–5. IEEE.

[9] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252

[10] https://web.inf.ufpr.br/vri/databases/phasicdopaminerelease/

[11] Joseph Redmon, Darknet-19: Open Source Neural Networks in C, http://pjreddie.com/darknet/, 2013-2016.

[12] V. Ghenescu, R. E. Mihaescu, S. Carata, M. T. Ghenescu, E. Barnoviciu and M. Chindea, "Face Detection and Recognition Based on General Purpose DNN Object Detector," 2018 International Symposium on Electronics and Telecommunications (ISETC), Timisoara, Romania, 2018, pp. 1-4, doi: 10.1109/ISETC.2018.8583861.

[13] Lin, M., Chen, Q., Yan, S.: Network In Network. arXiv:1312.4400 [cs] (2014).

[14] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE, Boston (2015). https://doi.org/10.1109/CVPR.2015.7298594

[15] Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs] (2015)

[16] Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. arXiv:1612.08242 [cs] (2016)

[17] ImageNet. http://www.image-net.org

[18] Benali Amjoud A., Amrouch M. (2020) Convolutional Neural Networks Backbones for Object Detection. In: El Moataz A., Mammass D., Mansouri A., Nouboud F. (eds) Image and Signal Processing. ICISP 2020. Lecture Notes in Computer Science, vol 12119. Springer, Cham. https://doi.org/10.1007/978-3-030-51935-3_30

[19] Suki Lau, Image Augmentation for Deep Learning, Medium: Towards Data Science, 2017.

[20] J. Brownlee, A Gentle Introduction to k-fold Cross-Validation, in Statistics, Machine Learning Mastery, 2018.

Manipulating GPS Signals to Determine the Launch Location of Drones in Rescue Mode

Hosam Alamleh
Department of Computer Science
University of North Carolina at Wilmington
Wilmington, NC
hosam.amleh@gmail.com

Nicholas Roy
Department of Computer Science
University of North Carolina at Wilmington
Wilmington, NC
ncroy7@gmail.com

***Abstract*—The use of drones for civilian purposes has recently grown. Drones are used for recreational, commercial, and governmental purposes. Today's drones can fly over a long range and many of them are equipped with cameras and a GPS chip to allow users to control them over a long range. This creates privacy and security concerns. For example, drones flying in restricted areas, or using the camera to spy on individuals and entities This paper addresses this issue and proposes a method to respond to violating drones by exploiting the lack of source authentication in GPS systems. In this paper, we propose a technique to ground and find the launch location of violating drones. This is done by invoking the rescue mode, then, employing Radio frequency software-defined radios to broadcast manipulated GPS signals following a novel algorithm that not only grounds violating drones but also find the drone's launch location, which is what unique about our approach. In this paper, we present the proposed method. Then, we perform both hardware tests and a simulation to demonstrate the proposed system's performance.**

***Index Terms*—drone, GPS, spoofing, bisection, algorithm**

I. Introduction

Recently there has been an increase in the use of drones by individuals. A drone is an aircraft without a human pilot on board but instead uses a ground-based controller. Drones are used by recreational flyers, enthusiasts, photographers, public safety, and governmental entities. The number of drones is on the increase from previous years. As of January 2021, there were 1,782,479 drones registered in the United States by the FAA. 71 percent of registrations were for recreational purposes. 27 percent of registrations were for commercial operation with 15 percent of those Americans having never flown a drone before [1]. The global drone market reached 22 billion dollars in 2020 with the majority of production being done in the United States and China[2]. Despite a large number of drones are airborne, privacy violations by drones are oftentimes overlooked. Modern-day drones are equipped with a camera and can fly incredibly far from where the user is piloting them. Drones can also legally fly above private property and photograph/record with little to no repercussions because they are treated as aircraft in the eyes of the law [3]. This can create what feels like an invasion of privacy for some who do not wish there to be a drone recording them or what is around them. On the other hand, drones create a security threat, as some drones can fly into restricted areas such as airports or sensitive facilities. An area where drones are not allowed to fly is known as " no drone zone". Despite the existence of some techniques that can ground violating drones, it is often difficult to catch the drone's pilot. Without any information on the pilot or where the drone came from, it allows for repeated offenses by individuals for violations.

Many of today's drones are equipped with a GPS chip. This allows the controller to track and manage drones over a long-range. GPS is the Global Positioning System, and is a global navigation satellite system that provides location, velocity, and time synchronization. It is used in products all across the world with the most common being cars. A utilization of GPS in drones is supporting the rescue mode or return to home functionalities. These functionalities enable the drone to return to the launch location if the video feed or the controller signal are lost. Despite, the functionalities enabled by GPS, GPS location can be exploited since there is no source authentication procedure in place to verify the source of GPS signals [4]. In other words, a GPS chip does not verify if the GPS signal came from a satellite or other malicious entities. GPS manipulation can be done by transmitting spoofed GPS signals using a radio frequency software-defined radio (RF-SDR). In this paper, we propose a method to ground violating GPS-enabled civilian drones in rescue mode and find the launch location of a violating drone. This is done by utilizing an RF-SDR to transmit manipulated GPS signals and using our algorithm to use the drone's response to the transmitted signals to find the launch location of the drone.

II. Background

Many modern-day solutions for dealing with such a problem would be physically taking the drone down using force, but there could be legal retaliation from the drones' pilot [5]. There is no one "correct" approach to solving the problem of eliminating a drone since there can be multiple solutions. One solution is to transmit jamming signals [6][7]. However, drones are equipped with rescue mode which enables the drone to return to the launch location making this only a short-term solution. Moreover, there have been techniques to detect jamming attacks on drones [8][9]. Other solution

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

includes sending a spoofed signal that overcomes the original controller signal [10]. However, such attacks can be overcome by utilizing cryptography to source authenticate the controller [11].Another method used exploits the optical sensors that are built into some types of drones. The reasoning for attacking the optical sensors was to make it appear that the drone was drifting and needed to remain steady using its built-in functions. Through projecting images onto the ground level, the drone cannot differentiate between what the actual surfaces are versus what is being generated [12]. This has more noticeable constraints in order to work as intended that have been noted by the creators of this study. The first is that the ground can be changed enough that the projected image seems natural to the drone, which heavily limits where this method would work. The next piece of information that must be known is how the drone will respond to the image, or more specifically the algorithm used. While information can be inferred about how the drone will avoid a potential crash, actions such as turns, accelerating, and magnitude changes can not be as easily predicted.

Other research work proposes targeting the network security of Wi-Fi-enabled drones [13][14]. These systems aim to break into the Wi-Fi system to hack into the drone and possibly control it. In fact, even extremely secure networking such as WPA2 can be broken through a third-party source by forcing a key reinstallation [15]. However, Wi-Fi attacks take a long time and might not be practical in case of a security violation. Other research works proposed spoofing GPS locations to take over unmanned aircraft vehicles (drones) and control their movements [16][17][18]. The spoofed GPS signals are used as an attack to confuse the original receiver and lock them out of locating the true coordinates While this is an efficient method to use, it has some drawbacks that are apparent with it. The first is that this gives no information about where the drone originated from, or who would have launched it. It is useful for the current time but still allows for more potential situations that would be similar in the future from the same location and launcher. As seen from the related work, none of the approaches attempt to determine the launch location of the drone, which is what unique about this work. The next section demonstrates the proposed method.

III. Method

In this section, we discuss the method used by the proposed system. In order to be able to determine where the launch occurred, the current location is used to evoke the rescue mode in the drone or GPS rescue mode in the case of a GPS-enabled drone. GPS-enabled rescue mode is intended to bring the drone autonomously in case of an emergency such as loss of video or radio link [19]. Therefore, rescue mode can be invoked by transmitting jamming signals on the frequency of the radio link or the video link. After the rescue mode is in action, the drone is going to return to the launch location at a constant speed and with bearing as shown in equation (1):

$$\theta = atan\frac{sin\Delta\lambda.cos\phi_2}{cos\phi_1.sin\phi_2 - sin\phi_1 cos\phi_2.cos\Delta\lambda} \tag{1}$$

Where $\Delta\lambda$ is the difference in longitude between the current drone's coordinates and the launch coordinates. ϕ_1 is the latitude of the current drone's coordinates. ϕ_2 is the latitude of the current drone's coordinates.

Consequently, the $x-$component and the $y-$component of the drone's speed is going to be:

$$\begin{aligned} x_{speed} &= A.cos\theta \\ y_{speed} &= A.sin\theta \end{aligned} \tag{2}$$

Where A is the constant speed the drone travels in when rescue mode is in action.

For equation (1), the current coordinates of the drone are known or can be obtained. The goal here is to find the coordinates of the launch location. In the proposed method, iterative launch coordinates longitude and latitude guesses are used to find the value of these coordinates. These values cause the $x-$component and the $y-$component in equation (2) to go to zero. In other words, the coordinate values that lead to that are the drone's launch coordinates. This is because in rescue mode the drone is going to stop moving once it arrives at launch coordinates. In the proposed method, the coordinates guesses are broadcasted by using a manipulated GPS signal through an RF-SDR. This is possible because unlike smartphones [20] a drone's GPS chip does not verify the authenticity of GPS signals it receives. Once the Drone receives the manipulated GPS signals, it recalculates its current GPS location and adjusts the travel bearing accordingly as in equation (2). This change in the bearing is going to be observed and is used to calculate the next guess. This process continues until the $x-$component and the $y-$component of the drone speed become zero. Once the components are discovered that means so are the launch coordinates.

It is important to minimize the number of coordinates guesses in the algorithm to decrease the time required for the proposed system to retrieve the drone's launch location. Therefore, an efficient root-finding method is required. In this paper, the bisection method is employed. The bisection method is a root-finding method that repeatedly bisects the interval defined by these values and then selecting the sub-interval in which the function changes sign, and therefore must contain a root. [21]. In the proposed method, the bisection method is applied to find the root of both equations in (2). The interval is determined by defining the maximum range from which the drone is launched from. Usually, civilian drones have a limited range measured in miles. This is because the distance the drone controller's signal can travel is limited as the radio signals attenuate with distance. These limits are going to be the initial interval for the bisection method. Then the interval is updated based on how the drone changes the travel bearing once it receives the manipulated GPS signal. A step by step explanation of the proposed is as follows:

1) The rescue mode is invoked by transmitting jamming signals either on the drone's controller frequency or the video feed frequency.

2) Once rescue mode is in action, the drone is going to return to its launch location, using a constant speed A and bearing θ as shown in (1).
3) The maximum longitude and latitude and the minimum Longitude and latitude for the launch location are determined based on the current location and the maximum drone's range.
4) The bisection algorithm starts with the interval [Minimum longitude, Maximum longitude] for the $x-$component of the speed and the interval [Minimum latitude, Maximum latitude] for the $y-$component of the speed. It is known that there is a root for equations in (2) in these intervals.
5) Manipulated GPS signals are transmitted according to the bisection algorithm according to the intervals above for first coordinates guess.
6) Once the drone receives the manipulated signals, it recalculates its current GPS location. This changes current longitude and current latitude. Consequently, a new θ is calculated as in equation (1) impacting the $x-$component and the $y-$component of speed in (2).
7) Based on the change of the values of the $x-$component and the $y-$component of speed (whether they increase or decrease). The bisection interval is adjusted. Then, new coordinates guess is transmitted.
8) Bisection intervals are adjusted according to the changes in the $x-$component and the $y-$component of speed. Based on the new interval, the bisection algorithm generates new guesses. This process continues until the $x-$component and the $y-$ are zero or very close to 0. The coordinates guess that leads to zeroing the $x-$component and the $y-$component of speed are the drone's launch coordinates.

As seen from the steps above, several guesses can lead to finding the drone's launch coordinates. Moreover, the correct guess causes the drone to land. As at this point the drone is deceived that it is at the launch location. The next section shows the experiments conducted to inspect the proposed method's performance.

IV. Experiments and Results

In this section, we present the experiments that were performed to analyze the proposed method's performance. Then, the experiment's results are analyzed and discussed. In this paper, we performed two experiments. The first, experiment inspects hardware performance for the RF-SDR that transmits manipulated GPS performance and for a drone's chipset. The second is a simulation to measure the proposed system's performance in determining a drone's launch coordinates.

A. Hardware Performance

The purpose of this experiment is to show a proof of concept on how manipulated GPS signals can lead the drone to update its rescue mode travel bearings. In this experiment, The RF-SDR used is the HackRF[22]. This RF-SDR was used to transmit manipulated GPS signals[23]. An Arduino chipset with a GPS-chip was used for the drone. Arduino GPS drone firmware was installed[24] on the Arduino Chipset. Arduino data was then logged into an SD. The logged data shows the Arduino calculating a cheated GPS location and adjusting bearing as shown in the method. This experiment was performed on the drone's chipset only (not a full drone). Moreover, the hack RF antenna and the chipset were in close proximity to allow transmitting the manipulated GPS signals at low power to avoid impacting the surroundings.

B. Algorithm Performance

The purpose of this experiment is to simulate the performance of the proposed method to find a drone's launch coordinates. The drone movement in rescue mode is simulated by writing a python function that takes transmitted manipulated GPS signal as an input and returns the bearing of the travel. Then, a second function was written that simulates the proposed method. First, it determines the bisection interval, for simulation purposes, the maximum range of the drone was considered 1 mile. Accordingly, the maximum and minimum longitude was determined using Haversine formula[25]. In the second function, the bisection algorithm calculates what GPS coordinates to be broadcast by the RF-SDR. Consequently, the transmitted coordinates are plugged into the drone rescue mode python function which returns the updated bearing. Then, the $x-$component and the $y-$component of the speed are calculated, and based on these changes the next bisection interval is calculated and transmitted. This algorithm iterates until the difference between the guessed launch coordinates and the actual launch coordinates is less than an error of 1 percent. Since the bisection method is a root finding algorithm, the run-time for the implemented function is $O(log_2 N)$, where N is the range of the drone. Fig. 1 shows how the algorithm closes into the launch coordinates as it iterates. Fig. 2. shows how the $x-$component and the $y-$component go to zero as the algorithm iterates. As can be seen from Fig. 1, the algorithm closes in the launch location as it iterates. In other words, it calculates the possible launch location in every iteration, with the accuracy of the calculation increasing as it iterates. This is important as it can allow law enforcement to provide a faster response to violators. As can be seen in Fig. 2, in the simulation, it took seven guesses to get the $x-$component and the $y-$component of the speed close to zero. An error of 1 percent was tolerated to a decrease in the number of iteration. The proposed algorithm still finds the launch coordinates in high accuracy even with this error included, as it gets the launch area determined within inches. Moreover, it is possible to get higher accuracy with more iterations but it would come at the cost of using more time which may not always be needed.

V. Conclusion

The drone market is expanding exponentially each year with more civilians purchasing drones. With the lack of regulations that monitor malicious drone-flyers behavior, privacy and security concerns are raised. Such violations can be in the form

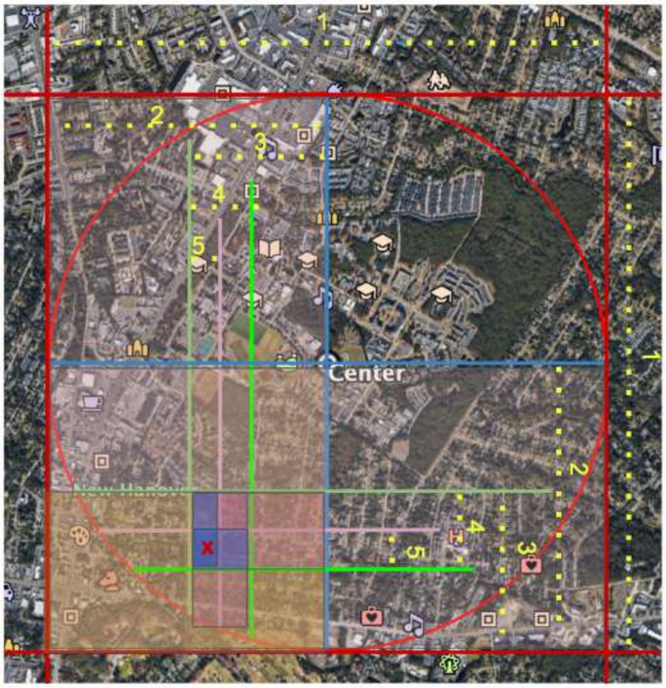


Fig. 1. Closing into the launch location as the algorithm iterates

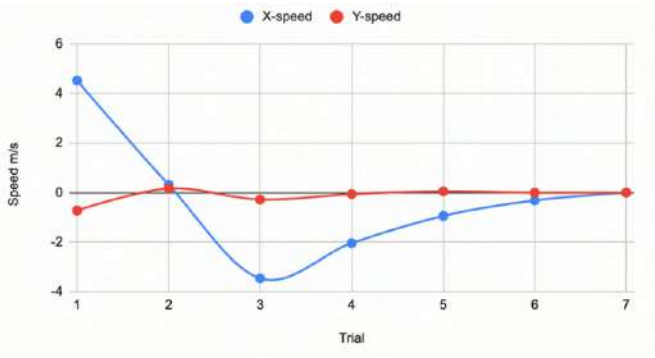


Fig. 2. The $x-$component and the $y-$component of the drone's speed going to 0 as the bisection algorithm iterates

of flying into restricted areas or violating people's privacy. This paper addresses this issue by proposing a scheme to counter violating drones. The proposed approach exploits the lack of source authentication in GPS systems and utilizes RF-SDR to transmit manipulated GPS signals according to the proposed algorithm to ground drones in rescue mode. Moreover, the proposed algorithm finds the drone's launch location, which is what is unique about our approach. In this paper, we explained the proposed algorithm's operation step-by-step, then ran hardware and simulation tests to validate the proposed system's performance. Future work will include working to minimize the number of guesses by the algorithm. Moreover, it will include testing the proposed method on flying drones by utilizing RF-SDR and directional antennas.

References

[1] Phillybyair. 2021. Ultimate List of Drone Stats for 2021 retreived from https://www.phillybyair.com/blog/drone-stats/

[2] DroneII. 2021. The Drone Market Size 2020-2025: 5 Key Takeaways. retreived from https://droneii.com/the-drone-market-size-2020-2025-5-key-takeaways

[3] Federal Aviation Administration. 2016. Federal Aviation Administration Part 107 guidelines.

[4] K. Wesson, M. Rothlisberger, and T. Humphreys(2011). Practical cryptographic civil GPS signal authentication.

[5] Electronic Code of Federal Regulations. 2021. Title 14 Aeronautics and Space retreived from https://www.ecfr.gov/ecfrbrowse/Title14/

[6] Y. Yang, K. Li, J. Li, H. Zhu, Y. Zhang and K. Huang, "Low-Cost, High-Power Jamming Transmitter Based on Magnetron," in IEEE Transactions on Electron Devices, vol. 67, no. 7, pp. 2912-2918, July 2020, doi: 10.1109/TED.2020.2992980.

[7] W. -C. Jin, K. Kim and J. -W. Choi, "Robust Jamming Algorithm for Location-Based UAV Jamming System," 2019 IEEE Asia-Pacific Microwave Conference (APMC), Singapore, 2019, pp. 1581-1583, doi: 10.1109/APMC46564.2019.9038440.

[8] M. Sliti, W. Abdallah and N. Boudriga, "Jamming Attack Detection in Optical UAV Networks," 2018 20th International Conference on Transparent Optical Networks (ICTON), Bucharest, Romania, 2018, pp. 1-5, doi: 10.1109/ICTON.2018.8473921.

[9] J. Farlik, M. Kratky and J. Casar, "Detectability and jamming of small UAVs by commercially available low-cost means," 2016 International Conference on Communications (COMM), Bucharest, Romania, 2016, pp. 327-330, doi: 10.1109/ICComm.2016.7528287.

[10] M. Donatti, F. Frazatto, L. Manera, T. Teramoto and E. Neger, "Radio frequency spoofing system to take over law-breaking drones," 2016 IEEE MTT-S Latin America Microwave Conference (LAMC), Puerto Vallarta, Mexico, 2016, pp. 1-3, doi: 10.1109/LAMC.2016.7851290.

[11] J. H. Cheon et al., "Toward a Secure Drone System: Flying With Real-Time Homomorphic Authenticated Encryption," in IEEE Access, vol. 6, pp. 24325-24339, 2018, doi: 10.1109/ACCESS.2018.2819189.

[12] Davidson D, Wu H, Jellinek R, et al. (2016) Controlling UAVs with sensor input spoofing attacks[C]// Usenix Conference on Offensive Technologies. USENIX Association

[13] J. Gordon, V. Kraj, J. H. Hwang and A. Raja, "A Security Assessment for Consumer WiFi Drones," 2019 IEEE International Conference on Industrial Internet (ICII), Orlando, FL, USA, 2019, pp. 1-5, doi: 10.1109/ICII.2019.00011.

[14] E. Vattapparamban, İ. Güvenç, A. İ. Yurekli, K. Akkaya and S. Uluağaç, "Drones for smart cities: Issues in cybersecurity, privacy, and public safety," 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, Cyprus, 2016, pp. 216-221, doi: 10.1109/IWCMC.2016.7577060.

[15] Mathy Vanhoef and Frank Piessens. 2017. Key Reinstallation Attacks: Forcing Nonce Reuse in WPA2. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17). Association for Computing Machinery, New York, NY, USA, 1313–1328

[16] Andrew J. Kerns, Daniel P. Shepard, Jahshan A. Bhatti, and Todd E. Humphreys. 2014. Unmanned aircraft capture and control via GPS spoofing. J. Field Robot. 31, 4 (2014)

[17] J. Gaspar, R. Ferreira, P. Sebastião and N. Souto, "Capture of UAVs Through GPS Spoofing," 2018 Global Wireless Summit (GWS), Chiang Rai, Thailand, 2018, pp. 21-26, doi: 10.1109/GWS.2018.8686727.

[18] S. P. Arteaga, L. A. M. Hernández, G. S. Pérez, A. L. S. Orozco and L. J. G. Villalba, "Analysis of the GPS Spoofing Vulnerability in the Drone 3DR Solo," in IEEE Access, vol. 7, pp. 51782-51789, 2019, doi: 10.1109/ACCESS.2019.2911526.

[19] Betaflight. 2020. GPS Rescue Mode. retreived from https://github.com/betaflight/betaflight/wiki/GPS-rescue-mode

[20] H. Alamleh and A. A. S. AlQahtani, "A Cheat-proof System To Validate GPS Location Data," 2020 IEEE International Conference on Electro Information Technology (EIT), Chicago, IL, USA, 2020, pp. 190-193, doi: 10.1109/EIT48999.2020.9208243.

[21] R. Burden, D. Faires. 1985, "2.1 The Bisection Algorithm", Numerical Analysis (3rd ed.), PWS Publishers, ISBN 0-87150-857-5

[22] Great Scott gadgets. HackRF Technical information. retrieved from https://greatscottgadgets.com/hackrf/

[23] Osqzss. 2015.GPS-SDR-SIM. Retreived from https://github.com/osqzss/gps-sdr-sim

[24] Elementguy. Arduino Drone With GPS. retreived from https://www.instructables.com/Arduino-Drone-With-GPS/

[25] V. Brummelen, G. Robert. 2013. Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry. Princeton University Press. ISBN 9780691148922. 0691148929.

Mitigating Remote Code Execution Vulnerabilities: A Study on Tomcat and Android Security Updates

Stephen Bier[1], Brian Fajardo[2], Obinna Ezeadum[3], German Guzman[4], Kazi Zakia Sultana[5], Vaibhav Anu[6]
Department of Computer Science
Montclair State University
Montclair, New Jersey, USA
{biers1[1], fajardob1[2], ezeadumo1[3], guzmang3[4], sultanak[5], anuv[6]}@montclair.edu

***Abstract*—The security of web-applications has become increasingly important in recent years as their popularity has grown exponentially. More and more web-based enterprise applications deal with sensitive personal and private information, which, if compromised, can not only lead to system downtime, but can also cause mean millions of dollars in damages to the organization. It is critical to protect web-applications from the constant onslaught of hacker attacks. Remote Code Execution (RCE) attacks are one of the most prominent security threats for software systems, especially Java-based systems. In the current study, we have studied the security update reports for RCE vulnerabilities published by two Java-based projects: Apache Tomcat and Android. We analyzed and categorized the code-fixes (i.e., patches/updates) that were applied to mitigate/fix fifty-one (51) RCE vulnerabilities in the two above-mentioned Java projects. Our analysis showed that a significant majority of the RCE vulnerabilities found in Java projects can be mitigated with just five (5) types/categories of code-fixes. Overall, our goal was to study RCE vulnerabilities in an effort to provide programmers with a handy list of code-fixes, thus making it easier for them to effectively mitigate known RCE vulnerabilities in their own Java-based applications.**

Keywords—software security, software engineering, vulnerabilities, remote code execution, open source software

I. Introduction

Software security bugs, also known as vulnerabilities, continue to be an important and potentially the most expensive issue affecting all aspects of our cyber society. There has been significant research effort toward preventing vulnerabilities from occurring in the first place, as well as toward automatically discovering vulnerabilities, but so far these results remain fairly limited [15, 16].

Remote Code Execution (RCE) has been recognized as one of the most harmful threats for web applications [1]. Although RCE is a special kind of cross-site scripting attack, RCE attacks have some variants including requiring state consideration of both server and client, both string and non-string manipulation of client inputs, and involving multiple requests to more than one server-side scripts [1]. Static analysis tools can be potentially used for detecting vulnerabilities [17, 18, 19]. Static code analysis tools locate vulnerabilities within source code using data flow analysis or taint analysis techniques [20]. As RCE attacks mostly depend on path conditions and involve both string and non-string operations, most static analysis tools fail to detect RCE attacks as they follow context free grammar and model only string operations [21, 22]. As a result, the false positive rates are high in those tools. On the other hand, the existing literatures do not focus on how RCE vulnerabilities have been resolved in real world applications so that developers can have a handy list of techniques to mitigate those problems.

In this paper, we focus on identifying the various ways developers (i.e., programmers) mitigate/fix RCE vulnerabilities that are reported in Java-based software systems (fixing a reported or known security vulnerability is more commonly referred to as security update or patch and is generally done by changing/adding/deleting lines of code).

In a sense, the primary objective of this study is to identify the most common types of code changes (i.e., updates/patches) that are applied by programmers when RCE vulnerabilities are reported in their software. To meet our objective, we reviewed different systems that publish security update reports and determine if there are any similarities in the RCE vulnerabilities and the updates that were implemented to fix them. In our research, we reviewed security updates reports for Apache Tomcat and Android. The major contributions of our work have been stated below:

1. The study is conducted on two major Java-based systems: Apache Tomcat and Android. The study has identified the most frequently used mitigation techniques for fixing RCE vulnerabilities that can be exemplary for Java based software developers.
2. We anticipate the findings in this study may be of assistance to the developers in avoiding frequent programming mistakes that can lead to RCE attacks.
3. The common security updates discussed in this paper will help the developers to mitigate (or fix) RCE issues and thus reduce the likelihood of RCE attacks in the future.

In Section II, we discuss some related works to our study. Section III describes our research methodology. Section IV focuses on data analysis and results. In Section V, we discuss the limitations of our work. Section VI provides a brief discussion on the implications of our findings and finally Section VII concludes the paper with some future plan.

II. Related Work

In this section, we highlight some existing works that focused on remote code execution (RCE) vulnerability analysis and detection.

Remote Code Execution is considered as a special kind of Cross Site Scripting (XSS) attacks [1]. Like XSS and SQL injection attacks, RCE occurs when invalid client-side inputs

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

are undesirably converted to scripts and executed [1]. Although researchers have already put significant efforts on identifying and mitigating XSS and SQL injections vulnerabilities [2-7], RCE vulnerability got very little attention due to its unique characteristics [1]. Zheng et al. [1] proposed a path and context sensitive inter procedural static analysis to detect RCE vulnerabilities in PHP scripts. They devised a novel algorithm featuring both string and non-string behavior of a program and successfully could detect RCE vulnerabilities in PHP scripts with less false positive rates [1]. In another study [8], the authors assessed the multi-variant code execution technique to prevent the execution of malicious code. The idea of multi-variant code execution is detecting any malicious attempt during run-time. While running two or more slightly different variants of the same program in lockstep on a multiprocessor, the variants are monitored and any divergence from the regular behavior raises an alarm indicating the possible anomaly. The trade-off between security and performance is the major limitation of this approach [8]. Hannes et al. [9] studied expert opinions on how three variable (i) non-executable memory, (ii) access and (iii) exploits for High or Medium vulnerabilities as defined by the Common Vulnerability Scoring System contribute to the successful remote code execution attacks. Both access and the severity of the exploited vulnerability were perceived as important by the experts; non-executable memory was not seen as relevant to RCE according to the study [9]. In [10], the authors presented a case study on RCE vulnerability and analyzed different types of RCE and their impact on applications. Another paper [13] proposed a new mechanism for trusted code remote execution. The method creates a trusted platform integrating the identity authentication, platform authentication and behavior authentication based on trusted computing technology, remote attestation and trusted behavior for remote code execution [13]. There are some other research studies on remote code execution which focused on remote code execution vulnerabilities in specific domains or platforms [11, 12, 14].

Overall, there is a shortage of research on modeling of discovered security vulnerabilities to capture how and why an implementation fails to achieve the desired level of security. This paper analyzes some real vulnerable code and their fixes so that programmers can be aware of those frequently happened programming mistakes and are aware of their possible mitigation techniques.

More specifically, we focused on Java-based applications and identified common code changes/fixes that are used to mitigate RCE vulnerabilities (which has not been investigated in earlier research). Most of the previous studies either devised techniques to prevent RCE or detect RCE during runtime. Those studies lack in highlighting some common programming practices that are used by developers to fix RCE vulnerabilities. In our study, we present RCE updates/fixes so that developers can be guided during the maintenance phase and can ensure future software releases are secure.

III. Methodology

This section describes the Research Questions (RQs) and the data collection process for this study.

A. Research Questions

The following research questions were formulated to guide the data collection for this study:

RQ1: Do software systems suffer from Remote Code Execution (RCE) vulnerabilities more frequently when compared to the other types of security vulnerabilities?

RQ2: What types of patches (i.e., code-fixes) are usually added to mitigate the known RCE vulnerabilities in Java-based software systems?

B. Data Collection

The following paragraphs describe the data that we collected to answer the two research questions (RQs).

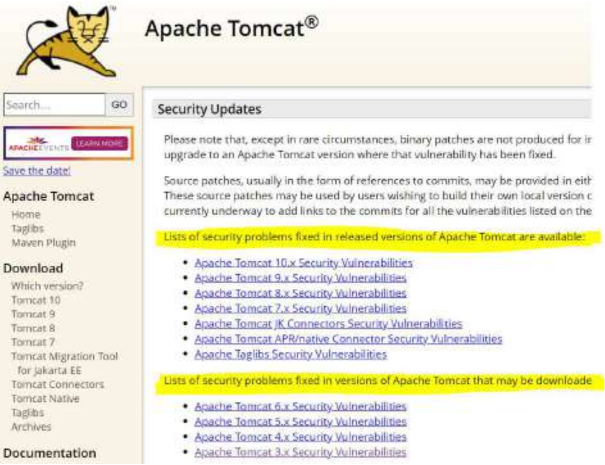
To answer *RQ1*, we collected the vulnerability-counts for the last 5 years (2015 to 2019) for the most common types of vulnerabilities (RCE, Denial of Service, Overflow, XSS, SQL injection) reported to a vulnerability datasource called CVE Details (https://www.cvedetails.com/).

With respect to *RQ2*, we focused specifically on collecting information about Remote Code Execution (RCE) vulnerabilities reported in Java-based software projects. Furthermore, we wanted to collect information about *how programmers fix the RCE vulnerabilities* reported in their Java-based software systems. Many open-source software projects publish *security update reports* on their project websites (for example, the security reports for Apache Tomcat are publicly available and can be found here: http://tomcat.apache.org/security.html).

We identified two such open-source Java-based software projects: ***Apache Tomcat*** (mentioned above) and ***Android*** (https://source.android.com/security/bulletin). An overview of the steps taken to collect the code-fixes (i.e., patched) applied to fix RCE vulnerabilities reported in each of the two systems is provided below:

1) Apache Tomcat Data Collection

To gather the data from Tomcat, we first went to the Tomcat's Security Reports page: http://tomcat.apache.org/security.html. This page displays a list of the known security vulnerabilities for each version of Tomcat as illustrated in Fig. 1. Tomcat Release 3.x was selected and all the Remote Code Execution (RCE) updates within this release were identified and reviewed. This was the most time and effort intensive step of our data collection process. Please note that our goal was to collect information regarding what kind of updates (i.e., patches or code-fixes) are applied by the programmers to known RCE vulnerabilities.



Security Updates

Please note that, except in rare circumstances, binary patches are not produced for ir upgrade to an Apache Tomcat version where that vulnerability has been fixed.

Source patches, usually in the form of references to commits, may be provided in eitl These source patches may be used by users wishing to build their own local version c currently underway to add links to the commits for all the vulnerabilities listed on the

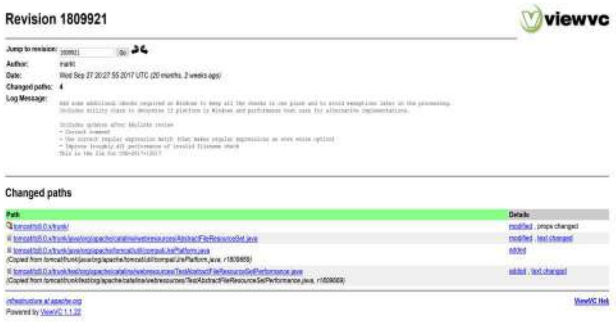
Lists of security problems fixed in released versions of Apache Tomcat are available:

- Apache Tomcat 10.x Security Vulnerabilities
- Apache Tomcat 9.x Security Vulnerabilities
- Apache Tomcat 8.x Security Vulnerabilities
- Apache Tomcat 7.x Security Vulnerabilities
- Apache Tomcat JK Connectors Security Vulnerabilities
- Apache Tomcat APR/native Connector Security Vulnerabilities
- Apache Taglibs Security Vulnerabilities

Lists of security problems fixed in versions of Apache Tomcat that may be downloade

- Apache Tomcat 6.x Security Vulnerabilities
- Apache Tomcat 5.x Security Vulnerabilities
- Apache Tomcat 4.x Security Vulnerabilities
- Apache Tomcat 3.x Security Vulnerabilities

Fig.1. Tomcat: Security Updates



Revision 1809921

Jump to revision: 1809921 Go

Author: markt

Date: Wed Sep 27 20:27:55 2017 UTC (20 months, 2 weeks ago)

Changed paths: 4

Log Message:

```
Add some additional checks required on Windows to keep all the checks in one place and to avoid exceptions later in the processing.
Includes utility class to determine if platform is Windows and performance test case for alternative implementations.

Includes updates after kkolinko review
- Correct comment
- Use correct regular expression match that makes regular expressions as even more optimal
- Improve (roughly x2) performance of invalid filename check
This is the fix for CVE-2017-12617
```

Changed paths

Path	Details
tomcat/tc8.0.x/trunk/	modified , props changed
tomcat/tc8.0.x/trunk/java/org/apache/catalina/webresources/AbstractFileResourceSet.java	modified , text changed
tomcat/tc8.0.x/trunk/java/org/apache/tomcat/util/compat/JrePlatform.java (Copied from tomcat/trunk/java/org/apache/tomcat/util/compat/JrePlatform.java, r1809888)	added
tomcat/tc8.0.x/trunk/test/org/apache/catalina/webresources/TestAbstractFileResourceSetPerformance.java (Copied from tomcat/trunk/test/org/apache/catalina/webresources/TestAbstractFileResourceSetPerformance.java, r1809909)	added , text changed

infrastructure at apache.org

Powered by ViewVC 1.1.22

Fig.2. Tomcat: Security Update Details for a Sample RCE Vulnerability found in Tomcat

Therefore, for each RCE update found during our search, the revision number was selected as presented in the information page for that particular revision. Fig. 2 contains the information page for Revision 1809921. As can be seen in Fig. 2, each revision page contains a "path changed" link which in turn contains the line(s) of code that were added and/or removed to fix/mitigate the reported RCE vulnerability.

Table I presents a sample set of RCE vulnerabilities in Apache Tomcat that were identified and investigated during this study.

TABLE I. TOMCAT: SAMPLE SET OF RCE VULNERABILITY PATCHES/UPDATES THAT WERE FOUND AND STUDIED

Common Vulnerabilities and Exposures No. (CVE No.)	Affected Versions	Fixed version
2013-4444	7	7.0.39
2016-8735	7, 8, 9	9.0.0.M12
2017-12615	7	7.0.80
2017-12617	7, 8, 9	9.0.0.M15
2019-0232	7, 8, 9	9.0.17

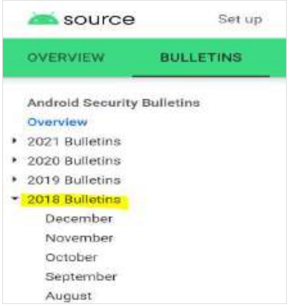
2) Android Data Collection

Similar to Tomcat, for Android we reviewed the Security Bulletins page of the Android website (https://source.android.com/security/bulletin). From here we selected a year and month from the dropdown on the left side of the webpage as shown in Fig. 3. Next, we searched for RCE updates (see Fig. 4) and clicked on the selected reference number which brought up the information page for that specific update. The information page for a sample RCE update is shown in Fig. 5. By selecting the "diff" link (highlighted in yellow in Fig. 5), a page containing the code that was added and/or changed to fix/mitigate the reported RCE vulnerability was displayed.

Following the data collection process described above, we collected a total of fifty-one (51) RCE updates/patches (including both the systems, Tomcat and Android). We analyzed these patches to understand if there are certain frequently used patterns in these RCE updates/patches. The data analysis conducted using the above-mentioned 51 RCE updates/patches is presented in Section IV. B.

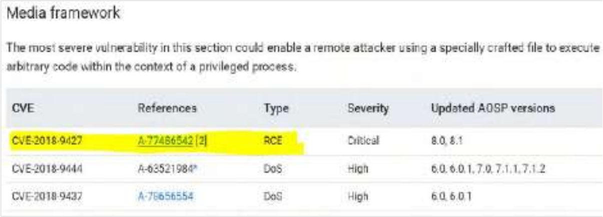
IV. DATA ANALYSIS AND RESULTS

This section presents the results and findings obtained from analyzing the data collected during this study. This section is organized around the two research questions (RQs) that were described in Section III.A.



Android Security Bulletins
- Overview
- 2021 Bulletins
- 2020 Bulletins
- 2019 Bulletins
- 2018 Bulletins
  - December
  - November
  - October
  - September
  - August

Fig.3. Android: Security Update Reports (Listed by Year)

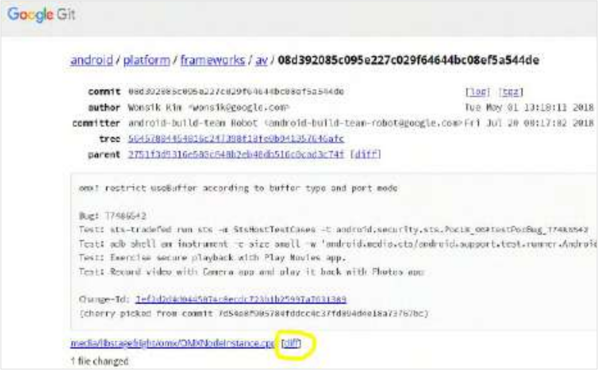


Media framework

The most severe vulnerability in this section could enable a remote attacker using a specially crafted file to execute arbitrary code within the context of a privileged process.

CVE	References	Type	Severity	Updated AOSP versions
CVE-2018-9427	A-77486542 [2]	RCE	Critical	8.0, 8.1
CVE-2018-9444	A-63521984*	DoS	High	6.0, 6.0.1, 7.0, 7.1.1, 7.1.2
CVE-2018-9437	A-78656554	DoS	High	6.0, 6.0.1

Fig.4. Android: Identifying RCE Vulnerabilities and their Respective Updates/Patches



android / platform / frameworks / av / 08d392085c095e227c029f64644bc08ef5a544de

```
   commit 08d392085c095e227c029f64644bc08ef5a544de                         [log] [tgz]
   author Wonsik Kim <wonsik@google.com>                                   Tue May 01 13:18:11 2018
committer android-build-team Robot <android-build-team-robot@google.com>  Fri Jul 20 08:17:02 2018
     tree 56457884454816c247398f18fe0b941357646afc
   parent 2751f3d9316e583c648b2eb48db516c0cad3c74f [diff]
```

```
omx: restrict useBuffer according to buffer type and port mode

Bug: 77486542
Test: sts-tradefed run sts -m StsHostTestCases -t android.security.sts.Poc18_06#testPocBug_77486542
Test: adb shell am instrument -e size small -w 'android.media.cts/android.support.test.runner.Androi
Test: Exercise secure playback with Play Movies app.
Test: Record video with Camera app and play it back with Photos app

Change-Id: Ief2d2d4d04458T4c8ecdcT23b1b25997a7031389
(cherry picked from commit 7d54e8f905784fddcc4c37fd894d4e18a73767bc)
```

media/libstagefright/omx/OMXNodeInstance.cpp [diff]

1 file changed

Fig.5. Android: Security Update Details for a Sample RCE Vulnerability found in Android

A. *RQ1: Do software systems suffer from Remote Code Execution (RCE) vulnerabilities more frequently when compared to the other types of security vulnerabilities?*

As mentioned before, we hypothesized that RCE vulnerabilities are the most frequently found vulnerabilities in software systems. In order to evaluate our hypothesis, we collected the vulnerability-count data from the "CVEDetails.com" datasource. This datasource receives its vulnerability data through National Vulnerability Database (NVD) xml feeds provided by NIST (National Institute of Standards and Technology). We collected the vulnerability-count data for the top-5 vulnerability types for the recent five (5) years, i.e., from the year 2015 to 2019 (please note that currently the CVEDetails datasource has vulnerability-count data till the year 2019).

Fig. 6 provides an overview of the data analyzed for RQ1. As can be seen in Fig. 6, RCE vulnerabilities were reported more frequently than other types of vulnerabilities during the years 2015, 2018, and 2019. Even during the other years (2016 and 2017), RCE vulnerabilities remained in the top-2 most reported vulnerabilities.

Furthermore, in the year 2019, the count of reported RCE vulnerabilities (2277) was significantly higher than the count of next most reported vulnerability (1593 Cross-Site Scripting

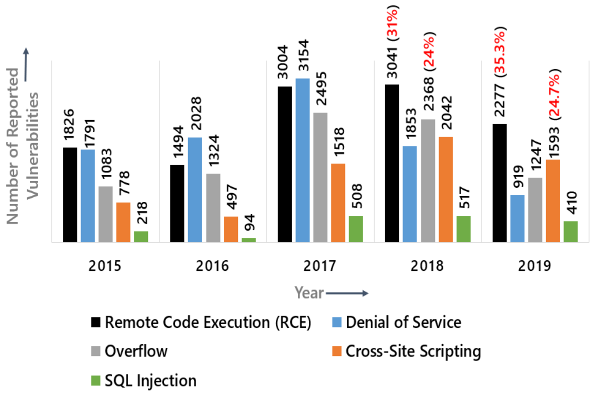


Fig. 6. Vulnerability-count by Type (for five recent years)

vulnerabilities were reported). Therefore, 35.3% of all the vulnerabilities reported during the year 2019 were of the type RCE (as can be seen in Fig. 6).

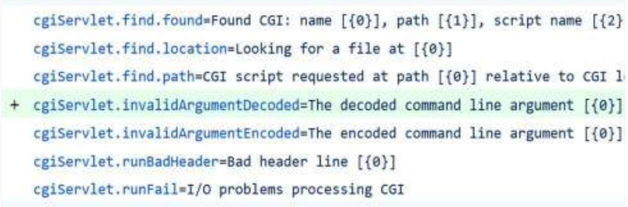
The data analysis described above and displayed in Fig. 6 clearly shows that software systems often suffer from RCE vulnerabilities more frequently than the other types of vulnerabilities. This in turn leads to a frequent need for programmers to fix RCE vulnerabilities through adding/editing lines of code in their software systems (i.e., adding updates/patches to mitigate the reported RCE vulnerabilities). This motivated us to identify some common patterns that are used by programmers when they are trying to fix the RCE vulnerabilities (with a focus on RCE vulnerabilities in Java-based software systems). The next section describes some of the patterns that we identified during this study.

B. *RQ2: What types of patches (i.e., code-fixes) are usually added to mitigate the known RCE vulnerabilities in Java-based software systems?*

As described in Section III.B (Data Collection), we identified a total of fifty-one (51) patches/updates that were made to fix RCE vulnerabilities in two Java-based software projects (Tomcat and Android). These patches are essentially changes/edits that made to lines of code to mitigate or fix a reported security vulnerability. Our primary goal with RQ2 was to identify and list some code-fix patterns that were used frequently to mitigate reported RCE vulnerabilities.

For a study such as ours, projects such as Tomcat and Android are a great resource as they highlight exactly what code changes/edits were made in order to fix a vulnerability. As an example, in Fig. 7, in order to fix the vulnerability titled CVE-2019-0232, the following variable was added: cgiServlet.invalidArgumentDecoded.

Similar to the process described above, we analyzed the updates/patches (i.e., code changes) that were applied to all fifty-one (51) RCE vulnerabilities that were part of this study.



```
  cgiServlet.find.found=Found CGI: name [{0}], path [{1}], script name [{2}
  cgiServlet.find.location=Looking for a file at [{0}]
  cgiServlet.find.path=CGI script requested at path [{0}] relative to CGI l
+ cgiServlet.invalidArgumentDecoded=The decoded command line argument [{0}]
  cgiServlet.invalidArgumentEncoded=The encoded command line argument [{0}]
  cgiServlet.runBadHeader=Bad header line [{0}]
  cgiServlet.runFail=I/O problems processing CGI
```

Fig. 7. Code Change (or Patch) Applied to Fix/Mitigate an RCE Vulnerability (CVE-2019-0232) in Apache Tomcat

Next, in order to find if there were similarities (i.e., patterns) between the coding changes that were made by the programmers in order to fix the RCE vulnerabilities, we converted the code-changes (patches/updates) into pseudocode. After converting the code-changes into pseudocode, we found that the code-changes (patches) could be classified into five (5) categories or patterns.

The five types of code-changes or updates (identified as a result of our analysis of 51 RCE vulnerabilities) that can be used to fix or mitigate a majority of RCE vulnerabilities are described as follows:

RCE Update Type 1 *– Check if the Packet Size is a Positive Integer*: For this update the programmer added an If Statement to check that the packet size was a positive integer (i.e., not negative or zero). That is, check if the source buffer contained enough bytes to copy the packet and check that the packet size does not exceed the destination buffer. The pseudocode for this type of update is shown in Table II.

TABLE II. RCE UPDATE TYPE 1

Pseudocode for RCE Update Type 1

```
if ((size <= 0) || ((read - sizeof(var1) - sizeof(var2)) < size) || (sizeof(msg)
< size)) {
        return -1;
}
```

RCE Update Type 2 *– Checking for the Proper Variable Size*: Another common update made by the programmers was that they checked for the proper size. In this If statement, they check if the variable is greater than the max size. And if the variable is greater than the max size, then set the variable to the max size. The pseudocode for this type of update is shown in Table III.

TABLE III. RCE UPDATE TYPE 2

Pseudocode for RCE Update Type 2

```
if (result->num_val > MAX_ATTR_SIZE) {
   errorWriteLog;
   result->num_val = MAX_ATTR_SIZE;
}
```

RCE Update Type 3 *– Applying an Offset*: In many reported RCE vulnerabilities, it was found that the buffer is not properly calculated causing a memory overflow. To fix this, they adjusted the calculation by dividing the offset. The pseudocode for this type of update is shown in Table IV.

TABLE IV. RCE UPDATE TYPE 3

Pseudocode for RCE Update Type 3

```
display->buffer = buffer + (offset / FACTOR);
```

RCE Update Type 4 – In another commonly used patch/update for RCE vulnerabilities, the programmer moved the If Statement to the top. The intention is to run the fail-check before creating a new class and assigning the size. The pseudocode for this type of update is shown in Table V.

RCE Update Type 5 *– If statement to prevent out of bounds in the function*: In another common patch/update, the programmer added an If statement to check validity of pSettings->noOfPatches to prevent out of bounds in the function, which can also cause the memory size to be negative.

The pseudocode for this type of update (i.e., Update Type 5) is shown in Table VI.

TABLE V. RCE UPDATE TYPE 4

Pseudocode for RCE Update Type 4

```
status = function();

    if (status != SUCCESS) {
      variable1 = NULL;
      return ERROR;
    }
variable1 = new class;
variable1->size = sizeof(object);
```

TABLE VI. RCE UPDATE TYPE 5

Pseudocode for RCE Update Type 5

```
if (noOfPatches > 0) {
  int target = array[x].targetStartBand + array[x].numBandsInPatch;

  int size = (64 - target) * sizeof(FIXP_DBL);
  if (!useLP) {
    for (i = startSample; i < stopSampleClear; i++) {
      function1(&array2[i][target], size);
      function1(&array3[i][target], size);
    }
  } else
    for (i = startSample; i < stopSampleClear; i++) {
      function1(&array2[i][target], size);
    }
}
```

As is evident from the above-mentioned five commonly used updates/patches, most of the code updates was to account for changes in size of boundaries and buffers. When the size was not properly accounted for, it caused errors in the application which in turn leads to a potential for bad actors to perpetrate a Remote Code Execution (RCE) attack.

Overall, we believe that the five update types that we have identified can provide a good starting point for programmers when they are trying to fix/mitigate reported RCE vulnerabilities in their Java-based software systems. We anticipate that a list of commonly used updates/patches (such as the one presented in this study) can improve the efficiency of programmers when they are trying to determine the best way to fix vulnerable code in their software system.

V. THREATS TO VALIDITY

In this section, we describe the major threats to the validity of the results found in this study.

One major validity threat is to the generalizability of our results. This is because we have studied RCE vulnerabilities and their respective updates/patches in a limited number of systems (two systems, Tomcat and Android). Owing to this, even though we have been able to locate some viable fixes (i.e., updates/patches) for RCE, they may not resolve each and every RCE vulnerability. Through our study we saw that many instances of RCE vulnerabilities can be mitigated by fixing buffering and boundary issues. At this time, our research is limited to Tomcat and Android, and thus there are likely other instances of RCE vulnerabilities and their respective updates/patches that we have not come across in our research.

Another limitation arises from the unavailability of public vulnerability datasets for software projects. Most software projects do not make their vulnerabilities and related fixes public (with a few exceptions such as Android and Tomcat). The scarcity of vulnerable dataset makes any vulnerability related research challenging.

The authors also note that our findings do not guarantee prevention against an RCE attack from a malicious actor (i.e., attacker). Our goal is simply to provide a readily usable list of updates/patches that can be potentially employed for fixing RCE vulnerabilities. The final decision about using the most appropriate update/patch has to be made by the programmer by conducting a thorough evaluation of the vulnerability they are trying to fix.

VI. DISCUSSION ON IMPLICATION OF RESULTS

In this section we provide a brief discussion about the implications of our findings on programming practices that lead to injection of RCE vulnerabilities. Although, our main goal in this study was to identify what kind of code-changes (i.e., updates/patches) are commonly used to fix RCE vulnerabilities, our data analysis also highlighted some weaknesses or issues in coding practices (when the software systems are being developed). The paragraphs below provide a discussion related to such bad coding/programming practices that lead to injection of RCE vulnerabilities when the software is being developed.

Overall, we found that the vulnerability landscape for remote code execution (RCE) needs to be approached with a persistent and analytical approach. We must not only rely on advisories but also correlate the weakness types and attack vectors that are associated with each vulnerability type. Having such insight is meaningful in making informed decisions as well as prioritize each vulnerability based on their risk factor. Although not every RCE instance will have the same weakness type, we learned that some weakness types still correlate with the root causes that were found for the associated vulnerability. Our research was able to successfully identify a root cause (size of boundaries and buffers) that frequently leads to injection of RCE vulnerabilities.

Our research has shown that the opportunity for RCE vulnerabilities can be reduced by simply ensuring that buffers and boundaries are developed with proper sizing. With this is mind, developers can develop more efficient code and avoid at least some of the on-going RCE attacks being deployed by hackers worldwide. This research has also highlighted that some cognitive issues such as carelessness or inattention when coding can lead to vulnerability injection. We intend to integrate existing research [23, 24, 25, 26] on human cognition and human error in our future research on vulnerability prevention and vulnerability mitigation. Section VII further highlights our future research directions.

VII. CONCLUSION AND FUTURE WORK

We have conducted a detailed analysis of the updates/patches (i.e., code-changes) that were applied by programmers to mitigate/fix fifty-one (51) RCE vulnerabilities reported in two Java-based software projects: Apache Tomcat and Android.

Based on our analysis, we proposed a list of five common updates/patches (see Table II through Table VI) that can be used to mitigate or fix a significant majority of RCE vulnerabilities in Java-based systems. We believe that our findings about these common RCE updates/patches can be handy and readily-usable when programmers are trying to determine ways or means to fix RCE vulnerabilities in their own system. Therefore, we anticipate that our list of common RCE updates (shown in Tables II to VI) will help in reducing the time that is required by programmers to fix RCE vulnerabilities that have been reported in their system. To our knowledge, this is the first study of its kind that has focused on analyzing RCE vulnerabilities and their relevant updates/patches.

The results from this initial investigation are anticipated to be beneficial in reducing RCE attacks and hence the results motivate further research in the area. We plan to extend our research to other programming languages and systems to determine if such update (i.e., code-fix) patterns exist in systems coded in languages such as Python, PHP, C#, etc. Once this research is extended to other languages and systems, the natural evolution of this research is to study more vulnerabilities such as Elevation of Privilege, Information Disclosure, and SQL Injection in the future. In closing, our intent is to continue to learn about the nature of vulnerabilities and how to mitigate/fix them so that we may enhance our research to help prevent future exploits or attacks.

ACKNOWLEDGMENT

The co-authors on this paper, Dr. Kazi Zakia Sultana and Dr. Vaibhav Anu have grant funding from the National Science Foundation (NSF) of the USA through an REU grant.

REFERENCES

[1] Y. Zheng and X. Zhang, “Path sensitive static analysis of web applications for remote code execution vulnerability detection,” *2013 35th International Conference on Software Engineering (ICSE)*, San Francisco, CA, USA, 2013, pp. 652-661

[2] D. Bates, A. Barth and C. Jackson, “Regular expressions considered harmful in client-side XSS filters,” *In Proceedings of the 19th international conference on World wide web (WWW '10)*. NC, USA, 2010, pp. 91–100.

[3] M. V. Gundy and H. Chen, “Noncespaces: Using Randomization to Enforce Information Flow Tracking and Thwart Cross-Site Scripting Attacks,” *In Proceedings of the Network and Distributed System Security Symposium*, San Diego, California, USA, 2009.

[4] W. Halfond and A. Orso, “Preventing SQL injection attacks using AMNESIA,” *In Proceedings of the 28th international conference on Software engineering*, Shanghai, China, 2006, pp. 795-798.

[5] M. T. Louw and V. N. Venkatakrishnan, “Blueprint: Robust Prevention of Cross-site Scripting Attacks for Existing Browsers,” *2009 30th IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 2009, pp. 331-346.

[6] Y. Nadji, P. Saxena and D. Song, “Document Structure Integrity: A Robust Basis for Cross-site Scripting Defense,” *In Proceedings of the Network and Distributed System Security Symposium*, San Diego, California, USA, 2009.

[7] G. Wassermann and Z. Su, “Sound and Precise Analysis of Web Applications for Injection Vulnerabilities,” *In Proceedings of the 28th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '07)*, San Diego, California, USA, 2007, pp. 32-41.

[8] Todd Jackson, Babak Salamat, Gregor Wagner, Christian Wimmer, and Michael Franz, “On the effectiveness of multi-variant program execution for vulnerability detection and prevention,” *In Proceedings of the 6th International Workshop on Security Measurements and Metrics (MetriSec '10)*, Bolzano, Italy, 2010, pp. 1-8.

[9] H. Holm, T. Sommestad, U. Franke and M. Ekstedt. “Success Rate of Remote Code Execution Attacks - Expert Assessments and Observations,” *Journal of Universal Computer Science*, vol. 18, pp. 732-749, 2012.

[10] S. Biswas, M. Sohel, M. Sajal, Md. Mizanur, T. Afrin, T. Bhuiyan, and M. Hassan, “A Study on Remote Code Execution Vulnerability in Web Applications,” *International Conference on Cyber Security and Computer Science (ICONCS’18)*, Oct 18-20, 2018, Safranbolu, Turkey.

[11] Q. H. Mahmoud, D. Kauling and S. Zanin, “Hidden android permissions: Remote code execution and shell access using a live wallpaper,” *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, 2017, pp. 599- 600.

[12] S. Mohammad and S. Pourdavar, “Penetration test: A case study on remote command execution security hole,” *2010 Fifth International Conference on Digital Information Management (ICDIM)*, Thunder Bay, ON, 2010, pp. 412-416.

[13] L. Zhang, H. Zhang, X. Zhang and L. Chen, “A New Mechanism for Trusted Code Remote Execution,” *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, Heilongjiang, 2007, pp. 574-578.

[14] M. Carlisle and B. Fagin, “IRONSIDES: DNS with no single-packet denial of service or remote code execution vulnerabilities,” *2012 IEEE Global Communications Conference (GLOBECOM)*, Anaheim, CA, 2012, pp. 839-844.

[15] D. Votipka, R. Stevens, E. M. Redmiles, J. Hu, and M. L. Mazurek, “Hackers vs. Testers: A Comparison of SoftwareVulnerability Discovery Processes,” *2018 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2018, pp. 374-391.

[16] A. Austin and L. Williams, “One Technique Is Not Enough: A Comparison of Vulnerability Discovery Techniques,” *2011 International Symposium on Empirical Software Engineering and Measurement,* Banff, AB, Canada, 2011, pp. 97-106.

[17] A. Kiezun, V. Ganesh, P. J. Guo, P. Hooimeijer and M. D. Ernst. “HAMPI: a solver for string constraints,” *ACM Trans. Softw. Eng. Methodol.*, vol. 21, no. 4, Article 25, 2013.

[18] W. Halfond, S. Anand and A. Orso. “Precise Interface Identification to Improve Testing and Analysis of Web Applications,” *In Proceedings of the eighteenth international symposium on Software testing and analysis (ISSTA '09),* Chicago, IL, USA, 2009, pp. 285-296.

[19] Dinis Cruz, “OWASP O2 Platform - Open Platform for Automating Application Security Knowledge and Workflows”, *Web Application Security Conference*, 2010, pp. 5-5.

[20] S. Tyagi and K. Kumar, “Evaluation of Static Web Vulnerability Analysis Tools,” *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan, India, 2018, pp. 1-6.

[21] Y. Xie and A. Aiken. “Saturn: A scalable framework for error detection using Boolean satisfiability,” In ACM Trans. Program. Lang. Syst. May, 2007, vol. 29, no. 3, pp. 16–es.

[22] M. Das, S. Lerner, M. Seigel. “ESP: path-sensitive program verification in polynomial time,” *In Proceedings of the ACM SIGPLAN 2002 conference on Programming language design and implementation (PLDI '02),* 2002, Berlin, Germany, pp. 57-68.

[23] V. Anu, G. Walia, W. Hu, J. C. Carver and G. Bradshaw, "Using a Cognitive Psychology Perspective on Errors to Improve Requirements Quality: An Empirical Investigation," *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*, Ottawa, ON, Canada, 2016, pp. 65-76.

[24] W. Hu, J.C. Carver, V. Anu et al., "Using human error information for error prevention", *Empir. Software Eng.*, vol. 23, pp. 3768-3800, 2018.

[25] V. Anu, G. Walia, W. Hu, J. C. Carver and G. Bradshaw, "Issues and Opportunities for Human Error-Based Requirements Inspections: An Exploratory Study," *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Toronto, ON, Canada, 2017, pp. 460-465.

[26] V. Anu, K. Z. Sultana and B. K. Samanthula, "A Human Error Based Approach to Understanding Programmer-Induced Software Vulnerabilities," *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, Coimbra, Portugal, 2020, pp. 49-54.

Energy Performance Analysis of a Differential Wheeled Mobile Robot with Fuzzy Logic Controller

Said Fadlo*, Abdelhafid Ait Elmahjoub†and Nabila Rabbah‡
dept.Structural Engineering, Intelligent Systems, and Electrical Energy
ENSAM, Hassan II University
Casablanca, Morocco
Email: *fadlosaid@gmail.com, †aitelmahjoub@gmail.com, ‡nabila_rabbah@yahoo.fr

***Abstract*—Very few models have been proposed for estimating the energy consumption of mobile robots based on fuzzy controller methods. In comparison to the classical controllers, using this approach offers many advantages. In this paper, a complete energy model has been developed in the Simulink tool to illustrate that. The estimated energy consumption is compared with the typical controllers. This comparison shows that energy losses in based-fuzzy Differential Wheeled Mobile Robot are 2.51 % per actuator less than the others. This model is more accurate and helps to optimize the consumption of power in mobile robots.**

***Index Terms*—Energy model, Fuzzy Controllers, Simulation, Differential Wheeled Mobile Robot.**

I. Introduction

Nowadays, mobile robots are commonly used in education, military, space, medicine, rescuing, agriculture, mining, entertainment, and many more. In these fields, mobile robots are expected to have good energy autonomy in order to perform long-range, long-term, and complex missions. These missions require a good energy model to predict the energy consumption of the robot with acceptable accuracy.

The Energy Consumption (EC) studies on mechatronic systems, especially mobile robots, are categorized into two aspects: hardware and software. Studies at the hardware level consider the EC characteristics of single-board microcontroller, sensor, DC motor drivers, communication system, and actuators to evaluate consumed energy or to design more energy-saving hardware components [1]. On the other hand, there are many studies in EC estimation and optimization of software for computer systems [2]–[4]. In this context, fuzzy set theory is a mathematical method that allows designers to achieve a lot of advances. For instance, the fuzzy logic-based model for Li-ion battery, considering state of charge and temperature effect on parameters is common-used for its simplicity and suitability for control and energy estimation [5]. The solutions are mature with years of development. Combining those strategies with the optimization of the mobile robot movements lead to design systems globally efficient.

Many researchers have been focused on designing models that can reduce energy consumption of mobile robots. Most of these works are based on the optimization of robot movement. Carabin et al [6], presented a survey of existing techniques based on Richiedei et al [7] classification. A. Stefek et al [8], firstly concluded, after reviewing many pieces of research, that these studies refer only to continuous or smooth curves, where the energy consumption of the robot is not mentioned. Secondly, it is difficult for the robot to track the given path smoothly in an uncertain or harsh environment. But,recent works [9], show that there is a great scope in applying newly developed algorithms such as Shuffled frog leaping algorithm, Cuckoo search algorithm, Invasive Weed Optimization, Bat Algorithm, Harmony Search Algorithm, Differential Evolution Algorithm, Bacterial foraging optimization Algorithm , Artificial bee colony Algorithm and Firefly Algorithm for navigation in an unknown complex environment in the presence of maximum uncertainty and these can be used to develop new kinds of hybrid approaches.

In this paper, we adopted a comparison approach to investigate the energy consumption for different controllers in a mobile robot. In this case, the losses contributions of DC motors, and motion strategies are taken into account to perform an accurate model.

II. Methodology

The original aim of our work is to identify the relationship between the control method adopted for the mobile robot and energy-efficiency improvement opportunities.

To illustrate this approach, firstly a total energy consumption model is adopted in the section II.A. Secondly, we based the controllers design, simulation, and comparison for tow driving Direct Current (DC) motor actuators of the mobile robot, on a study conducted by Shamshiri et al in [10]. In this study three control methods were used for the robot motion: (i) a common PID controller, (ii) a lead-lag compensator filter, (iii) a fuzzy logic controller. The control objective was the angular rate of the rotation shaft by varying the applied input voltage. More information about the control criteria can be found in the work cited above. However, we reported in section IV some results that helped us to design the controllers.

A. Total Energy Model

The total energy consumption of the mobile robot is divided into four parts and represented by equation:

$$E_{Tot} = E_{DC} + E_G + E_K + E_f \tag{1}$$

978-1-6654-4067-7/21/$31.00 © 2021 IEEE

Where E_{DC} term symbolize the Energy losses in DC Motors, E_G the energy losses in gearhead, E_K the kinetic losses, and E_f the energy losses due to friction.

B. Mobile robot kinematic Energy

To develop an energy model, a kinematic model of the robot is needed. A Wheeled Differential Drive Mobile Robot $(WDDMR)$ is used in this work. It has two driven wheels that are attached to DC motors and a rear castor which is added for balancing. The rotation of the wheels produces linear and angular motions of the robot. The kinematic model is given in Fig 1. Where r is drive wheel radius, b symbolize the axle length, x and y denote the position of the center of axle, ϕ is the angle between the robot axle and X- axis, $\dot{\theta}_L$ and $\dot{\theta}_R$ are the angular velocities of the wheels, which are related to the linear v and angular ω velocities of the robot by the following equations:

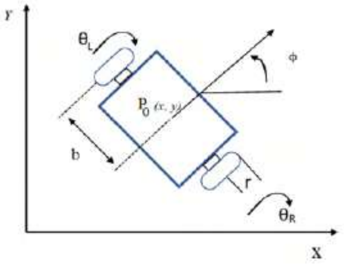


Fig. 1. Schematic diagram of the mobile robot.

$$v = \frac{r\left(\dot{\theta}_R+\dot{\theta}_L\right)}{2} \tag{2}$$

$$\omega = \frac{r\left(\dot{\theta}_R-\dot{\theta}_L\right)}{2b} \tag{3}$$

So the kinetic energy loss equation, can be expressed as same as in [11]:

$$E_k = \frac{1}{2}\left(mv\left(t\right)^2 + I\omega\left(t\right)^2\right) \tag{4}$$

Where m is the mass, and I is the moment of inertia of the robot.

In this study, we only focus on the kinetic energy losses of the robot.We have already studied the other energy losses in a previous work [12].

C. Modeling Energy DC motor

The DC motor behavior is characterized using an equivalent circuit model. Fig. 2 shows the DC motor circuit with torque and rotor angle consideration. The differential equations that describe the dynamic model are:

$$\begin{cases} L\frac{di}{dt} + Ri + K_\omega\dot{\theta} = V \\ I_s\ddot{\theta} + K_t i + f\dot{\theta} + \tau = 0 \end{cases} \tag{5}$$

Where:
V and i are the armature voltage and current.
R and L are the armature resistance and inductance.
f is the viscous friction coefficient.
τ is the dynamic load applied to the motor.
K_t is the motor torque constant.
K_ω is the voltage constant.
I_s is the motor shaft inertia.
$\theta = [\theta_R\theta_L]$ are the angular positions of the wheels.

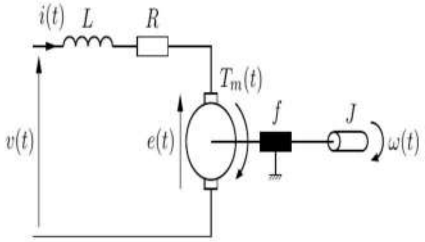


Fig. 2. Schematic diagram of the DC motor.

Using equation (5) we can describe the behavior of the motor speed for given voltage by the open-loop transfer function (6) in laplace domain with voltage $V(s)$ as input and shaft speed $\dot{\theta}\,(s)$ as output [13]:

$$\frac{\dot{\theta}\,(s)}{V\,(s)} = \frac{\frac{K_t}{I_sL}}{s^2 + \frac{(I_sR + fL)\,s}{I_sL} + \frac{fR + K_eK_t}{I_sL}} \tag{6}$$

We can deduce the relationship between the reference speed $\dot{\theta}_{ref}$ and the output speed $\dot{\theta}_{out}$ with a constant gain of K as:

$$\frac{\dot{\theta}_{out}\,(s)}{\dot{\theta}_{ref}\,(s)} = \frac{\frac{KK_t}{I_sL}}{s^2 + \frac{(I_sR + fL)\,s}{I_sL} + \frac{KK_t + fR + K_eK_t}{I_sL}} \tag{7}$$

III. Model Implementation

In order to estimate the energy for the mobile robot, a simulation was performed with SIMULINK. The result is displayed in Fig. 3. The structure of the model is:

- the upper branch shows DC Motor with Fuzzy Controller.
- In the middle we have DC Motor with Lead-lag Compensator filter.
- the lower branch shows DC Motor with PID Controller.

IV. Model Results

A. Energy model results

To set up the energy model that is represented in Fig. 3, the mobile and motor parameters of the mobile robot present in [10], [14] were used. These parameters are shown in the tables (table I, II and III). A typical structure of a fuzzy logic controller is shown in Fig. 3,Fig. 4 and Fig. 5. The final transfer function of the complete lead-lag compensator filter is:

$$G_{lead-lag}\,(s) = \frac{1.233s^2 + 25.88s + 98.6}{s^2 + 64.55s + 3.64} \tag{8}$$

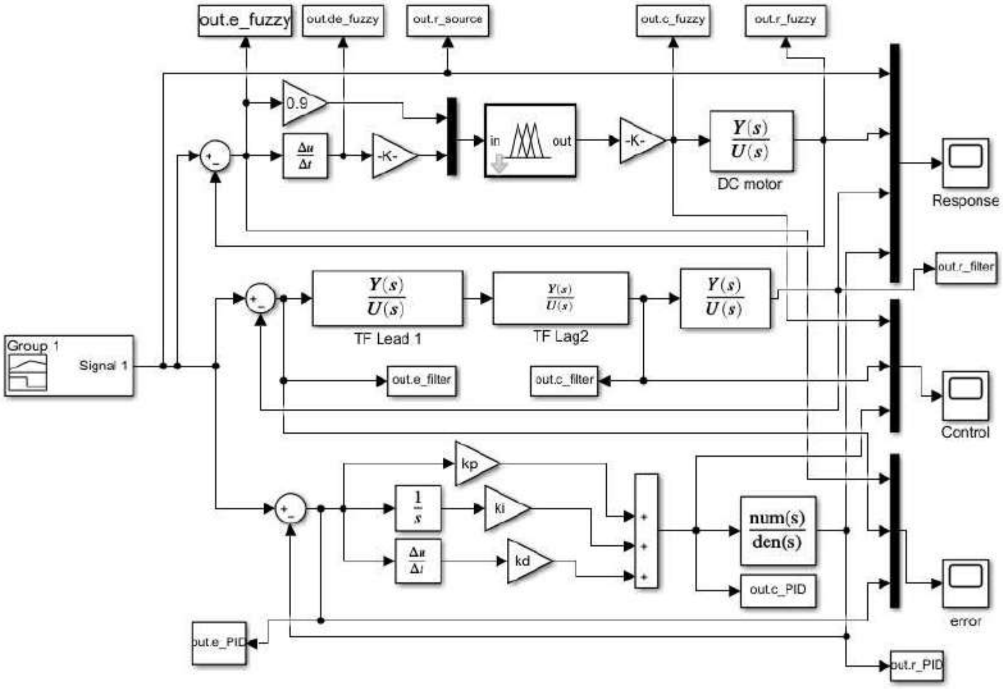


Fig. 3. Blocks diagram of controllers for DC motor speed

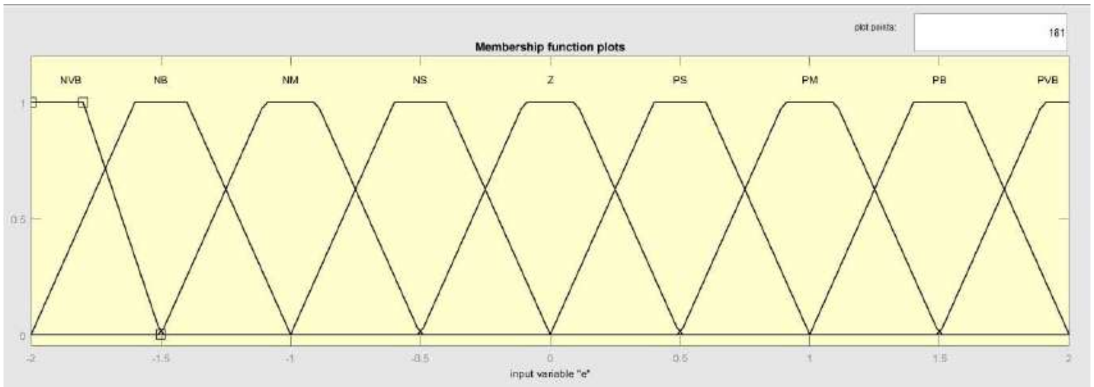


Fig. 4. Fuzzy logic first input variable, error

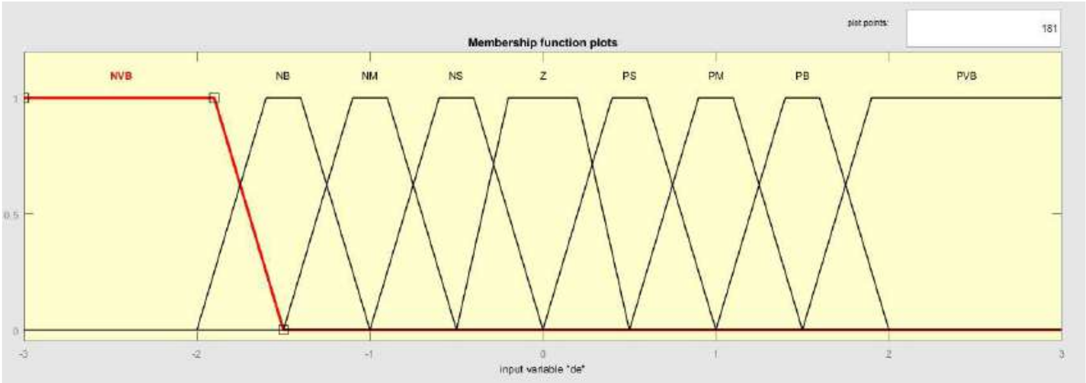


Fig. 5. Fuzzy logic second input variable, change of error

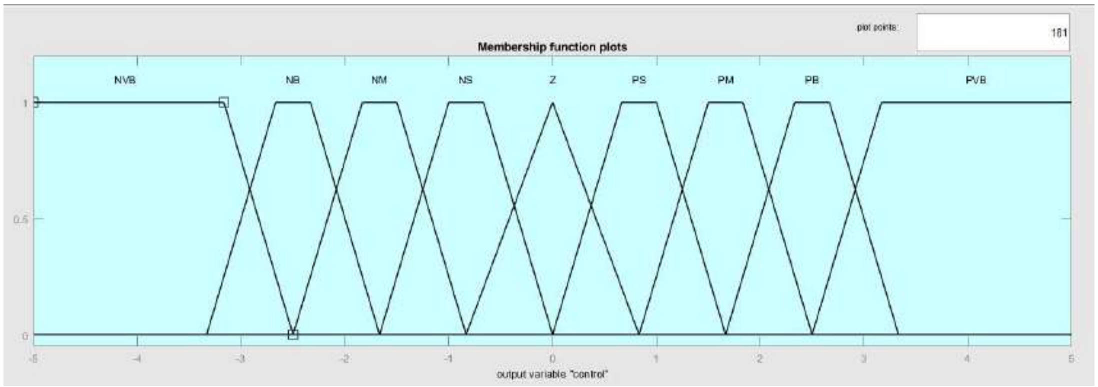


Fig. 6. Fuzzy logic output variable, control

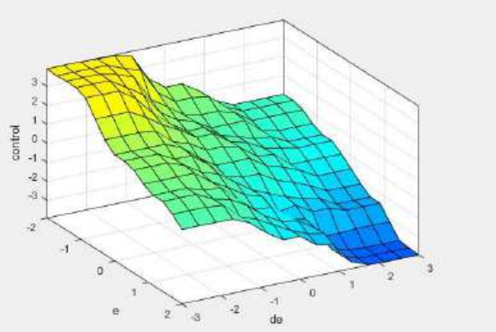


Fig. 7. Fuzzy logic rule surface

TABLE I
MOTOR PARAMETER OF THE WDDMR

Parameters	Values	Parameters	Values
R	0.7 Ω	K$_t$	0.88 Nm/A
f	0.035 Nm/(rad/s)	K$_\omega$	0.88(rad/s)/V
I$_s$	$0.0713 Kgm^2$	K	0.01

In this section, we present simulation to estimate the total energy consumption using as input a step velocity profile. The simulation tests were performed by driving the robot in a straight line on a flat surface. The robot accelerated from 0 to the desired velocity $(1m/s)$. It maintained this velocity for 1s and 10s respectively. As output, we have the angular velocity of the wheels. Fig. 8 shows the velocity obtained by the model, that were put in equation (4) to calculate total kinetic energy loss as shown in Fig.9 , Fig. 10 and Fig. 11. From the result shown in figure 8, we can affirm that the PID controller provides the smallest energy loss during the beginning of the movement, and this is due to the smallest overshot time. However, if the motion takes more than 10 seconds as in Fig. 9 and Fig. 11, the output trajectory provided by the fuzzy logic controller converges to steady-state faster with smaller tracking errors leading to energy saving up to 2.5% per motor.

TABLE II
MOTOR PARAMETER OF THE THE WDDMR

Parameters	Values	Parameters	Values
r	0.095 m	m$_c$	6.04 Kg
b	0.165 m	mω	1.48 Kg

TABLE III
PID CONTROLLER PARAMETER

K$_P$	K$_I$	K$_D$
0.6844	0.5975	0.0119

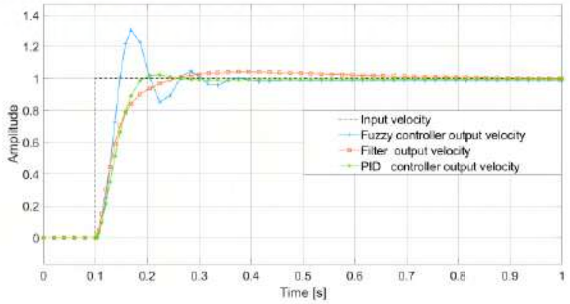


Fig. 8. Kinetic energy loss of the robot over at the end of motion robot

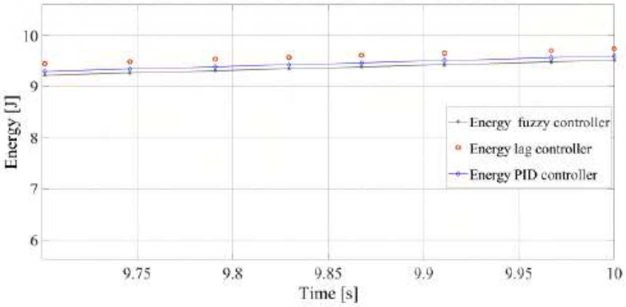


Fig. 9. Kinetic energy loss of the robot at the end of motion robot

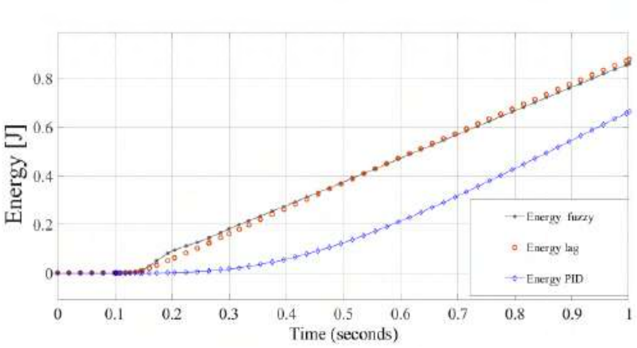


Fig. 10. Kinetic energy loss of the robot over 1 second

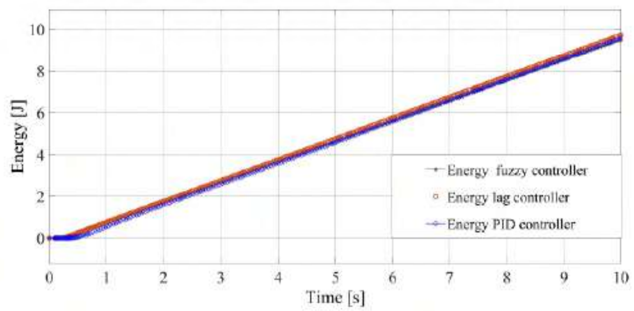


Fig. 11. Total kinetic energy loss of the robot over 10 seconds

V. Conclusion

A comparison of PID,Fuzzy controller and Lead-lag compensator filter for a differential wheeled drive robot mobile was performed using SIMULINK. An energy model is presented based on the dynamic motor and mobile robot parameters. Our model shows that we can achieve up to 2.51% per motor when the fuzzy controller is used. In this work, we have studied the control motion algorithm influence on energy consumption, we conclude that by implementing the fuzz controller in WDDMR we can improve the saving energy with more efficiency. In our future work, we plan to incorporate new kinds of hybrid approaches for controller algorithms into the model to provide a better energy consumption estimation.

References

[1] Y. H. Jung, B. Park, J. U. Kim, and T.-i. Kim, "Bioinspired electronics for artificial sensory systems," *Advanced Materials*, vol. 31, no. 34, p. 1803637.

[2] M. Bazzaz, M. Salehi, and A. Ejlali, "An accurate instruction-level energy estimation model and tool for embedded systems," *Instrumentation and Measurement, IEEE Transactions on*, vol. 62, pp. 1927–1934, 07 2013.

[3] M. Krunic, M. Popovic, V. Krunic, and N. Četić, "Energy consumption estimation for embedded applications," *Elektronika ir Elektrotechnika*, vol. 22, 06 2016.

[4] F. Poursafaei, M. Bazzaz, and A. Ejlali, "Npam: Nvm-aware page allocation for multi-core embedded systems," *IEEE Transactions on Computers*, vol. PP, pp. 1–1, 05 2017.

[5] D. Jiani, L. Zhitao, W. Youyi, and W. Changyun, "A fuzzy logic-based model for li-ion battery with soc and temperature effect," in *11th IEEE International Conference on Control Automation (ICCA)*, June 2014, pp. 1333–1338.

[6] G. Carabin, E. Wehrle, and R.Vidoni, "A review on energy-saving optimization methods for robotic and automatic systems," *Robotics*, vol. 196, p. 39, 2017.

[7] D. Richiedei and A. Trevisani, "Analytical computation of the energy-efficient optimal planning in rest-to-rest motion of constant inertia systems," *Mechatronics*, vol. 39, pp. 147–159, 2016.

[8] A. Stefek, T. V. Pham, V. Krivanek, and K. L. Pham, "Energy comparison of controllers used for a differential drive wheeled mobile robot," *IEEE Access*, vol. 8, pp. 170 915–170 927, 2020.

[9] B. Patle, L. Ganesh Babu, A. Pandey, D. Parhi, and A. Jagadeesh, "A review: On path planning strategies for navigation of mobile robot," *Defence Technology*, vol. 15, no. 4, pp. 582–606, 2019.

[10] R. Shamshiri and W. I. Wan Ismail, "Design and simulation of control systems for a field survey mobile robot platform," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, pp. 2307–2315, 08 2013.

[11] A. Trzynadlowski, "Energy optimization of a certain class of incremental motion dc drives industrial electronics," *IEEE Transactions on*, vol. 35, no. 1, pp. 60–66, Feb 1988.

[12] S. Fadlo, N. Rabbah, and A. A. Elmahjoub, "Energy modeling for a differential guide mobile robot using simscape," presented at the 3rd IEEE International Symposium on Advanced Electrical and Communication Technologies ISAECT, November 25–27, 2020.

[13] R. Dorf and R. Bishop, *Modern Control Systems, 13th Edition*, 01 2017.

[14] J. M.Brossog, M.Bornschlegl, "Reducing the energy consumption of industrial robots in manufacturing systems," *Int. J. Adv. Manuf. Technol*, vol. 78, p. 1315–1328, 2015.

An Efficient Multihoming Scheme to Support Seamless Handover in SDN–based Network Mobility

Jae Won Lim, Tahira Mahboob, and *Min Young Chung
Department of Electrical and Computer Engineering, Sungkyunkwan University
2066 Seobu–ro, Jangan–gu, Suwon–si, Gyeonggi–do, 16419, Republic of Korea
{jwlim0724, tahira, mychung}@skku.edu

***Abstract*—Multihoming technology is widely used to contribute reliability and performance improvement in network mobility by providing multiple connections to mobile routers. The mobile router may have to choose a suitable connection among available connections for the mobile node. In this paper, we propose a handover method for network mobility that utilizes the multihoming feature of a mobile router in SDN–based network mobility (SDN–MNEMO). We newly introduce an access router (AR) selection method to select the optimal AR for seamless handover in multihoming network mobility. The optimal AR is selected considering throughput of the backhaul links and downlink bit rate information of the available ARs in the mobility network. Using an experimental testbed, we present a thorough handover performance evaluation of our proposed scheme, and compare it with a conventional SDN–NEMO scheme. The results show that the proposed scheme outperforms the conventional scheme while achieving seamless handover. The handover delay in the conventional scheme is 7.15s compared to 72ms in the proposed scheme. Also, average throughput for the mobile node reached up to 950Kbps and it remained constant during the handover for the proposed scheme.**

***Index Terms*—Software–defined networking (SDN), multihoming, seamless handover, network mobility**

I. Introduction

Recently there has been a massive growth in size of mobile networks and number of mobile devices. In addition to increase in mobile applications, mobile consumers have been increasing every year [1]. Due to geographic location mobility, mobile devices may lose connectivity by changing their association to another access point. Researchers have proposed several hosts and network–based systems for mobility management. Internet engineering task force (IETF) has worked on mobile IPv6 (MIPv6) [2] for handling mobility in networks. In their proposal, a host–based system supports mobility of mobile devices. To improve management of mobility function in a network, proxy mobile IPv6 (PMIPv6) and MIPv6 were introduced.

Since the handover is done between mobile device and mobile access gateway (MAG), mobile devices must support features required for handling mobility. In PMIPv6, tunnelling has been used to transmit network traffic between the local mobility anchor (LMA) and an MAG [3]. Tunneling has been used to encapsulate and de–capsulate packets of network traffic at each endpoint of the tunnel. However, tunneling requires higher bandwidth and can cause performance degradation, such as longer delays. To overcome these issues, researchers have worked on group mobility scenarios. In order to provide group IP mobility management to users, IETF has worked to standardize a mobile network mobility protocol referred to as network mobility (NEMO) [4]. NEMO supports features for mobility of an entire network via mobile router (MR).

Multihoming is a method of connecting a host or network to more than one network. It has been used to increase stability of a network and improve network performance [5]. In an existing studies using multihoming, the authors performed handover by comparing the layer 2 signal strength after each care–of–address (CoA) was assigned to two interfaces [6]. Multihoming was required to make multiple connections that enables effective control to connects to multiple networks. Multihoming has been supported by using software defined networks (SDN). SDN provides easier management of an efficient and adaptable network architecture built supported by open source API Open–Flow. Open–Flow has been introduced as a standard communication interface defined between control and forwarding layer. Open–Flow has been developed to suit the high bandwidth and dynamic characteristics demands of current applications. It separates control and forwarding functions of the network to enable direct programming of network control and abstracts the underlying infrastructure for application and network services.

In conventional multihoming NEMO, two WiFi interfaces were considered at the MR to connect to multiple ARs for seamless handover. In general, the ARs for the handover are selected by the MR considering the received–signal–strength information (RSSI) of the available ARs. However, since the RSSI–based selection method only considers signal strength, it is difficult to provide a network environment suitable for a service because the network state of the corresponding AR is not considered. The corresponding AR may be overloaded and the mobile nodes (MNs) connected to it may receive degraded service after handover. To overcome this problem, we propose a new AR selection method.

[*] Min Young Chung is the corresponding author.

978–1–6654–4067–7/21/$31.00 ©2021 IEEE

In this paper, we propose a SDN–based network mobility architecture, SDN–MNEMO. Our proposed scheme provides seamless handover of MR by selecting the optimal AR considering the downlink bit rate information and throughput of the backhaul links of the available ARs in the mobility network. Rest of the paper is organized as follows. In Section 2 we briefly describe network mobility basic support protocol (NEMO–BSP) and SDN–NEMO. In Section 3, we explain our proposed scheme. In Section 4, we describe performance evaluation and in Section 5 we conclude the paper.

II. Background

A. NEMO–BSP

To support group mobility, MIPv6 was extended to NEMO–BSP [4]. In this scheme, a mobile router (MR) was installed in public or private transport vehicles. MR connects to an access router (AR) of a network domain in order to support seamless internet connectivity for MNs. MR works as an intermediary between the AR and active MNs of AR's network. MR provides continuous monitoring of attachment and de–attachment of MNs that connect and disconnect from the network. Multiple MNs can connect to a MR and a MR can connect to different ARs simultaneously to support network group mobility.

In a mobility scenario, users travelling in a vehicle and receiving service, connect to the internet via a MR installed in the vehicle. Consequently, in the case of a handover, only the MR performs binding update instead of all these users travelling in that group. This reduces the amount of communications between the MNs and the network. In such a case, MNs are not involved to supporting mobility which reduce processing at these devices. In addition to this, group mobility for the MNs reduced the complexity of network providers to implement the network. However, all network traffic was transmitted using tunneling between home agent (HA) and MNs. Therefore, network may experience congestion at the HA. This may degrade the network performance due to single point of ingress/egress for the network domain.

B. SDN–NEMO

SDN–based solution was introduced to support distributed mobility management (DMM) for NEMO [7]. Several advantages such as global network view, centralized network control, flexible routing methods, dynamic data plane configuration and management, of SDN were leveraged to support DMM. SDN–based network mobility architecture overcame several existing problems such as traffic transmisison without tunneling and centralized mobility management. Also, the mobility management features were no longer required at the AR.

In SDN–NEMO, the control plane is decoupled from the data plane. It provides centralized network control similar to efficient PMIPv6–based distributed network mobility management (EPD–NEMO). The data plane was constructed in a distributed architecture. The main communicaton protocol used between the control and data plane was the south–bound API such as Open–Flow. The data plane consists of software switches such as OpenVswitch (OVS) installed on the ARs that can be dynamically configured by the SDN controller at the control plane. Furthermore, network mobility status of MNs and MRs is tracked based on DMM. The solution provides seamless IP connectivity for MNs connected to the network. The mobility management features have been implemented at the application plane. The applications on the application were controlled by the SDN controller.

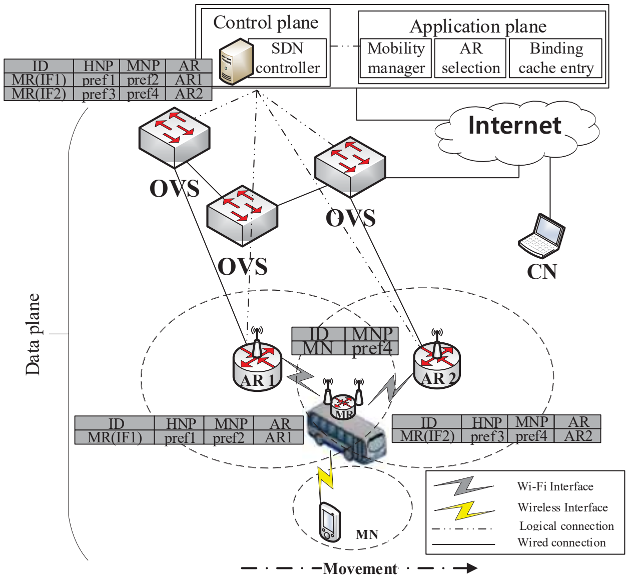


Fig. 1. An example of SDN–MNEMO

III. Proposed scheme

In this section, we describe the proposed system architecture, the attachment procedure of MR and MN, handover procedure of MR, and handover procedure of MN.

A. System architecture

An example architecture of SDN–MNEMO is shown in Fig. 1. The architecture consists of the control plane, the data plane and the application plane. The control plane consists of the controller. The data plane consists of the OpenVSwitch (OvS), access router (AR), mobile router (MR), mobile node (MN) and corresponding node (CN). The application plane consists of the mobility manager module, the AR selection module, and the binding cache entry. The mobility manager is responsible for providing mobility support for the MNs and MR. The AR selection module selects the optimal AR from available ARs in the mobility network. The binding cache entry is a database that stores the IP address of ARs, MR, and MN. The SDN controller allocates IP to MR and MN, and stores and manages mobility session information and location of MN and the gateway. Here, the gateway is referred to as the MR.

AR acts as an Open–Flow support switch that exchanges and manages Open–Flow messages with the controller. MR is used to transmit all traffic from MNs to/from the network.

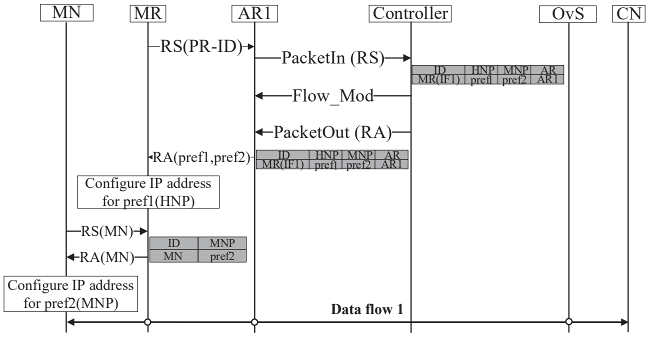


Fig. 2. Attachment procedure of MR and MN

Also, MR provides internet service to the MNs connected to it that is communicating within its coverage area. It also handles all signalling data and procedures for all MNs connected to it. MR can be connected to AR1 and AR2 through two WiFi interfaces. MN has a wireless interface to connect with the MR. It can use WiFi communication link to communicate with a CN. When the mobile vehicle moves from coverage region of an AR, i.e., AR1 to the coverage region of another AR, i.e., AR2, as indicated in the Fig. 1, handover is required for communication link of the MN.

In the proposed architecture, we newly introduce an access router (AR) selection method to select the optimal AR for seamless handover in multihoming NEMO. The optimal AR is selected considering the downlink bit rate information and throughput of the backhaul links of the available ARs. We refer to the available ARs as those in the coverage region of the MR. The information of the selected AR is sent to the MR via the AR. To handle the seamless handover for the MN, we consider two WiFi interfaces at the MR such as IF1 and IF2. When IP address allocation request is received at the controller from MR via AR, the controller allocates two IP addresses: one to the MR and other to the MN. The controller transmits Open–Flow "PacketOut" message containing the RA message to the AR. The RA message contains the IP addresses of the MR and MN. AR receives this "PacketOut" message and extracts the RA message and it then transmits RA message to the MR.

B. Attachment procedure of MR and MN

The attachement procedure of MR and MN is presented in Fig. 2. When a MR (IF1) connects to AR1 for the first time, it transmits a router solicitation (RS) to AR1 [8]. This RS message contains its own MR–ID value and a flag 'N'. This flag 'N' is sent to request a subnet prefix for the MN. On receiving this message, AR1 transmits 'PacketIn' message including RS message to the controller. The controller that received the 'PacketIn' message allocates HNP (Home Network Prefix) and MNP (Mobile Network Prefix), which is the subnet prefix of MR (IF1). It then transmits 'Packetout' message including router advertisement (RA) message to AR1. At the same time, the controller updates the flow table by sending the 'Flow_Mod' message to AR1. AR1 receives this 'PacketOut' message and transmits the RA message to the MR.

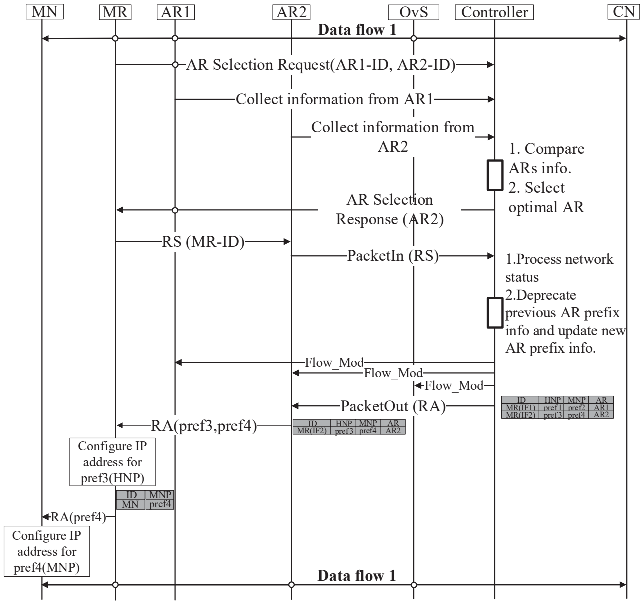

MN
MR
AR1
AR2
OvS
Controller
CN
Data flow 1
AR Selection Request(AR1-ID, AR2-ID)
Collect information from AR1
Collect information from AR2
1. Compare ARs info.
2. Select optimal AR
AR Selection Response (AR2)
RS (MR-ID)
PacketIn (RS)
1.Process network status
2.Deprecate previous AR prefix info and update new AR prefix info.
Flow_Mod
Flow_Mod
Flow_Mod
PacketOut (RA)
ID | HNP | MNP | AR
MR(IF1) | pref1 | pref2 | AR1
MR(IF2) | pref3 | pref4 | AR2
RA(pref3,pref4)
ID | HNP | MNP | AR
MR(IF2) | pref3 | pref4 | AR2
Configure IP address for pref3(HNP)
ID | MNP
MN | pref4
RA(pref4)
Configure IP address for pref4(MNP)
Data flow 1

Fig. 3. Handover procedure of MR

The MR that receives the RA message configures its own IP using HNP (pref1). When the MN accesses the MR, the pre-allocated IP address from the controller, i.e., MNP (pref2), is notified to the MN via a RA message. Consequently, MN can communicate with a CN by transmitting data from AR1 via MR.

C. Handover procedure of MR

A handover is required when the mobile vehicle moves away from AR1 and is in the coverage area of AR2. The handover procedure of MR to a new AR is presented in Fig. 3. During mobility, the MR approaches the area that overlaps AR1 and AR2. The AR selection module receives the network status information, such as bit rate for downlink and throughput of backhaul link of available ARs via controller. It then processes the received information to select the optimal AR with the highest bit rate and throughput values.

Once the AR is selected, the information of the selected AR, i.e., AR–ID, is transmitted in a AR Selection Response message to the MR. The MR then transmits an RS message containing MR–ID to the AR2. AR2 transmits a Open–Flow "PacketIn" message to the controller containing the RS message. This information is transmitted to the mobility manager. The mobility manager extracts the binding cache entry information of MR and MN from the binding cache entry module.

The mobility manager compares current network status, i.e., prefix information of MR and MN. If there is some change in the network status, the mobility manager transmits this information to the controller. The controller receives this information and transmits the RA message to the MR through AR2 using Open–Flow "PacketOut" message. When the MR receives the RA message, the prefix in AR1 is set to depre-

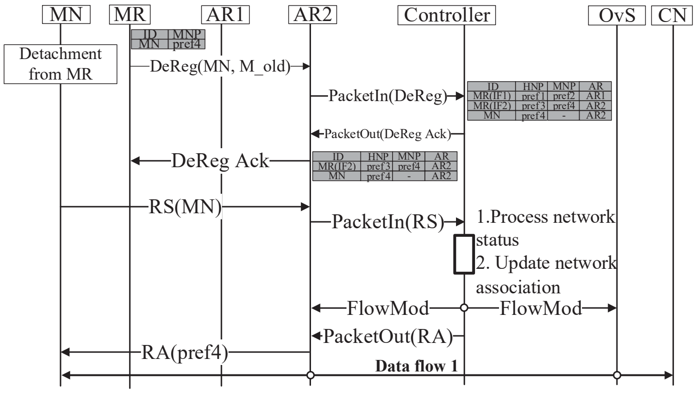


Fig. 4. Handover procedure of MN

cated, and the new prefix for use in AR2 is provided within the RA message. Consequently, the new prefix, i.e., pref4, used for AR2 enables session continuity, i.e., the MN doesnot experience session break for the existing communication. This enables route optimization and seamless connectivity for the communications for of MNs and MR during handover.

D. Handover procedure of MN

When the MN accessing the MR in a mobile vehicle leaves, it needs handover to a new AR of the same network provider. The procedure when the MN leaves the MR and hands over to the new AR is shown in Fig. 4. Since the MN is travelling in a group mobility scenario and MN is connected to the MR, we assume when the MR handovers to AR2, MN handovers along with. For mobility management, when the MN leaves the MR area, the MR sends a 'DeRegister message' to AR2 of the disconnected MN's ID and previous IP address. Upon receiving this, AR2 sends a 'PacketIn message' to the controller. Controller registers MN's information in binding caching entry. Following this, when MN is connected to AR2, it transmits an RS message which contains its ID and previous address. Upon receiving this, AR2 transmits the RS message to the controller through the 'PacketIn message'. Controller compares MN's prefix with existing prefix information, if it is the same prefix as AR2's MNP. In addition, the controller provides the optimal path by updating the flow table of AR2 and OVS through the 'Flow_Mod messages'.

IV. Performance evaluation

Performance of the proposed scheme has been evaluated using an experimental testbed. Wireshark packet analyzer version 2.6.10 and iPerf version 2.0.5 tools have been used to obtain the results. The experiment was repeated 30 times.

A. Testbed environment

To evaluate the performance of SDN–MNEMO, we build an experimental test bed with the ONOS controller and java–based application layer modules, i.e., AR selection module, mobility manager module, and the binding cache entry module. We configure desktop computer having Ubuntu operating system version 18.04 LTS 64 bits, as an SDN controller by installing ONOS Drake version 1.3.0 [9] on it. Also, the application layer modules are implemented on the desktop computer.For the data plane three OVS, two ARs with WLAN APs[1], and MR as shown in Fig. 5. OF–SW refers to an OVS switch and OF–AR refers to an AR. OF–SW and OF–AR enable the functions of the Open–Flow protocol by installing the OVS inside the WLAN AP. It is noted that the ONOS controller has a fixed global IP address, so OVS can connect with it via internet.

Ubuntu mate 18.04 version is installed on the open source Raspberry pi 3 which serves as an AR [10]. The AR has a OVS and WLAN AP installed on it . Open–Flow protocol version 1.3 is used as the communication protocol in the test bed [11]. We use Wireless8265 Intel dual chipset to create wireless interfaces at the MR. We use iptime N600UA WLAN USB adapters to create two WiFi interfaces at MR and AR. Raspberry pi is used to create OF–ARs by installing hostapd package in them [12]. We use laptops with Ubuntu 18.0.4 OS to work as MN and CN.

B. Experimental results

CN operates as the server and MN as the client. An UDP session is established between the client and the server. UDP transmission speed between MN and CN is set as 1MBps, and traffic is generated using the iPerf application. CN receives data traffic from the MN. In this paper, throughput is defined as number of successful bytes received at the client per unit time and is given as:

$$T_h = \frac{B_{t_{n+\Delta}} - B_{t_n}}{t_{n+\Delta} - t_n}, \tag{1}$$

where $B_{t_{n+\Delta}}$ and B_{t_n} are the bytes received at time $t_{n+\Delta}$ and bytes received at t_n. The throughput of CN for SDM–MNEMO for a sample pass during handover between AR1 and AR2 is shown in Fig. 6. In the conventional scheme, it is observed during handover that in the interval [25s,32s] of experiment time, throughput goes to 0. The throughput for the proposed seamless handover strategy reduces but does not drop the connection.

Fig. 7 shows the delay experienced by a MN during handover from AR1 to AR2 for a sample pass. Handover delay H_d is measured using:

$$H_d = P_{AR2_f} - P_{AR1_l}, \tag{2}$$

where P_{AR2_f} and P_{AR1_l} are the first packet received at AR2 after handover and last packet received from AR1 when the handover started. MN connects with CN using a TCP connection at transmission rate of 1MBps using the iPerf application. Data captured form CN using wireshark packet analyzer is analyzed to measure the handover delay. It is noted that the delay experienced by the MN from time 10s to 25s is nearly similar. It is because the test environment is identical for the proposed and conventional scheme and there is no interference

[1] Wireless APs operate on 802.11n

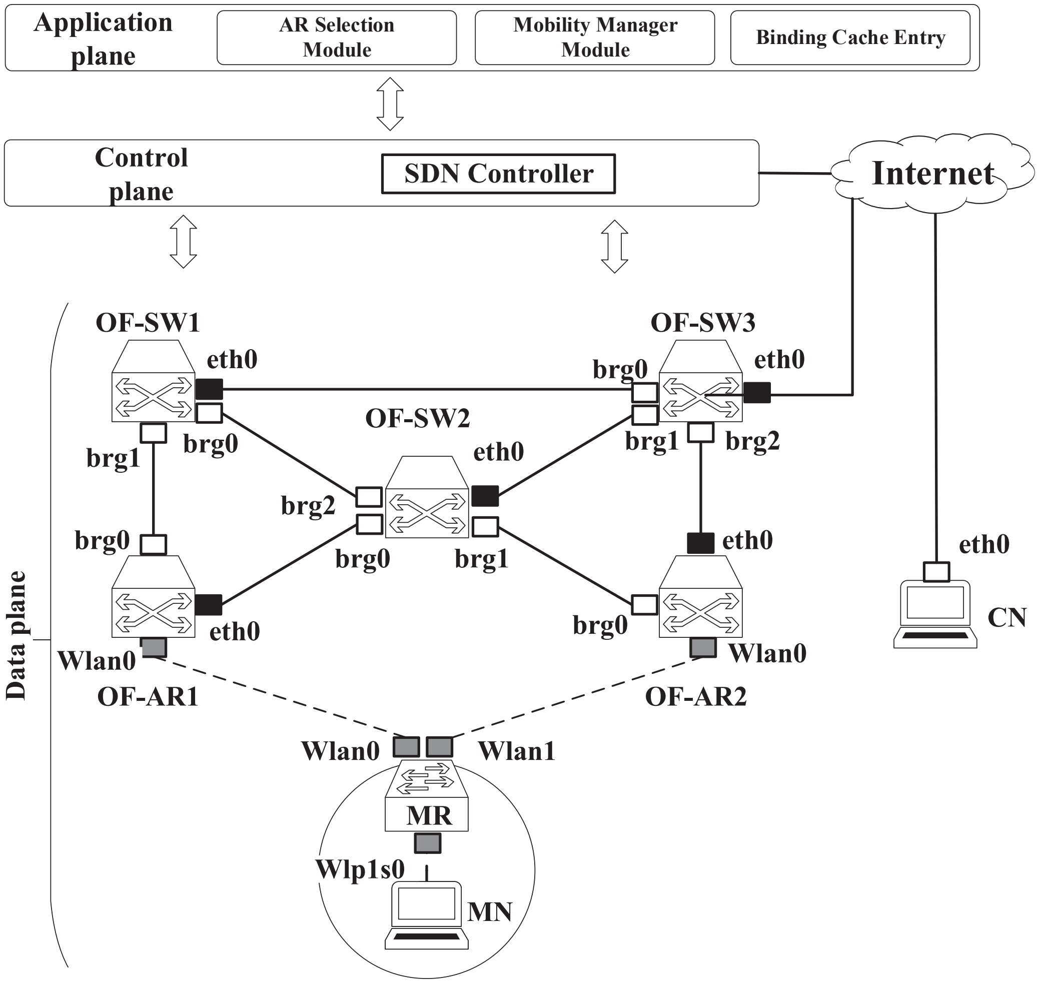


Fig. 5. Test bed architecture.

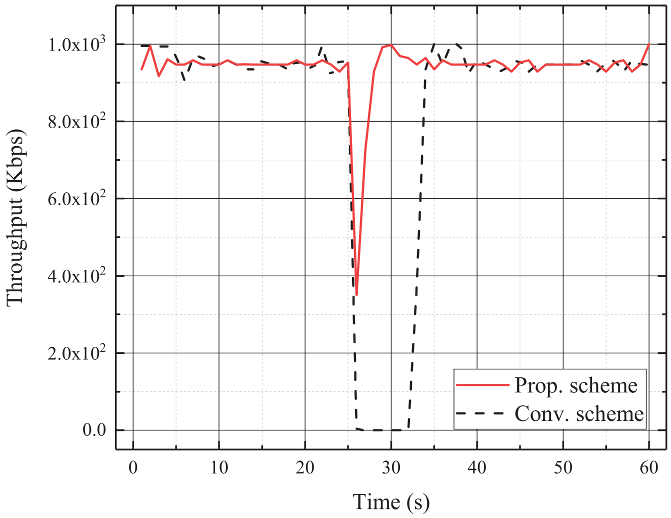


Fig. 6. Throughput of MN during handover

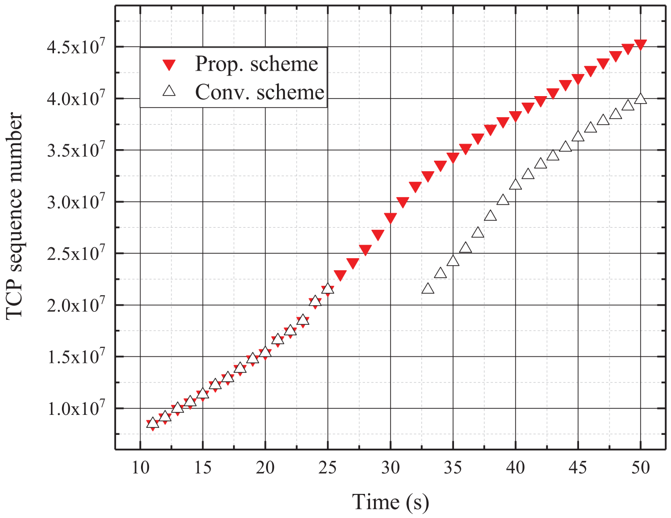


Fig. 7. Delay experienced by MN during handover

from any other access point/router. Consequently, there is insignificant difference in delay values from time 10s to 25s. The handover occurs at time t=25s approximately. At time t = 25.0125s, the CN receives the last packet with sequence number 21472393 via AR1, and at t = 25.0845s the first packet of sequence number 21473841 arrives at the CN via AR2. Therefore, the handover delay for the proposed scheme is 72msec. In comparison, the conventional scheme experiences handover delay of 7.15s. Also, the proposed scheme provides seamless handover, whereas in case of conventional scheme, the MN loses connectivity during the handover.

V. Conclusion

In this paper, we proposed a SDN–based MR multihoming scheme for seamless handover of MR and AR in a mobility scenario. We introduced an AR selection module at the control plane that selects an AR from available ARs for handover considering the downlink bit rate information and throughput of the backhaul links of the available ARs in the mobility network. An experimental testbed was implemented to measure the performance of the proposed scheme in terms of the handover throughput and delay. The proposed scheme outperformed the conventional scheme and the results showed that the handover delays for conventional scheme and proposed scheme were 7.15s and 72ms, respectively. For the conventional scheme, the connection was dropped and throughput went to zero during the handover. However, in the proposed scheme, average throughput for the mobile node reached up to 950Kbps and it remained constant during the handover. In case of mobility, the new prefix information of access router is transmitted to the mobile router, whereas the old prefix information is set to be deprecated. This enables session continuity and the throughput is maintained during handover.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (2015-0-00567, Development of Access Technology Agnostic Next-Generation Networking Technology for Wired-Wireless Converged).

Min Young Chung is the corresponding author.

References

[1] A. Yegin, J. Park, K. Kweon, and J. Lee,“Terminal-centric distribution and orchestration of IP mobility for 5G networks,” *IEEE Communications Magazine*, vol. 52, no.11, pp. 86-92, Nov. 2014.

[2] D. Johnson, C. Perkins, & J. Arkko, “Mobility support in IPv6., Mobility Support in IPv6”, *IETF RFC 6275*, July 2011. (http://www.ietf.org).

[3] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, & B. Patil, Proxy mobile ipv6. 2008.

[4] V. Devarapalli, R. Wakikawa, A. Petrescu, P. Thubert, “Network Mobility (NEMO) basic support protocol,” *IETF RFC 3963*, Jan. 2005. (http://www/ietf.org).

[5] C. Ng, E. Paik, T. Ernst, & M. Bagnulo, “Analysis of Multihoming in Network Mobility Support,” *IETF RFC 4980*, Oct. 2007. (http://www/ietf.org).

[6] M. S. Hossain, M. Atiquzzaman, & W. Ivancic, “Performance evaluation of multihomed NEMO,” in *Proc. IEEE International Conference on Communications. (ICC)*, Ottawa, ON, Canada, pp. 5429–5433, Jun. 2012.

[7] P. Sornlertlamvanich, T. Ang-Chuan, S. Sae-Wong,T. Kamolphiwong, & Kamolphiwong, “SDN-based Network Mobility”, in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Phuket, Thailand, pp. 1-6, Oct. 2016.

[8] B. K. Bae, M. Shin, & M. Y. Chung, “An Efficient Scheme for Supporting Network Mobility in Partially Distributed Mobility Management”, in *Proc. TENCON 2018-2018 IEEE Region 10 Conference*, Jeju, South Korea, pp. 1039-1044, 2018.

[9] ON.LAB, “Introducing ONOS - a SDN network operating system for Service Providers”, *White paper*, 2014.

[10] Raspberry Pi, Accessed on May 11, 2020. [Online]. Available: https://www.raspberrypi.org/.

[11] Open vSwitch, Accessed on May 15, 2020. [Online]. Available: http://openvswitch.org/.

[12] hostapd: IEEE 802.11 AP, IEEE 802.1X/WPA/WPA2/EAP/RADIUS Authenticator, Accessed on June 11, 2020. [Online]. Available: https://w1.fi/hostapd/.

Privacy-Preserving Zero-effort Class Attendance Tracking System

Aidan Shene
Department of Computer Science
University of North Carolina Wilmington
Wilmington, NC
aidanshene@gmail.com

Jake Aldridge
Department of Computer Science
University of North Carolina Wilmington
Wilmington, NC
jakealdridge6274@gmail.com

Hosam Alamleh
Department of Computer Science
University of North Carolina Wilmington
Wilmington, NC
hosam.amleh@gmail.com

***Abstract*—Student attendance tracking is a vital process in education. This process can be tiring and time-consuming. We believe it is possible to automate it using technologies available in the educational infrastructure and user smartphones. Today smartphones can sense several types of signals over the air using radio frequency technologies (e.g., Wi-Fi, Bluetooth, cellular signals, etc.). Furthermore, smartphones receive broadcast messages from transmitting entities and can measure the received signal strength. We believe that these signals can be utilized in the context of classroom attendance tracking, primarily because they can indicate the location of a user's device. The proposed system aims to have student smartphones in the classroom generate "location proofs" based on the radio frequency fingerprints scanned by their devices, which are later used to verify their locations. In this paper, we propose the utilization of Wi-Fi access points in buildings on school campuses in conjunction with the instructor and student smartphones to build a zero-effort and privacy-preserving attendance tracking system. Our system is unique as it does not require any effort from users in the system. Moreover, it is privacy-preserving, as the App server has no information about user identities nor class locations.**

***Index Terms*—Class attendance, tracking, Wi-Fi, Privacy, zero-effort**

I. Introduction

Attendance tracking is necessary for both students and instructors. If students miss lessons they lose out on information essential to their learning that may be the basis for future classes. On the other hand, instructors also suffer. It is difficult for instructors to determine whether the course material is being taught adequately without student feedback. Studies indicate a strong positive correlation between attendance and grades, with mandatory attendance policies having an additional positive effect [1]. One randomized experiment conducted in a class of 114 students yielded an improvement of 9.4 to 18.0 percent on exams, depending on the frequency of attendance [2]. This observation makes streamlining attendance tracking and incentivization paramount in enhancing students' overall academic performance. Traditional attendance tracking methods often exhaust a substantial portion of class time, preventing instructors from maximizing instruction time. Additionally, conventional strategies are particularly susceptible to dishonesty from students regarding truthfulness about attendance. The drawbacks of traditional methodology call for a more modern and foolproof approach towards the matter, especially with the technology available today in the educational infrastructure (e.g., wireless networks) and user devices (e.g., laptops, tablets, and smartphones).

Today, there has been increasing utilization of smartphones. According to Pew research center [3], 81 percent of Americans own smartphones. Smartphones can sense several types of signals over the air using divergent radio frequency technologies (e.g., Wi-Fi, Bluetooth, cellular signals, etc.). Furthermore, smartphones receive broadcast messages from transmitting nodes (e.g., network access points, cellular phone towers, etc.), in which it receives connectivity information about the transmitting nodes. Also, smartphones can measure the received signal strength of the broadcast messages transmitted b these entities. With the availability of these technologies in the classroom, we believe it is plausible to exploit them for class attendance. In this paper, we explore the feasibility of using existing technology in educational infrastructure to build an automated, zero-effort, and privacy-preserving class attendance system. The proposed system utilizes the Received Signal Strength Indicators (RSSI) measured in dbm that instructor and student smartphones collect from nearby network access points to determine room-level proximity between the student devices and the instructor device. Then, this room-level proximity determination is used to track student attendance.

II. Background

This section surveys previous attempts at utilizing technology to automate attendance tracking in educational settings. Hoo and Ibrahim [4] discussed using biometric recognition systems such as the iris, voice, fingerprint, and face to validate user information through systems that do not suffer from the same drawbacks as other techniques. However, their system requires integrating potentially expensive and complex hardware into an educational environment. Moreover, there are also significant security and privacy concerns associated with collecting biometric data from users. Another project proposes using location and voiceprint [5]. However, their system requires users to speak to verify their identity, which might cause distractions in the classroom. Other systems propose using QR codes [6], RFID [7], [8], and a secret code [9]. However, these systems need extensive interaction from students. Another technique utilizes an android application in which the user stores the GPS location and radius of a room [10]. Once the users enter the room they are clocked in and clocked out when they leave using GPS on their android devices. However, GPS suffers from performance limitation indoors and it would be challenging to use GPS-based systems in educational buildings as many of these buildings are multi-story.

Other research projects propose using Bluetooth and Wi-Fi. These technologies are a popular method for indoor localization. There are many research projects that produce indoor location for Wi-Fi using Wi-Fi fingerprint [11], [12], Bluetooth fingerprint [13], Wi-Fi beacons [14], [15], [16], and Bluetooth beacon [17]. Some of these systems were used in the context of attendance tracking using Wi-Fi [18] and Bluetooth [19]. However, these systems did not take any actions to ensure user privacy. Moreover, using Bluetooth beacons requires additional hardware that may not be available in the infrastructure. The related work demonstrates the difficulty of balancing user privacy and retaining the ability to verify the authenticity of attendance data without requiring extra effort from participants. Our approach utilizes previously integrated technology, does not send to the server any information that could lead to the session location, and perform attendance tracking without any effort from students nor instructors (we call it a zero-effort system).

III. Method

In this section, we discuss the system design and operations. As shown in Fig. 1., The proposed system consists of user devices and the application server. Then, we have the Wi-Fi access points in the infrastructure. The user devices scan Wi-Fi access points in the area. Then, the MAC addresses of the Wi-Fi access points are hashed and then uploaded along with the measurements to the application server. The application server compares the measurements and decides whether student devices are within room level proximity to the instructor device. Based on that, the attendance decisions are made and sent to the instructor device. Each element of the system is discussed in detail in the following subsections:

A. User device

In the proposed system, user devices are the student and the instructor smartphones with the mobile application installed. The mobile application has two types of users: the students and the instructor. Each student user has a unique userID, generated by hashing the student name with the course number used as a salt. During the class, the mobile application performs a Wi-Fi scan which returns the MAC addresses of nearby access points and their measured RSSIs. The devices then send the location proof to the application server along with the userID of the student. Location proof is shown in (1):

$$||_{i=0}^{n} Hash(MAC_{Api}).RSSI_{Api} \quad (1)$$

Where $||$ denotes concatenation. MAC_{Api} is the MAC address of access point i. $RSSI_{Api}$ is the RSSI measured for access point (dbm)i.

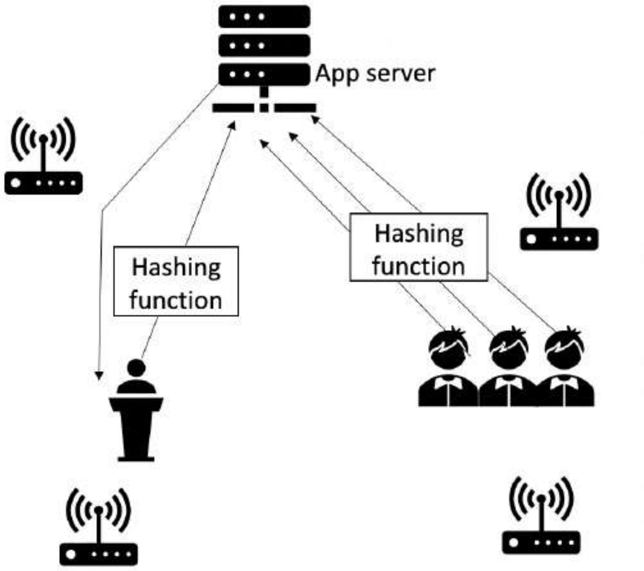


Fig. 1. System architecture

After the application server receives the data an attendance decision is made. The application server sends the decision to the instructor device where the userID is matched with the student name.

B. App server

As discussed in the previous section, the location proof is sent to the application server. The application server calculates the Euclidian distance [20] between every student datum and the instructor data as shown in (2):

$$D = \sum_{i=0}^{n} (InstructorRSSI_{Api} - StudentRSSI_{Api})^2 (2)$$

Where D is the Euclidian distance. If D is below a certain defined threshold, the attendance decision is present. If D is above this threshold, the attendance decision is absent. Finally, the application server relays userIDs and attendance decisions to the instructor device where userIDs are matched with student names.

The proposed system is zero-effort because no extra effort is required from the student to record attendance. All that it requires is the mobile app running in the background. Moreover, the proposed system is privacy-preserving as not the student names nor the MAC addresses of the access points are known by the application server. Thus, the server has no information about the identity of the students nor where the class is taking place.

IV. Experiment and Results

To test the proposed system, an experiment was conducted. The experiment took place inside a classroom in a campus building. We collected data from a total of 15 different locations in the building. Nine from inside the classroom and six from other locations in the building as shown in Fig. 2. Data collection was conducted using an android device with an application that scans for Wi-Fi access points and uploads the result to a server. A data point was collected every minute for one hour to provide sufficient data to analyze. To determine whether the student device is in the attendance range, the instructor device was set as a reference point. In the experiment, the instructor podium was used to collect the instructor data. The collected data was parsed in Python using the Pandas library DataFrame class. RSSI values for each access point MAC address were sorted for each data set. The top ten MAC addresses by RSSI were taken from the reference set. These MAC addresses were found in the other data sets along with their top ten RSSI values.

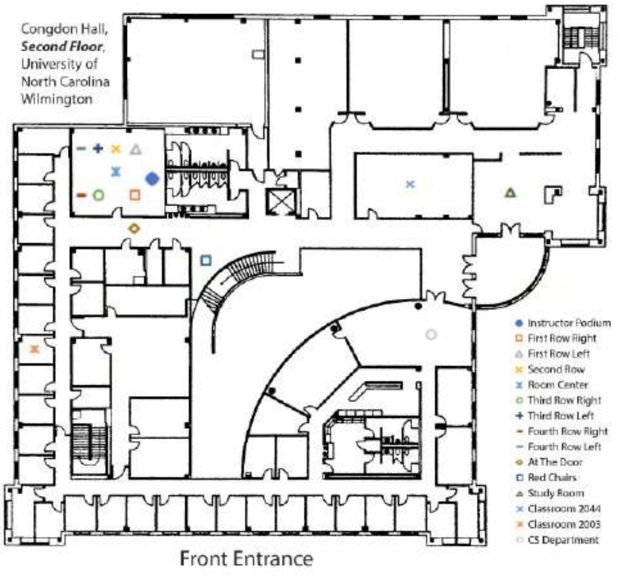


Fig. 2. Collected test points locations

After data collection, the ten access points with the highest average RSSIs were selected from the instructor podium as the reference set. Corresponding averages for the access points from the other datasets were aggregated into a single table to train the classifier. Visualization of the recorded data is shown in Fig. 3. Then, the Euclidean distance was calculated for each data set and reference point as in (2). The Euclidian distance between each test location and the reference set was calculated. An RSSI threshold value of 56.5 was found to provide the most accurate results. This threshold represents the maximum Euclidean distance value a data set can return to be classified as ”inside” the room. If a value over 56.5 is returned, the data is classified as ”outside” of the room.

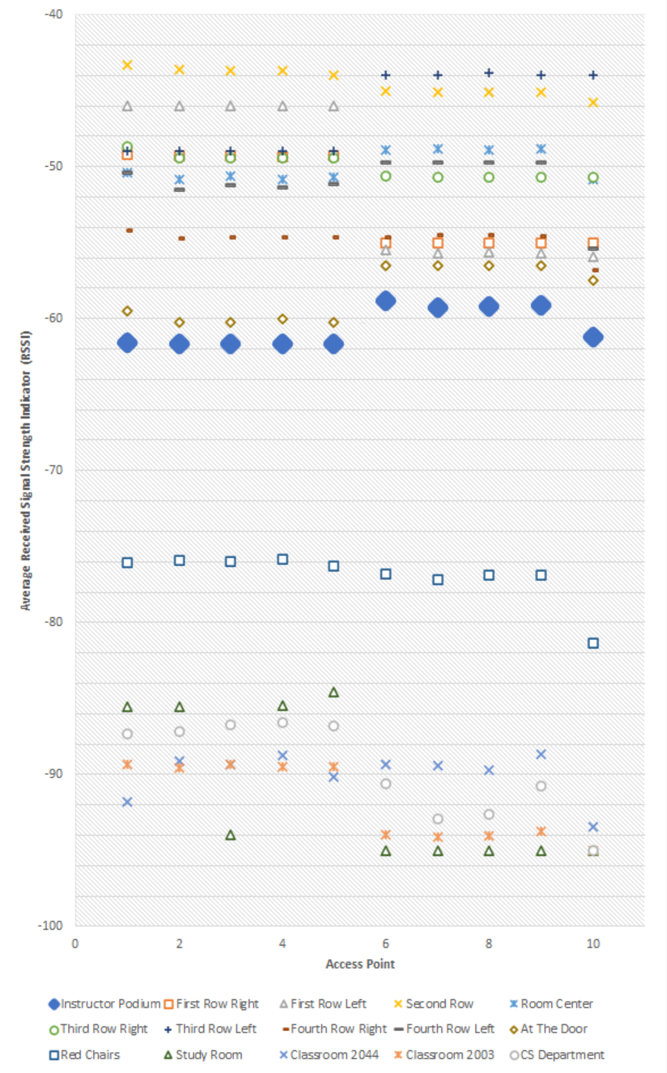


Fig. 3. Average RSSI for datapoints collected at different locations

To test our classifier, 198 test data points were collected. The Euclidean distance was calculated between each of the data points and the reference set. A threshold of 56.5 is considered to classify if the data point is inside or outside the classroom. The test results are shown in Fig. 4. Error reporting on the classification demonstrates that predictions for locations truly inside of the classroom were correct 100 percent of the time, while predictions for locations truly outside of the classroom were correct 98.75 percent of the time. The accounted for error is a result of the measurements recorded from the location at the door, which triggered a false positive in the classification due to its proximity to the classroom itself and the instructor podium.

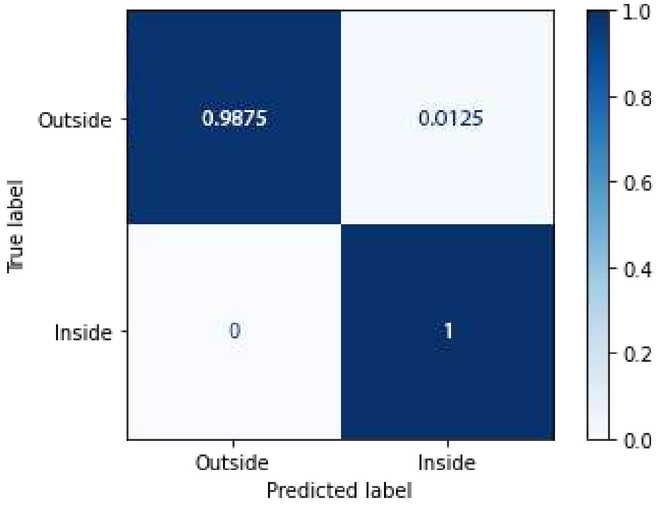


Fig. 4. Test results

V. Conclusion

Student attendance tracking is important in delivering quality education. To optimize the use of the class time we propose a system to automate class attendance tracking utilizing technologies available in the educational infrastructure and user smartphones. In the proposed system, student smartphones in the classroom generate “location proofs” based on the Wi-Fi RSSI fingerprints scanned by the user devices. The proofs are later used to verify user locations. Verification is completed by comparing the data from student devices with the data measured by the instructor device. Our system is unique as it does not require any effort from participating users, as scan and upload of data can happen in the background without any user involvement. Moreover, it is privacy-preserving, as the App server has no information about user identities nor class locations. Privacy is maintained by hashing the student names and access point MAC addresses and salting the student names before sending data to the App server.

References

[1] Chen, Jennjou, and Tsui-Fang Lin. “Class Attendance and Exam Performance: A Randomized Experiment.” The Journal of Economic Education, vol. 39, no. 3, 2008, pp. 213–227. JSTOR, www.jstor.org/stable/43608749. Accessed 2 Feb. 2021

[2] Credé M, Roch SG, Kieszczynka UM. Class Attendance in College: A Meta-Analytic Review of the Relationship of Class Attendance With Grades and Student Characteristics. Review of Educational Research. 2010;80(2):272-295. doi:10.3102/0034654310362998

[3] Pew Research center. 2020. Mobile fact sheet. retreived from https://www.pewresearch.org/internet/fact-sheet/mobile/

[4] Hoo, Seng Chun, and Haidi Ibrahim. “Biometric-Based Attendance Tracking System for Education Sectors: A Literature Survey on Hardware Requirements.” Journal of Sensors, 15 Sept. 2019, pp. 1–25., doi:10.1155/2019/7410478.

[5] S. Yang, Y. Song, H. Ren and X. Huang, ”An automated student attendance tracking system based on voiceprint and location,” 2016 11th International Conference on Computer Science & Education (ICCSE), Nagoya, Japan, 2016, pp. 214-219, doi: 10.1109/ICCSE.2016.7581583.

[6] A. Nuhi, A. Memeti, F. Imeri and B. Cico, ”Smart Attendance System using QR Code,” 2020 9th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2020, pp. 1-4, doi: 10.1109/MECO49872.2020.9134225.

[7] D. Eridani and E. D. Widianto, ”Simulation of attendance application on campus based on RFID (radio frequency identifiation),” 2015 2nd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 2015, pp. 460-463, doi: 10.1109/ICITACEE.2015.7437850. C

[8] . Sai Krisha, N. Sumanth and C. Raghava Prasad, ”RFID based student monitoring and attendance tracking system,” 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 2013, pp. 1-5, doi: 10.1109/ICCCNT.2013.6726702. C

[9] T. J. Zhi, Z. Ibrahim and H. Aris, ”Effective and efficient attendance tracking system using secret code,” Proceedings of the 6th International Conference on Information Technology and Multimedia, Putrajaya, Malaysia, 2014, pp. 108-112, doi: 10.1109/ICIMU.2014.7066613.

[10] Sultana, Shermin, et al. “A SMART, LOCATION BASED TIME AND ATTENDANCE TRACKING SYSTEM USING ANDROID APPLICATION.” International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol. 5, no. 1, Feb. 2015, pp. 1–5., doi:10.5121/ijcseit.2015.5101.

[11] A. A. S. AlQahtani, H. Alamleh and J. Gourd, ”0EISUA: Zero Effort Indoor Secure User Authentication,” in IEEE Access, vol. 8, pp. 79069-79078, 2020, doi: 10.1109/ACCESS.2020.2990604.

[12] Wang, W., Chen, Y. and Zhang, Q., 2016. Privacy-Preserving Location Authentication in Wi-Fi Networks Using Fine-Grained Physical Layer Signatures. IEEE Transactions on Wireless Communications, 15(2), pp.1218-122

[13] Q. Zhang, M. D’souza, U. Balogh and V. Smallbon, ”Efficient BLE Fingerprinting through UWB Sensors for Indoor Localization,” 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, 2019, pp. 140-143, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00065.

[14] Huseynov, E. and Seigneur, J., 2015. WiFiOTP: Pervasive two-factor authentication using Wi-Fi SSID broadcasts. In ITU Kaleidoscope: Trust in the Information Society (K-2015). pp. 1-8.

[15] Pandey, S., Anjum, F., Kim, B. and Agrawal, P., 2006. A low-cost robust localization scheme for wlan. In WICON ’06 Proceedings of the 2nd annual international workshop on Wireless internet.

[16] H. Alamleh and J. Gourd, ”Unobtrusive Location-based Access Control Utilizing Existing IEEE 802.11 Infrastructure,” 2020 IEEE Microwave Theory and Techniques in Wireless Communications (MTTW), Riga, Latvia, 2020, pp. 157-162, doi: 10.1109/MTTW51045.2020.9245032.

[17] Vaščák, J. and Savko, I., 2018. Radio Beacons in Indoor Navigation. In World Symposium on Digital Intelligence for Systems and Machines (DISA).

[18] S. Anand, K. Bijlani, S. Suresh and P. Praphul, ”Attendance Monitoring in Classroom Using Smartphone & Wi-Fi Fingerprinting,” 2016 IEEE Eighth International Conference on Technology for Education (T4E), Mumbai, India, 2016, pp. 62-67, doi: 10.1109/T4E.2016.021.

[19] B. Zorić, M. Dudjak, D. Bajer and G. Martinović, ”Design and development of a smart attendance management system with Bluetooth low energy beacons,” 2019 Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, 2019, pp. 86-91, doi: 10.1109/ZINC.2019.8769433.

[20] Kapil, A. K. (2018, March 18). Euclidean distance [Illustration]. KNN- DISTANCE METRICS. https://www.datavedas.com/knn-distance-metrics/

A Radiation-Hardened Wide Spectral Response 4T CMOS Image Sensor Pixel Employing a Through-Silicon Via based Photodetector

Abdollah Pil-Ali
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
apilali@uwaterloo.ca

Razieh Khalili
Faculty of Sciences
University of Guilan
Rasht, Iran

S. Mohammad Reza Safaee
Electrical and Computer Engineering Faculty
McGill University
Montreal, Canada

Mohammad Azim Karami
School of Electrical Engineering
Iran University of Science and Technology
Tehran, Iran

***Abstract*—A novel radiation-hardened 4-transistor (4T) complementary metal-oxide-semiconductor (CMOS) image sensor pixel is proposed, employing a through-silicon-via (TSV) based metal-semiconductor-metal (MSM) photodetector as the photosensitive element. The presented pixel structure is designed and simulated in 0.18 μm CMOS technology. It is shown that deploying the TSV-based MSM photodetector decreases the pixel sensitive area to 3 μm^2 compared to conventional planar MSM ones, and enhances both the lower and higher wavelengths of the spectral response characteristic of the pixel as well. The designed pixel has a 2.75 μm radius circular structure, the dark current of 77 fA at 1 V reverse bias (at room temperature), a conversion gain of 116 μV.e^{-1}, a dynamic range of 58.62 dB, and an approximately linear spectral response in the range of 300 to 800 nm wavelengths of the incident light spectrum. The proposed pixel structure, enabled with a p+ guard ring, also demonstrates insignificant dark current degradation under gamma ray radiation of up to 1 Mrad.**

***Index Terms*—4T CMOS Image Sensor, CMOS IS, Wide Spectral Response, Radiation Hardened, Radiation Tolerant, Ionizing Radiation, Total Ionizing Dose, TID, p+ guard ring, Enclosed Layout, TSV, MSM.**

I. Introduction

A wide spectral response in parallel with high quantum efficiency (QE) performance is one of the key specifications of CMOS image sensor pixels required by a variety of applications such as astronomical, medical, as well as scientific imaging [1]–[3]. Due to the long absorption depth of light in silicon, which is near 10 μm for 800 nm wavelength of incident light [4], conventional pn-junction and planar metal-semiconductor-metal (MSM) silicon-based photodetectors (PDs) have the limitation to achieve a high spectral response at the whole visible spectrum simultaneously [5], [6]. In order to obtain a wide spectral response with high QE pixel, several attempts with different corresponding pixel structures have been made [7]–[11]. Since concurrently achieving both latter objectives in silicon-based PDs needs a relatively deep sensitive area, in [10] a lateral p-i-n photodiode was proposed deploying trenches to build a deep space-charge-region in the substrate depth, which could demonstrate an extended depletion region far below the photodetector's surface. On the other side, there are various individual attempts to enhance the spectral response of MSM photodetector implemented on silicon substrate [5], [6], [11]–[13]; however, these designs suffer from incompatibility with standard CMOS technology. In [14], a different type of MSM photodetector was proposed using modern TSV technology. This metal-trench-semiconductor photodetector (MTS-PD) creates a deep space-charge-region, where a high QE in UV to NIR range can be attained. It is noteworthy that this trench-like structure has been used in many applications such as metal-oxide-semiconductor capacitor [15].

Application of electronic devices under radiation environments could lead to displacement damage (DD) and total ionizing dose (TID) effects. These effect may potentially degrade electronic devices' performance [16]. The superiority of MSM photodetectors compared to pn-junction based ones has been demonstrated in terms of higher tolerance to TID effect under radiation environments [17], [18]. Deploying proper radiation-hardening design such as guard ring structures would help minimizing TID effect [19], although a combination of an MSM-based PD with the latter techniques could boost the performance of the radiation-hardened design.

In this paper, a 4-transistor (4T) CMOS image sensor pixel enabled by the MTS-PD as the sensitive area is proposed to reach a wider spectral response. Employing metallic-filled TSV-based PD to create a deep Schottky junction inside the silicon substrate enlarges the space-charge-region, and affords the absorption of low energy (i.e. high wavelength) photons as well.Furthermore, a simple p+ guard ring which surrounds the whole pixel design, along with the MTS-PD structure,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

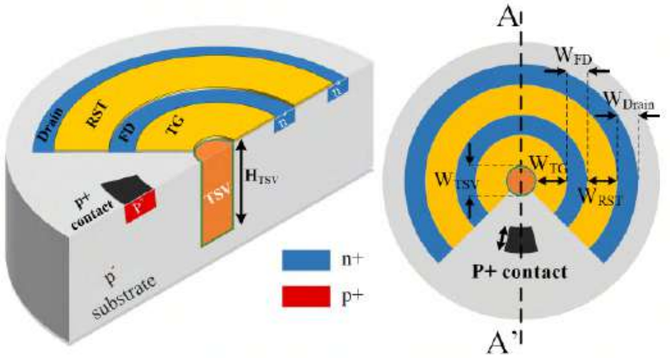
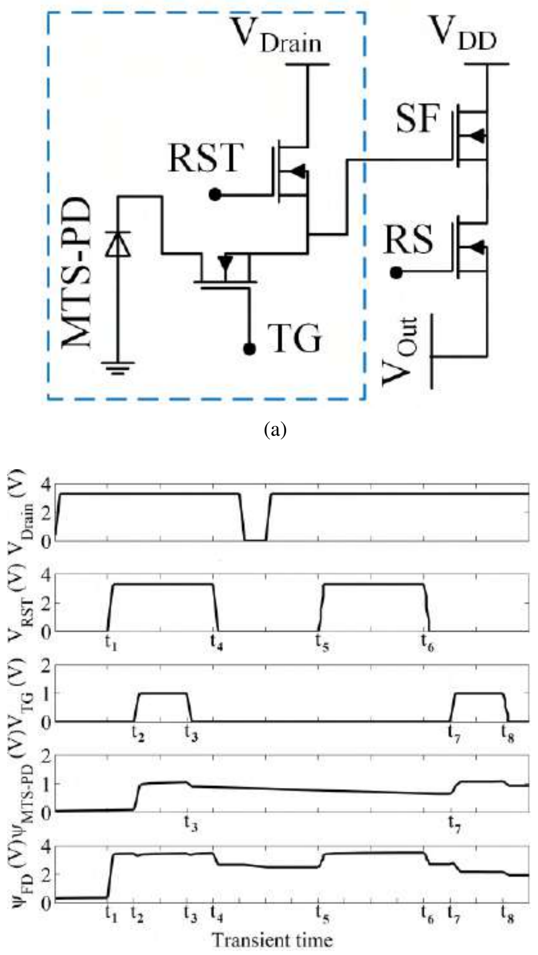


Fig. 1: Illustration of the proposed pixel structure. Top view (right), and perspective cross-sectional schematic view (left) along the AA' cutline of the proposed pixel structure.

demonstrates a radiation tolerance of up to 1 Mrad of gamma ray radiation.

II. Pixel Structure

Fig 1 illustrates the top and perspective schematic view of the proposed pixel including the photosensitive element - which is the TSV-based Schottky junction, floating diffusion (FD), transfer gate (TG), and the reset transistor (RST). The pixel is designed based on 0.18 μm standard CMOS process technology parameters, and simulations are performed using SILVACO TCAD simulator [20]. In order to make a Schottky junction inside the p-type silicon substrate according to the standard CMOS technology, Copper (Cu) can be used as a conductive metal to fill the TSV. The trench can be made, employing reactive ion etching (RIE) process, inside the substrate to etch the silicon for 10 μm in depth and 0.5 μm in diameter [21]. Creating deeper trenches are also possible using deep-RIE (Bosch DRIE process). Prior to filling the trench with Cu, the interior part of the trench wall should be covered with 50 nm of Titanium Nitride as a diffusive barrier to prevent Cu diffusion into the silicon, through a chemical vapor deposition (CVD) process [21]. For deeper trenches, this step can be done through an atomic layer deposition (ALD) process to ensure a complete step coverage of the trench interior parts. Following the MTS-PD design [14], the pixel structure is designed and simulated based on a cylindrical geometry which reduces the edge electric field distortion, and consequently produces a radial electric field around the Schottky contact (a radial deep space-charge-region). TG and RST are created using poly-silicon as a transparent metallic contact both with 1 μm width and a 270 degrees of a circle sector shape. The FD and the drain of the reset transistor (Drain) are defined sequentially with 10^{+19} atoms.cm^{-3} of n^{+} doping and metal contact. The p-type silicon substrate contact placement is designed through a 30 degrees sector shape, as illustrated in Fig 1 and is named p^{+} contact. The main geometrical dimensions of the proposed pixel are as W_{TSV}=0.5 μm, H_{TSV}=10 μm, W_{TG}=W_{RST}=1 μm, and W_{FD}=W_{Drain}=0.25 μm.



(a)

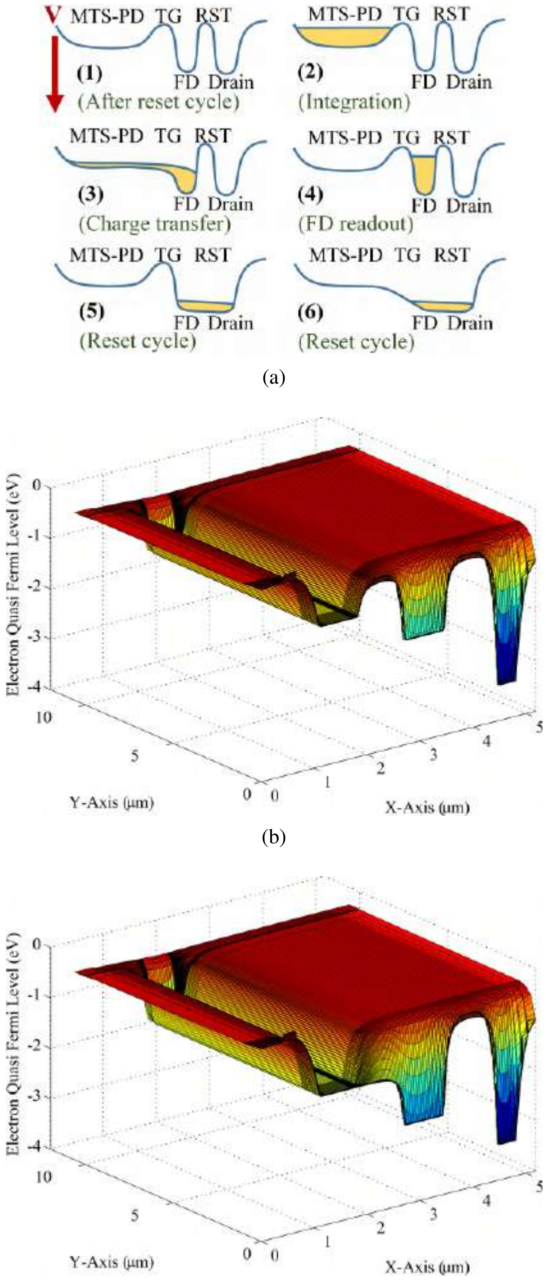
(b)

Fig. 2: (a) Circuit schematic of the proposed pixel; the dashed-line area which consists of sensitive area (MTS-PD), TG, FD, and RST is the circuit diagram of proposed structure. (b) Readout timing diagram illustration for one full cycle of pixel operation, including reset, illumination, and readout. The voltage of drain (V_{Drain}), the RST gate (V_{RST}), the transfer gate (V_{TG}), and the potential of MTS-PD (ψ_{MTS-PD}) and FG (ψ_{FD}) are illustrated in the timing diagram. The device has been illuminated between t_3 and t_7.

III. Results and discussion

A. Readout Timing Diagram

Fig 2(a) illustrates the circuit schematic and Fig 2(b) is a representation of the readout timing diagram in a full cycle of the pixel operation. During the reset cycle between t_1 and t_4, the RST and TG are both set to the high digital voltage level of 3.3 V, so that the MTS-PD potential well could be discharged from any charges stored during previous phases. Then, as the pixel is illuminated, the photo-generated carriers collection is triggered as soon as the TG sets to the low digital voltage level – from t_3 to t_7. During the illumination process and before it ends, a reset cycle will be conducted again — between t_5 and t_6 — in order to make the FD empty



(a)

(b)

(c)

Fig. 3: (a) Illustration of a complete cycle of one pixel operation in six steps including: (1) a prepared pixel after a reset cycle, (2) illumination and photo-generated charges storage, (3) transformation of charges to FD-node, (4) reading the FD-node, and (5, 6) a reset cycle; (b) Electron Quasi Fermi level when TG is off and pixel is ready to collect photo-generated charges; (c) Electron Quasi Fermi level when TG is on and photo-generated charges are being transferred to FD

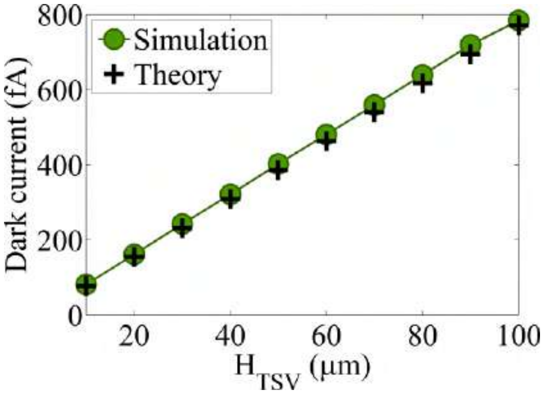


Fig. 4: Dark current evolution for different TSV depths. A linear increment of dark current versus the TSV depth is observed.

of any charges and to prepare this node for the next part of the readout cycle. Following this step, between t_7 and t_8, the stored photo-generated charges in the MTS-PD potential well will be transferred to the FD by lowering the TG potential barrier by setting the TG to a high-level value of V_{TG}, equal to 1.3 V. Albeit the high-level of TG is chosen to decrease the TG potential barrier but not to affect the potential level of the MTS-PD well, an approximately 0.15 V reduction in the MTS-PD potential level occurs due to the TG electric field effect. Simulation results of Fig 2 is based on 0.01 W-cm^{-2} of incident light power during 100 μsec illumination – between t_3 and t_7 [22]. The electrical potential drop in the MTS-PD under this incident light power is around 0.25 V, which is in correspondence with 0.76 V of electrical potential drop in the FD after transferring the stored photo-generated charges to the FD node. Fig 3(a) is an illustration of different steps during one full cycle of pixel operation including reset, illumination, and the FD-readout based on the readout timing diagram in Fig 2. Fig 3(b) and (c) show the 3D simulation results of the Electron Quasi Fermi level at two different scenarios, when TG is off (b) and on (c). The 3D simulation results demonstrate how lowering TG could facilitate the photo-generated charges transfer into the FD node through surface areas.

B. *Dark Current*

Simulation and verification of dark current for the designed MTS-PD has been previously investigated in [14]. Fig 4 illustrates the simulation and theoretical results of dark current evolution for different TSV depths. By the increase of the TSV depth from 10 μm to 100 μm, the dark current rises from 77 fA to 780 fA at 1 V reverse bias and 300 K temperature.

C. *Capacitance of MTS-PD and Floating Diffusion*

The dynamic range and the conversion gain calculation requires the exact capacitance value of MTS-PD and FD-node. Simulation results, showing the capacitance of the MTS-PD and FD-node, are illustrated in Fig 5. A 10 μm deep TSV shows a capacitance of 6.77 fF at 1 V reverse bias; the FD-node demonstrates a capacitance of 1.31 fF at 3.3 V reverse bias with 1 μm width and 270 degrees of a circle sector shape.

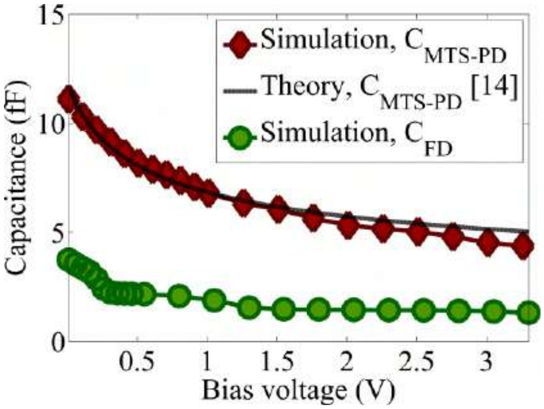


Fig. 5: The MTS-PD capacitance value attained by both simulation and theoretical analysis [14] as well as the simulation result of FD capacitance node under different bias voltage conditions.

Theoretical derivations to verify the MTS-PD capacitance are stated in (1) and (2):

$$C_{MTS-PD} = \frac{2\pi\varepsilon_{Si}L}{Ln\left(\frac{W+R_M}{R_M}\right)} \qquad (F) \tag{1}$$

$$W = \sqrt{\frac{2\varepsilon_{Si}\left(V_{bi} - V_a - {KT}/{q}\right)}{qN_A}} \qquad (m) \tag{2}$$

where C$_{MTS-PD}$ is MTS-PD capacitance; ϵ_{Si} is the silicon permittivity (due to standard CMOS usage); W is the depletion width around the TSV; V_{bi} is the built-in potential; V_a is the applied reverse bias; K is Boltzmann's constant; T is the temperature in Kelvin; q is the electron charge, N_A is the semiconductor acceptor concentration density; and L and R_M are the depth and radius of the TSV [4].

Enlarging the TSV depth increases its capacitance value and provides a greater charge storage well, which consequently leads to an enhancement in the spectral response at higher wavelengths. On the other hand, larger FD-node capacitance would result in a lower conversion gain value. Accordingly, from the design perspective, the FD-node capacitance value is determined based on the MTS-PD capacitance (small enough, 1.31 fF at 3.3V reverse bias, to obtain a considerable conversion gain, 116 μV-e$^-$); since higher conversion gain is advantageous toward minimizing the signal to noise ratio [23].

D. Conversion Gain

Conversion gain which presents the image sensor sensitivity is defined as the output generated signal voltage per photogenerated charge, and can be formulated as (3):

$$CG = \frac{q}{C_{FD}} \qquad (\mu V/e^-) \tag{3}$$

where C$_{FD}$ is the FD capacitance and q is the electron charge [24]. Since the C$_{FD}$ as shown in Fig 5, at 3.3 V reverse bias is calculated as 1.31 fF, the conversion gain will be 116 μV/e$^-$.

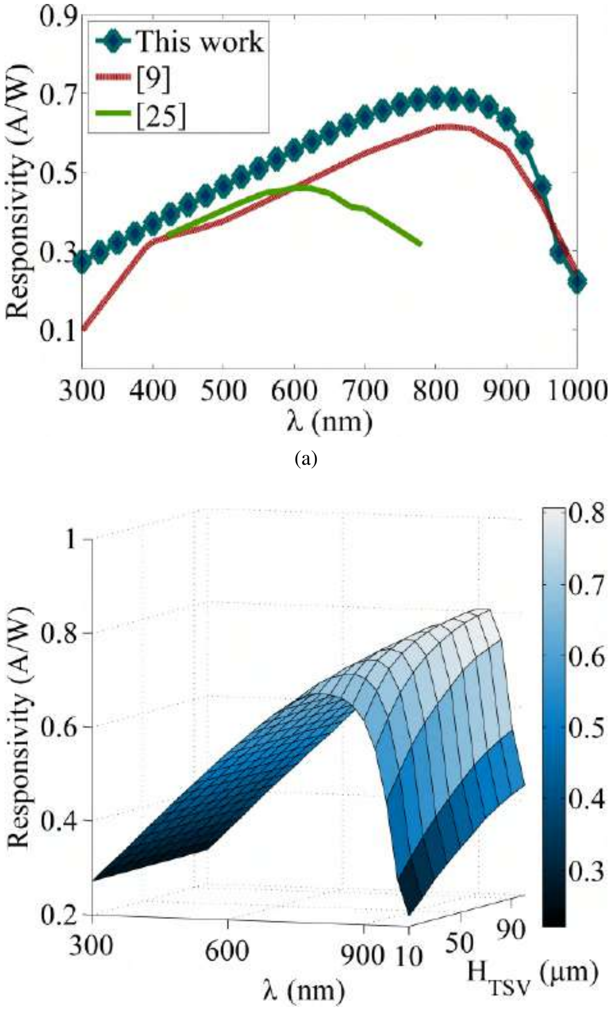


Fig. 6: (a) Spectral response derivation of the proposed pixel with 10 μm height of the TSV in comparison to [9] and [25]. (b) Simulation result of spectral response for different TSV depths (H$_{TSV}$) over different wavelengths of incident light (λ); as the TSV depth increases, the spectral response of higher wavelengths would be enhanced.

E. Spectral Response

Spectral response, or responsivity, of the proposed pixel–which is the ratio of photo-generated current to incident light power [4]–is compared with [9] and [25], two of the high spectral response image sensors implemented in CMOS standard technology. The result of this comparison is illustrated in Fig 6(a). The proposed pixel with a 10 μm deep TSV shows higher spectral response over the whole visible wavelengths; furthermore, TSV deepening causes an enhancement at higher wavelengths of spectral response because of the uniform depletion region extension within the silicon depth. Simulation result of the spectral response for different TSV depths (H$_{TSV}$) is also illustrated in Fig 6(b). In order to obtain such a result, several simulations with different depths of the

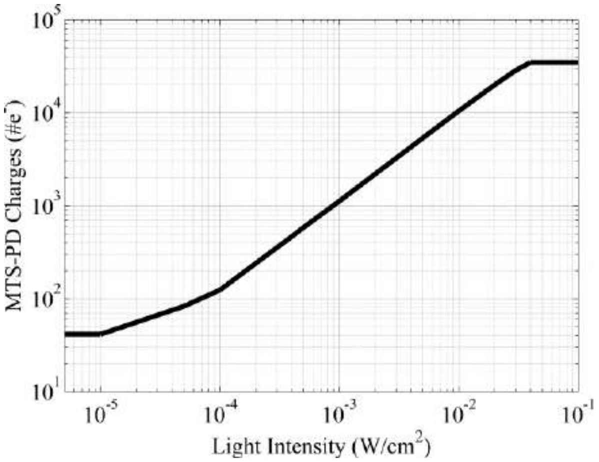
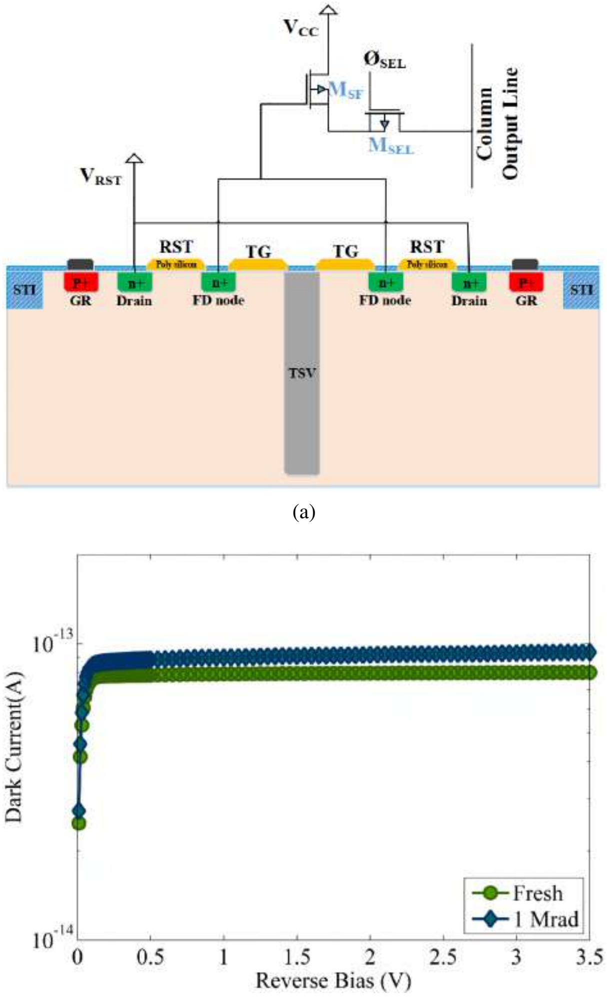


Fig. 7: Illustration of the photo-generated charges as a function of light intensity. The difference between minimum, dark current, and maximum, saturated-well of photodetector, is defined as the dynamic range.



(a)

(b)

Fig. 8: (a) Illustration of p+ guard ring (GR) around the pixel, (b) simulation results of radiation effect on the dark current of a pixel before (fresh) and after exposure of 1 Mrad gamma ray radiation.

TSV are performed while the photo-generated current is swept for each simulation over a range of incident light wavelengths between 300 to 1000 nm. The linearity behavior of the spectral response over the range of 300 to 800 nm wavelength is due to the radial deep space-charge-region, created by the MTS-PD and extended from the surface to the H_{TSV}; hence any photo-generated charges inside the space-charge-region can be collected by the engineered Schottky contact (MTS-PD). The proposed design demonstrates a responsivity of 0.27 $A.W^{-1}$ at 300 nm wavelength which increases linearly to 0.68 $A.W^{-1}$ at 800 nm wavelength for a 10 μm deep TSV. Making the TSV deeper to 100 μm inside the substrate enhances the spectral response of higher wavelengths, where the linearity behavior extends to 900 nm wavelength – the spectral response at this wavelength reaches 0.8 $A.W^{-1}$.

F. Dynamic Range

Dynamic range in image sensors is defined as the ratio of the maximum photo-generated current to the non-illuminated current (i.e. dark current) [26]. Fig 7 shows the number of photo-generated charges versus light intensity. From the simulation perspective, dynamic range is attained by changing light intensity from the dark to the state as which the MTS-PD well is saturated with photo-generated charges. Eventually, the difference between the minimum stored charge in the MTS-PD—related to the dark current—and saturated-well charges is translated as light intensity and a 58.62 dB dynamic range is derived for the proposed pixel through the simulation.

G. Radiation-Hardening

According to [17], [18], an MSM-based design on its own demonstrates radiation tolerance to a reasonable extent. However, in order to isolate a pixel from its neighboring pixels and to minimize the TID effect, a p+ guard ring is considered in our design. This p+ guard ring design has been previously investigated in details in [19]. This p+ guard ring not only acts as the substrate contact, but also protects the pixel from any leakage current presents outside of the pixel–originating mainly from oxide defects and shallow-trench isolation (STI) defects due to the TID effect. Fig 8(a) illustrates the p+ guard ring (GR) around the pixel design in a cross-sectional view of one pixel. Simulation results in SILVACO demonstrates a radiation tolerance with insignificant dark current increase under gamma ray radiation of up to 1 Mrad (Fig 8(b)).

IV. Conclusion

A new type of 4T CMOS image sensor employing a TSV based MSM photodetector as the photosensitive element is designed and simulated in this work. To the best of the authors' knowledge, this is the first time a TSV-based photodetector is used to realize a complete pixel structure. Table I represents the comparison of the proposed pixel simulation results with two other structures. The dark current of the proposed pixel

TABLE I: COMPARISON OF THE SPECIFICATION AND PERFORMANCE OF THE PROPOSED PIXEL

Specification	**Ref [27]**	**Ref [11]**	**This Work**
Technology	0.18 μm CMOS	na	0.18 μm CMOS
Pixel Size (μm^2)	41	na	28
Active Area (μm^2)	na	143000	3
Responsivity (A/W)	na	0.24 @ 300 nm 0.66 @ 800 nm	0.27 @ 300 nm 0.68 @ 800 nm
Conversion Gain (dB)	na	na	58.62
Full Well Capacity (e^-)	40000	na	35000
Dark Current	50 e^-	In the range of nano-ampere	77 fA @ 1 V (41 e^-)

(77 fA) is slightly higher than state-of-the-art designs, which is because of using an MSM-junction based photodetector rather than a pn-junction based one. A semi-linear behavior spectral response of more than 0.27 $A.W^{-1}$ at the wavelength spectrum from 300 to 800 nm is achieved within 3 μm^2 active area using the TSV based photodetector. The proposed pixel shows a conversion gain and dynamic range of 116 $\mu V/e^{-1}$ and 58.62 dB, respectively. Furthermore, deploying a p+ guard ring design, along with the metallic-filled TSV based photodetector, demonstrates negligible dark current increase under simulated gamma ray radiation of up to 1 Mrad. More importantly, this pixel architecture is compatible with standard CMOS technology. The output of this work could be a preliminary result toward realization and fabrication of a 4T CMOS image sensor pixel using TSV based MSM photodetectors for various applications, such as astronomical or medical, where the image sensor is supposed to work under radiation.

References

[1] J. Goy, B. Courtois, J. M. Karam, and F. Pressecq, "Design of an aps cmos image sensor for space applications using standard cad tools and cmos technology," in *Design, Test, Integration, and Packaging of MEMS/MOEMS*, vol. 4019. International Society for Optics and Photonics, 2000, pp. 145–152.

[2] K. S. Karim, Y. K. Vygranenko, A. Avila-Munoz, D. A. Striakhilev, A. Nathan, S. Germann, J. A. Rowlands, G. Belev, C. Koughia, R. Johanson *et al.*, "Active pixel image sensor for large-area medical imaging," in *Medical Imaging 2003: Physics of Medical Imaging*, vol. 5030. International Society for Optics and Photonics, 2003, pp. 38–47.

[3] K. Yoon, C. Kim, B. Lee, and D. Lee, "Single-chip cmos image sensor for mobile applications," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 12, pp. 1839–1845, 2002.

[4] S. M. Sze, Y. Li, and K. K. Ng, *Physics of semiconductor devices*. John wiley & sons, 2021.

[5] J. Y. Ho and K. Wong, "Bandwidth enhancement in silicon metal-semiconductor-metal photodetector by trench formation," *IEEE Photonics Technology Letters*, vol. 8, no. 8, pp. 1064–1066, 1996.

[6] L.-H. Laih, T.-C. Chang, Y.-A. Chen, W.-C. Tsay, and J.-W. Hong, "Characteristics of msm photodetectors with trench electrodes on p-type si wafer," *IEEE Transactions on Electron Devices*, vol. 45, no. 9, pp. 2018–2023, 1998.

[7] C. Yin and D. Hu, "High quantum efficiency p/sup+/-pi-n/sup-/-n/sup+/silicon photodiode," *IEE Proceedings J-Optoelectronics*, vol. 137, no. 3, pp. 171–173, 1990.

[8] K. De Munck, J. Bogaerts, D. Tezcan, P. De Moor, S. Sedky, and C. Van Hoof, "Backside thinned cmos imagers with high broadband quantum efficiency realised using new integration process," *Electronics Letters*, vol. 44, no. 1, pp. 50–52, 2008.

[9] K. De Munck, P. Ramachandra Rao, K. Minoglou, J. De Vos, D. Sabuncuoglu Tezcan, and P. De Moor, "Backside illuminated hybrid fpa achieving low cross-talk combined with high qe," in *IEEE Int. Image Sensor Workshop*, 2011, pp. 146–148.

[10] M. Yang, K. Rim, D. L. Rogers, J. D. Schaub, J. J. Welser, D. M. Kuchta, D. C. Boyd, F. Rodier, P. A. Rabidoux, J. T. Marsh, A. Ticknor, Q. Yang, A. Upham, and S. Ramac, "A high-speed, high-sensitivity silicon lateral trench photodetector," *IEEE Electron Device Letters*, vol. 23, no. 7, pp. 395–397, 2002.

[11] E. Budianu, M. Purica, F. Iacomi, C. Baban, P. Prepelita, and E. Manea, "Silicon metal-semiconductor–metal photodetector with zinc oxide transparent conducting electrodes," *Thin Solid Films*, vol. 516, no. 7, pp. 1629–1633, 2008.

[12] S. Yuanjie, J. Yadong, W. Zhiming, and Z. Guodong, "Spectral response of metal-semiconductor-metal photodetector based on black silicon," *Energy Procedia*, vol. 12, pp. 615–619, 2011.

[13] Y. Su, S. Li, Z. Wu, Y. Yang, Y. Jiang, J. Jiang, Z. Xiao, P. Zhang, and T. Zhang, "High responsivity msm black silicon photodetector," *Materials Science in Semiconductor Processing*, vol. 16, no. 3, pp. 619–624, 2013.

[14] A. Pil-Ali and M. A. Karami, "Through silicon via based metal-semiconductor-metal photodetector in cmos technology," *Optical and Quantum Electronics*, vol. 48, no. 1, p. 13, 2016.

[15] M. Zulkifeli, S. Sabki, S. Jamuar, S. Taking, and N. Azmi, "Effect on different geometric dimensions of metal-oxide-semiconductor capacitor by using tcad simulation," in *Electronic Design (ICED), 2016 3rd International Conference on*. IEEE, 2016, pp. 44–47.

[16] M. A. Karami, A. Pil-Ali, and M. R. Safaee, "Multistable defect characterization in proton irradiated single-photon avalanche diodes," *Optical and Quantum Electronics*, vol. 47, no. 7, pp. 2155–2160, 2015.

[17] H. C. Chiamori, C. Angadi, A. Suria, A. Shankar, M. Hou, S. Bhattacharya, and D. G. Senesky, "Effects of radiation and temperature on gallium nitride (gan) metal-semiconductor-metal ultraviolet photodetectors," in *Sensors for Extreme Harsh Environments*, vol. 9113. International Society for Optics and Photonics, 2014, p. 911304.

[18] H. C. Chiamori, R. Miller, A. Suria, N. Broad, and D. G. Senesky, "Irradiation effects of graphene-enhanced gallium nitride (gan) metal-semiconductor-metal (msm) ultraviolet photodetectors," in *Sensors for Extreme Harsh Environments II*, vol. 9491. International Society for Optics and Photonics, 2015, p. 949107.

[19] K. H. Irani, A. Pil-Ali, and M. A. Karami, "A new guard ring for radiation induced noise reduction in photodiodes implemented in 0.18 μm cmos technology," *Optical and Quantum Electronics*, vol. 49, no. 9, pp. 1–10, 2017.

[20] M. S. Lee and H. C. Lee, "Dummy gate-assisted n-mosfet layout for total ionizing dose mitigation," *IEEE Transactions on Nuclear Science*, vol. 60, no. 4, pp. 3084–3091, Aug. 2013.

[21] L. Wilson, "International technology roadmap for semiconductors (itrs)," *Semiconductor Industry Association*, 2013.

[22] S. Kavadias, B. Dierickx, D. Scheffer, A. Alaerts, D. Uwaerts, and J. Bogaerts, "A logarithmic response cmos image sensor with on-chip calibration," *IEEE Journal of Solid-state circuits*, vol. 35, no. 8, pp. 1146–1152, 2000.

[23] B. Beecken and E. Fossum, "Determination of the conversion gain and the accuracy of its measurement for detector elements and arrays," *Applied optics*, vol. 35, no. 19, pp. 3471–3477, 1996.

[24] X. Guo, X. Qi, and J. G. Harris, "A time-to-first-spike cmos image sensor," *IEEE Sensors Journal*, vol. 7, no. 8, pp. 1165–1175, 2007.

[25] S. Yu, Z. Ping, X. Jiangtao, G. Zhiyuan, and X. Chao, "Full well capacity and quantum efficiency optimization for small size backside illuminated cmos image pixels with a new photodiode structure," *Journal of Semiconductors*, vol. 33, no. 12, p. 124006, 2012. [Online]. Available: http://stacks.iop.org/1674-4926/33/i=12/a=124006

[26] M. Sasaki, M. Mase, S. Kawahito, and Y. Tadokoro, "A wide-dynamic-range cmos image sensor based on multiple short exposure-time readout with multiple-resolution column-parallel adc," *IEEE Sensors journal*, vol. 7, no. 1, pp. 151–158, 2007.

[27] M. Kobayashi, M. Johnson, Y. Wada, H. Tsuboi, H. Takada, K. Togo, T. Kishi, H. Takahashi, T. Ichikawa, and S. Inoue, "A low noise and high sensitivity image sensor with imaging and phase-difference detection af in all pixels," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 2, pp. 123–128, 2016.

Telemetry System for 2D Flow Map using Ultrasonic Velocity Profiler

Zeliang Zhang
Department of Mechanical Engineering
Tokyo Institute of Technology
Tokyo, Japan
zhang.z.ao@m.titech.ac.jp

Munkhbat Batsaikhan
Laboratory for Advanced Nuclear Energy
Tokyo Institute of Technology
Tokyo, Japan
batsaikhan.m.aa@m.titech.ac.jp

Wongsakorn Wongsaroj
Department od Mechanical Engineering
Tokyo Institute of Technology
Tokyo, Japan
wongsaroj.w.aa@m.titech.ac.jp

Naruki Shoji
Department of Mechanical Engineering
Tokyo Institute of Technology
Tokyo, Japan
shoji.n.aa@m.titech.ac.jp

Hideharu Takahashi
Laboratory for Adavanced Nuclear Energy,
Tokyo Institute of Technology
Tokyo, Japan
htakahashi@lane.iir.titech.ac.jp

Hiroshige Kikura
Laboratory for Adavanced Nuclear Energy,
Tokyo Institute of Technology
Tokyo, Japan
kikura@lane.iir.titech.ac.jp

***Abstract*— In the decommissioning of the Fukushima Daiichi Nuclear Power Plant (FDNPP), the Japanese government has drawn mid- and long-term roadmap towards 1F decommissioning. However, the decommissioning work of the three reactors has faced difficulties due to the lack of realistic information about the damaged cores such as distribution of fuel debris. Since the radiation levels inside the reactor buildings have been too high for human access, remote detection, which combines robot and measurement technique, is required. In this study, we proposed a remote controlling method. The experiment has been conducted remotely controlling the ultrasonic measurement. Finally, the experiment data and results are successfully obtained through remote control.**

Keywords—decommissioning of Fukushima Daiichi nuclear power plant, remote control, UVP method,

I. Introduction

A severe accident at TEPCO's Fukushima Dai-ichi Nuclear Power Plant (FDNPP) occurred due to a strong earthquake and massive tsunami that struck the eastern region of Japan on 11 March 2011. During this accident, the reactor core of Units 1, 2, and 3 of the 1F melted, and fuel debris formed in the reactor pressure vessel (RPV) and primary containment vessel (PCV). The Japanese government has drawn mid- and long-term roadmap towards 1F decommissioning [1]. In order to accelerate the decommissioning of 1F, information on fuel debris and contaminated water leakage becomes critical. Because the radioactivity inside the reactor is too high for human access, the detection and the location of fuel debris cannot be done by human beings. For this reason, researchers have proposed various methods and approaches to survey the interior of the reactor and to improve the understanding of the reactor conditions, such as accident sequence analysis [2], [3], detection of fuel debris [4] [5], internal inspection with mobile robots [6]. Meanwhile, ultrasound technique was applied to detect of fuel debris information and contaminated water leakage point [7] [8]. Ultrasound has advantages that other technologies cannot match in highly radioactive, turbid, and low luminosity environments. Among the ultrasound techniques, ultrasound velocity profiler (UVP) is noted for its ability to measure the velocity distribution of fluids in real-time. Thus, remotely control the UVP measurement in the reactor is a promising method of detecting the leakage point in the RPV and PCV.

The UVP method has already been combined with PIV [9]. In this study, the UVP experiment to detect the flow velocity near the leakage point and thus to get information about the leakage point was conducted, and PIV measurement was conducted to compare with UVP . Then, the UVP sensors were attached to the robot. The UVP measurement was conducted remotely by controlling the robot. The results demonstrate that remote control is possible for UVP flow velocity measurement experiments and a concept that combines experiment with remote control.

II. Experiment Method

A. Ultrasonic Velocity Profiler

The UVP method is based on pulsed ultrasound echography. Ultrasound pulse beam is emitted in-line by the element of the transducer and will be reflected from the surface of tracer particles. The reflected pulse is called echo signal, and it is Doppler-shifted because of the velocity of the tracer particles. Using the Doppler-shift frequency's method, the velocity of the tracer particle can be obtained through analyzing sequence of tracer particles' reflection. During the measurement, ultrasonic pulses are emitted at intervals, and each pulse is emitted multiple times to achieve the Doppler-shift frequency. This is called pulse repetition frequency. Generally, the number of pulses at an interval is 128. Y. Takeda [10] had explained the one-dimensional principle of UVP in his previous work. In this paper, it was expanded to the 2D environment. But the principle is same with the one-dimensional case.

In two-dimensional velocity profile, it required at least two measure lines interact and two velocity along the measure line can be synthesized into a velocity vector [11]. For instance, as shown in Fig. 1, the velocity components u₁ and u₂ are intersected at an intersection point, they are measured along the measure lines emitted by sectorial array transducers. In each measurement line, the calculation is based on the principle of one-dimensional UVP measurement method. Then, based on the incident angle of each transducer, the velocity components are synthesized into the velocity vector u of the intersection point. The velocity along the axis is expressed by (1) and (2):

$$u_1 = V_x \sin\alpha + V_y \cos\alpha \quad (1)$$

$$u_2 = -V_x \sin\alpha + V_y \cos\alpha \quad (2)$$

Then, the velocity along the axis can be expressed as:

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

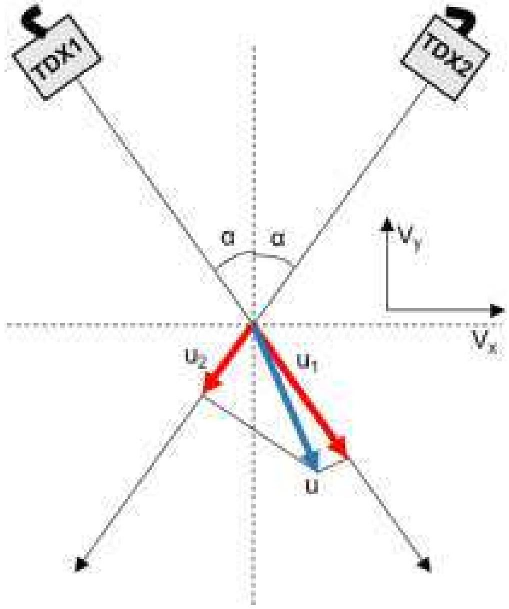


Fig. 1. Basic principle of vector reconstruction using two transducers.

$$\boldsymbol{u} = \begin{pmatrix} V_x \\ V_y \end{pmatrix} = \begin{pmatrix} \dfrac{u_1 - u_2}{2\sin\alpha} \\ \dfrac{u_1 + u_2}{2\cos\alpha} \end{pmatrix} \quad (3)$$

B. Robot System

The robot is utilized in carrying UVP sensors and controlling the position of it. In this case, a four-wheel and 3 degrees of freedom wheel-arm robot was applied. The applied robot is shown in Fig. 2. The robot arm tip can be operated from -180° to 180° around the vertical axis. The sensor unit is mounted via a horizontal compensation mechanism at the end of the hand. The ultrasonic sensor position was controlled by a motor driver (EPOS2 24/5, Maxon) using a DC motor (RE25 ϕ25 24V 20W, Maxon). The position of the sensor is calculated from the rotary encoder attached to the DC motor. The wheel body robot has EC motor (EC 25 φ 60 24 V 100 W, Maxon) mounted on each wheel, and independent control is possible by using each motor controller (EPOS2 24/5, Maxon). Using the robot can achieve high mobility and remote controllable of the UVP measurement.

C. Remote Control Method

The internet of thing (IoT) is a system that connects devices, mechanical machines and digital devices with unique identifiers (UIDs) [12] It can transfer data without any interaction from human-human or human-computer. Based on this technology, the remote control system can be built in any place [13] [14]. In this experiment, devices were controlled by the controller on the PC or cell phone through the set Wi-Fi signal with the internet services provider (ISP). The structure is shown in Fig. 4. ISP provides internet services that connect data sent through router devices (e.g. Dial, DSL, and wireless modems). The wireless modems was chosen in this study. Computers, tablets, and smartphones are responsible for remote control of the internet via 3G/4G or routers that communicate with the ISP.

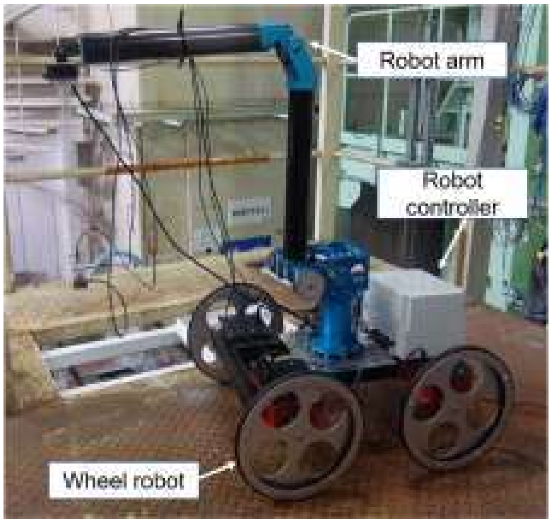


Fig. 2. Wheel and arm robot system.

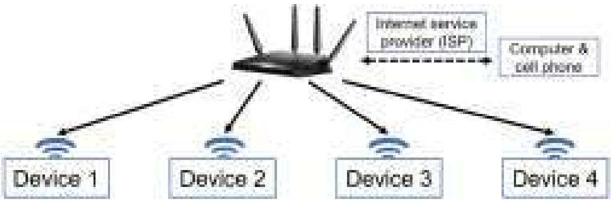


Fig. 3. Schematic diagram of communication between devices, Wi-Fi, and users.

III. Experimental Setup

A. UVP and PIV Experiment Setup

A simultaneous measurement was conducted by UVP and PIV. PIV measurement is aiming to compare with and validate the UVP data. UVP-DUO monitor with a multiplexer (from Met-Flow SA, Switzerland) were used in the UVP experiment. The Fluid container is a rectangular acrylic box with a size of 1200 mm×450 mm×450 mm. it has good transparency for the PIV measurement. Tap water was the measured fluid in this experiment, and the height of it is 290 mm from the bottom. Two sectorial array transducers from Japan Probe (shown in Fig. 4) were placed into the water and with a height of 140 mm, the distance between two transducers are 90 mm, and they are placed above the leakage point of the fluid container. The PIV experiment setup consists of a High-Speed Camera (HSC, FastCam mini AX50 type, 170K-M-8GB, Photron) and laser sheet (Raypower 450, Dantec), the laser sheet was placed above two sectorial transducers and in the middle of it, high speed camera was placed perpendicular to the fluid container to record the image sequence of the fluid. The experiment setup is shown in Fig. 5. The flow was controlled by the control value and the flow rate is displayed by the flowmeter, the flow rate was controlled in the 10L/min and the accuracy of it is ±5%, flow condition is shown in Table I.

The main parameters of UVP measurement are shown in Table II. the HSC camera provids a resolution of 896×512 pixels 250 fps and 12000 frames. The PIV result was performed in MATLAB to calculate the average velocity vectors from all sequences. The parameters of PIV measurement are shown in Table III.

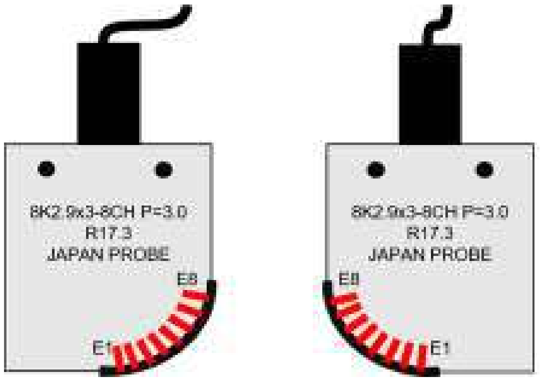


Fig. 4. Schematic diagram of the ultrasonic sectorial sensor

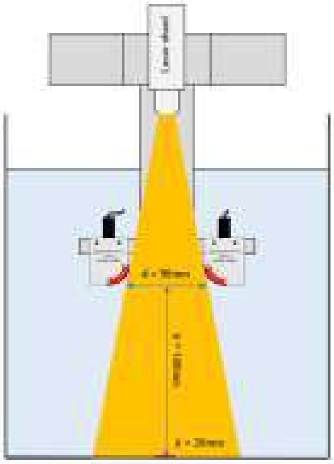


Fig. 5. Schematic diagram of experimental conditions.

TABLE I. EXPERIMENT CONDITION

Experiment conditions	Value
System pressure	Atmospheric
Water temperature (℃)	29±1
Flow rate [L/min]	10±1
Nylon particle size [μm]	80
Distance between transducers [mm]	90
Distance from the bottom wall to transducer surface [mm]	140

TABLE II. PARAMETERS OF UVP MEASUREMENT

Parameters	value
Number of channels	178
Number of profiles	12000
Maximum depth [mm]	129.96
Channel width [mm]	0.72
Channel distance [mm]	0.72
Repetitions	96
Cycles number	8
Time of capture one sample [ms]	32

TABLE III. PARAMETERS OF PIV MEASUREMENT

Parameters	value
Frame per second (fps)	250
Shooting Duration [s]	49.892
Number of frames	12000
Shutter speed	1/2000
Spatial resolution [pixels]	896×512

B. *Remote Control Measurement Setup*

Fig. 6 illustrates the schematic diagram of the remote control experiment (the X, Y, and Z axis are shown in Fig. 6, Y-axis is perpendicular to the direction of the paper surface). The system contains three parts: UVP measurement part, robot part and remote controlling part. To keep the result's consistency and to compare it, the UVP setup is kept same with first experiment. The remote control part is divided into two parts. Firstly, we set a computer in the room where is far away from the experiment site, we call this computer as operating computer. Secondly, another computer which we called experimental computer is placed in the experiment site, it is used to connect all the experimental equipment, receive commands from the operating computer, and will communicate the commands to each experimental device. The ISP between operating computer and experimental computer is the key point of this experiment. In this case, we use a commercial software Team viewer to build up the bridge between two computers. In order to remotely control the robot and UVP measurement. a Wi-Fi router is placed in the experiment site, experimental computer, UVP-DUO multiplexer and robot are connected by accessing the Wi-Fi. In this case, a control panel that integrate the UVP-DUO switch and robot switch had been designed and used to control the switch of the UVP-DUO multiplexer and robot remotely. Fig. 7 illustrates the control panel on the experimental computer. On the other hand, UVP-DUO and robot connected with experimental computer through USB, the operating software for both are integrated into the experimental computer to simultaneously control both UVP measurement and robot movement. The operating program of the robot is shown in Fig. 8. In addition to the experimental robot and UVP measurement apparatus, there are two USB cameras have been prepared in the experimental site, two cameras are directly connected to the experimental computer through USB, the experimental computer has already installed the software can simultaneously view the two cameras real-time shooting and video recording.

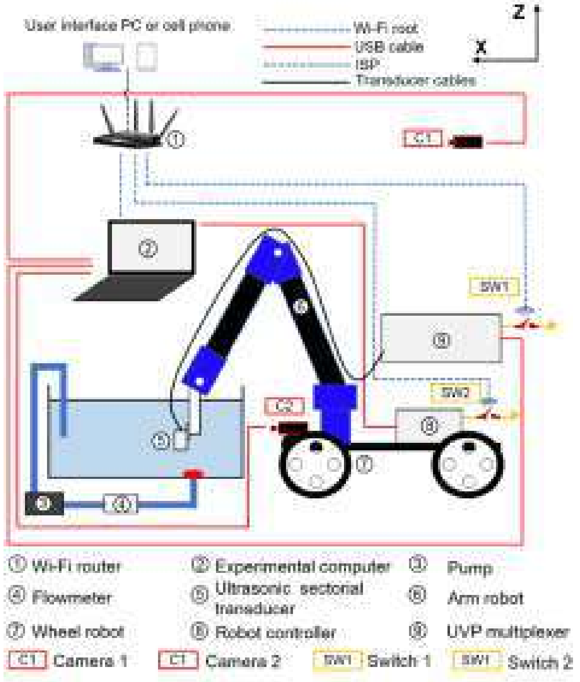


Fig. 6. Schematic diagram of remote control UVP measurement.

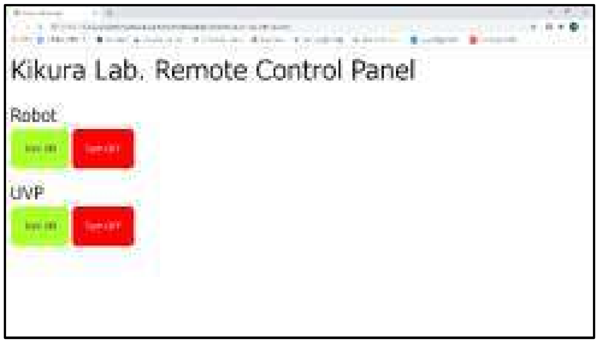


Fig. 7. The remote control panel of UVP and robot.

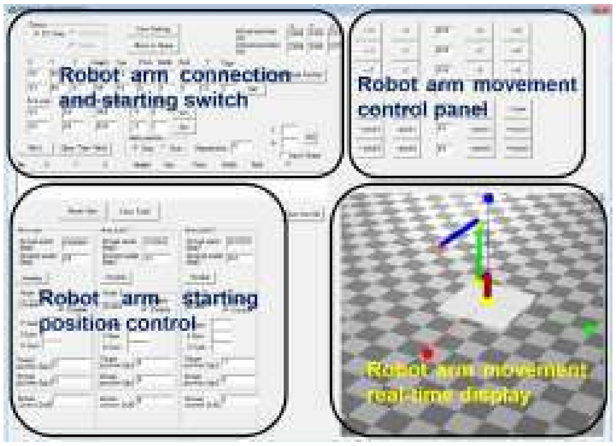

Robot arm connection and starting switch
Robot arm movement control panel
Robot arm starting position control
Robot arm movement real-time display

Fig. 8. Robot operating program.

IV. Result and Discussion

A. UVP and PIV Measurement results

Unlike the one-dimensional velocity profile of UVP, the two-dimensional velocity profile measurement is based on the intersection point generated by two interacting ultrasonic pulses. Based on the calculation of the coordinates, a grid map of the UVP 2D velocity profile was drawn, and the generated intersection point locations are shown in Fig. 9. Flow maps measured by UVP and PIV were generated using TECPLOT software. The PIV images were preprocessed by PIVlab [15] in MATLAB. UVP results and PIV results were generated and compared, respectively. To calibrate the magnitude of the measurements, a calibrated PIV image was used to combine the UVP measurement locations with the PIV measurement locations, as shown in Fig. 10.

At the same time, because the sectorial array transducers have 16 measurement lines in total at this experiment (shown in Fig. 9), the upper part of the measurement line didn't show the enough information of the velocity profile. The reason is that those lines are too far away from the leakage point, (seeing the line emitted by E7, E8 and E15, E16 in Fig. 9). After comparing the profile between E1 to element E8, the result of E2, E3, E4 and E5 was decided to represent two-dimensional velocity profile (also symmetrically, E10, E11, E12, E13 was chosen) . Thus, the result from element 2 to element 5 of each side are selected as a UVP measurement result. Meanwhile, because of the flow rate is 10L/min, the PIV result is also obvious only in the region where near the leakage point. Seeing that, the region near the leakage point had been chosen as the overlapped UVP and PIV profile.

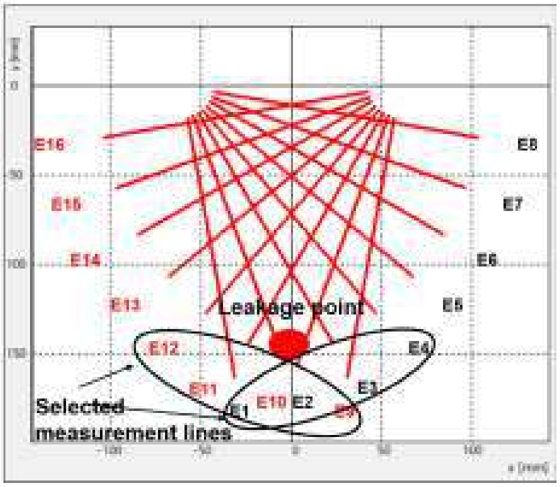
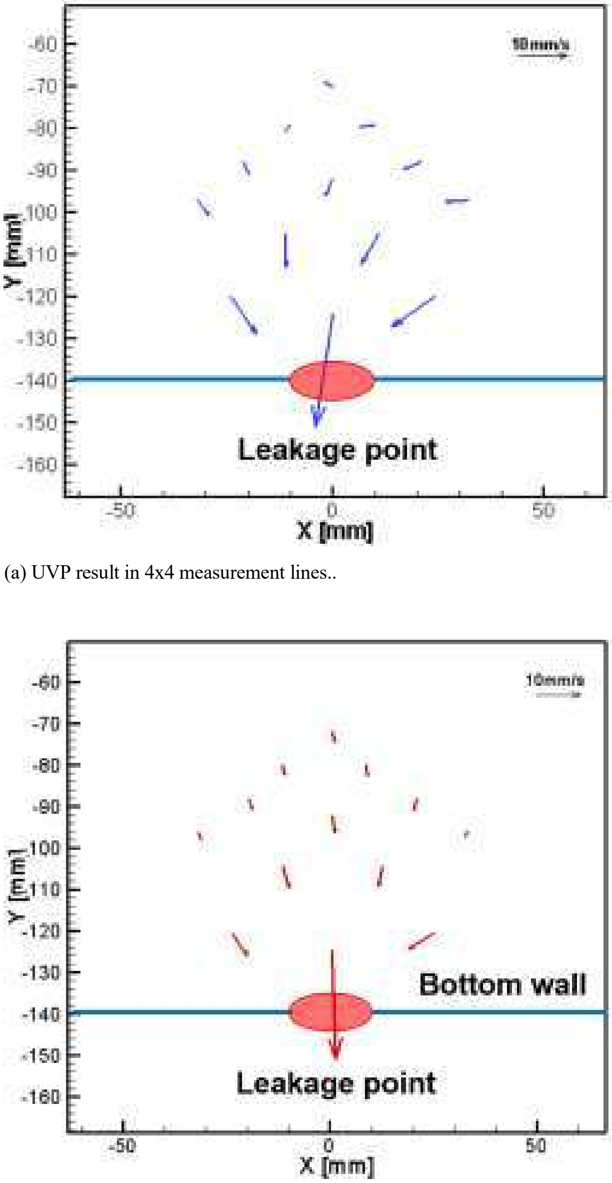


Fig. 9. Measurement grid map of sectorial array transducer.



(a) UVP result in 4x4 measurement lines..

(b) PIV result in 4x4 measurement lines.

Fig. 10. Result of UVP and PIV measurement without remote control.

By comparing (a) and (b) in Fig. 10, we can observe that the result of UVP and PIV are overlapped well when closing to the leakage point. But when in the region where is further from the leakage point, the data between PIV and UVP start to generate the difference. The reason for this phenomenon is that the number of measurement repetitions required for the high accuracy of UVP measurements is sacrificed in the experiment to maintain the consistency of PIV and UVP results. It lead to a certain amount of errors in the UVP data. However, the high degree of agreement between the two results at locations close to the leakage point illustrates the usability of UVP measurements in this environment. Also, provides a direction for future enhancement of UVP measurement.

B. Remote Control UVP Measurement Results

Fig. 11 shows a screenshot of the remote controlling process using Team Viewer. In this experiment, most of the parameters of UVP measurement were kept same with first experiment, only depth of UVP sectorial sensors changed from 140 mm from the bottom wall to 120 mm from the bottom wall. The Wi-Fi control panel is remotely switched on in the operating computer through Team Viewer to power on the robot and UVP.

During the experiment, two cameras were set in different positions to observe different parameters in the experiment. Camera one was set behind the experiment site to capture the motion of the robot arm, and camera two was set at the front of the wheeled robot to capture the position of the simulated leak point in relation to the UVP sectorial transducer. On the other hand, the control panel of the robot and the software of UVP are integrated into one screen in order to achieve the simultaneous real-time remote control. The UVP sectorial transducers are attached on the tip of the robot arm and be carried to the 120 mm from the bottom wall. Then, multiple UVP measurements were performed at different (X, Y) values (move 50 mm on Y-axis) while keeping the Z-axis values constant to verify the validity of the remote UVP measurements and to observe the flow velocity distribution at different locations above the simulated leakage point (shown in Fig. 12). The result is shown in Fig. 13.

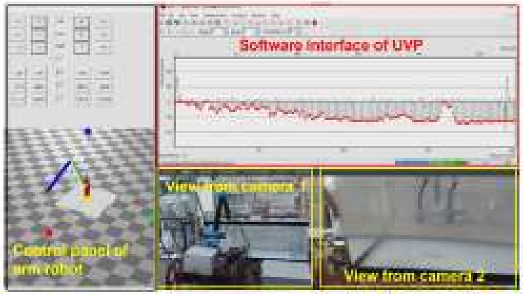


Fig. 11. Screenshot of remote controlling system.

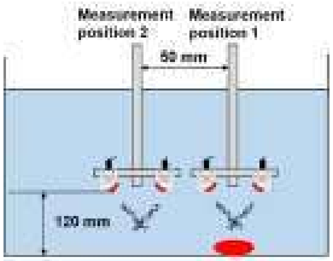
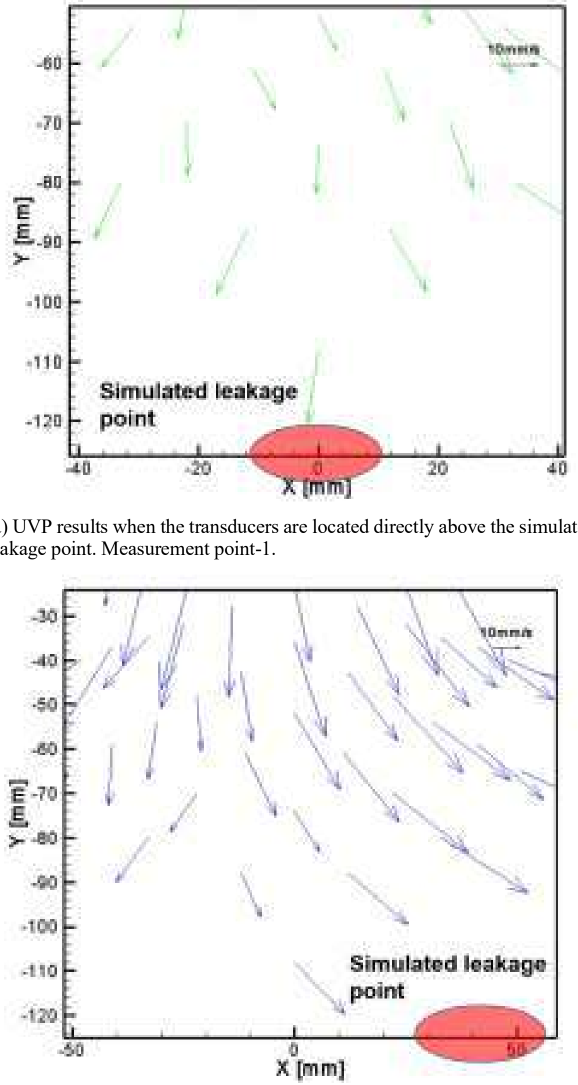


Fig. 12. Schematic diagram of UVP measurement in different position.



(a) UVP results when the transducers are located directly above the simulated leakage point. Measurement point-1.

(b) UVP result when the transducers are located upper left (50 mm from the leakage point) of the simulated leakage point. Measurement point-2.

Figure 13. Results of remote UVP measurement in different position.

From the results, we can find that when UVP sectorial transducers are located at different positions above the simulated leak, the results are completely different. When transducers are located directly above the simulated leakage point, the velocity profile in the measurement results points to the leak directly below; When transducers are located at the top left of the simulated leakage point, the velocity profile points to the simulated leak at the bottom right, The results demonstrate that remote control can be successfully applied to the UVP measurement.

However, as shown in the figure, there is still some errors that exist in the measurement results, which is related to two points: One is because the UVP measurement experiment was conducted in order to obtain the results quickly, the number of repetitions in the selection of measurement parameters sacrificed, thus reducing the accuracy of the data, and the second is because the UVP sectorial transducer connected with UVP-DUO through UVP transducer cable when the experiment was conducted. Compared with the vertical downward gravity of the transducers, the drag force from the gravity of the cable is relatively large, resulting the position of the transducer changed from perpendicular to the water surface to a certain angle with the water surface. These two problems also provide the direction for future improvement.

V. CONCLUSION

In this work, a telemetry system has been proposed to detect leakage point. A mock-up experiment was conducted in a laboratory scale tank. The experiment of UVP combined with PIV was first conducted and validated the UVP result. Then, the remote UVP measurement was conducted at a total telemetry environment. The results of the UVP flow map were obtained successfully through the remote control. At the same time, the experimental results showed some errors, which either occurred due to insufficient experimental conditions or improperly set experimental parameters, and in general, this provides a direction for future research.

REFERENCE

[1] Tokyo Electric Power Company Holdings Corporation, “Progress of Medium- and Long- Term Roadmap” (2019). http://www.tepco.co.jp/decommission/information/committee/roadmap_progress/pdf/2019/d191219_05-j.pdf (Reference date: March 4, 2021).

[2] Y. Sibamoto, K. Moriyama, Y. Maruyama, and T. Yonomoto. A simple mass and heat balance model for estimating plant conditions during the Fukushima Dai-ichi NPP accident. Japan Nuclear Science Technology, 2012, vol. 49, pp. 768-781.

[3] Y.Yamanaka, S. Mizokami, M. Watanabe, and T. Honda, Update of the first TEPCO MAAP accident analysis of units 1, 2 and 3 at Fukushima Daiichi Nuclear Power Station, Nuclear Technology, 2014, vol. 186, pp. 263-279.

[4] A. Sari, and et al., Investigation of fission chamber response in the frame of fuel debris localization measurements at Fukushima Daiichi, Radiation Measurements, 2020, vol. 130.

[5] T. Nagatani, et al, Characterization Study of Four Candidate Technologies for Nuclear Material Quantification in Fuel Debris at Fukushima Daiichi Nuclear Power Station, Energy Procedia, 2017, vol. 131, pp. 258-263.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] H. Kikura, T. Kawachi, T. Ihara, Study on ultrasonic measurement for determination of leakage from reactor vessel and debris inspection, The 11th National Conference on Nuclear Science and Technology, 2015.

[8] T. Kawachi, H. Takahashi, H. Kikura, A study on two-dimensional vector flow mapping by Echo-PIV with total focusing method, WIT Transactions on Engineering Sciences, 2018, vol. 120, pp.275-286.

[9] M. Batsaikhan, A. Hamdani, H. Kikura, Velocity measurement on two-phase air bubble column flow using array ultrasonic velocity profiler, Multiohase Flow: Theory and Applications, 2018, vol. 6 (1), pp. 86-97.

[10] Y. Takeda, Development of an ultrasound velocity profile monitor, Nuclear Engineering and Design, 1991, vol. 126 (2), pp. 277-284.

[11] Y. Takeda, and et. al., Ultrasonic Doppler Velocity Profiler for Fluid Flow: Springer: Tokyo, Japan, 2012.

[12] L. Atzori, A. Iera, and G. Morabito, The internet of Things: A survey, Computer Networks, 2010, vol. 54, pp. 2787-2805.

[13] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, The industrial internet of things (IIoT): An analysis framework, Computers in Industry, 2018, vol. 101, pp. 1-12.

[14] S. Dey, A. Roy, and S. Das, Home automation using internet of thing, IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 20-22 Oct2016, New York, NY, USA.

[15] W. Thielicke, E. J. Stamhuis, PIVlab – towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB, Journal of Open Research Software, 2014, vol 2 (1).

A Review: Predicting Student Success at Various Levels of their Learning Journey in a Science Programme

Judith Goodness Khanyisa Mabunda
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1489219@students.wits.ac.za

Ashwini Jadhav
Faculty of Science
The University of the Witwatersrand
Johannesburg, South Africa
ashwini.jadhav@wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—This paper examines how features affect student persistence or dropout at South African higher education institutions, based on three previous studies. In the previous studies, high school grades were used as a valid predictor of student success. The quality of a high school's learning environment has an effect on almost every aspect of higher education success. Students who are better prepared coming out of high school are ideally suited to do well in higher education institutions, who they are, how much money they have, and where they go don't matter. This review aims to identify effective features that warrant student success from high school grades and choice of academic courses during registration in higher education. The following questions are used to guide this review: How can we define student success? Which features should we focus on? Which models work? Based on data mining techniques such as machine learning models that the previous studies have used to predict student success, it has been revealed that the most important features that influence student success in a Computer Science programme are Prior Computer experience, Mathematics, English from High school and the choice of a course.**

***Index Terms*—Student success, Prediction, Higher Education, Machine Learning**

I. Introduction

Student success is significant in South African higher education because it is commonly used as a metric for the institution's performance. At-risk students would have a much greater chance of thriving if they are detected early and preventative measures are taken. High school grades have long been a strong predictor of student success in postsecondary institutions [10]. Students who complete four years of mathematics, science, and English in high school have an 87 percent chance of graduating from college, compared to a 62 percent chance for those who do not complete the coursework [1], [14]. Machine learning algorithms have been widely used for prediction in recent years.

Admission to science programmes in South African universities, requires that students perform well in both pure mathematics and English as confirmed by [12], [15]. However, studies that applied machine learning techniques show that pure mathematics and English cannot be the only predictors of student success in higher educational institutions [11]. The effective and efficient application of machine learning techniques entail many decisions, ranging from how to define student success, through which student features to focus on, up to which machine learning method is more appropriate to the given problem.

II. How can we define success?

The authors in each paper have outlined what drives student success, in simple words, they have defined what student success is. [6] defined student success as dependent on biographical and enrolment observations as features that ensures efficient student performance at a South African higher educational institution. This study argues that there exists characteristics, attributes, and features in a student profile that can accurately predict the student's performance from the first year of registration until qualifying, providing a contribution that suggests a support and supplementary mechanism to the current university Admission Point Score (APS) system, which has evidently been struggling, generating between 25.7% and 32.2% minimum time (3-year) graduates in South Africa for the academic years 2000 to 2017 [6].

[9] believes that student enrolment and biographical data are rich sources of information that can help universities and staff solve a number of problems, such as identifying at-risk students, restricting student intake, and changing course content to understand and assist disadvantaged students.

[11] predicted student success for first year computer science students based on three features and believes that prior computer experience is required for students to perform well in a computer science course.

III. Which Features Should we focus on?

The authors considered different features and looked closely at the extent to which these features impact student success.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

[11] considered two groups when investigating the effect that prior computer experience had on the final first year computer science results. Students in group 1 had no prior computer experience. Students in group 2 had prior computer experience. The paper further outlined all three features that were looked at when predicting success for first year computer science students. The features include: Final grade 12 results such as Pure Mathematics as a predictor (this is a known predictor of performance for computer science as confirmed by [15]), Past computer experience as a predictor and Language performance as a predictor. The features are shown in TABLE 1.

On the other hand, paper [9] considered school quintile, high school grades, Naional Banchmark Test scores as well as the major a student chose when registering at university as predictors when predicting the Success of First Year University Students. From the results it was concluded that a student's undergraduate major, school quintile, Life Sciences and Mathematics marks in matric have a significant effect on their chance of completing their studies. The features of this paper are outlined in Fig. 1.

Some features used by [6] are the same as the ones used by [9]. But some are different. TABLE I listed all the features that [6] used to predict student performance at each year of study until qualifying, for students at a South African higher education institution. Fig. 2 represents features used throughout in [6].

TABLE I: A table presenting the various features used for prediction of student success in [11].

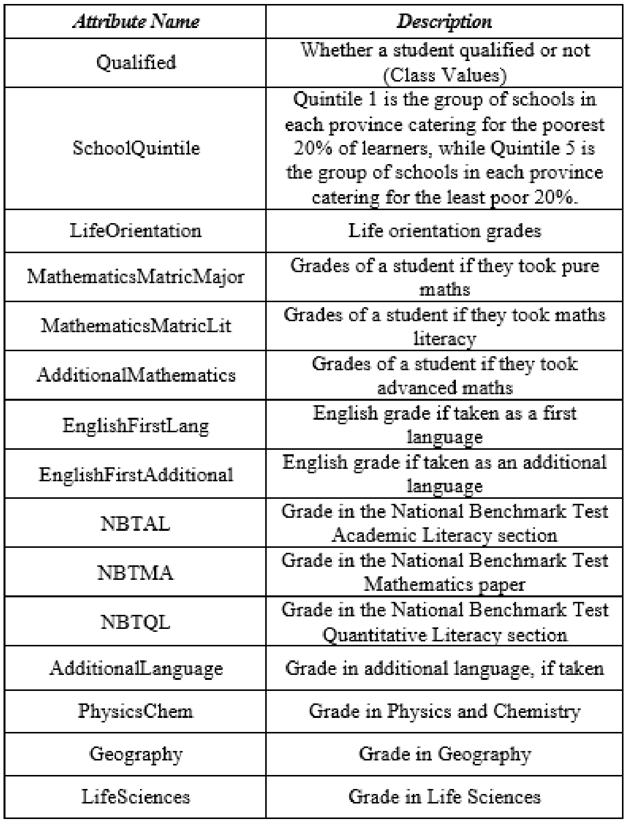
Features
Mathematics as a predictor
Past computer experience as a predictor
Language performance as a predictor

IV. Which Models work?

To achieve their objectives and answer research problem statement and questions that are set out for their papers to answer, the authors considered different models that are suitable and best possible approaches for their given problems.

[11] used four models. Linear regression was used to predict the true results of students, the outcome produced from the model is a real number. The classification models such as Logistic regression, Naive Bayes and Decision tree were used to predict a student's results as either a PASS or FAIL. The four models are used to identify which features are important in higher educational institution success to first year computer science students.

The three classification models performed very well in group 2 as they resulted in highest accuracy than in group 1. Although the classification models performed better in group 2, the difference in accuracies between group 1 and group 2 was not much when using the logistic regression model and Naive Bayes model. The differences for these two models between group 1 and group 2 are 2.65 and 0.37 percents respectively. The difference in accuracies when using the decision tree classifier model was greater between the two groups as compared to using the logistic regression model and Naive Bayes model. The difference was 8.51 percent when using the gini index and 13.2 percent when using entropy in decision tree model.

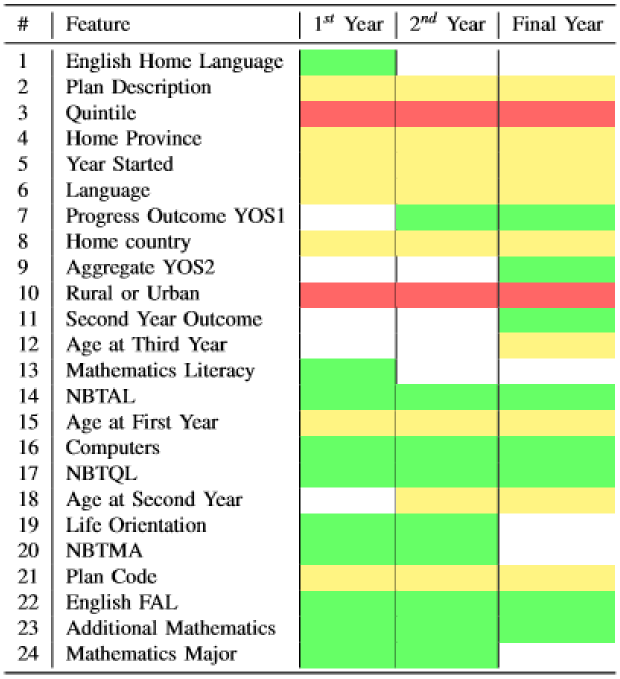


Attribute Name	*Description*
Qualified	Whether a student qualified or not (Class Values)
SchoolQuintile	Quintile 1 is the group of schools in each province catering for the poorest 20% of learners, while Quintile 5 is the group of schools in each province catering for the least poor 20%.
LifeOrientation	Life orientation grades
MathematicsMatricMajor	Grades of a student if they took pure maths
MathematicsMatricLit	Grades of a student if they took maths literacy
AdditionalMathematics	Grades of a student if they took advanced maths
EnglishFirstLang	English grade if taken as a first language
EnglishFirstAdditional	English grade if taken as an additional language
NBTAL	Grade in the National Benchmark Test Academic Literacy section
NBTMA	Grade in the National Benchmark Test Mathematics paper
NBTQL	Grade in the National Benchmark Test Quantitative Literacy section
AdditionalLanguage	Grade in additional language, if taken
PhysicsChem	Grade in Physics and Chemistry
Geography	Grade in Geography
LifeSciences	Grade in Life Sciences

Fig. 1: A figure presenting the various features used for prediction in [9].

Results from the hypothesis test and confidence interval test showed that the students in group 2 outperformed the students in group 1. Based on these results, it is worth considering past computer experience as an additional criterion to studying computer science. However, we should not ignore students without prior computer experience considering the fact that the accuracies produced from both group 1 and group 2 differed by a small percentage when using both the logistic regression model (difference of 2.65 percent) and Naive Bayes model (difference of 0.37 percent).

On the other hand, through six machine learning models like Random Forest, Logistic Model Trees (LMT), Decision Trees (J48), Sequential Minimal Optimization (SMO), Multinomial Logistic Regression and Naive Bayes, the authors predicted students' first, second, and final year outcomes based on synthetic dataset. In the 1st, 2nd and final years of student enrollment, all of these models performed outstandingly in predicting student success, however, amongst these models, Random Forests performed better with accuracy of 94.40%,



#	Feature	1^{st} Year	2^{nd} Year	Final Year
1	English Home Language	🟩		
2	Plan Description	🟨	🟨	🟨
3	Quintile	🟥	🟥	🟥
4	Home Province	🟨	🟨	🟨
5	Year Started	🟨	🟨	🟨
6	Language	🟨	🟨	🟨
7	Progress Outcome YOS1		🟩	🟩
8	Home country	🟨	🟨	🟨
9	Aggregate YOS2			🟩
10	Rural or Urban	🟥	🟥	🟥
11	Second Year Outcome			🟩
12	Age at Third Year			🟨
13	Mathematics Literacy	🟩		
14	NBTAL	🟩	🟩	🟩
15	Age at First Year	🟨	🟨	🟨
16	Computers	🟩	🟩	🟩
17	NBTQL	🟩	🟩	🟩
18	Age at Second Year		🟨	🟨
19	Life Orientation	🟩	🟩	
20	NBTMA	🟩	🟩	
21	Plan Code	🟨	🟨	🟨
22	English FAL	🟩	🟩	🟩
23	Additional Mathematics	🟩	🟩	🟩
24	Mathematics Major	🟩	🟩	

Fig. 2: A figure presenting the various features used for classification in [6]. The table sorts the features according to whether they were used for the prediction of the students' 1st Year Outcome, 2nd Year Outcome, or Final Year Outcome.

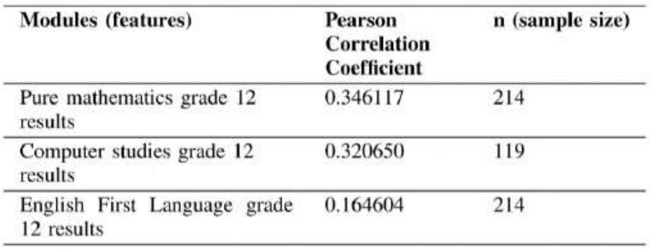
93.70% and 95.45% respectively [6].

Another study aimed at investigating which features of a student best predict whether they will graduate so as to identify vulnerable students and offer them crucial assistance applied six machine models such as Bootstrap Aggregating (Bagging), Bayesian Network, Logistic Regression, Multilayer Perceptron (MLP), K-Nearest Neighbours (KNN) and Random Forests. Bagging was found to be the most effective model for these features, correctly classifying 75.97% of the data. Random Forests followed closely, correctly classifying 75.57%, while KNN came in last with 64.83% [9].

These models are briefly described below.

- Bayesian Network - A probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).
- Logistic Regression - used to model the probability of a certain class or event existing such as pass/fail, win/lose.
- Multilayer Perceptron (MLP) - is a class of feedforward artificial neural network. The Perceptron consists of an input layer and an output layer which are fully connected.
- K-Nearest Neighbours (KNN) - K-Nearest Neighbour is a non-parametric classification algorithm. This algorithm works by finding the distances between a new data point and all the labelled datasets in the data, it reads through the whole dataset to find out the k nearest neighbours closest to the new data point, then the votes for the most frequent label in classification will be the class for the new data point [8].
- Bootstrap Aggregating (Bagging) – a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.
- Random Forests - Random forest algorithm is a supervised classification model. This model creates the forest with number of trees [3]. Random forest classifier works well with missing values. It does not overfit the model when there are more trees in the forest.
- Naïve Bayes (NB) - has been widely used in text classification. Given a set of labelled data, NB often uses a parameter learning method called Frequency Estimate (FE), which estimates word probabilities by computing appropriate frequencies from data [13].
- Linear Regression - models a relation between a dependent variable and one or more independent variable.

Random Forest followed by Logistic model trees, Multinomial Logistic Regression, Sequential Minimal Optimization and Decesion tree are suitable and best possible approaches for predicting student success.



Modules (features)	Pearson Correlation Coefficient	n (sample size)
Pure mathematics grade 12 results	0.346117	214
Computer studies grade 12 results	0.320650	119
English First Language grade 12 results	0.164604	214

Fig. 3: This figure shows the result of the linear regression from [11].

TABLE II summarises and compares the accuracies of each model that the authors obtained when conducting their studies.

V. Discussion

All of these indicators of student success have been studied in the literature to varying degrees, and there is widespread consensus on their significance. How each model performed for any given feature indicates whether the feature is very important in influencing student success. While other studies have concentrated on Mathematics and English as the best predictors of student success in a Computer Science program, the findings in [11] indicate that previous computer experience is also important for student success in higher educational institutions.

On the other hand, [9] demonstrate that school quintile, NBT grades, a course or major a student registered for, life sciences and mathematics from high school matters. [6] has as well focused on high school and intra-university grades and individual attributes such as home language, home province, age, etc.

In other studies, a handful of additional elements of student success have emerged, representing new dimensions, and

TABLE II: The accuracy of each model.

Models	Paper1 [11]		Paper2 [6]			Paper3 [9]
	Group 1	Group 2	Year1	Year2	Year3	After
Logistic Regression	59.85%	62.5%	-	-	-	75.48%
Decision Tree	60.24%	79.10%	-	-	-	-
Naive Bayes	60.95%	61.32%	83.95%	83.40%	84.40%	-
Decision Tree (J48)	-	-	87.55%	86.20%	91.45%	-
Random Forest	-	-	94.40%	93.70%	95.45%	75.57%
Sequential Minimal Optimization (SMO)	-	-	87.25%	84.5%	89.20%	-
Multinomial Logistic Regression	-	-	87.80%	86.20%	90.70%	-
Logistic Model Trees (LMT)	-	-	91.90%	91.75%	93.15%	-
Bayesian Network	-	-	-	-	-	74.12%
Multilayer Perceptron (MLP)	-	-	-	-	-	75.09%
K-Nearest Neighbours (KNN)	-	-	-	-	-	64.83%
Bootstrap Aggregating (Bagging)	-	-	-	-	-	75.97%

variations on common indicators. Examples of such indicators are theoretical perspectives (such as Sociological Perspectives, Organizational Perspectives, Psychological Perspectives, Cultural Perspectives and Economic Perspectives). The indicators also includes student background characteristics, precollege experiences, and enrolment patterns. Student engagement such as Student behaviors, activities, and experiences in post secondary education [5].

The importance of student engagement can be divided into two categories. The first is the amount of time and effort students put into their studies and other educational pursuits. ”A person’s level of commitment to the learning process has a huge effect on learning.” [2]. The second category of student engagement is how the school allocates resources and organizes the curriculum, other learning opportunities, and support services to enable students to participate in behaviors that contribute to positive interactions and results such as persistence, satisfaction, learning, and graduation [4]. As [7] concluded, individual effort and participation in academic, interpersonal, and extracurricular offerings on a campus decide the effect of higher eductaional institution.

VI. Conclusion

From this review, it’s concluded that there are more influential features on student success other than high school grades and choice of course during registration. Everyone that tries to predict student success has to closely at various features that might just make a very huge difference in ensuring that the results are correct and efficient to make decisions. The best prediction models should be considered when predicting student success. Different students perform differently based on different features, that is why it is very vital that various features be considered when predicting student success.

Acknowledgment

This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835).

References

[1] Clifford Adelman. *Answers in the tool box: Academic intensity, attendance patterns, and bachelor’s degree attainment.* US Department of Education, Office of Educational Research and Improvement, 1999.

[2] PA Alexander and PK Murphy. The research base for apa’s learner-centered psychological principles’, paper presented at the. In *Annual Meeting of the American Educational Research Association*, 1994.

[3] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[4] George D Kuh. Assessing what really matters to student learning inside the national survey of student engagement. *Change: The magazine of higher learning*, 33(3):10–17, 2001.

[5] George D Kuh, Jillian L Kinzie, Jennifer A Buckley, Brian K Bridges, and John C Hayek. *What matters to student success: A review of the literature*, volume 8. National Postsecondary Education Cooperative Washington, DC, 2006.

[6] Ndiatenda Ndou, Ritesh Ajoodha, and Ashwini Jadhav. Educational data-mining to determine student success at higher education institutions. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–8. IEEE, 2020.

[7] Ernest T Pascarella and Patrick T Terenzini. *How College Affects Students: A Third Decade of Research. Volume 2.* ERIC, 2005.

[8] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[9] Nastassja Philippou, Ritesh Ajoodha, and Ashwini Jadhav. Using machine learning techniques and matric grades to predict the success of first year university students. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–5. IEEE, 2020.

[10] Gary R Pike and Joseph L Saupe. Does high school matter? an analysis of three methods of predicting first-year grades. *Research in higher education*, 43(2):187–207, 2002.

[11] Thabo Ramaano, Ritesh Ajoodha, and Ashwini Jadhav. Different models relating prior computer experience with performance in first year computer science.

[12] Sarah Rauchas, Benjamin Rosman, George Konidaris, and Ian Sanders. Language performance at high school and success in first year computer science. *ACM SIGCSE Bulletin*, 38(1):398–402, 2006.

[13] Jiang Su, Jelber S Shirab, and Stan Matwin. Large scale text classification using semi-supervised multinomial naive bayes. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 97–104. Citeseer, 2011.

[14] Edward C Warburton, Rosio Bugarin, and Anne-Marie Nunez. Bridging the gap: Academic preparation and postsecondary success of first-generation students. statistical analysis report. postsecondary education descriptive analysis reports. 2001.

[15] Laurie Honour Werth. Predicting student performance in a beginning computer science class. *ACM SIGCSE Bulletin*, 18(1):138–143, 1986.

A Design of Greenhouse Monitoring System Based on Low-Cost Mesh Wi-Fi Wireless Sensor Network

*Note: Sub-titles are not captured in Xplore and should not be used

1st Tung Cao Pham
ECE (of Aff.)
Vietnamese German University (of Aff.)
Ho Chi Minh City, Viet Nam
eeit2014-tung.pc@student.vgu.edu.vn

2nd Hien Bich Vo
ECE (of Aff.)
Vietnamese German University (of Aff.)
Ho Chi Minh City, Viet Nam
hien.vb@vgu.edu.vn

3rd Nhu Quang Tran
ECE (of Aff.)
Vietnamese German University (of Aff.)
Ho Chi Minh City, Viet Nam
nhu.tq@vgu.edu.vn

***Abstract*—Agricultural environment monitoring has become a crucial field of control and protection providing a real-time system with the physical world. This paper is proposed the design and implementation of a wireless agricultural monitoring system platform based on the Mesh Wi-Fi Networking with low-cost and resource-limited hardware-ESP8266 Platform. The system includes a root, 3 stations-repeaters, and 6 nodes, each node has three sensors to measure parameters: soil moisture and soil temperature, the environmental temperature and environmental humidity of the farm, and light intensity. The collected data are transmitted from the nodes to the root and updated on the cloud database and displayed on the dashboard. The website based on the MEAN stack will be built for the owners to track real-time information and save historical data.**

***Index Terms*—Smart Agriculture, Internet of Things, Wi-Fi Mesh, MQTT, MEAN Stack, Web Application.**

I. Introduction

The Internet of Things (IoT) is one of the dynamic trends in the technology world, besides Artificial Intelligence (AI). IoT is becoming an explosion in the modern world and being a backbone for future technologies. Ericsson predicts the number of IoT devices reach 18 billion by 2022 [1]. Wireless Sensor Networks (WSNs) are key components of the Internet of Things (IoT). It is a network of small and inexpensive sensor nodes, that monitor and communicates with each other through radio signals, a controller, and a communication system. The integration of WSNs with IoT has huge potential applications, from smart homes to smart cities, remote healthcare, energy and water control, precision agriculture... [2], [3].

Precision Agriculture (PA) is a new method of farming by combining technology with traditional farming practices [4], [5], [6], [7] to achieve the maximum profitability of farm.

So we propose a greenhouse with an IoT system that can help farmers or growers monitor with required parameters of their farm with the primary objective of this project is to integrate and develop new applications so as to advance precision agriculture.

This paper is organized as Section 2 describes the Materials and Methodologies, Section 3 illustrates the Results with the implementation, and Section 4 includes a Conclusion.

II. Materials and methods

Figure.1 shows the system architecture organization. There are three main elements: (1) sensor nodes that measure the environmental information; (2) root-gateway, and (3) the server, which receives, stores, and displays the data.

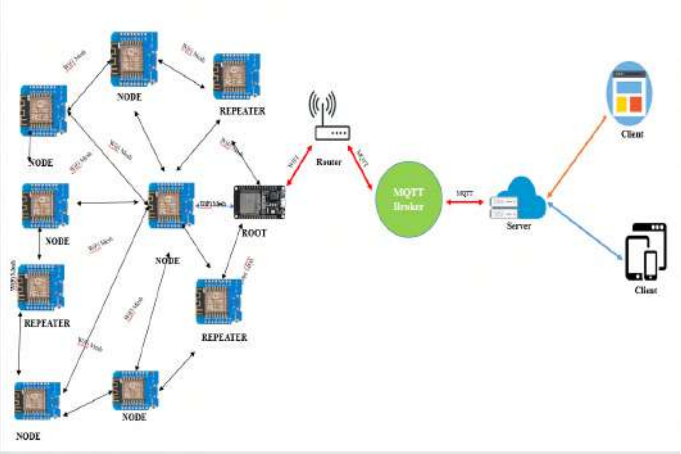


Fig. 1. System Architecture Overview

A. System Hardware

The following points describe the hardware used to implement the system.

1) Sensor Nodes:

- Soil Moisture Sensor: SHT-30 (manufactured by Sensirion) sensor has been chosen for moisture and temperature of the soil. It is encapsulated in a sintered metal mesh encasing so as to protect electronics when the sensor is submersible in the water in a short time-below 1 hour. The SHT30 sensor has an excellent ±2 percentage relative humidity and ±0.5°C accuracy for most uses. This sensor has an Inter-Integrated Circuit or I2C interface protocol and the 1-meter long cable [8].

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

- Temperature and Air Humidity Sensor: The DHT22 is a low-cost temperature and humidity sensor with a single wire digital interface. So, it interfaces with the microcontroller using I/O pins. The sensor is calibrated and does not require extra components so it can easily get right to measuring relative humidity and temperature. This sensor has ±2 percentage relative humidity (max ±5 percentage RH) and ±0.5°C in temperature accuracy [9].

- Light Intensity Sensor: The BH1750 FVI Ambient Light sensor is produced by Rohm Semiconductor, supports 16 bits unsigned integer output over I2C bus interface, includes a highly sensitive photodiode, a low noise amplifier, and the output data is directly a digital signal and no need to calculate complicatedly. The unit of output data is Lux (Lx). When objects which are lighted in homogeneous get the 1 Lx luminous flux in one square meter, their light intensity is 1 Lx [10]. Sometimes to take good advantage of the illuminant, a reflector is added to the illuminant. So that there will be a more luminous flux in some directions and it can increase the illumination of the target surface [11]. It is possible to detect a wide range and high resolution from 1 to 65535 Lux. If an error occurs, the negative value will be got.

- Processor and wireless communication module: The Wemos D1 Mini ESP8266-12 board has been chosen in this project for sensor node hardware because it is programmed in a well-known environment (PlatformIO [12] IDE and Arduino framework), which is used by a wide community of developers that freely share libraries and resources for programming. The WeMos D1 mini is a mini Wi-Fi integrated board, supports IEEE 802.11n2009 or 802.11n standard and works as a 2.4 GHz ISM band. It is based on the ESP8266EX microcontroller. Besides the Wi-Fi functionalities, ESP8266EX also integrates an enhanced version of Tensilica's L106 Diamond series 32-bit processor and on-chip SRAM. The chip runs at 80MHz and reaches a maximum clock speed of 160 MHz. It can be interfaced with external sensors and other devices through the GPIOs.

  The D1 Mini has 4MB Flash memory, a PCB antenna. It uses external SPI flash to store user programs and supports up to 16 MB memory capacity theoretically [13], small dimensions, and 11 digital IO pins, support many communication protocols like I2C protocol (Inter-Integrate Circuit), UART protocol (Universal Asynchronous Transmitter Receiver), ADC, and quite many peripherals [14]. The board has a typical current consumption of about 56 mA when receiving and 170 mA when transmitting data. The module also is used in connectivity applications, data collection, and control via Wi-Fi protocol [15].

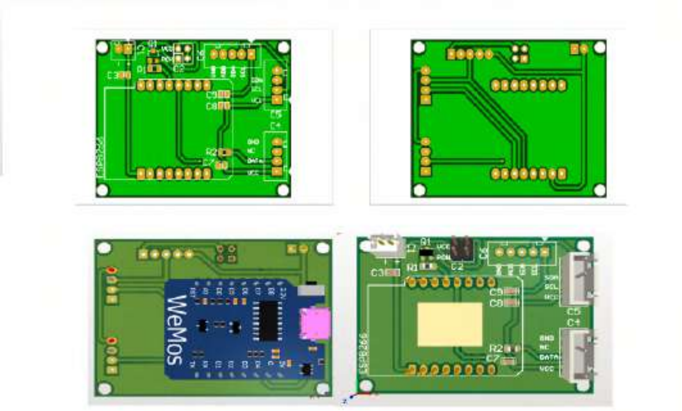


Fig. 2. PCB Connection Layout and 3D View

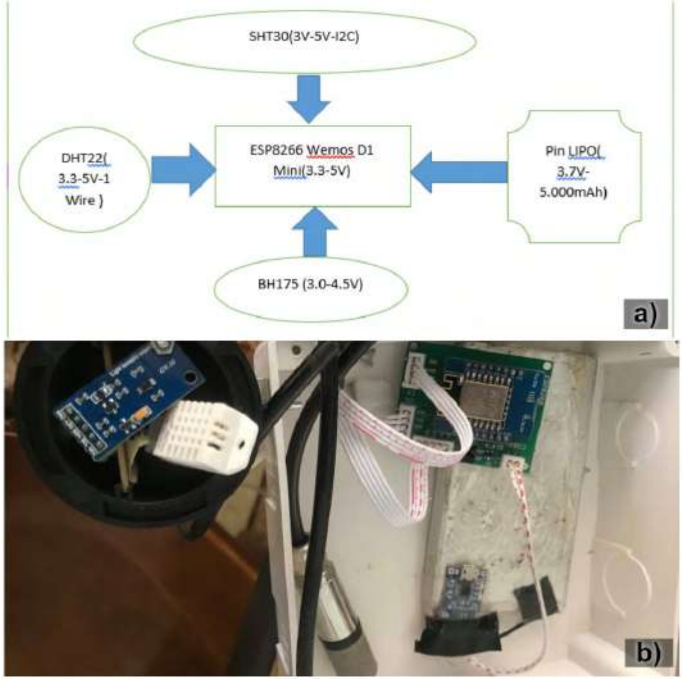


Fig. 3. a)High level block diagram of Sensor Node. b) Sensor Node

2) Gateway: In this system, we use ESP32-WROOM-32 development kit as a root node or gateway. It collects multi-sensor data simultaneously from sensor nodes then transmits it to the cloud in JavaScript Object Notation-JSON format. The ESP32-WROOM-32 is advertised as a low-cost, low-power System on Chip architecture with integrated CPU(s), RAM, Wi-Fi, and Bluetooth 4.2 capabilities on a single chip [16]. At the core, there is a dual-core Tensilica Xtensa LX6 microprocessor that runs up to 240MHz, 4MB of flash, and a PCB antenna.

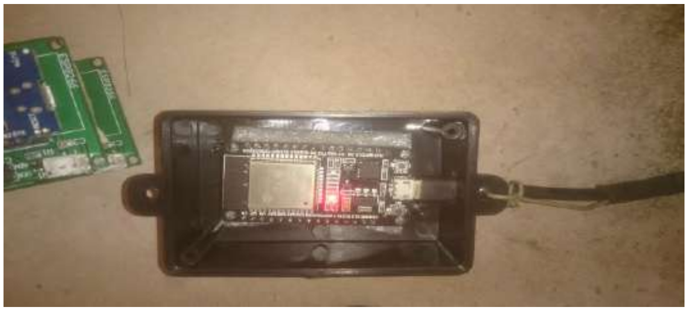


Fig. 4. Gateway

3) The Server: In the server section, there are two main elements: Message Queuing Telemetry Transport Protocol-MQTT protocol and MEAN Stack.

*MQTT Protocol: is a machine to machine or “Internet of Things” connectivity protocol on top of TCP/IP meaning it supports the event-driven exchange of messages through wireless networks. It allows extremely lightweight publish/subscribe messaging transport. MQTT protocol has clients-broker architecture and is used for the data transfer between the electronic system and the cloud. MQTT protocol has better performance than HTTP in IoT applications [17].

*Web developers have been used the LAMP open-source tool stack [18] (consisting of Linux Operating System, the Apache Web Server, MySQL as a database, and PHP as the scripting language). However, new applications faster and continually roll out enhancements and still ensuring that the application is highly available and can be scaled appropriately when needed by using the MEAN service Stack or just MEAN. MEAN stands for MongoDB, ExpressJS, AngularJS, and Node.JS [19].

MEAN also provide both client and server-side for interactive web applications. All four tools of MEAN are heavily based on JavaScript language. So each component in the MEAN full stack speaks the language of JavaScript Object Notation (JSON).

- Node.js: Node.js or just Node is an open-source, Javascript runtime environment built around Google’s V8 JavaScript engine and implemented in C++. Node provides a high-performance, asynchronous event-based server [20]. It uses an event-driven, non-blocking I/O model which makes it lightweight, efficient, and excellent for data-intensive real-time applications that run across shared devices [21]. Node.js is the most crucial tool of the MEAN stack.
- Express.js: is the framework built on the underlying capability of Node.js. It provides a web application server framework. This framework also provides a wrapper around a lower-level Node interface: giving the Web developers, a convenient means to handle routing, URL, and HTTP operations (such as GET and POST). Express.js module facilitates a simplified as well as a lean solution than implementing directly Node.
- MongoDB: MongoDB is an open-source, crossplatform database that is written in C++. It stores data in the keyvalue pair, using binary data type like JavaScript Object Notation (JSON), MongoDB provides persistence for application data and is designed with both scalability and developer agility in mind. And it bridges the gap between keyvalue stores, which are fast and scalable, and relational databases, which have rich functionality. Instead of storing data in rows and columns as one would with a relational database, so being a document-oriented NoSQL Database [22].
- Angular.js or Angular is a front-end web app framework, maintained by Google. The framework runs JavaScript code in the user’s browser and provide provides a client-side framework for MVC (Model-View-Controller) [23]. This allows the application User Interface to be dynamic.

B. System Software

1) Sensor Nodes: The sensor node is programmed in C++ language through PlatformIO IDE [12] and Arduino framework. After the node is powered up, the serial UART module starts, The if condition is used, if the node is a defined node, all sensors are initialized, and then will read physical data from the environment. If not, the node will move to the next task-Mesh Network initialization. It means that, in Mesh Wi-Fi Network, both sensor nodes and repeaters can piggyback the data of others. The sensing period is equal to 5 minutes.

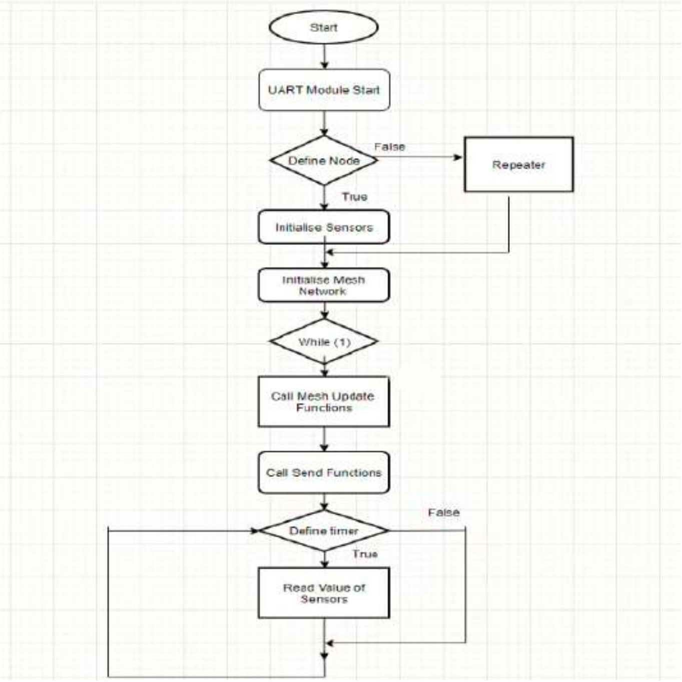


Fig. 5. Software Flow Diagram of Sensor Node and Repeater

2) The Gateway: The selected hardware to operate as a gateway has been the ESP32-WROOM-32 module, and its mission is to bridge between the sensor nodes and the server application. The ESP32-WROOM-32 module gateway is programmed in C++ language through PlatformIO IDE and Arduino framework. We set up the gateway into Access Point (AP) so as to the sensor nodes and repeaters can connect via the Mesh Wi-Fi network. The gateway connects to the cloud via an Internet connection, supplied by Tp-link M7350 4g LTE Mobile WiFi Wireless Router/Hotspot [24].

The hub initiates its Mesh Wi-Fi interface, after powering up, and seeks for available nodes at 2.4GHz, the radio frequency spectrum toed. When the setup is completed, the gateway requests commands through the TCP socket stored in the server. On the contrary, if the node is available, it will read and store the sensor's information coming from the other nodes through Mesh Wi-Fi. Afterward, the gateway will send an answer to the node with user-defined information related to the sensors and will send the information of nodes to the server through the MQTT protocol. Finally, it will update the server and database with the data of nodes.

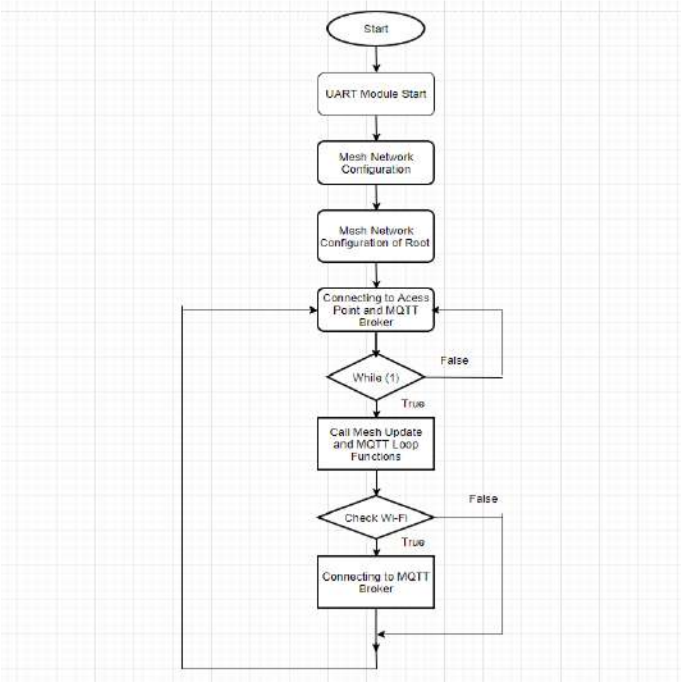

Start
UART Module Start
Mesh Network Configuration
Mesh Network Configuration of Root
Connecting to Acess Point and MQTT Broker
While (1)
False
True
Call Mesh Update and MQTT Loop Functions
Check Wi-Fi
False
True
Connecting to MQTT Broker

Fig. 6. Software Flow Diagram of Gateway

3) The server: In this project, the server is implemented in a Virtual private server (VPS) with a MEAN service stack application, which has three layers: an MQTT Broker, the Web Page, and the NoSQL database.

Because we use Mosquitto MQTT Broker [25], "Mosquitto" is an MQTT Broker developed by Eclipse Organization, so the VPS must be installed in Mosquitto MQTT Broker library. The other library, as well as the frameworks, have to be installed: mongoose, Socket.IO [26], Express.js, and MongoDB. The installing process can be conducted via Plugin Remote-SSH, which can search in Microsoft Visual Studio. After done debugging and organizing, all of the coding files and library is upload to the BKhost hosting service. The Bkhost is a VPS hosting seller, our VPS Hosting address is http://103.130.212.132:8000/

The web page (Application) can be divided into a Front-end and Back-end. Angular.js with Bootstrap packet [27] provides for the client-side framework. The server-side or back-end side is used to query databases for information and processing any logic that a web application requires. Node.js is a server-side scripting language that is used to interact with the database as well as allows a server to a high number of connections while handling multiple requests.

III. Results and Discussion

The sensor nodes have been tested inside the Vietnamese German University campus - an ideal greenhouse, with three floors, covered by the metallic structure as well as concrete walls, which have over 5000-meter square surface. In our ideal greenhouse, the study is not much affected by natural reasons like wind, dust, rain. The nodes are installed at a height of 100 cm above the ground while the height of repeaters is 150cm. The gateway is put on the first floor. The distance between the sensor nodes and the other nodes or repeaters is 15m. And the distance between the gateway and nearest node (light of sight) is 100m.

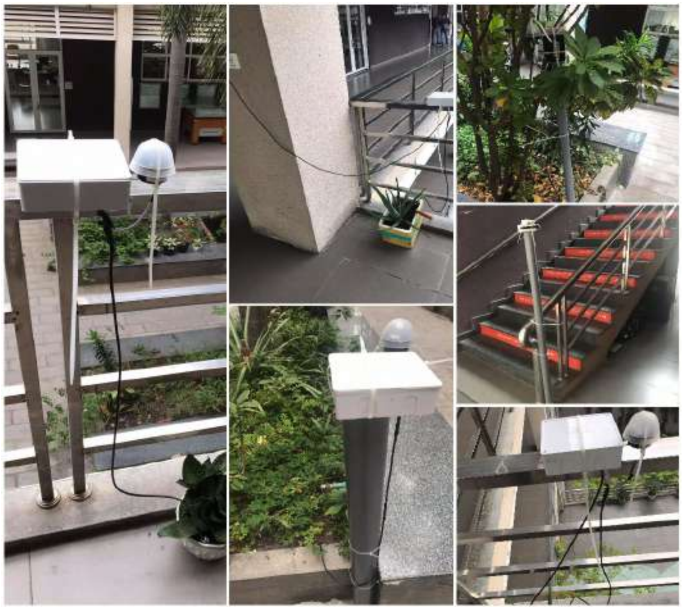


Fig. 7. Sensor Nodes and Repeaters Installation inside Vietnamese German University Campus

The results of the implementation of our system are shown:

1) Mesh Wi-Fi Network: The Fig.8. shows the Network topology of the system.

2) Power Consumption: This section demonstrates some data about battery tests. In this calculation, we performed the current measurement through the sensor node. We also calculated the total current consumption of the sensor node and found out how long the battery could last. The gateway is not included in this test because It is powered by the

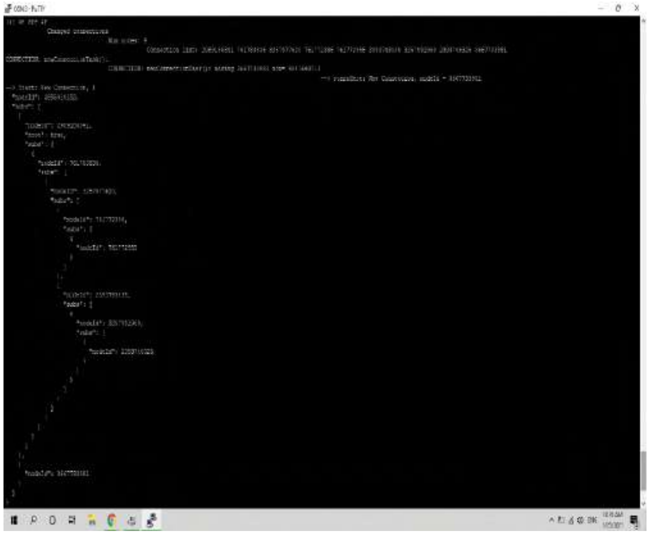


Fig. 8. The Topology of Mesh Network

wall outlet. The sensing device is powered by a small 3.7V Lithium Polymer (LiPo) battery, at a full charge voltage is 4.2V. The average power of node, measured, is 281mW (It is up to the distance of nodes). In our test, the line of sight distance of the two nodes is 15m, the current consumption is 74mA. Even when a node is not transmitting data, it has current consumption is about 70 mA [28] just because the Wi-Fi mode is always on so as to maintain the Mesh network. It also means that the current consumption of repeater is 70mA. However, it is negligible because according to [15] the current consumption of ESP8266 WEMOS D1 is 170mA in data transmission mode.

The battery life is calculated by the equation. It is very harmful to use an exhausted battery. In the battery, we attached a TP4056-Micro USB 5V 1A Lithium Battery Charger with Protection to protect it. The ideal time of the battery is calculated by 5000mAh/74mA = 67hours. However, in real testing, the time just lasts over 55 hours or within 2,5 days.

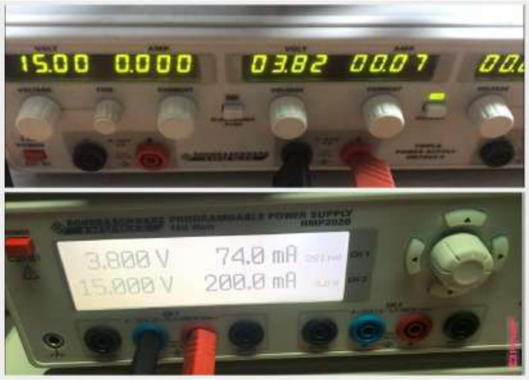
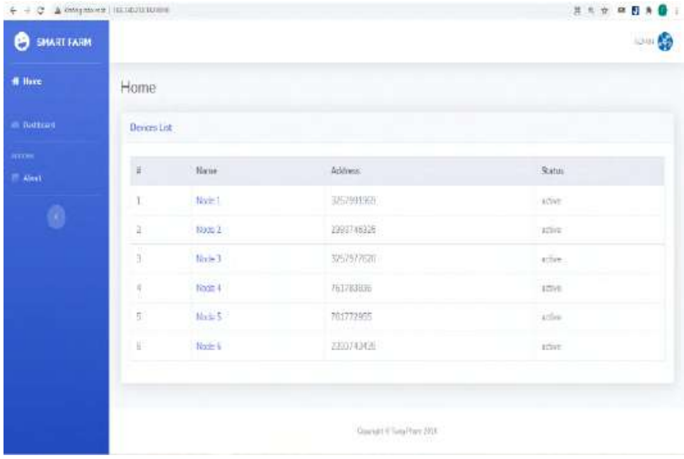


Fig. 9. Power Consumption of a Node

3) Functionality: The app's main screen is the map as shown in Fig.10. The monitoring platform server is available at the following URL: http://103.130.212.132:8000/. The main screen is the entry page the user sees when accessing the website. In there, there is a device list of nodes, node addresses, and the status activity as shown in Fig.10.



Home

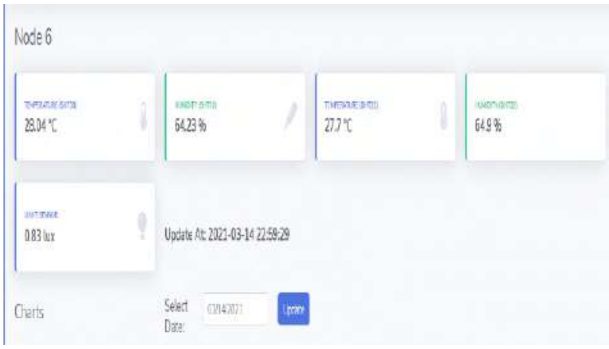
Devices List

#	Name	Address	Status
1	Node 1	3257991909	active
2	Node 2	2993746328	active
3	Node 3	3257577620	active
4	Node 4	761783836	active
5	Node 5	761772955	active
6	Node 6	2320743428	active

Fig. 10. The Homepage

When a node is clicked, the farmer can see more detailed information about the real-time data of its node. The real-time data is automatically updated without a refreshing web page. As shown in Fig.11, there are parameters and meaning:

- TEMPERATURE (SHT30) - The Value of soil temperature
- HUMIDITY (SHT30) - The Value of soil humidity
- TEMPERATURE (DHT22) - The Value of environmental temperature
- HUMIDITY (DHT22) - The Value of environmental humidity
- LIGHT SENSOR - The Value of environmental light intensity



Node 6

TEMPERATURE (SHT30)
28.04 °C

HUMIDITY (SHT30)
64.23 %

TEMPERATURE (DHT22)
27.7 °C

HUMIDITY (DHT22)
64.9 %

LIGHT SENSOR
0.83 lux

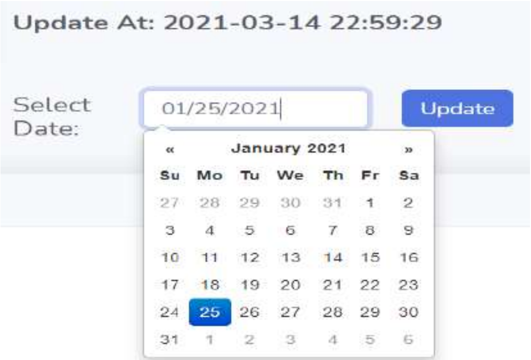
Update At: 2021-03-14 22:59:29

Charts

Select Date: 03/14/2021

Fig. 11. Real-time data and detailed information of a sensor node.

To see the historic data of the node in that day as well as the data of previous days, the users click the select date and click update. The activities can be described in Fig.12 and 13 respectively.



Update At: 2021-03-14 22:59:29

Select Date: 01/25/2021 Update

« January 2021 »

Su	Mo	Tu	We	Th	Fr	Sa
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Fig. 12. The Date Selection

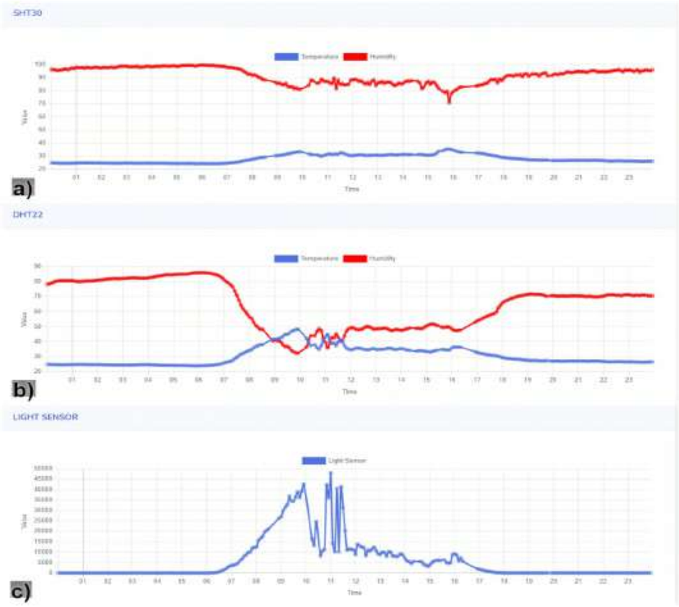


Fig. 13. The Graphs with sensor node data from 24-hour period

IV. Conclusions

In this paper, a cost-effective and highly reliable system with high communication speed is implemented by using a mesh Wi-Fi wireless sensor network. Wi-Fi Mesh Network and the MEAN stack web development framework play a significant role in the monitoring system.

This work opens up new doors for future studies as well as developed new tasks for the foreseeable future. The first main goal of the system is to deploy in a near scenario of an agriculture application. Next, the necessary tasks will be carried in the near future, referred to as short-term (ST) like Powering system, Irrigation system whereas long-term (LT) tasks like Security and Privacy, Edge Analytic should be researched more to enhance the productivity of the field.

References

[1] Ericsson, 2021, Connectivity is the foundation of IoT, accessed: March, https://www.ericsson.com/en/internet-of-things/iot-connectivity

[2] Pico-Valencia, J. A. Holgado-Terriza and X. Qui n onez-Ku, ”A BriefSurvey of the Main Internet-Based Approaches. An Outlook from the Internet of Things Perspective,” 2020 3rd International Conf.(ICICT), San Jose, CA, USA,2020, pp. 536-542, DOI: 10.1109/ICICT50521.2020.00091

[3] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, M.Palaniswami, ”Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions. Future Generation Computer Systems”, 29, 1645-1660. DOI: https://doi.org/10.1016/j.future.2013.01.010

[4] V.Grimblatt, ”IoT for Agribusiness: An overview,” 2020 IEEE 11th (LAS-CAS), San Jose, Costa Rica, 2020, pp.1-4, DOI:10.1109/LAS-CAS45839.2020.9068986.

[5] T. Guo, and W. Zhong, ”Design and implementation of the span greenhouse agriculture Internet of Things system,” in Proc. of IEEE International Conf. on Fluid Power and Mechatronics (FPM), 2015. DOI: 10.1109/FPM.2015.7337148

[6] M. Ryu, J. Yun, T. Miao, I.Y. Ahn, S.C. Choi, and J. Kim, “Design and implementation of a connected farm for smart farming system,” in Proc.of IEEE SENSORS, 2015.

[7] T. Qiu, H. Xiao, and P. Zhou, ”Framework and case studies of intelligence monitoring platform in facility agriculture ecosystem,” in Proc. of IEEE International Conference on Agro-Geoinformatics, 2013.

[8] Adafruit, 2021, SHT-30 Mesh-protected Weather-proof Temperature/Humidity Sensor-1M Cable, accessed: March https://www.adafruit.com/product/4099

[9] Mouser, 2021, DHT11, DHT22 and AM2302 Sensors, accessed: March https://www.mouser.com/datasheet/2/737/dht-932870.pdf

[10] DFrobot, 2021, Light Sensor-BH1750, accessed: March https://www.dfrobot.com/product-531.

[11] Wang Yanhui, Ji Xiaofei, 2021, ”The design of greenhouse lighting control system”, May 2015. DOI: 10.1109/CCDC.2015.7162363

[12] PlatformIO, 2021, What is PlatformIO?, accessed: March https://docs.platformio.org/en/latest/what-is-platformio.html

[13] Espressif, 2021, Espressif technical-documents, accessed: March https://www.espressif.com/en/support/documents/technical-documents

[14] ESP8266 Pinout Reference, 2021, Which GPIO pins should you use?, accessed: March https://randomnerdtutorials.com/esp8266-pinout-reference-gpios/

[15] Wemos, 2021, Wemos D1 mini documentation, accessed: March https://www.wemos.cc/en/latest/d1/d1-mini.html

[16] Espressif-ESP32 Series, 2021, ESP32 Series Datasheet, accessed: March https://www.espressif.com/sites/default/files/documentation/esp32-datasheet-en.pdf

[17] T.Yokotani and Y.Sasaki, ”Comparison with HTTP and MQTT on required network resources for IoT,” In (ICCEREC), 2016. DOI: 10.1109/ICCEREC.2016.7814989

[18] G.Lawton, 2021, ”LAMP lights enterprise development efforts”, IEEE Computer vol. 38, pp. 18-20, September 2005. DOI: 10.1109/MC.2005.304

[19] MEAN.io. (2015) MEAN—Full-Stack JavaScript UsingMongoDB, Express, AngularJS, and Node.js. accessed: March, http://mean.io/

[20] Stefan Tilkov and Steve Vinoski. ”Node.js: Using JavaScript to Build High Performance Network Programs”. vol. 14, pp. 80-83. Nov. 2010. DOI: 10.1109/MIC.2010.145

[21] K. I. D. K. I. K.Chaniotis And N. D.Tselikas, ”Is Node.js a viable option for building modern web applications? A performance evaluation study,” SPRINGER, Computing, VOL.97, PP. 1023-1044, OCT., 2015.

[22] ”The Modern Application Stack”. URL: https://www.mongodb.com/blog/post/the-modern-application-stack-part-1-introducing-the-mean-stack

[23] A. Leff and J. T. Rayfield. ”Web-application development using the Model/View/Controller design pattern”. Nov. 2001. DOI: 10.1109/EDOC.2001.950428

[24] Tp-link, 2021, Tp-link M7350 4g LTE Mobile WiFi Wireless Router/Hotspot Support to 15 Devices, accessed: March,https://www.amazon.com/Tp-link-M7350-Wireless-HotspotSupport/dp/B07NQKBDSJ

[25] Roger.A.Light. ”Mosquitto: server and client implementation of the MQTT protocol”. JOSS. May. 2017. DOI:10.21105/joss.00265

[26] Socket.io, 2021, accessed: March, https://www.npmjs.com/package/socket.io

[27] SB Admin 2, 2021, accessed: March,https://startbootstrap.com/theme/sb-admin-2

[28] Espressif-ESP8266 Series, 2021, ESP8266 Power Consumption, accessed: March https://bbs.espressif.com/viewtopic.php?t=133.

A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review

Deyby Huamanchahua
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, N.L., México
a00816108@itesm.mx

Diana Rosales-Gurmendi
Department of Mechatronics Engineering
Universidad Continental
Huancayo, Perú
72569080@continental.edu.pe

Yerson Taza-Aquino
Department of Mechatronics Engineering
Universidad Continental
Huancayo, Perú
74239368@continental.edu.pe

Dalma Valverde-Alania
Department of Industrial Engineering
Universidad Continental
Huancayo, Perú
71784357@continental.edu.pe

Miguel Cama-Iriarte
Department of Electronics Engineering
Universidad Nacional Tecnológica de Lima Sur
Lima, Perú
2013200524@untels.edu.pe

Adriana Vargas-Martinez
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, N.L., México
adriana.vargas.mtz@tec.mx

Ricardo A. Ramirez-Mendoza
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, N.L., México
ricardo.ramirez@tec.mx

Abstract*- *There are many prostheses focused on people with an amputation below the elbow. There is also limited information for its development and construction, many studies provide information that helps the researcher, but these are scattered, each one proposing different characteristics and solutions. Based on this, the objective is to provide the researcher with a structured matrix that integrates different studies related to hand prosthesis, this will allow him to evaluate alternatives where he will be able to choose, analyze the study or characteristics that better contribute to his research topic. As methods, it used specialized search engines that helped to structure the matrix. Currently there is little information on prostheses based on encephalographic signals, so it was also incorporated into the matrix, likewise in the proposals it considers from 2 to 15 DoF (degrees of freedom), actuators used, and the classification of prostheses that are active, passive, electric, myoelectric and hybrid. In conclusion, the article will provide information to people who want to create a superficial extension below the elbow and make use of the control of electromyographic signals that are extracted from muscle contractions.

Keywords: Electromyography - Encephalography – Upper-Limb - Prosthesis - Degrees of freedom

I. Introduction

Currently, the development and creation of prostheses for people with upper-limb amputations are progressive because they are required for the advantages they can provide to the patient. But, although it knows that some prostheses are developed with a direct approach to the clinical problem and others created from the patient's specifications, then the problem arises to know if the artificial extension acquired meets the requirements that the patient wishes to obtain.

A very particular problem that affected people present are the injuries produced below the elbow, it is estimated that worldwide there are at least 10 million amputees, of which 30% (3 million) are amputations of the arm and hand, of which 2.4 million are people living in developing countries. [1] Given the problem, hand prostheses should be analyzed with respective importance since a well-developed prosthesis substantially improves the affected person's quality of life. On the other hand, it can see already models of feedback prosthesis as the myoelectric prosthesis (Otto Bock) and the most marketed which is the bionic prosthesis I-Limby. Each one is specifically differentiated by its characteristics, but despite the advances that hand prostheses show day by day, there are still limited functions that do not achieve the real performance of the hand in its totality. Any study can be achieved if it has a good base of relevant data. In general, most of the studies, literature reviews of hand prosthesis propose individual solutions considering separate characteristics of which it could be observed that independently each one contains very important information and these organized together can contribute a lot for the development of a hand prosthesis.

3D printing was seen as a new trend due to its low cost. This is the case of Ruiz et al.[2] who created a 3D prototype using electromyography, focused on the field of rehabilitation and with greater accessibility, Piazza et al. [3] showed the possibility of realizing prosthetic systems from a hybrid and electrical solution, using inputs from a shoulder harness to control 19 degrees of freedom (19 DoF), creating the SoftHand Pro-H, On the other hand, Castillo et al. [4] designed and fabricated a hand prosthesis with anthropometric and anthropomorphic characteristics extracted from radiographs, and Cognolato et al [5] created an efficient myoelectric control system based on the Myo bracelet and used a gesture classifier to control it.

Finally, there is already a prosthesis focused on upper-limb amputees who work in offices and want to resume work at the computer can do so thanks to this design where Lopez et al. [6] proposed a 3D prosthesis controlled by EMG signals, which facilitates the use of keyboards for development. Likewise, if it analyzes and evaluates the characteristics, it will provide complete information for people who want to develop a hand prosthesis. Therefore, this research aims to provide a structured

978-1-6654-4067-7 / 21 / $ 31.00 © 2021 IEEE

matrix with information from different studies that contain relevant data and provide true data. For the development of the matrix, it considered studies that were carried out in a certain period and with this, it has current information, it details the control inputs used in each study, the degrees of freedom they develop, the type of loss of the affected people, type of prosthesis, the actuators they used and finally the state of development in which it is.

II. Data Collection Methodology

The search was carried out in the specialized search engines "Google Schola", "IEEE Explore", "PubMed" and "ERIC" also open access articles from journals such as "Science", "Redalyc" as well as collaboration and dissemination platforms such as "ResearchGate", "Elsevier" and "Dialnet" were examined. The publication period of the selected articles spans from December 2016 to May 2020. Since the review aims to analyze publications that consider the design of upper-limb prostheses, either the control system or the physical implementation, the search terms used were: "prosthetic hand", "design" and "DoF" (degrees of free). A total of 83 articles were collected, 28 in Spanish and 55 in English. The result of the reading, analysis, interpretation, and integration of the information from each source is shown in the text presented here.

III. Publications Review

The reviewed publications of the last 5 years on upper-limb prostheses are shown in Table 1 in detail, considering the most important aspects of the publication.

TABLE I. ROBOTIC PROSTHESIS FOR UPPER-LIMB

Name / Ref.	Control input	DOF	Loss Type	Prosthesis type	Actuators	TRL
Martínez A.[7]	EMG	5(fingers)	ATR	Myoelectrical	S	7
Gretsch KF.[8]	IMU	10(fingers)	ATR	Hybrid	S	7
Zhao H [9]	Light, Force	5	ADM	Electrical	AS	7
Rodríguez ME [10]	EMG	2	ADM	Myoelectrical	S	7
Bennett [11]	IMU, EMG	1(wrist)	ATR	Myoelectrical	DC	7
Daniel B [12]	EMG	10	ADM	Myoelectrical	DC	7
Rodríguez [13]	EMG	2	ATR	Myoelectrical	S	7
García C [14]	Fuerza	15	N.E.	Active	S	7
Yepez MA [15]	N.E.	9	ATR	Mechanical	S	3
Armas AE [16]	EMG	15	ADM-ATR	Myoelectrical	S	4
Xiangxin L. [17]	sEMG-EEG	N.E.	ATH	Myoelectrical	N.E.	3
Ayats [18]	N.E.	15	ADM	Mechanical	DC	4
Alvial P. [19]	N.E.	2	APM	Mechanical	N.E.	4
Geethanjali [20]	sEMG	15	ATR	Myoelectrical	DC	3
Piazza [3]	sEMG	7	ATR	Hybrid	ML	9
Bandara [21]	N.E.	15	ATH	Active	DC	3
Mustafa N [22]	sEMG	N.E.	ATR	Myoelectrical	DC	3
Castillo [4]	EEG	19	ADM	Hybrid	MC	9
Controzzi M. [23]	EMG, Force, Position	4	ATR	Myoelectrical	DC	7
Proaño [24]	N.E.	3	ATR	Active	MC	7
Silva [25]	sEMG	N.E.	ATR	Myoelectrical	S	7
Lee K. [26]	sEMG -Pressure	N.E.	ATR	Myoelectrical	N.E.	3
Ortega [27]	Force - Position	4	ADM	Hybrid	DC-S	7
Abbasi SH [28]	N.E.	5	ATR	Active	DC-S	3
Fourie R [29]	sEMG	N.E.	ADM	Active	MC	9
PonPriya P [30]	sEMG	3	ATR	Myoelectrical	DC	7
Wattanasiri [31]	N.E.	1	ADM	Hybrid	DC	7

Vignesh [32]	EMG	5	ADM	Hybrid	DC-S	7
Benatti [33]	EMG	N.E.	ATR	Myoelectrical	DC	3
Krasoulis [34]	sEMG - IMU	N.E.	ATR	Myoelectrical	N.E.	3
Rodríguez [35]	N.E.	5(fingers)	ADM	Mechanical	N.E.	7
Rodríguez [36]	N.E.	15	ADM	Hybrid	S	7
Kai Xu [37]	N.E.	8 (fingers)	ATMC	Active	N.E.	7
Yu Wu [38]	EIT	1(wrist)	N.E.	Electrical	N.E.	7
Cognolato M. [5]	sEMG - IMU	15(fingers)	ADM - ATR	Myoelectrical	S	9
Hurtado P [39]	sEMG - IMU	14(fingers)	ATR	Myoelectrical	S	7
Gonzalez [40]	EMG	10(fingers), 1(wrist)	ADM	Myoelectrical	S	7
Vujaklija I [41]	EMG	19	ATC, ATR y ATH	Hybrid	DC	9
Semasinghe CL [42]	N.E.	15	ATR	Hybrid	S	7
Zhang [43]	EMG	10	ADM	Hybrid	DC	7
Kanitz [44]	EMG	N.E.	ADM - ATR	Myoelectrical	N.E.	3
Lizarraga [45]	EMG - IMU	19(fingers)	N.E.	Myoelectrical	N.E.	4
Atasoy [46]	EMG - IMU	24	ADM	Hybrid	DC	3
Harada [47]	EMG	2	ADM - ATR	Myoelectrical	N.E.	7
Hussain [48]	EMG - IMU	15	ADM	Myoelectrical	S	7
Ismail [49]	sEMG	5	ATR	Myoelectrical	ML	7
Teng [50]	EEG	5	ADM - ATR	Hybrid	ML	7
Marasco [51]	EMG	N.E.	ATH	Myoelectrical	N.E.	3
Hahne [52]	EMG	2	ATR bilateral y DC	Myoelectrical	N.E.	7
O'Brien [53]	Force, Position	15	ADM	Hybrid	DC	4
Galileo Hand , Fujiwara [54]	Light	21	N.E.	Virtual manipulator	N.E.	3
Bejarano [55]	sEMG	N.E.	ATR	Myoelectrical	MS	9
Javier R [56]	EMG - IMU	2	ATR	Myoelectrical	S	7
Vargas [57]	EMG	5	ATR	Myoelectrical	S	7
Vargas [58]	EMG	5	ATR	Myoelectrical	ML	7
Cuellar [59]	N.E.	5	ATR	Myoelectrical	N.E.	7
López [6]	sEMG	15	ADM	Myoelectrical	S	9
Beninati [60]	EMG - IMU	5(fingers), 1(wrist)	ATR	Myoelectrical	ML	7
Williams [61]	EMG	5(fingers), 1(wrist)	ATR	Myoelectrical	N.E.	3
Alvarado [62]	sEMG	N.E.	ATR	Myoelectrical	MS	7
Valadez [63]	EMG	5(fingers)	ATR	Myoelectrical	S	4
Nesrine A [64]	EMG	5(fingers)	ATR	Myoelectrical	N.E.	3
Prakash [65]	EMG	5(fingers),1(wrist)	ATR	Myoelectrical	S	7
Iwatsukia [66]	EMG	N.E.	ATR	Myoelectrical	S	3
Ku [67]	EMG	16 (fingers)	ATR	Myoelectrical	N.E.	9
Pérez [68]	EMG-IMU	N.E.	N.E.	Myoelectrical	N.E.	9

Areiza [69]	Pulse	14 (fingers)	ADM	Electrical	N.E.	7
Li [70]	EMG	N.E.	ADH - ATH	Myoelectrical	N.E.	7
Ruiz [2]	EMG-gestures	N.E.	ADM	Myoelectrical	MS	7
Parajuli [71]	EMG	N.E.	N.E.	Myoelectrical	N.E.	3
Linares [72]	EMG	10(fingers),3(wrist)	ATR	Myoelectrical	S	7
Salazar [73]	N.E.	N.E.	DPM	Mechanical	S	7
Alkhatib [74]	N.E.	N.E.	ADM	Passive	N.E.	7
Cruz [75]	sEEG	N.E.	ATR	Myoelectrical	S	7
Cruz A[76]	N.E.	N.E.	ATR	Myoelectrical	S	7
Gonzalez [77]	sEMG	N.E.	ATR	Myoelectrical	S	7
Shibanoki [78]	EMG - IMU	N.E.	ATR	Myoelectrical	N.E.	7
Mohammadi [79]	EMG	N.E.	ATR	Myoelectrical	DC	7
Higa [80]	EMG -Force	N.E.	N.E.	Myoelectrical	S	4
Sherif [81]	EMG	9	ATR	Myoelectrical	ML	7
Ojeda [82]	EMG	5(fingers)	ATC	Myoelectrical	S	6
Alturkistani [83]	N.E.	6(fingers)	ATMC	Passive	N.E.	7

Note: Abbreviation: EMG: Electromyography, EEG: Electroencephalography, sEMG: Surface electromyography, EIT: Electrical impedance tomography, IMU: Internal Measurement Unit, ATR: Transradial amputation, ADM: Wrist disarticulation amputation, APM: Partial hand amputation, ATH: Transhumeral amputation, DC: Congenital deficiency, ADH: Amputation of the disarticulated shoulder, DPM: Partial deformation of the hand, ATC: Transcarpal amputation, ATMC: Transmetacarpal amputation, ML: linear actuator, DC: DC motor, MC: micromotor, MS: micro-servo. , AS soft actuator, TRL 1: Basic research, TRL 2: Technology formulation, TRL 3: Applied research, TRL 4: Small scale development, TRL 5: Full-scale development, TRL 6: Validated system in a simulated environment, TRL 7: Validated system in a real environment, N. E.: Not specified

A. Control inputs

It is possible to use different signals as device control inputs, depending on the type of control strategy to be used.

In Table II. Among the most used signals are electromyography (EMG) signals with a 47.6% share. These signals are used for the control of upper-limb prostheses. It is called (EMG) because the signals are captured from muscle contractions. The graph also shows that 13.4% of the EMG signals are complemented by inertial sensors (IMU). For example, commercial devices such as the Myo armband bracelets use sensors on their control card that read EMG and inertial signals (IMU). These devices have been used for the control of prostheses in various investigations. 84] On the other hand, other prostheses use other types of sensors complementary to EMG, such as those that measure force, EIT, pulse, EMG-EEG, EMG-pressure, EMG-gestures occupying 1.2% and those of light, force-position with 2.4%, but to date, they register an infrequent use and finally there are unspecified control inputs with 20.7%.

TABLE II. DISTRIBUTION PER CONTROL INPUT

Control input types	Quantity	
	Frequency of use	Percentage %
EMG	39	47.6%
Not specified	17	20.7%
EMG - IMU	11	13.4%
EEG	3	3.7%
Force – Position	2	2.4%
Light	2	2.4%
EMG–Force-Position	1	1.2%
EMG-Gestures	1	1.2%
EMG-Force	1	1.2%
EMG-Pressure	1	1.2%
EMG-EEG	1	1.2%
Pulse	1	1.2%
Force	1	1.2%
EIT	1	1.2%

B. Degrees of freedom (DOF)

According to the model proposed by Quiang Zhan [85] the human hand has 21 degrees of freedom [DoF]. The index, middle, ring, and little finger consist of 3 phalanges and 4 DoF, the thumb of 2 phalanges and 51 DoF, and the wrist of 3 DoF. See Fig1.

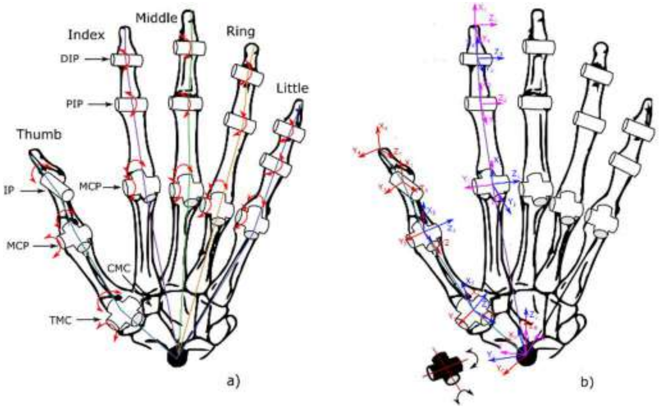


Fig. 1. Qiang Zhan 21 DoF model of the human hand. a) structural model and b) kinematic model Source: Zhan Q, Zhang C, and Xu Q. Measurement and Description of Human Hand Movement [2017].

It is considered ideal that a hand prosthesis manages to represent the total DoF that make up the hand and wrist, but generally, among the prototypes and designs proposed by various authors referenced in the research, they are mainly focused on representing between 2 to 15 DoF of the hand to perform the opening and closing movements of the fingers, among other movements. Likewise, the degrees of freedom of the wrist given to the prosthesis is necessary; most of the proposed prototypes consider between 1 to 2 DoF of the wrist. In the market, there are more advanced prostheses whose mechanism deploys the 3 DoF of the wrist, giving greater versatility of the prosthesis to perform more precise tasks, but they have a less accessible cost.

C. *Loss type*

There are different types of upper-limb prostheses in the biomedical market for users with congenital malformations or who have suffered a loss of the entire limb or specific segments.

In Table III. In the studies reviewed in this research, it was found that most researchers developed their prosthesis proposals oriented at 47.6% for users with trans-radial amputation (ATR) and 24.4% for users with amputation with wrist disarticulation (ADM). It was also found for users with other types of loss such as ADH (amputation with shoulder disarticulation) and ATH (transhumeral) with the participation of 6.1% as well as only ATR and ATMC (transmetacarpal) with 2.4%. Prostheses adaptable to their condition were also developed, but in smaller numbers, such as APM, ATC, DPM, ADH-ATH, ATR-DC, ATC-ATR-ATH, and 1.2%, and finally, there is a percentage of unspecified type of loss with 9.8%.

TABLE III. DISTRIBUTION BY LOSS TYPE

Loss Type	Quantity	
	Frequency	Percentage %
ATR	39	47.6%
ADM	20	24.4%
Not specified	8	9.8%
ADM-ATR	5	6.1%
ATMC	2	2.4%
ATH	2	2.4%
APM	1	1.2%
ATC	1	1.2%
DPM	1	1.2%
ADH-ATH	1	1.2%
ATR-DC	1	1.2%
ATC-ATR-ATH	1	1.2%

D. *Movement type*

From a physiological point of view, the hand represents the upper-limb's effector that forms its logistic support and enables it to adopt the most appropriate position for a specific action. [86] It should be noted that the hand is a complex limb; therefore, it is essential that the design of a prosthesis takes into consideration the movements and articular functions of the limb. According to the articular physiology of the wrist, Fig. 2 a) shows the terminology of the movements that the joint executes; similarly, Fig. 2 b) shows the terminology of the movements of the hand and fingers.

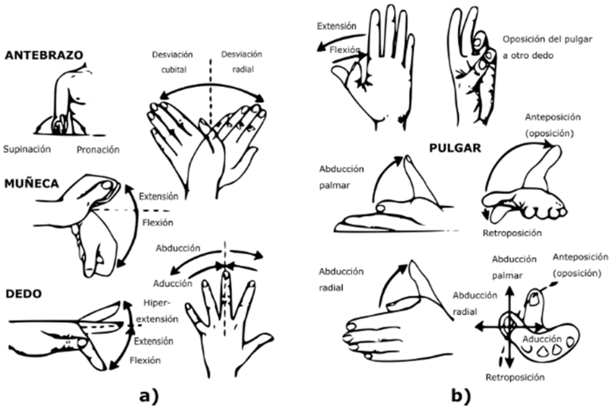


Fig. 2. a) Terminology of wrist and finger movements, b) Terminology of hand and finger movements. Sources: "The Hand: Examination and Diagnosis"(p. 11) by American Society for Surgery of the Hand.

Concerning the articles reviewed in this research in the last 5 years, the researchers considered some of the hand's movements and versatility in the development of the design and construction of the upper-limb prosthesis proposals. Table IV shows the most performed movements and functions that were granted to the prostheses.

TABLE IV. MOST PERFORMED MOVEMENTS AND FUNCTIONS

Upper-limb segment	Movement type
Hand	Opening and closing Clamp Hook grip Cylindrical grip Thumb abduction and adduction
Wrist	Pronation Supination Ulnar deviation Radial deviation Flexion Extension

In the upper-limb prostheses proposed in the research, the researchers represented up to 2DoF of the 3DoF that make up the wrist joint. This shows a limitation in their proposals. Among the most represented movements in the design and construction of the prostheses were the pronation-supination and flexion-extension movements.

E. *Prosthesis type*

In this area, 7 types of prostheses were found: myoelectric, mechanical, active, hybrid, electric, passive, and virtual.

Table V. Myoelectric prostheses with a percentage of 62.2% are the most applied since they currently have aesthetics, strength, and speed of grip, for myoelectric control, is based on the muscles' electrical signals (EMG). [87] The mechanical prostheses with a percentage of 7.3%, unlike the myoelectric ones, do not use any type of sensor for their activation. The control applied is using movements of the elbow, shoulders, or chest.

The active prostheses have a percentage of 7.3%, this type of prosthesis is defined because they have movement, unlike the passive prostheses with a 2.4% that do not generate any kind of movement its application is aesthetic and hide the injury that has in the upper-limbs, also shown with 15.9% to hybrid prostheses this type of prosthesis is practically the combination of myoelectric and mechanical prosthesis making it more versatile. [88]

Finally, it is observed that the types of electrical and virtual prosthesis occupying 3.7% and 1.2% values are minimal because its application, unlike myoelectric, hybrid, and mechanical prostheses, does not have the requirements to replace the operation of a human upper-limb.

TABLE V. DISTRIBUTION BY TYPE OF PROSTHESIS

Prosthesis Type	**Quantity**	
	Frecuencia	**Percentage %**
Myoelectrical	51	62.2%
Hybrid	13	15.9%
Mechanical	16	7.3%
Active	8	7.3%
Electrical	4	3.7%
Passive	2	2.4%
Virtual	1	1.2%

F. *Actuators*

In this area, a review of the types of actuators used for the development of prostheses, which are important to be able to mimic the movements of each joint that make up a human arm, 5 types of actuators were identified: linear actuator (ML), motors (DC), servomotor (S), micromotor (MC), micro servo (MS) and the conjugation of the DC-S motor.

In Table VI. The servomotors have a percentage of 32.9% and are the most used because of the ease of controlling these actuators since they have delimited rotations and are designed to move several degrees and remain stable in that position. DC motors have a percentage of 17.1%; these actuators have magnets inside generating a magnetic field to be powered by electricity-generating rotary movements. It is also observed that linear actuators have a percentage of 7.3%, placing them as the third most used actuator for the development of these robotic devices, the most relevant characteristic of these actuators is the movement in a straight line. Finally, the MS, DC-S actuators are found with 3.7%, which are not so widely used for the development of these robotic devices. And among others that could not be specified in the studies.

TABLE VI. DISTRIBUTION BY ACTUATOR TYPE

Prosthesis Type	**Quantity**	
	Frequency	**Percentage %**
S	27	31.7%
Not specified	26	32.9%
DC	14	17.1%
ML	6	7.3%
MC	3	3.7%
MS	3	3.7%
DC-S	3	3.7%

G. *Prosthesis development status*

In this area, a review of the status of prostheses such as prototype, design, study, simulation, comparison, and commercial prostheses was carried out.

Table VII. It shows the more significant number of studies that are reached with prototyping obtaining a percentage of 60.2% and 14.5% of the research remains in studies contributing new methods of prosthesis development. The 9.6% of the research has a validation of the functioning of its prototypes, and they are exposed in the market for its commercialization. It was also found research that reached to make the design of its prototypes with 9.6% of the total of the reviewed research. Furthermore, it found research that made comparisons of new control methods occupying 2.4% and simulations with 3.6% contributing to improving future research related to the creation of these robotic devices.

TABLE VII. PROPORTION BY STAGE OF DEVELOPMENT

Types of Advance	**Quantity**	
	Frequency	**Percentage %**
Prototype	50	60.2%
Studies	12	14.5%
Commercial Prosthesis	8	9.6%
Design	8	9.6%
Simulation	3	3.6%
Comparison	2	2.4%

IV. Conclusion

Since the number of users requesting the use of an upper-limb prosthesis due to traumatic amputation or malformation has been increasing in the last 5 years, numerous research groups have made proposals for upper-limb prostheses usually controlled with myoelectric signals and position signals from inertial sensors (IMU), with remarkable results. Similarly, proposals have developed from 2 to 15 DoF for finger movement in addition to wrist movement, but are still in the development stage.

The proposals are oriented to users with an amputation below the elbow since their control is based on acquiring muscle signals from the forearm. It should be noted that with these devices the user will be able to perform multiple grips, perhaps not all the movements made by the hand, but the necessary ones. The article reviewed various investigations regarding designs and prototypes of upper-limb prostheses, several aspects need to be improved such as the cost of development, ergonomics to achieve better adaptability with the upper-limb in question.

A lot of research has been done on material types. However, most of the devices did not focus on a special material that will have contact with the patient's residual limb as this is important in order not to generate injuries when using these robotic devices. There is still room for improvement and further development of more prostheses that come closer to the actual functioning of a human upper-limb.

REFERENCES

[1] "Amputee Coalition," Limb Loss Statistics, 2016.

[2] J. P. Ruiz Cea, C. A. Mejia Romo, A. S. Ortega Ravelo and B. A. Diaz Legaria, "Diseño, modelado y construcción de una mano robótica 3D empleando electromiografía," LaSalle, 2019.

[3] C. Piazza, M. Catalano, S. B. Godfrey, M. Rossi, G. Grioli and M. Bianchi, "El SoftHand Pro-H," IEEEXplore, 2017.

[4] E. Castillo Castañeda and A. Bernardo Vásquez, "Diseño personalizado de una prótesis de mano considerando la antropometría de una mano real extraída de radiografía.," IEEEXplore, 2017.

[5] M. Cognolato, M. Atzori, C. Marchesini, S. Marangon, D. Faccio, C. Tiengo, F. Bassetto, R. Gassert, N. Petrone and H. Muller, "Multifunction control and evaluation of a 3D printed hand prosthesis," bioRxiv, Octubre 2018.

[6] E. E. López López, R. Martínez Méndez and A. Vilchis González, "Diseño de una prótesis de mano para uso en teclados con interfaz sEMG," BIOMEDICA, vol. 8, no. 1, mayo 2019.

[7] A. Martínez Miguel, S. A. Vargas Pérez, E. Gómez Merlín, M. Arias Montiel, E. Lugo González and R. Miranda Luna, "Control de Movimiento de una Mano Robótica Mediante Señales Electromiográficas," 2016.

[8] K. Gretsch, H. Lather, K. Peddada, C. Deeken, L. Wall and C. Goldfarb, "Development of novel 3D-printed robotic prosthetic for transradial amputees," Prosthetics and Orthotics International, vol. 40, no. 3, pp. 400-403, 2016.

[9] H. Zhao, K. O'Brien, S. Li and R. Shepherd, "Optoelectronically innervated soft prosthetic hand via stretchable optical waveguides," Sci Robot, vol. 1, no. 1, diciembre 2016.

[10] M. E. Rodriguez Garcia, G. Dorantes Mendez and M. O. Mendoza Gutierrez, "Desarrollo de una Prótesis para Desarticulado de Muñeca Controlada por Señales de Electromiografía.," Revista mexicana de ingeniería biomédica, vol. 38, no. 3, pp. 602-620, 2017.

[11] D. A. Bennett and M. Goldfarb, "IMU-Based Wrist Rotation Control of a Transradial Myoelectric Prosthesis," IEEE Trans Neural Syst Rehabil Eng, vol. 26, no. 2, pp. 419-427, Febrero 2018.

[12] D. B, M. D, G. B and E. D, "Diseño de un Electromiógrafo Implementado Sobre una Prótesis de Mano," in Memorias del Congreso Nacional de Ingeniería Biomédica, 2017.

[13] M. E. Rodriguez Garcia, G. Dorantes Méndez and M. O. Mendoza Gutierrez, "Development of a Myoelectric-Controlled Prosthesis for Transradial Amputees," Mexican Journal of Biomedical Engineering, vol. 38, no. 3, pp. 602-620, septiembre 2017.

[14] C. García, B. Osuna and F. Martínez, "Esquema para el Control de una Mano Robotica Antropomorfa," Bogota, 2017.

[15] M. D. Yépez Rosero and C. A. Villareal Bolaños, "Diseño mecánico de un prototipo de prótesis de mano," 2017.

[16] A. E. Armas Alvarez, A. K. López Castañeda, I. Uriarte, M. A. Díaz and N. A. Barboza, "Control de modelo de prótesis de mano por señal mioeléctrica," in Memorias del Congreso Nacional de Ingeniería Biomédica, 2017.

[17] X. Li, O. W. Samuel, X. Zhang, H. Wang, P. Fang and G. Li, "A motion-classification strategy based on sEMG-EEG signal combination for upper-limb amputees," Journal of NeuroEngineering and Rehabilitation, vol. 14, no. 2, 2017.

[18] M. Ayats and R. Suarez, "Diseño de una prótesis de mano adaptable al crecimiento," In XXXVIII Jornadas de Automática, pp. 664-671, 2017.

[19] P. Alvial, G. Bravo, p. Bustos, G. Moreno, R. Alfaro, R. Cancino and J. Zagal, "Quantitative functional evaluation of a 3D-printed silicone-embedded prosthesis for partial hand amputation: A case report," Journal of Hand Therapy, vol. 31, no. 1, 2017.

[20] P. Geethanjali, "A mechatronics platform to study prosthetic hand control using EMG signals," Australas Phys Eng Sci Med, vol. 39, no. 3, 2016.

[21] D. Bandara, R. Gopura, K. Hemapala and K. Kiguchi, "Development of a multi-DoF transhumeral robotic arm prosthesis," Medical Engineering & Physics, vol. 48, pp. 131-141, 2017.

[22] M. Nadi and R. Midha, "Myoelectric functional hand prosthesis for total brachial plexus injury," Journal of Neurosurgery, 2017.

[23] M. Controzzi, F. Clemente, D. Barone, A. Ghionzoli and C. Cipriani, "los SSSA-MyHand: un diestro ligero," IEEE TNSRE, vol. 25, no. 5, pp. 459 - 468, 2017.

[24] A. Proaño Rosero, A. Lastre, K. Esparza and D. Zurita, "Parametrización de prótesis de mano usando el diseño asistido por computadora," lauinvestiga, vol. 4, no. 1, 2017.

[25] A. Silva - Moreno and E. Lucas Torres, "Design of a Customized Myoelectric Hand Prosthesis.," 2017.

[26] K. Lee, H. Bin, K. Kim, S. Y. Ahn, B.-O. Kim and S.-K. Bok, "Funciones manuales de mioeléctrico e impreso en 3D Prótesis con sensor de presión. Un estudio comparativo," Annals of Rehabilitation Medicine, vol. 41, no. 5, pp. 875-880., 2017.

[27] O. Ortega, "Diseño del sistema de control de un prototipo de prótesis de mano," 2017.

[28] S. Abbasi and Mahmood, "Bond Graph Modelling of a Customized Anthropomorphic Prosthetic Hand with LQR Control Synthesis," in International Multi-topic Conference (INMIC).

[29] R. Fourie and R. Stopforth, "The mechanical design of a biologically inspired prosthetic hand, the touch hand 3," in 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), 2017.

[30] P. PonPriya and E. Priya, "Design and control of prosthetic hand using myoelectric signal," in 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), 2017.

[31] P. Wattanasiri, P. Tangporprasert and C. Virulsri, "Design of Multi-Grip Patterns Prosthetic Hand with Single Actuator," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 6, 2017.

[32] T. Vignesh, P. Karthikeyan and S. Sridevi, "Modeling and Trajectory Generation of Bionic Hand for Dexterous Task," in INCOS, 2017.

[33] S. Benatti, B. Milosevic, . E. Farella, . E. Gruppioni and L. Benini, "A Prosthetic Hand Body Area Controller Based on Efficient Pattern Recognition Control Strategies," Sensors (Basel), vol. 17, no. 4, 2017.

[34] A. Krasoulis, I. Kyranou, M. Erden, K. Nazarpour and S. Vijayakumar, "Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements," Journal of NeuroEngineering and Rehabilitation, vol. 14, no. 71, 2017.

[35] V. Rodriguez and J. Salaña, "Prótesis en impresión 3D de bajo costo"Hand to Hand"," 2018.

[36] R. Rodriguez, J. Aroca Trujillo, D. Delgado and R. Sagaro Zamora, "Diseño e implementación de una prótesis de mano robótica antropomórfica subactuada," in AmITIC 2018, David, 2018.

[37] K. Xu, H. Liu, Z. Zhang and X. Zhu, "Wrist-Powered Partial Hand Prosthesis Using a Continuum Whiffle Tree Mechanism: A Case Study," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 3, 2018.

[38] Y. Wu, D. Jiang, X. Liu, R. Bayford and A. Demosthenous, "A Human-Machine Interface Using Electrical Impedance Tomography for Hand Prosthesis Control," IEEE Trans Biomed Circuits Syst, vol. 12, no. 6, 2018.

[39] P. Hurtado Manzanera, D. Luviano Cruz, L. Vidal Portilla and L. A. García Villalba, "Diseño y construcción de un prototipo de prótesis mioeléctrica," Mundo Fesc, vol. 15, no. 8, pp. 14-25, 2018.

[40] G. González Badillo, D. I. Torres Urestí, V. Espinoza López and G. Guerrero Mora, "“Control de una prótesis de mano fabricada por impresión 3D utilizando señales electromiográficas y lógica difusa," SOMIM, 2018.

[41] I. Vujaklija and D. Farina, "3D printed upper limb prosthetics," Expert Review of Medical Devices, 2018.

[42] C. Semasinghe, R. Ranaweera, J. Prasanna, H. Kandamby, D. Madusanka and R. Gopura, "HyPro: A Multi-DoF Hybrid-Powered Transradial," Journal of Robotics, 2018.

[43] T. Zhang, L. Jiang and H. Liu, "Design and Functional Evaluation of a Dexterous Myoelectric Hand Prosthesis with Biomimetic Tactile Sensor," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 7, pp. 1391 - 1399, 2018.

[44] G. Kanitz, C. Cipriani and B. Edin, "Classification of transient myoelectric signals for the control of multi-grasp hand prostheses," IEEE Trans Neural Syst Rehabil Eng, vol. 26, no. 9, pp. 1756-1764, septiembre 2018.

[45] J. Lizrraga Rodrguez, L. Lechuga Gutierrez and E. Bayro Corrochano, "Spike Quaternion Neural Networks Control for a Hand prosthesis," in 2018 IEEE Latin American Conference on Computational Intelligence (LACCI), 2018.

[46] A. Atasoy, E. Toptay, S. Kuchimov, S. Gulfize, M. Turpcu, E. Kaplanoglu, B. Guclu and M. Ozkan, "Biomechanical Design of an Anthropomorphic Prosthetic Hand," in 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob), 2018.

[47] T. Harada, S. Togo, Y. Jiang and H. Yokoi, "Development of Myoelectric Prosthetic Hand Control System Using Mobile Terminal," in IEEE International Conference on Intelligence and Safety for Robotics, Shenyang, 2018.

[48] I. Hussain, Z. Iqbal, M. Malvezzi, L. Seneviratne, D. Gan and D. Prattichizzo, "Modeling and Prototyping of a Soft Prosthetic Hand Exploiting Joint Compliance and Modularity," in 2018 IEEE International Conference on Robotics and Biomimetics, Kuala Lumpur, 2018.

[49] R. Ismail, M. Ariyanto, W. Caesarendra, G. Wijaya and A. Suriyanto, "Development of Myoelectric Prosthetic Hand based on Arduino IDE and Visual C# for Trans-radial Amputee in Indonesia," in International Conference on Applied Engineering (ICAE), Batam, 2018.

[50] Z. Teng, G. Xu, R. Liang, M. Li, S. Zhang, J. Chen and C. Han, "Design of an Underactuated Prosthetic Hand with Flexible Multi-Joint Fingers and EEG-Based Control," in 2018 IEEE International Conference on Cyborg and Bionic Systems, Shenzhen, 2018.

[51] P. Marasco, J. Hebert, J. Sensinger, C. Shell, J. Schofield, Z. Thumser, R. Nataraj, D. Beckler, M. Dawson and D. Blustein, "Illusory movement perception improves motor control for prosthetic hands," SCIENCE TRANSLATIONAL MEDICINE, vol. 10, no. 432, 2018.

[52] J. Hahne, M. Schweisfurth, M. Koppe and D. Farina, "Simultaneous control of multiple functions of bionic hand prostheses: Performance and robustness in end users," SCIENCE ROBOTICS, vol. 3, no. 19, 2018.

[53] K. O'Brien, P. Xu, D. Levine, C. Aubin, H.-J. Yang, M. Xiao, L. Wiesner and R. Shepherd, "Elastomeric passive transmission for autonomous force-velocity adaptation applied to 3D-printed prosthetics," Science Robotics, vol. 3, no. 23, 2018.

[54] E. Fujiwara, Y. T. Wu, C. Suzuki, D. Guedes de Andrade, A. Neto and E. Rohmer, "Optical Fiber Force Myography Sensor for Applications in Prosthetic Hand Control," in 2018 IEEE 15th International Workshop on Advanced Motion Control (AMC), Tokio, 2018.

[55] J. Bejarano and D. Barrera, "Desarrollo de una prótesis mioeléctrica de miembro superior con amputación transradial por medio del uso de tecnologías 3D," Infometric, vol. 1, no. 2, 2018.

[56] J. Patiño, Y. Acevedo and D. Albarracín, "Diseño y construcción de un prototipo de prótesis mioeléctrica de antebrazo y mano con dos grados de libertad," en INVESTIGACIÓN FORMATIVA EN INGENIERÍA, Segunda ed., Medellin, IAI, 2018, pp. 190-196.

[57] O. Vargas, O. Flor, F. Suárez and C. Chimbo, "Construcción y pruebas de funcionamiento de un prototipo robótico para prótesis humana," Espirales, vol. 4, no. 32, 2020.

[58] O. Vargas and O. Flor, "Diseño de un prototipo robótico de mano y antebrazo diestro para prótesis," Universidad Ciencia Y Tecnología, vol. 24, no. 96, pp. 27-34, 2020.

[59] J. Cuellar, G. Smit, P. Breedveld, A. Zadpoor and D. Plettenburg, "Functional evaluation of a non-assembly 3D-printed hand prosthesis," Proc IMechE Part H, vol. 233, no. 11, 2019.

[60] G. Beninati and V. Sanguineti, "A dynamic model of hand movements for proportional myoelectric control of a hand prosthesis*," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, 2019.

[61] H. Williams, Q. Boser, P. Pilarski, C. Chapman, A. Vette and J. Hebert, "Hand Function Kinematics when using a Simulated Myoelectric Prosthesis," in 019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, 2019.

[62] V. Alvarado, J. Sánchez, J. Carlos, E. Chihuan and C. De La Cruz, "Adquisición de señales EMG con electrodos secos para el control de movimiento de dedos en una prótesis robótica fabricada en una impresora 3D," Ingeniare, vol. 27, no. 3, 2019.

[63] A. Valadez Palacios, D. Luviano Cruz, F. García Luna and L. A. GarcÌa Villalba, "Diseño y Construcción de una Protesis de Mano Controlada por Medio de un Sensor Mioeléctrico," 2019, pp. 350 - 362.

[64] N. Amor, G. Rasool, N. Bouaynaya and R. Shterenbergc, "Constrained particle filtering for movement identification in forearm prosthesis," Signal Processing, vol. 161, 2019.

[65] A. Prakash, S. Sharma and N. Sharma, "A compact-sized surface EMG sensor for myoelectric hand prosthesis," Biomed. Eng. Lett, vol. 9, p. 467–479, Agosto 2019.

[66] K. Iwatsukia, M. Hoshiyamab, S. Oyama, S. Shimoda and H. Hirata, "Magnetoencephalographic evaluation for the myoelectric hand prosthesis with tacitlearning system," NeuroReh., vol. 44, pp. 19-23, 2019.

[67] I. Ku, G. Lee, C. Park, J. Lee and E. Jeong, "Clinical outcomes of a low-cost single-channel myoelectric-interface three-dimensional hand prosthesis," Arch Plast Surg, vol. 46, no. 4, Julio 2019.

[68] A. Pérez Rodríguez, M. Domínguez Morales, Á. Jiménez Fernández and A. Linares Barranco, MYO ARM: prótesis robótica con sensado emg y entrenamiento con redes neuronales, 2019, pp. 47-56.

[69] M. Dayneth Areiza, J. Mendoza, A. Saavedra and J. R. Serracín Pittí, "Músculos artificiales y optomiografía aplicados a una mano robótica," Revista de Iniciación Científica, vol. 5, no. 2, 2019.

[70] G. Li, O. W. Samuel, C. Lin, M. G. Asogbon, P. Fang and P. O. Idowu, "Realizing Efficient EMG-Based Prosthetic Control Strategy," Advances in Experimental Medicine and Biology, November 2019.

[71] N. Parajuli, N. Screenivasan, P. Bifulco, M. Cesarelli, S. Savino and V. Niola, "Real-Time EMG Based Pattern Recognition Control For Hand Prostheses:A review on Existing Methods. Challenges and Future Implementation," vol. 19, no. 4596, 2019.

[72] A. Linares and D. Rosas, "Desarrollo de prótesis electromecánica de miembro superior," Revista de Ingeniería Biomédica y Biotecnología, vol. 3, no. 10, Diciembre 2019.

[73] E. Salazar Cueva, A. Hidalgo Oñate, T. Berrazueta Espin, J. Freire Samaniego and B. Chavez Rios, "Prótesis antropomórfica multifuncional para pacientes con deformaciones en la mano," Revista Arbitrada Interdisciplinaria de Ciencias de la Salud. SALUD Y VIDA, vol. 3, no. 6, Julio - Diciembre 2019.

[74] F. Alkhatib and E. Mahdi, "Design and Analysis of Flexible Joints for a Robust 3D Printed Prosthetic Hand," In 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, Canadá, 2019.

[75] A. Cruz Rodríguez, "Construcción de una prótesis de mano, controlada a través de señales EEG obtenidas del lóbulo frontal, a través del análisis de una unidad Mindflex," 2016.

[76] I. Gutierrez, D. Calderón and E. Gonzales, "Construcción de prótesis robótica de mano para personas con amputación debajo del codo.," Revista Aristas: Investigación Básica y Aplicada, vol. 7, no. 14, 2019.

[77] Gonzalez Daza J and N. Lievano Guerrero, "prototipo de prótesis mioeléctrica activa para mano," 2019.

[78] T. Shibanoki and K. Jin, "A 3D-printable Prosthetic Hand Considering Dual-arm Operation.," in In 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), Kyoto, Japan, 2020.

[79] A. Mohammadi, J. Lavranos, H. Zhou, R. Multu, G. Alici and Y. Tan, "A practical 3D-printed soft robotic prosthetic hand with multi-articulating capabilities.," PLoS ONE, vol. 15, no. 5, 2020.

[80] J. Higa and B. Ojeda, "Diseño de filtros de señales electromiográficas para control de prótesis de mano.," Innovation & Development In Engineering And Applied Sciences (IDEAS), vol. 2, Enero 2020.

[81] S. Said, I. Boulkaibet, M. Sheilkh, A. Karar and S. Alkork, "Machine-Learning-Based Muscle Control of a3D-Printed Bionic Arm," Sensors, vol. 20, no. 3144, June 2020.

[82] B. Ojeda, "Prótesis de mano para personas con amputación transcarpiana.," Innovation & Development In Engineering And Applied Sciences (IDEAS), vol. 2, no. 1, Junio 2020.

[83] R. Alturkistani, S. Devasahayam, R. Thomas, E. Colombini and C. Cifuentes, "Affordable passive 3D-printed prosthesis for persons with partial hand amputation," Prosthet Orthot Int, vol. 44, no. 2, April 2020.

[84] P. A. Hurtado Manzaneraa, D. Luviano Cruz, L. Vidal Portilla and L. A. García Villalba, "Diseño y construcción de un prototipo de prótesis mioeléctrica," Mundo Fesc, vol. 8, no. 15, 2018.

[85] Q. Zhan, C. Zhang and Q. Xu, "Measurement and Description of Human Hand Movement," MATEC Web of Conferences, 2017.

[86] A. I. Kapandji, Fisiología articular: esquemas comenmdos de mecánica humana, Sexta ed., Madrid: Mádica Panamericana, 2006, p. 208.

[87] K. Schoepp, M. Dawson, J. Schofield, J. Carey and J. Herbert, "Design and Integration of an Inexpensive Wearable Mechanotactile Feedback System for Myoelectric Prostheses.," IEEE Journal of Translational Engineering in Health and Medicine.

[88] A. Ravida, S. Barootchi, M. Tattan, J. Carey, M. Saleh and J. Gargallo, "Clinical outcomes and cost effectiveness of computer-guided versus conventional implant-retained hybrid prosthesis: A long-term retrospective analysis of treatment protocols.," Periodontology, vol. 89, no. 9, 2018.

Mechatronic Exoskeletons for Lower-Limb Rehabilitation: An Innovative Review

Deyby Huamanchahua
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, N.L., México
a00816108@itesm.mx

Yerson Taza-Aquino
Department of Mechatronics Engineering
Universidad Continental
Huancayo, Perú
74239368@continental.edu.pe

Jhon Figueroa-Bados
Department of Mechatronics Engineering
Universidad Continental
Huancayo, Perú
71873208@continental.edu.pe

Jason Alanya-Villanueva
Mechatronic Engineering
Universidad Continental
Huancayo, Perú
73492658@continental.edu.pe

Adriana Vargas-Martinez
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, N.L., México
adriana.vargas.mtz@tec.mx

Ricardo A. Ramirez-Mendoza
School of Engineering and Sciences
Tecnologico de Monterrey
Monterrey, N.L., México
ricardo.ramirez@tec.mx

***Abstract*— Studies on the development of lower-limb exoskeletons began in the 1960s. However, these robotic devices' technological progress has been plodding, despite being studied for a long time. Therefore, the purpose of this article is to document a systematic review of the trends in the application of components that are part of the development of exoskeletons applied to the rehabilitation of lower-limbs. The objective is to provide the researcher with a structured matrix that integrates these components related to lower-limb exoskeletons development. As methods, it used different databases and specialized search engines that helped collect research from 2016 to 2020. The process of filtering and selecting the reviewed research was carried out, leaving 90 selected. Finally, it is concluded that there is still room for improvement and continue development of lower-limb exoskeletons that contribute to the rehabilitation process and are closer to the human lower-limb's real functioning.**

***Keywords*— exoskeleton, lower-limbs, rehabilitation, TRL**

I. INTRODUCTION

Robotics is in development and growth; in recent years robotic structures such as exoskeletons are being created and manufactured to contribute and help human limbs in different tasks. Exoskeletons are mechanical structures attached to the human limbs to amplify or increase the user's strength. The main applications are in industries, military technology, and medicine. The exoskeleton can be applied to perform rehabilitation therapies for the upper and lower-limbs due to some type of disease or accident that produces a temporary loss of muscle activity [1].

The history of the development of exoskeletons began in the 60s with the name of Man Amplifier created by Neil Mizen; in 1965 the Hardiman Suit project led by General Electric in the United States was carried out; this Motorized exoskeleton was used to manipulate materials [2]. In 1980 Jeffrey Moore stated that the exoskeleton can help increase human performance for military purposes [3].

The lower-limb exoskeletons began to develop in the 1960s; the reasons for their development were muscle weakness. This robotic mechanism helped patients with spinal cord injury to regain their gait. Rehabilitation technologies for people with locomotion disabilities have been developed over the years in several institutions and universities around the world. Nowadays, there are different robotic devices that provide autonomy to people with walking difficulties, among which is the exoskeleton for lower-limbs.

People with some type of disability face common problems that are reflected in the deterioration of their quality of life, since in most cases they must depend economically and socially on someone for their integration into society. According to the World report on disability, in the year 2019, it is estimated that more than one billion people live with some type of disability, or about 15% of the world's population.

The main reasons for the disabilities can be divided into two groups: muscular diseases and diseases that affect the nervous system; in both cases, a treatment that favors the recovery of the affected parts must be chosen. It is crucial to perform rehabilitation therapy when you suffer from a temporary loss of the upper and lower-limbs. Research trends show that failure to perform good rehabilitation therapy results in the ultimate loss of limbs. Based on the literature, exoskeletons are being developed primarily for military, industrial, and medical purposes.

This article's objective is to systematically review the development and creation trends of lower-limb exoskeletons applied for rehabilitation where analysis of the characteristics that compose them such as type of loss, supported movements, degrees of freedom, material, actuators, and developmental stage of the exoskeleton. Finally, conclusions will be made based on recent trends shown by exoskeletons.

II. METHODOLOGY

For the development of this research, a search query was made in the database of "Scopus", "IEEE Explore", "Springer link", "PubMed", "Google Scholar", also access articles from journals such as "Redalyc", "Science" as well as collaboration and dissemination platforms such as "ResearchGate" and "Dialnet". Keywords were used such as an exoskeleton, lower-limbs, and control. Eight hundred eighty results were obtained and publications on lower-limb exoskeleton, movement

978-1-6654-4067-7 / 21 / $ 31.00 © 2021 IEEE

control, dynamic and kinematic analysis were excluded. This process resulted in an average of 90 publications that covered from 2016 to 2020.

III. REVIEW OF PUBLICATIONS

To simplify and structure the information. The lower-limb exoskeletons were classified according to the sensors implemented, degrees of freedom (DoF), type of loss of the lower-limbs, supported limbs of the exoskeleton, material used for manufacturing, actuators, and state of development of the exoskeleton [prototype or study].

The review describes the different designs of the mechanical and control parts of the lower-limb exoskeletons. Also, interface connection components connect the actuators with the human body.

TABLE I. EXOSKELETONS FOR REHABILITATION OF LOWER-LIMB

Reference	Kind of loss	Supported movements	Degrees of freedom (DoF)	Material	Actuators	TRL
Pais, C [4]	PMI	CRT	NE	Aluminum	AL	7
Lei, Z [5]	PMI	CR	4	Aluminum	AL	5
Ann, K [6]	PMI	R	1	Carbon fiber	NE	4
Tang, C [7]	PMI	CR	4	Aluminum	AL	5
Hernández, J [8]	LME	CR	4	Aluminum	AL	5
Zhou, L [9]	PMI	CR	2	Aluminum	Al	4
Tibaduiza, D [10]	AC	CR	2	Aluminum	AL	5
Peralta, V [11]	H	CRT	6	Aluminum	APP	4
Gudiño J [12]	PMI	CRT	6	Aluminum	APP	5
Maldonado, I [13]	PMI	CRT	6	Nylon	APP	5
Blanco, A [14]	AC	CRT	6	Aluminum	NE	5
Luna, L [15]	PMI	CRT	6	Aluminum	S	2
Mohammad, S [16]	PMI	CR	4	NE	NE	3
Wang, H [17]	PMI	CR	4	NE	NE	6
Wang, Y [18]	PMI	CR	4	NE	AL	1
Baquero, K [19]	AC	CRT	6	Aluminum	AL	5
Dunai, L [20]	PMI	CRT	6	Aluminum-PLA	AL	5
Alburqueque, C [21]	PMI	C	1	Aluminum-PLA	S	5
Zhang, X [22]	PMI	T	1	NE	SH	4
Kazemi, J [23]	PMI	CRT	4	Aluminum	AL	5
Elias, A [24]	PMI	R	3	Aluminum	AL	5
García, I [25]	PMI	CRT	3	Aluminum	S	5
Chen C [26]	PMI	CRT	14	NE	AL	7
Siddique, N [27]	DM	R	2	Aluminum-Carbon Fiber	AL	5
Dos Santos, W [28]	PMI	CRT	6	NE	AL	1
Elias, D [29]	PMI	CRT	3	NE	AL	1
Villena, G [30]	PMI	CRT	4	Aluminum	AN	4
Chicoma, C [31]	PMI	CRT	8	NE	AL-SH	2
Mineev, S [32]	PMI	CRT	4	NE	AL	2
Zheng, T [33]	P	CR	4	Aluminum	AL	7
Leal, A [34]	PMI	R	NE	NE	AL	1
Yue, C [35]	PMI	CRT	6	Aluminum	AL	5
Ruiming, L [36]	PMI	NE	6	NE	NE	2
Long, Y [37]	PMI	CRT	14	Aluminum	Ah-S-AL	7
Kawale, S [38]	PMI	CRT	12	NE	AH	5
Tanyildizi, A [39]	PMI	R	2	NE	NE	2
Menga, G [40]	PMI	CRT	6	Aluminum	AL	4
Villa, A [41]	PMI	CR	4	Aluminum	AL	5
Rajasekaran, V [42]	PMI	CRT	6	Aluminum	Al	5

Ling, F [43]	PMI	T	1	Aluminum	S	5
Mayorca, D [44]	PMI	R	1	NE	AL	4
Nunes, P [45]	AC	CRT	6	Aluminum	NE	1
Munadi, M [46]	AP	CRT	6	Aluminum	AL	6
Riesco, P [47]	PMI	CR	4	Aluminum	S	5
Villarejo, J [48]	PMI	CR	4	NE	NE	6
Torres, M [49]	AP	CR	4	Aluminum	S	6
Arnago, G [50]	PMI	CRT	6	Aluminum	APP	5
Lliguay, J [51]	PMI	CRT	6	Steel-Nylon	NE	5
Davila, D [52]	PMI	CRT	6	Steel	S	5
Alcivar, M [53]	PMI	CRT	6	PLA-Aluminum	AL	5
Sánches, C [54]	P	CRT	6	Aluminum-Acer	AL	6
Aya, P [55]	AC	CR	2	Aluminum-Steel	AL	5
Garcés, A [56]	P	CRT	6	NE	APP	6
Dos Santos, W [57]	AC	CRT	6	Aluminum	AL	5
López, R [58]	PMI	RT	4	Polypropylene-Aluminum	AE	4
Aguirre, E [59]	PMI	CRT	6	Aluminum	S	6
Begue, J [60]	PMI	CRT	6	Aluminum-PLA	AL	5
Tamburrino, B [61]	¨PMI	CR	2	Aluminum-Steel	AL	5
Mendoza, E [62]	PMI	CR	4	Aluminum	S	5
Chamnikar, A [63]	PMI	CRT	3	NE	NE	2
Romero, M [64]	PMI	CRT	6	Aluminum	S	4
Long, Y [65]	PMI	CRT	10	Aluminum	AH	5
Peñafiel, A [66]	PMI	CR	2	PLA	Al	5
Marquez, H [67]	PMI	CRT	4	Aluminum	AL	5
Mendoza, D [68]	E	CRT	3	NE	NE	4
Ávila, E [69]	P	CR	2	Steel	Al	5
Jim, D [70]	PMI	CRT	6	Aluminum	AL	5
Majeeda, A [71]	PMI	CRT	3	NE	NE	5
Núñez, K [72]	H	CRT	2	Aluminum	S	5
Velandia, C [73]	PMI	RT	2	NE	AL	4
Stopforth, R [74]	PMI	CRT	6	Aluminum	AL	5
Hongchul, K [75]	PMI	CRT	6	Aluminum	SH	5
Yi, L [76]	PMI	CRT	3	Aluminum	Al	5
Byunghun, C [77]	PMI	CRT	6	NE	AH-AL-AN	5
Belkadi, A [78]	PMI	R	2	NE	AL	2
Pagre, A [79]	PMI	R	3	NE	NE	2
Ajayi, M [80]	PMI	CR	7	NE	AL	2
Yang, P [81]	DM	C	1	NE	NE	1
Durandau, G [82]	PMI	CRT	6	NE	AL	1
Trincado, F [83]	PMI	CRT	6	NE	AL	2
Zakaria, A [84]	PMI	CRT	3	NE	AL	1
Nasiri, N [85]	PMI	CRT	4	NE	AL	2
Long, Y [86]	PMI	CR	2	Aluminum-Carbon Fiber	AL	5
Aguilar, H [87]	PMI	CR	4	Aluminum	AL	5
Velandia, C [88]	PMI	CRT	6	Aluminum	AL	5
Rincon, K [89]	PMI	CRT	6	MDF	S	4
Tovar, M [90]	PMI	CRT	14	PLA	AL	6
Zhu, A [91]	PMI	CRT	6	Aluminum	NE	5
Durandau, G [92]	PMI	CRT	6	NE	Al	2
Wu, J [93]	AC	CRT	3	Aluminum	AL	5

Note: Abbreviation: AC: Strokes, H: Hypertrophy, PMI: Loss of lower-limb movement, P: Paraplegia, LME: Spinal cord injury, E: Sclerosis, DM: Muscle deficiency, AP: Leg amputation, CR: hip and knee, CT: Hip and ankle, RT: Knee and ankle, CRT: Hip, knee and ankle, R: Knee, C: Hip, T: Ankle, AL: Linear actuator, APP: Stepper Actuator, AN: Pneumatic Actuator, S: Servomotor, SH: Hydraulic Servo, AH: Hydraulic Actuator, AE: Elastic Actuator, TRL 1: Basic Research, TRL 2: Technology Formulation, TRL 3: Applied research, TRL 4: Small-scale development, TRL 5: Real-scale development, TRL 6: Valid system in simulated environment, TRL 7: Real-environment validated system, N.E.: Unspecified

A. *Type of loss*

Several types of loss were found in this group: loss of lower-limbs (PMI), muscular deficiency (DM), paraplegia (P), hypotrophy (H), cerebrovascular accidents (AC). Table II describes the characteristics of the types of losses mentioned.

TABLE II. TYPES OF LOSSES

DM (Muscle Deficiency)	Loss of strength in the muscles
P (Paraplegia)	Paraplegia reacts to the loss of control of the trunk and generates a paralysis of the muscles generating difficulty in mobilizing and walking [94].
H (Hypotrophy)	It is the decrease in muscle mass and this implies the loss of strength, generating the inability to stand.
AC (Cerebrovascular accident)	It is a neurological pathology more common in adults that generates severe disabilities such as muscle weakness, loss of volumetric movements, and causes difficulty in wandering [95].

It is crucial to consider the type of loss that the patient suffered and to be able to develop an appropriate exoskeleton to achieve rehabilitation since by not doing a good rehabilitation session it is very likely that the patient will lose the total mobility of their lower-limbs. As can be seen in Table III, most of the studies were focused on PMI with 78.9% [27,28,72], followed by AC with 7.8% [1,5,14] and the rest below 5%.

TABLE III. PERCENTAGE OF STUDIES IN FACIÓN TO THE TYPE OF LOSS

Muscle Deficiency	Amount	
	Frequency	Percentage
PMI	71	78,90%
AC	7	7,80%
Q	4	4,40%
H	2	2,20%
DM	2	2,20%
AP	1	1,10%
LME	1	1,10%

Note: AC: Strokes, H: Hypertrophy, PMI: Loss of lower-limb movement, P: Paraplegia, LME: Spinal cord injury, DM: Muscle deficiency, AP: Leg amputation

B. *Supported movements*

In this area, a review is carried out on the parts where rehabilitation is applied through the lower-limbs' exoskeletons. The lower-limbs count various movements are shown in Fig. 1; according to the reviewed research, applications were identified in the hip (C), ankle (T), knee-ankle (RT), hip-knee (CR), hip-knee-ankle (CRT), and knee-ankle (RT). In Table IV the application that stands out the most in the studies are the movements in CRT with 62.2% [33,37,45]. This reflects that the prototypes focus on performing a complete rehabilitation applying to the hip, knee, and ankle, which are the components of a human lower-limb. Furthermore, 23.3% focused on developing prototypes of movements in RC [54,83], and a percentage of 7.8% indicates that exoskeletons are applied to rehabilitate RT.

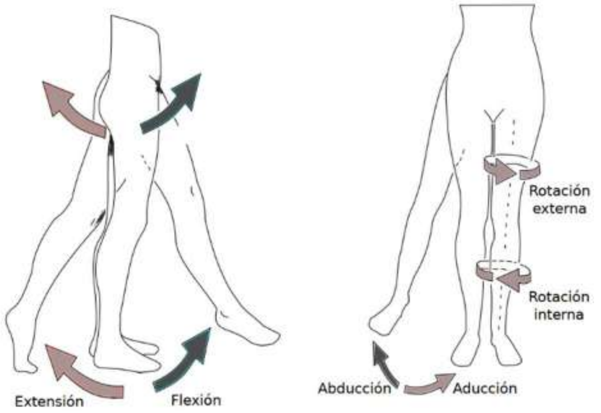


Fig. 1 Flexion / Extension, Abduction / Adduction and Internal / External Rotation movements. Source: K. L. Moore, A. F. Dalley, and A. M. Agur, Clinically Oriented Anatomy, Spain: Lippincott Williams & Wilkins, [2010]

The rest of the applications do not exceed 2.3% in Table IV, it is also observed that the most important supported movement is the knee since the mobility of the lower-limbs depends on that part of the body.

TABLE IV. PERCENTAGES OF STUDIES BASED ON SUPPORTED MOVEMENTS

Supported movements	Amount	
	Frequency	Percentage
CRT	56	62,20%
CR	21	23,30%
RT	9	10,00%
C	2	2,20%
T	2	2,20%

Note: CR: hip and knee, CT: Hip and ankle, RT: Knee and ankle, CRT: Hip, knee and ankle, R: Knee, C: Hip, T: Ankle

C. *Degrees of Freedom (DoF)*

In this area, a review is carried out on the degrees of freedom that exoskeleton prototypes have. The lower-limbs have a total of 8 DoF at the hip has 3 DoF as shown in Fig. 2, at the knee 2 DoF, and ankle 3 DoF as visualized in fig. 3 for each leg. It is considered ideal that a lower-limb exoskeleton has the total DoF that the lower-limbs makeup, Table V shows that the highest number of exoskeletons with a percentage of 38.2% have 6 DoF according to the reviewed research, with a percentage of 21.3% have 4 DOF followed by 13.5% with 2 DoF. The exoskeletons that have 3 DoF, according to the research reviewed with a percentage of 11.2%, are devices applied to a lower-limb. The image also shows exoskeletons with 1 DoF with a percentage of 6.7%; according to the research reviewed, these robotic devices consist of rehabilitating only a part of the lower-limb, be it ankle, knee, or hips. You can also find research on exoskeletons with 14 DoF occupying a 3.4% percentage, closer to the DoF that human limbs count.

Finally, with a percentage of 1.1%, investigations with 7 and 8 DoF were found according to the review carried out.

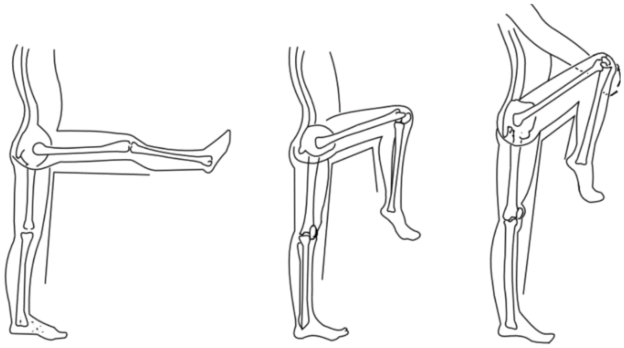


Fig. 2 Hip flexion increases by relaxing the hamstring muscles by flexing the knee. Passively the bending is greater. Source: R.C. Miralles Marrero, Clinical biomechanics of the locomotive system, Spain: MASSON, [2000]

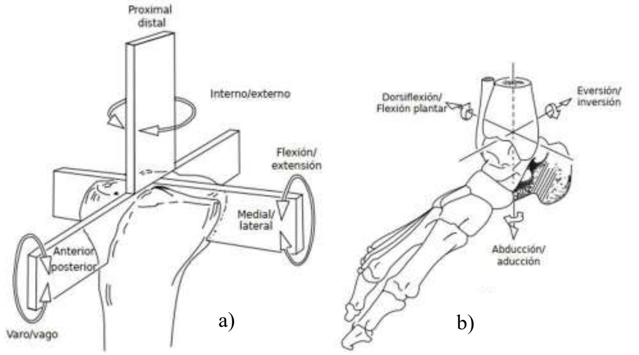


Fig. 3 (a) Representation and nomenclature of the six degrees of freedom of movement of the knee: anterior and posterior translation, medial/side translation and proximal/distal translation, flexion-extension rotation, internal-external rotation and b) Representation and nomenclature of the six degrees of knee freedom of movement: anterior and posterior translation, medial/side translation and proximal/distal translation, bend-extension rotation, internal-external rotation of varo-valgo. Source: M. Nordin and V. Frankel, BASIC BIOMECHANICS OF THE MUSCULOSKELETAL SYSTEM, Spain: McGRAW-HILL / INTERAMERICANA DE SPAIN, [2004]].

TABLE V. PERCENTAGE OF DOF OF LOWER-LIMB EXOSKELETONS

Degrees of freedom	Amount	
	Frequency	Percentage
1	6	6,70%
2	12	13,50%
3	10	11,20%
4	19	21,30%
6	34	38,20%
7	1	1,10%
8	1	1,10%
12	1	1,10%
14	3	3,40%
N.E.	2	2,20%

D. *Type of material*

In this area, a review is made of the type of material they tend to use to develop the exoskeleton. Table VI shows the various materials that are used for the construction of these robotic devices. 48.9% of the researches reviewed uses aluminum material, taking advantage of the properties it offers, unlike other materials, aluminum has a specific weight of one-third of the weight of the metal, and aluminum has a layer that protects it from rust; that is why most of the researches reviewed use aluminum for the creation of their prototypes [4]. As can be seen in the image for the development of the mechanical structures of exoskeletons, they combine other materials with PLA and aluminum with a percentage of 4.4%, aluminum, and steel with a percentage of 3.3%, carbon fiber and aluminum with a percentage of 2.2%, polypropylene, and aluminum with a percentage of 1.1%, steel and nylon with a percentage of 1.1% according to the research reviewed providing more versatility to your prototypes. The figure also shows the use of PLA in a percentage of 2.2%, it is a rarely used material, because the exoskeleton structure usually supports loads greater than what the PLA material can support, the use of steel is in the percentage 2.2% because this material is very heavy compared to aluminum and this generates muscle fatigue in patients when they undergo their rehabilitation treatment. Finally, materials such as MDF and nylon are rarely used for the development of exoskeletons due to their mechanical composition that does not provide the necessary support to support loads such as the weight of the patient.

TABLE VI. PERCENTAGE OF THE TYPE OF MATERIAL USED FOR THE DEVELOPMENT OF THE EXOSKELETAL STRUCTURE OF THE REVISED RESEARCH

Material	Amount	
	Frequency	Percentage
Aluminum	44	48,90%
PLA-Aluminium	4	4,40%
Aluminium-Steel	3	3,30%
PLA	2	2,20%
Steel	2	2,20%
Carbon-Aluminium Fiber	2	2,20%
MDF	1	1,10%
Nylon	1	1,10%
Polypropylene-Aluminum	1	1,10%
Carbon Fiber	1	1,10%
Steel-Nylon	1	1,10%
N.E.	28	31,10%

E. *Actuators*

The investigations developed solutions to imitating the movements with various actuators that do not affect the joints with sudden movements. Within these actuators, it was possible to identify linear actuators (AL), step by step (APP), hydraulic (AH), servomotors (S), and hydraulic servomotors (SH). In Table 7 it can be observed that the linear actuator stands out with a percentage of 54.4%, these actuators provide a greater force to perform the movements [30-32]. It is also possible to highlight the servomotors by restricting movements to delimited positions as the programming indicates and the patient requires [22, 25, 34]. In Table VII the stepper actuators are also shown with a percentage of 5.6%. And finally, the other actuators that do not exceed 2.2% are not as widely used for the development of these robotic devices.

TABLE VII. PERCENTAGES OF STUDIES BASED ON THE ACTUATOR EMPLOYED

Actuator	Amount	
	Frequency	Percentage
AL	49	54,40%
S	12	13,30%
APP	5	5,60%
SH	2	2,20%
OH	2	2,20%
AE	1	1,10%
AH-S-AL	1	1,10%
AH-AL-AN	1	1,10%
AL-SH	1	1,10%
AN	1	1,10%
N.E.	15	16,70%

Note: AL: Linear actuator, APP: Stepper actuator, AN: Pneumatic actuator, S: Servomotor, SH: Hydraulic servo, AH: Hydraulic actuator, AE: Elastic actuator.

F. State of exoskeleton development

Each investigation managed to determine different objectives to collaborate in developing an exoskeleton for the different types of rehabilitation reaching different states such as functional, prototypes, designs, and studies. Table VIII shows that the most significant number of studies managed to reach a prototyping state, managing to implement programming, mathematical analysis, and mechanical design in exoskeletons, which included 62.2% [34-60]. On the other hand, there were also investigations in charge of the analysis of the exoskeletons for the improvement of the functioning, being 24.4% [63, 66, 68] and with a percentage of 8.9% according to the reviewed investigations, the development of the exoskeletons remained in design by opening the possibility for other researchers to improve and prototype these proposed designs.

TABLE VIII. PERCENTAGES OF STUDIES BASED ON THE ACTUATOR EMPLOYED

State of development	Amount	
	Frequency	Percentage
Prototype	56	62,20%
Studies	22	24,40%
Design	8	8,90%
Functional	4	4,40%

IV. CONCLUSION

With population growth and a shortage of specialists, many researchers have come up with robotic devices to help with the rehabilitation process. The article reviewed various investigations regarding designs and prototypes of exoskeletons for lower-limbs applied to rehabilitation; several aspects must be improved, such as the cost of development, the ergonomics of the exoskeletal structure that is very important to consider so as not to generate injuries to the patients from prolonged use of these devices. There is still room to improve and continue developing more lower-limb exoskeletons that contribute to the rehabilitation process and are closer to the real function of a human lower-limb. These robotic devices are forecast to get closer and closer to ideal performance in the future. It is hoped that this review will help future projects and research to select the appropriate components for the development of new prototypes.

REFERENCES

[1] R. López, H. Aguilar, S. Salazar, R. Lozano, and J. A. Torres, “Modelado y control de un exoesqueleto para la rehabilitación de extremidad inferior con dos grados de libertad,” Revista iberoamericana de automática e informática industrial, vol. 11, no. 3, pp. 304–314, 2014.

[2] Anónimo. Inter empresas. [Online]. Exoesqueletos: la edad del 'hombre de hierro'.2015 [citado el 20 de agosto del 2020]. Disponible en: https://www.interempresas.net/Proteccion-laboral/Articulos/211884-Exoesqueletos-la-edad-del-hombre-de-hierro.html.

[3] Cisneros C. Maquinas específica EXOESQUELETOS. [Online]. Sites.google.com.2016. [citado el 20 de agosto del 2020]. Disponible en: https://sites.google.com/site/fgtce04equipo03tgigestion/funcionamiento-de-los exoesqueletos.

[4] C. Pais-Vieira, M. Allahdad, J. Neves-Amado, A. Perrotta, E. Morya,R. Moioli, E. Shapkova, and M. Pais-Vieira, “Method for positioning and rehabilitation training with the exoatlet® powered exoskeleton,”MethodsX, vol. 7, p. 100849, 2020.

[5] Zhang, W. Chen, Y. Chai, J. Wang, and J. Zhang, “Gait graph optimization: Generate variable gaits from one base gait for lower-limb rehabilitation exoskeleton robots,”arXiv preprint arXiv:2001.00728,2020.

[6] K. A. Witte and S. H. Collins, “Design of lower-limb exoskeletons and emulator systems,” in Wearable Robotics. Elsevier, 2020, pp. 251–274.

[7] T. Pan, C.-C. Chang, Y.-S. Yang, C.-K. Yen, Y.-H. Kao, and Y.-L.Shiue, “Development of MMG sensors using PVDF piezoelectric electro-spinning for lower limb rehabilitation exoskeleton,” Sensors and Actuators A: Physical, vol. 301, p. 111708, 2020.

[8] J. H. Hernandez, S. S. Cruz, R. Lopez-Gutierrez, A. Gonzalez-Mendoza, and R. Lozano, “Robust nonsingular fast terminal sliding-mode control for sit-to-stand task using a mobile lower limb exoskeleton,” Control Engineering Practice, vol. 101, p. 104496, 2020.

[9] L. Zhou, W. Chen, W. Chen, S. Bai, J. Zhang, and J. Wang, “Design of a passive lower limb exoskeleton for walking assistance with gravity compensation,” Mechanism and Machine Theory, vol. 150, p. 103840,2020.

[10] D. A. Tibaduiza Burgos, P.-A. Aya-Parra, and M. Anaya, “Exoesqueleto para rehabilitaci ón de miembro inferior con dos grados de libertad orientado a pacientes con accidentes cerebrovasculares,” INGE CUC,2019.

[11] V. G. P. Lugo, A. G. Betancourt, I. M. Panecatl, and R. E. L. Torres, “Exoesqueleto para hipotrofia en miembro inferior con asistencia de electroestimulación,”Dra. Lucia Marquez Perez Ing. Wendolın Jacinto Dıaz, p. 193.

[12] J. Gudiño-Lau, I. Rosales, S. Charre, J. Alcal ́a, M. Duran, D. Velez-Dıazet al., “Diseño y construcción de un exoesqueleto para rehabilitación,” XIKUA Boletín Científico de la Escuela Superior de Tlahuelilpan, vol. 7,no. 13, pp. 1–10, 2019.

[13] G. E. Maldonado Ibarra, “Desarrollo de un prototipo de andador-exoesqueleto de 6 grados de libertad para la rehabilitación física de miembros inferiores en infantes dentro del grupo de investigación en bioingeniería giebi.” 2019

[14] A. BLANCO-ORTEGA, D. PEREZ-VIGUERAS, E. ANTUNEZ-LEYVA, and J. COLIN-OCAMPO, “Controlador robusto para el seguimiento de trayectorias para un exoesqueleto de extremidades inferiores robust trajectory tracking controller for lower extremity exoskeleton,”Mecánica, vol. 3, no. 11, pp. 1–8, 2019.

[15] L. Luna, I. Garcia, M. Mendoza, G. Dorantes-Mendez, A. Mejia-Rodriguez, and I. Bonilla, “Emg-based kinematic impedance control ofa lower-limb exoskeleton,” inLatin American Conference on Biomedical Engineering. Springer, 2019, pp. 1494–1501

[16] M. S. Amiri, R. Ramli, and M. F. Ibrahim, “Hybrid design of pid controller for four dof lower limb exoskeleton,” Applied Mathematical Modelling, vol. 72, pp. 17–27, 2019.

[17] H. Wang, Y. Feng, X. Wang, L. Ren, J. Niu, and L. Vladareanu, “Re-tracted: Design and analysis of a spatial four-dof lower limb rehabilitation robot,”Journal of Fundamental and Applied Sciences, vol. 10, no. 4S,pp. 175–180, 2018.

[18] W. Yingxu, Z. Aibin, W. Hongling, Z. Pengcheng, X. Zhang, and C. Guangzhong, “Control of lower limb rehabilitation exoskeleton robotbased on cpg neural network,” in2019 16th International Conference on Ubiquitous Robots (UR). IEEE, 2019, pp. 678–682.

[19] K. C. Baquero Duarteet al., “Diseño de un estudio experimental para re habilitación de rodilla con exoesqueleto activo,” Ph.D. dissertation, Universidad del Rosario, 2019.

[20] L. Dunai, I. Lengua, G. Peris Fajarnes, and B. Defez Garcia, “Diseño de un exoesqueleto de extremidades inferiores,” DYNA Ingeniería e Industria, vol. 94, no. 3, pp. 297–303, 2019.

[21] C. A. Alburqueque Reyes and L. A. Rondón Gomez, “Diseño e implementación de un exoesqueleto para fisioterapia en pacientes con artrosis de rodilla en la clínica geriátrica militar de chorrillos,” 2019.

[22] X. Zhang, W. Jiang, Z. Li, and S. Song, “A hierarchical Lyapunov-based cascade adaptive control scheme for lower-limb exoskeleton,” European Journal of Control, vol. 50, pp. 198–208, 2019.

[23] J. Kazemi and S. Ozgoli, “Real-time walking pattern generation for a lower limb exoskeleton, implemented on the exposed robot,” Robotics and Autonomous Systems, vol. 116, pp. 1–23, 2019.

[24] A. Elias-Neto, A. C. Villa-Parra, T. Botelho, A. Frizera-Neto, and T. Bastos-Filho, "A robotic lower-limb exoskeleton for rehabilitation," in Latin American Conference on Biomedical Engineering.Springer,2019, pp. 1130–1136.

[25] L. García, L. Luna, M. Mendoza, A. Mejía-Rodríguez, I. Bonilla, and G. Dorantes-Mendez, "Development of a lower-limb exoskeleton for assistance of movements in the sagittal plane," in Latin American Conference on Biomedical Engineering. Springer, 2019, pp. 1023–1030.

[26] C.-F. Chen, Z.-J. Du, L. He, Y.-J. Shi, J.-Q. Wang, G.-Q. Xu, Y. Zhang,D.-M. Wu, and W. Dong, "Development and hybrid control of an electrically actuated lower limb exoskeleton for motion assistance," IEEE Access, vol. 7, pp. 169 107–169 122, 2019.

[27] N. Siddique, A. Saif, F. Imran, A. Kamran, U. S. Virk, I. Mahmood, A. Ali, N. Ahmad, and H. F. Maqbool, "Prototype development of an assistive lower limb exoskeleton," in2019 International Conference on Robotics and Automation in Industry (ICRAI). IEEE, 2019, pp. 1–6.

[28] W. M. dos Santos and A. A. Siqueira, "Design and control of a trans-parent lower limb exoskeleton," in International Symposium on Wearable Robotics. Springer, 2018, pp. 175–179.

[29] D. A. Elias, D. Cerna, C. Chicoma, and R. Mio, "Characteristics of a lower limb exoskeleton for gait and stair climbing therapies," in Interdisciplinary Applications of Kinematics. Springer, 2019, pp. 81–92.

[30] G. V. Prado, R. Yli-Peltola, and M. B. C. Sanchez, "Design and analysis of a lower limb exoskeleton for rehabilitation," in Interdisciplinary Applications of Kinematics. Springer, 2019, pp. 103–114.

[31] C. Chicoma, O. Cieza, E. Pujada, and D. A. Elias, "Modeling for the design of a lower limb exoskeleton for people with gait impairments," in Interdisciplinary Applications of Kinematics. Springer, 2019, pp. 129–139.

[32] S. Mineev, "Multimodal control system of active lower limb exoskeleton with feedback," in Proceedings of the Scientific-Practical Conference" Research and Development-2016". Springer, Cham, 2018, pp. 3–10.

[33] T. Zheng, Y. Zhu, Z. Zhang, S. Zhao, J. Chen, and J. Zhao, "Parametric gait online generation of a lower-limb exoskeleton for individuals with paraplegia," Journal of Bionic Engineering, vol. 15, no. 6, pp. 941–949,2018.

[34] A. G. Leal-Junior, A. Frizera, C. Marques, and M. J. Pontes, "Development of polymer optical fiber sensors for lower limb exoskeletons instrumentation," in International Symposium on Wearable Robotics. Springer, 2018, pp. 155–159.

[35] C. Yue, X. Lin, X. Zhang, J. Qiu, and H. Cheng, "Design and performance evaluation of a wearable sensing system for lower-limb exoskeleton," Applied bionics and biomechanics, vol. 2018, 2018.

[36] R. Luo, S. Sun, X. Zhao, Y. Zhang, and Y. Tang, "Adaptive CPG-based impedance control for assistive lower limb exoskeleton," in2018 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE,2018, pp. 685–690.

[37] Y. Long, Z.-j. Du, W.-d. Wang, L. He, X.-w. Mao, and W. Dong," Physical human-robot interaction estimation based control scheme fora hydraulically actuated exoskeleton designed for power amplification," Frontiers of Information Technology & Electronic Engineering, vol. 19,no. 9, pp. 1076–1085, 2018.

[38] S. S. Kawale and M. Sreekumar, "Design of a wearable lower body exoskeleton mechanism for shipbuilding industry," Procedia computer science, vol. 133, pp. 1021–1028, 2018.

[39] A. K. Tanyildizi, O. Yakut, and B. Tasar, "Mathematical modeling and control of lower extremity exoskeleton," 2018.

[40] G. Menga and M. Ghirardi, "Lower limb exoskeleton for rehabilitation with improved postural equilibrium,"Robotics, vol. 7, no. 2, p. 28, 20

[41] A. C. Villa-Parra, D. Delisle-Rodriguez, T. Botelho, J. J. V. Mayor,A. L. Delis, R. Carelli, A. Frizera Neto, and T. F. Bastos, "Control of a robotic knee exoskeleton for assistance and rehabilitation based on motion intention from semg,"Research on Biomedical Engineering, vol. 34, no. 3, pp. 198–210, 201

[42] V. Rajasekaran, E. López-Larraz, F. Trincado-Alonso, J. Aranda, L. Montesano, A. J. Del-Ama, and J. L. Pons, "Volition-adaptive control for gait training using wearable exoskeleton: preliminary tests with incomplete spinal cord injury individuals," Journal of neuro engineering and rehabilitation, vol. 15, no. 1, pp. 1–15, 2018.

[43] L.-F. Yeung and R. K.-Y. Tong, "Lower limb exoskeleton robot to facilitate the gait of stroke patients," Wearable Technology in Medicine and Health Care, vol. 91, 2018.

[44] H. H. G. Ch, D. Mayorca, and F. C. Gomez, "Control predictivo sujeto a restricciones de un motor dc presente en un exoesqueleto de miembro inferior," in Memorias del I Congreso Internacional de Bioingenierıa y Sistemas Inteligentes de Rehabilitación

[45] P. F. Nunes, W. M. dos Santos, and A. A. Siqueira, "Control strategy based on kinetic motor primitives for lower limbs exoskeletons," IFAC-Papers Online, vol. 51, no. 27, pp. 402–406, 2018.

[46] Munadi, M. Nasir, M. Ariyanto, N. Iskandar, and J. Setiawan, "Design and simulation of pid controller for lower limb exoskeleton robot," in AIP Conference Proceedings, vol. 1983, no. 1.AIP Publishing LLC, 2018,p. 060008.

[47] P. Riesco Gilet al., "Desarrollo del sistema para el estudio de exoesquele-tos de extremidades inferiores controlados mediante sensores," 2018.

[48] J. J. V. Mayor, N. J. Valencia-Jiménez, G. P. Arango-Hoyos, and E. F. Caicedo-Bravo, "Sistema de biofeedback para rehabilitación de marcha asistida por un exoesqueleto," Revista Ingeniería Biomédica, vol. 12, no. 24, pp. 47–57, 20

[49] M. Plaza Torres, A. Cifuentes, and F. Bernal Castillo, "Diseño de exoesqueletos para miembro inferior," Revista Cubana de Investigaciones Biomédicas, vol. 37, no. 2, pp. 95–104, 2018.

[50] L. A. A. Gomez, E. L. Gonzalez, M. A. Montiel, R. T. Herrera, and F. J. E. García, "Diseño de un prototipo de exoesqueleto para miembro inferior de infantes (design of a prototype exoskeleton for lower limbs of infants),"Pistas Educativas, vol. 42, no. 137, 2020.

[51] L. Calderon and J. Enrique, "Manufactura y pruebas de un prototipo de exoesqueleto para la rehabilitación física de miembros inferiores para el grupo de investigación y estudios de bioingeniería de la facultad de mecánica espoch." B.S. thesis, Escuela Superior Politécnica de Chimborazo, 2018.

[52] D. Davila Portals, "Rediseño del sistema mecánico del exoesqueleto pucp para rehabilitación de miembros inferiores."

[53] E. G. Alcivar Molinat al., "Desarrollo de algoritmos de control para un exoesqueleto rob ótico de 6 gdl," B.S. thesis, Espol, 2018.

[54] C. A. Sanchez Tapia, "Diseño y simulación de un prototipo de exoesqueleto de miembro inferior en la asistencia de la marcha para pacientes con paraplejia," 2018.

[55] P. A. Aya Parrat al., "Estudio anatómico y determinación de parámetros funcionales de un prototipo de exoesqueleto de miembro inferior con dos grados de libertad."

[56] A. E. Garces Beltran, "Diseño de un mecanismo del tipo exoesqueleto de miembros inferiores que permita reproducir patrones de movimiento,"2017.

[57] W. M. Dos Santos, S. L. Nogueira, G. C. de Oliveira, G. G. Pe ña, and A. A. Siqueira, "Design and evaluation of a modular lower limb exoskeleton for rehabilitation," in2017 International Conference on Rehabilitation Robotics (ICORR). IEEE, 2017, pp. 447–451.

[58] R. Lopez-Gutierrez, H. Aguilar-Sierra, S. Salazar, and R. Lozano, "Control adaptable en rutinas de rehabilitación pasiva utilizando elltio,"Revista mexicana de ingeniería biomédica, vol. 38, no. 2, pp. 458–478, 2017.

[59] E. E. Aguirre León and D. F. Cevallos Rodríguez, "Diseño mecánico estructural de un exoesqueleto orientado a la rehabilitación para extremidades inferiores de pacientes masculinos de edad productiva en la ciudad de Riobamba." B.S. thesis, Escuela Superior Politécnica de Chimborazo,2017.

[60] J. A. Begue Salcedo, W. I. Cobeña Minaya et al., "Diseño y construcción de un prototipo de la estructura mecánica de un exoesqueleto para rehabilitación de ni nos con discapacidad motora en extremidades inferiores," B.S. thesis, Espol, 2017.

[61] B. N. Tamburrino Cabrera, "Diseño y construcción de una pierna exoesquelética para la asistencia de la marcha," 2017.

[62] Mendoza Merchán E. Análisis y diseño de un prototipo de exoesqueleto para la rehabilitación pediátrica de los miembros inferiores, utilizando sistemas embebidos para el control del sistema y la interfaz de usuario. [Tesis pregrado]. Guayaquil: Universidad Católica de Santiago de Guayaquil; 2017. Disponible en: http://repositorio.ucsg.edu.ec/handle/3317/7729

[63] A. S. Chamnikar, G. Patil, M. Radmanesh, and M. Kumar, "Trajectory generation for a lower limb exoskeleton for sit-to-stand transition using a genetic algorithm," in Dynamic Systems and Control Conference, vol. 58271.American Society of Mechanical Engineers, 2017, p.V001T36A004.

[64] Romero M luisa. Desarrollo de un exoesqueleto de extremidades inferiores para rehabilitación [Licenciatura]. Universidad Michoacana de San Nicolás de Hidalgo; 2017.

[65] Y. Long, Z.-j. Du, W. Dong, and W.-d. Wang, "Human gait trajectory learning using online gaussian process for assistive lower limb exoskeleton," in Wearable Sensors and Robots. Springer, 2017, pp. 165–179.

[66] A. M. Peñafiel Tenorio, A. D. Santos Castañeda et al., "Implementación del sistema de control de movimiento de extremidades inferiores de exoesqueleto robótico usando un sistema embebido en fpg," B.S. thesis, Espol, 2019.

[67] H. Franco Marquez, "Diseño y construcción de un exoesqueleto para la asistencia en la marcha a pacientes con paraplejia flácida," 2017

[68] D. A. Mendoza Fuentes, "Modelamiento y control de un exoesqueleto de extremidades inferiores para pacientes con esclerosis lateral amiotrofica(ela) y esclerosis múltiple (em)," 2017.

[69] E. J.´Avila Palacios, "Diseño cad y análisis cae de una estructura de exoesqueleto para persona adulta con paraplejía," 2017.

[70] D. J. Hyun, H. Park, T. Ha, S. Park, and K. Jung, "Biomechanical design of an agile, electricity-powered lower-limb exoskeleton for weight-bearing assistance," Robotics and Autonomous Systems, vol. 95, pp. 181–195,2017.

[71] A. Majeed, Z. Taha, A. Abidin, M. Zakaria, I. Khairuddina, M. Razman, and Z. Mohamed, "The control of a lower limb exoskeleton for gaiter habilitation: a hybrid active force control approach," Procedia Computer Science, vol. 105, pp. 183–190, 2017.

[72] Núñes K. Análisis y diseño de un prototipo de exoesqueleto para la rehabilitación pediátrica de los miembros inferiores, utilizando sistemas embebidos para el control del sistema y la interfaz de usuario. [Ingeniero]. Universidad Católica de Santiago de Guayaquil; 2017.

[73] C. C. Velandia, D. A. Tibaduiza, and M. A. Vejar, "Proposal of novel model for a 2 dof exoskeleton for lower-limb rehabilitation," Robotics, vol. 6, no. 3, p. 20, 2017

[74] R. Stopforth, "Customizable rehabilitation lower limb exoskeleton system," International Journal of Advanced Robotic Systems, vol. 9, no. 4,p. 152, 2012.

[75] H. Kim, Y. J. Shin, and J. Kim, "Design and locomotion control of a hydraulic lower extremity exoskeleton for mobility augmentation," Mechatronics, vol. 46, pp. 32–45, 2017.

[76] Y. Long, Z. Du, C. Chen, W. Wang, L. He, X. Mao, G. Xu, G. Zhao, X. Li, and W. Dong, "Development and analysis of an electrically actuated lower extremity assistive exoskeleton," Journal of Bionic Engineering, vol. 14,no. 2, pp. 272–283, 2017.

[77] B. Choi, C. Seo, S. Lee, B. Kim, and D. Kim, "Swing control of a lower extremity exoskeleton using echo state networks," IFAC-Papers OnLine, vol. 50, no. 1, pp. 1328–1333, 2017.

[78] A. Belkadi, H. Oulhadj, Y. Touati, S. A. Khan, and B. Daachi, "On the robust pid adaptive controller for exoskeletons: A particle swarm optimization based approach," Applied Soft Computing, vol. 60, pp. 87–100, 2017.

[79] A. Page, N. Fahrat, V. Mata, A. Valera, M. Díaz, and M. Valles, "Biomechanical model of the lower limb for dynamic control of knee rehabilitation parallel robot," Gait & Posture, vol. 57, pp. 260–261, 2017.

[80] M. O. Ajayi, K. Djouani, and Y. Hamam, "Bounded control of an actuated lower-limb exoskeleton," Journal of Robotics, vol. 2017, 2017.

[81] P. Yang, G. Zhang, J. Wang, X. Wang, L. Zhang, and L. Chen, "Command filter backstepping sliding model control for lower-limb exoskeleton," Mathematical Problems in Engineering, vol. 2017, 2017.

[82] G. Durandau, M. Sartori, M. Bortole, J. C. Moreno, J. L. Pons, and D. Farina, "Real-time modeling for lower limb exoskeletons," in Wearable Robotics: Challenges and Trends. Springer, 2017, pp. 127–131.

[83] F. Trincado-Alonso, A. J. del Ama-Espinosa, G. Asín-Prieto, E. Piñuela-Martín, S. Perez-Nombela, A. Gil-Agudo, J. L. Pons, and J. C. Moreno, "Detection of subject's intention

to trigger transitions between sit, stand and walk with a lower limb exoskeleton," in Wearable Robotics: Challenges and Trends. Springer, 2017, pp. 249–253.

[84] A. Zakaria, A. A. Majeed, I. M. Khairuddin, and Z. Taha, "Kinematics analysis of a 3dof lower limb exoskeleton for gait rehabilitation: A preliminary investigation," in International Conference on Movement ,Health and Exercise. Springer, 2016, pp. 168–172.

[85] N. Mir-Nasiri, "Efficient lower limb exoskeleton for human motion assistance," in Wearable Robotics: Challenges and Trends.Springer,2017, pp. 293–297.

[86] Y. Long, Z. Du, L. Cong, W. Wang, Z. Zhang, and W. Dong, "Active disturbance rejection control based human gait tracking for lower extremity rehabilitation exoskeleton," ISA transactions, vol. 67, pp. 389–397, 2017.

[87] H. A. Sierra, "Control de un exoesqueleto para asistir en la bipedestación y la marcha de una persona," Tesis para obtener el grado de Doctoren Ciencias. Departamento de Control Automático de la Universidad Zacatenco, 2016

[88] C. C. V. Cárdenas, "Modelado, control y monitoreo de un exoesqueleto para asistir procesos de rehabilitación en miembro inferior," Ph.D. dissertation, Universidad Santo Tom´as, 2016.

[89] K. Rincón-Martínez, P. Vera-Tizatl, A. Luviano-Juárez, and I. Chairez, "Prototipo de movilizador robótico de miembros inferiores basado en el concepto de cuidado en el hogar, parte 1: Diseño mecánico e instrumentación," in Memorias del Congreso Nacional de Ingeniería Biomédica, vol. 3, no. 1, 2017, pp. 204–208.

[90] M.´A. Tovar Estrada, "Diseño de un exoesqueleto de miembros inferiores de 14 grados de libertad y su aplicación para emular la locomoción humana." Ph.D. dissertation, Universidad Autonoma de Nuevo León,2016.

[91] A. Zhu, S. He, D. He, and Y. Liu, "Conceptual design of customized lower limb exoskeleton rehabilitation robot based on axiomatic design," Procedia CIRP, vol. 53, pp. 219–224, 2016.

[92] G. Durandau, M. Sartori, M. Bortole, J. C. Moreno, J. L. Pons, and D. Farina, "Emg-driven models of human-machine interaction in individuals wearing the h2 exoskeleton," IFAC-papers online, vol. 49, no. 32, pp.200–203, 2016.

[93] J. Wu, J. Gao, R. Song, R. Li, Y. Li, and L. Jiang, "The design and control of a 3dof lower limb rehabilitation robot," Mechatronics, vol. 33,pp. 13–22, 2016.

[94] M. E. Moreno-Fergusson and M. C. d. P. A. Rey, "Cuerpo y corporalidad en la paraplejia: significado de los cambios," Avances en Enfermería, vol. 30, no. 1, pp. 82–94, 2012

[95] G. Gual Bonet, "¿Mejora la terapia de espejo la función motora de la extremidad inferior afectada tras sufrir un accidente cerebrovascular?

Assessment of Risk Estimates and Fatalities Involved with Covid-19

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Abhay Singh
Student – AIML
Great Learning
Hyderabad, India
abhayps@gmail.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Hyderabad, India
anwesh@greatlearning.in

Jayasurya M
Student – AIML
Great Learning
Hyderabad, India
mjayasurya20@gmail.com

Mukesh Kumar
Student – AIML
Great Learning
Hyderabad, India
muksjb@gmail.com

Krishna Chaitanya
Student – AIML
Great Learning
Hyderabad, India
krishna.kc69@gmail.com

Sumanth S
Student – AIML
Great Learning
Hyderabad, India
suramsumanth@gmail.com

***Abstract*-- The objective of this work is to understand COVID-19 spread and lethality across the world along with the factors affecting them. For this, we first studied the impact of COVID-19 spread across the world and then did a time series analysis to figure out the trend in cases and deaths across the world using FBPROPHET and ARIMA algorithms. We used the publicly available data until 21 January 2021(365 days) for building this model. The second part of this work involves predicting the case fatality rate through country specific indicators such as Socio-Economic, Health, Diet and Weather. To assess these factors, we would use Statistical Analysis and Feature Selection techniques to eliminate insignificant features and provide the most significant features that the healthcare professionals may focus on. In wake of delay in the advent of vaccines and eventual roll-out, this study aims to provide a tool that will assist the already overloaded healthcare system with its recommendations.**

Keywords -- COVID-19; Global; Pandemic; Time Series; Forecast; FBProphet; ARIMA; CFR.

I. INTRODUCTION

The year 2020 is riddled with a global pandemic named Covid-19 which has impacted every living form on this planet. Novel Coronavirus is part of the SARS family which was first identified in December 2019 in Wuhan city in the Hubei Province in China. This disease caused respiratory tract infection and, in some cases, also resulted in severe pneumonia. The disease is found to be extremely contagious by means of droplets and fomites resulting in worldwide pandemic. [1] WHO has announced it as a pandemic on March 11, 2020. [2] Since then this pandemic is spreading exponentially worldwide. As on Dec 06th 2020, current active cases are 6.6million, and 1.5million deaths are reported around the world. One of the most important ways to measure the burden of COVID-19 is mortality. [3] In the current outbreak of novel COVID-19, machine learning techniques have played a vital role in finding the patterns among various countries and thus helping epidemiologists and scientists alike in their research to overcome the disease.

The forecasting of the spread of the pandemic helps to inform governments and healthcare professionals what to expect and which measures to impose, and secondly, to motivate the wider public to adhere to the measures that were imposed to decelerate the spreading lest a regrettable scenario will unfold [4]. Case Fatality Rate is a metric to evaluate the mortality associated with a disease, which can be defined as the portion of confirmed cases leading to fatality. In the modified formula by WHO, the number of closed cases is replaced by number of deaths and number of recovered cases.

$$CFR\,(\%) = (No.of\ deaths/\ No.of\ closed\ cases) * 100. \qquad (1)$$

Current data indicate that, worldwide, case fatality rate (CFR, the ratio between number of deaths and number of confirmed cases) might be around 4%. However, at the country level, CFR ranges from 0 to more than 20%. There are many possible reasons for such a variation. With this study we attempt to model the spread of the pandemic[22] and help recommend the governments and health departments on the management and availability of health infrastructure for their subjects. In absence and or delay of the vaccine this effort will help contain the disease by means of providing targeted healthcare based on

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

individual's needs, mitigate the risks and substantiate the efforts to manage the health infrastructure.

II. LITERATURE REVIEW

Forecasting can predict the cases/deaths through the patterns associated with the time series data, this prediction provides a good understanding of the Spread of virus. This was studied using Prophet and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model forecasting model. [5] These models were able to generate forecast for confirmed cases for Canada, France, India, South Korea and UK. Even though Prophet being the procedure that is widely used for forecasting time series problems. It was not able to address the complex and varying patterns in COVID-19 Time series data, a SARIMA forecasting model is used which generates more accurate forecasts. These predictions will act as early warnings for government policy makers and help health care authorities effectively allocate resources. [5]

To take necessary actions, areas that are likely to be vulnerable should be known. Forecasting can identify the zones which have potential risk and is demonstrated in COVID-19 risk assessment in Counties of the USA [6]. Mann-Kendall and Sen's slope estimator trend analysis and homogeneity analysis (Pettitt's test) classified counties into 0-5 ranks, with 5 being 'very risky'.A Random Forest classifier was trained to classify the counties into risk zones. Socio-Economic status, household composition & disability, housing type & transportation, epidemiological factors and healthcare system factors of the counties were used for building the model and validated using Receiver Operating Characteristic (ROC)-Area Under the ROC Curve. The model achieved 90% accuracy (AUC = 0.90) during the training phase and 84% accuracy (AUC = 0.84) during the testing phase. This resulted to a Map of the USA with colors denoting the risk associated, thus enabling the officials to focus on region and factors which had led to high risk using the feature importance of the RF classifier.[6]

Forecasting[21] associated with case fatality and mortality rate can provide us valuable insights on not only with the spread of the pandemic but also handling the exposed population effectively. Forecasting deaths for different countries can capture the variable rates of fatality. To understand the irregularities pertaining to Case Fatalities for various countries, a research was conducted on Socio-Economic Factors and Health Indicators for a country's impact on the fatalities [7]. The authors have analyzed 16 potential factors which included GDP, Population, GNI, along with Health Indicators like Current Health Expenditure (CHE), Hospital beds per 10000 population, median age. Countries with at least 50 confirmed cases were selected for study at the time of research (14/03/2020). Pearson Correlation Coefficient was calculated for all factors in which 7 were observed to have very little impact on CFR. The Linear regression model built on the other 9 factors using the Ordinary Least Squares method, could capture 30.6% variability in CFR (Adjusted R-Square), GDP per capita, and the number of confirmed cases had the most impact. The best result was obtained by using the forward feature selection technique. Since only 47 countries had 50+ cases the data was insufficient, which had been the major challenge. [7]

Another interesting indicator that proved to have a significant effect on CFR was the Diet. The association between the global mortality rate of COVID-19 cases across different countries and dietary intake of different food groups was studied using an ecological study design [11]. The mortality rate is expressed in terms of CFR (Case-fatality rate). A total of ten food groups have been considered across 144 countries, data for which have been from Food and Agriculture Organization of the United Nations. The food intakes data was expressed in terms of kilocalories per person per day. The mathematical model is built using Bayesian regression model using Random-walk Metropolis-Hastings sampling. Results derived from the mathematical model's reports that COVID-19 case fatality rates were associated with food intake. The study suggests that nutritional factors available at country level could have a role in the mortality of COVID 19 pandemic.[11]

III. MATERIALS AND METHODS

The study is divided into 2 major steps, firstly understanding the spread and impact of this pandemic and then evaluating various factors such as the socio-economic features in the lethality of this global spread.

A. *Predicting the global cases trend(time series)*

Time Series is a series of observations taken at specified (mostly equal) time intervals. Analysis of time series helps us to predict future trends based on previously recorded values. In the Time series, we have only 2 variables, time & the variable we want to forecast, in this scenario the number of cases and number of deaths. To see how Covid-19 has impacted our world we first need to collect the past worldwide data through open-source channels such as the dataset made available by John Hopkins University. The dataset is updated on a daily basis web scraped from reliable worldwide web sources, for this research we are using the dataset until yesterday (21Jan2021). We have created 2 data frames based on countries and dates (for cases as well as deaths). We have bucketed all the cases into labels of different ranges and plot it for the entire world (As shown in below picture). Each color represents the impact of Covid-19 on each country based on cases reported.

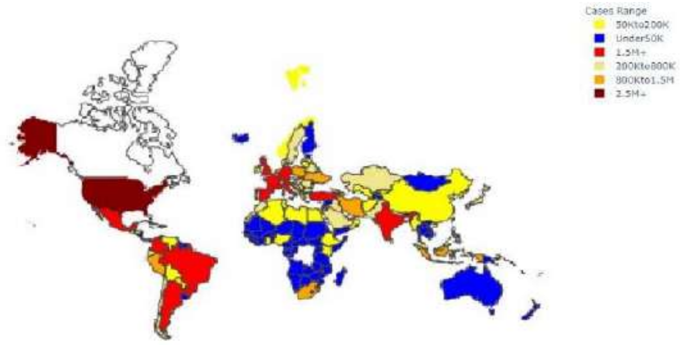


Fig 1. The map interprets the number of confirmed cases in each country (As per legend).

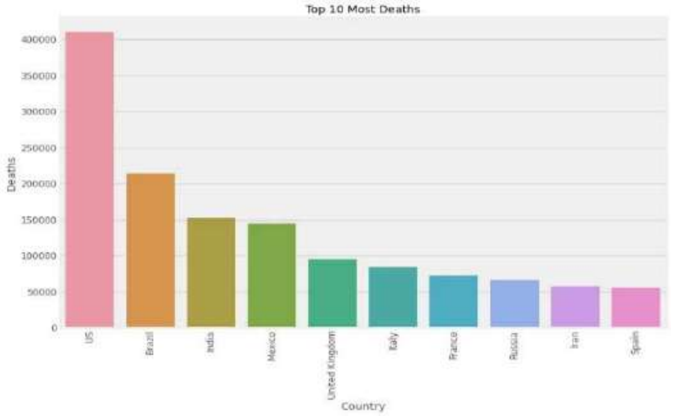


Fig 2. Histogram representation of top 10 countries based on deaths reported due to COVID_19.

For time series forecasting we have primarily used one of the 2 most popular algorithms.

1. Algorithm1- FBPROPHET:

FB Prophet is an open-source algorithm developed by Facebook for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. Just as any time series model we first validate our data with respect to the rolling mean.

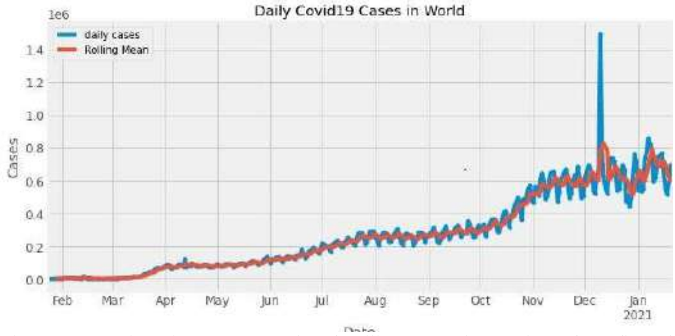


Fig 3. Comparison between actual and rolling mean in number of confirmed cases globally on daily basis.

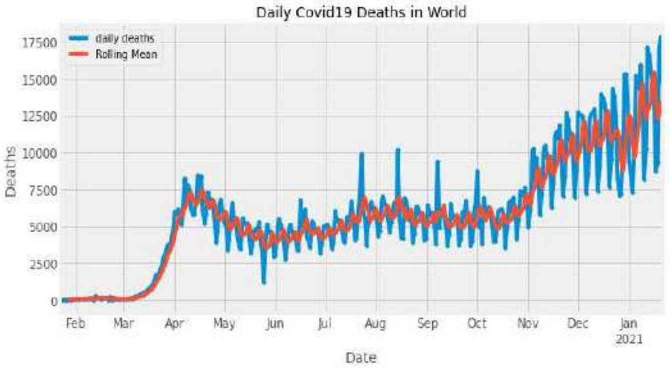


Fig 4. Comparison between actual and rolling mean in number of deaths due to covid-19 globally.

Before we look at the rolling mean one observation can be made with above graph on cases that during late December 2020 there was a sharp spike in cases observed mostly likely due to Christmas and US elections that saw mass gatherings leading to a surge in cases.

The rolling mean comparison is with the actual data that we train our model with and the moving average of that data (both confirmed cases and deaths), based on above graphs we can conclude that the rolling mean is neither overfitting nor deviating too much hence giving us the option for doing time series analysis.

2. Algorithm2 -: ARIMA:

After getting our predictions in one method we are following another algorithm named Auto Regressive Integrated Moving Average (ARIMA). ARIMA is a combination of 2 models AR (Auto Regressive) & MA (Moving Average). It has 3 hyperparameters – p (auto regressive lags), d (order of differentiation) and q (moving avg.) which respectively comes from the AR, I & MA components. The AR part is correlation between previous & current time periods. To smooth out the noise, the Moving Average part is used. The I part binds together the AR & MA parts. For the last model, ARIMA (1,1,1), a model with one AR term and one MA term is being applied to the variable

$$Z_t = X_t - X_{t-1} \qquad (2)$$

A first difference might be used to account for a linear trend in the data. The classical approach for fitting an ARIMA model is to follow the Box-Jenkins Methodology. Below is the summary of the ARIMA model developed for our modelling.

ARIMA basically involves 3 steps:

1. Model Identification: Use plots and summary statistics to identify trends, seasonality, and autoregression elements to get an idea of the amount of differencing and the size of the lag that will be required.

2. Parameter Estimation: Use a fitting procedure to find the coefficients of the regression model.

3. Model Checking: Use plots and statistical tests of the residual errors to determine the amount and type of temporal structure not captured by the model.

The process is repeated until either a desirable level of fit is achieved on the in-sample or out-of-sample observations (e.g., training or test datasets). The best fit is achieved using the lowest AIC value with respect to the best combination of p,d,q values.

To cross validate our predictions we will first split the data midway and then try to plot the predictions for the number of cases as well as deaths to see the accuracy we can achieve. Below mentioned graphs show the predicted value in red compared to the actual values in blue, showing us that our accuracy is pretty good in terms of predictions. We have split our daily data midway until June 2020.

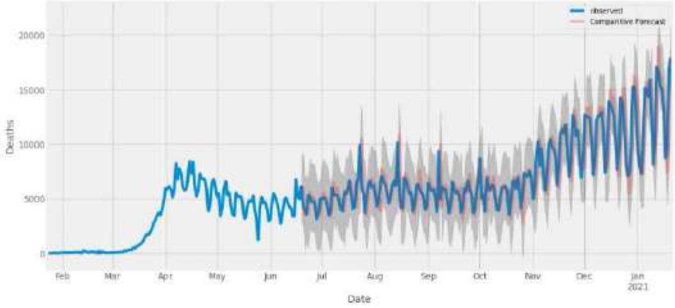


Fig 5. Cross validation between the actual data and predicted data in terms of deaths split midway over time.

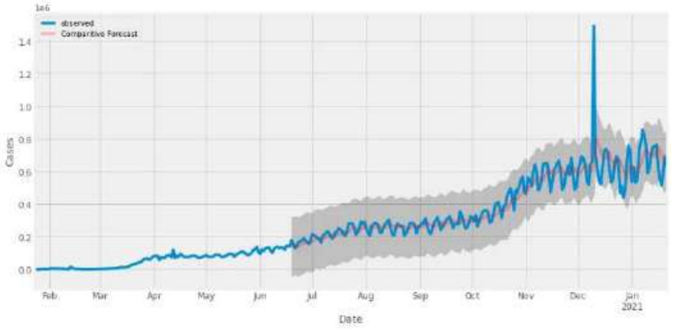


Fig 6. Cross validation between the actual data and predicted data in terms of cases split midway over time

B. *Analysis of factors impacting CFR*

The John Hopkins coronavirus dataset has been used to get the attributes ‘confirmed cases’, ‘recovered cases’ and ‘deaths’ until 15th Jan 2020 for every country. ‘WHO’ has recommended to calculate the CFR as ratio of deaths to sum of deaths and recovered cases, as active cases will not cause any bias in the study. Upon exploration of data, it has been observed that not all countries have followed the same standard to classify a case as recovered. For instance, in Belgium when the patient after treatment havingthe slightest of symptoms like headache were still considered as an active case. Therefore, having very low recovered cases leading to high CFRs. Hence the CFR for this study was calculated by ratio of deaths to sum of deaths and confirmed cases.

GDP, Population, Life Expectancy, GNI per capita, GDP per capita and Human capital Index (HCI) formed the socio-economic factors taken from the World Bank website. Current Health Expenditure (CHE) per capita, CHE as % of GDP, Hospital beds and number of doctors were extracted from WHO (World Health Organization). After the preprocessing and dropping countries with null values, the count came down to 136. Null values were not imputed as it will add a bias to the data.

As seen in research conducted previously in the Literature Review, CFR is treated as a regression problem. The model is expected to capture the maximum variability of the CFR for countries and in the process analyze the features which are able to explain this variability. R square is used as the metric to validate the model. Each attribute’s correlation with the CFR is studied as the univariate analysis. Feature Importance was evaluated with Recursive Feature Extraction (RFE) and Forward & Backward Feature Selection methods. Linear Regression and Random Forest regressor were built on the most important features.

In the next approach, the CFR was bucketed into two classes 0 and 1, where ‘0’ class indicates CFR less than 0.02 that is 2% of the confirmed cases lead to mortality. This approach gives the model to study the CFR by not precisely predicting the CFR. The aim is to study the features and not the precise prediction of CFR. Feature importance was evaluated the same way as the Regression, Logistic Regression and Random Forest Classifier were built to classify and CFR classes.

IV. Results

The FBProphet algorithms give us an estimated prediction in terms of both cases as well as deaths on a day-to-day basis that only suggest that the number of cases as well as deaths will rise going ahead with time.

However, the weekly trend shows a sharp drop during the weekends suggesting that increased number of gatherings and decreased number of testing during the weekends that lead to more cases being reported during weekdays.

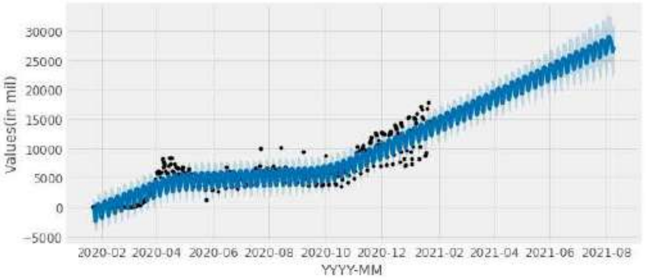


Fig 7. Getting the predictions of deaths (in thousands) for next 200 days (with black spots being the available data that was trained into model)

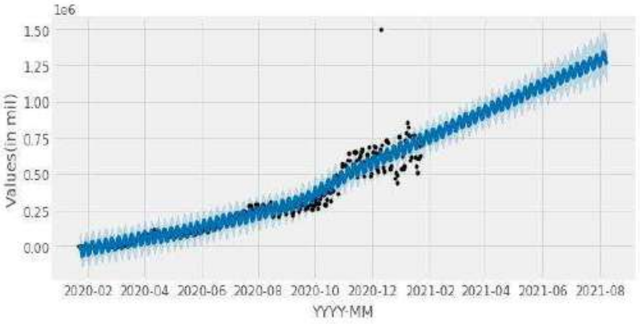


Fig 8. Getting the predictions of cases (in million) for next 200 days (with black spots being the available data that was trained into model)

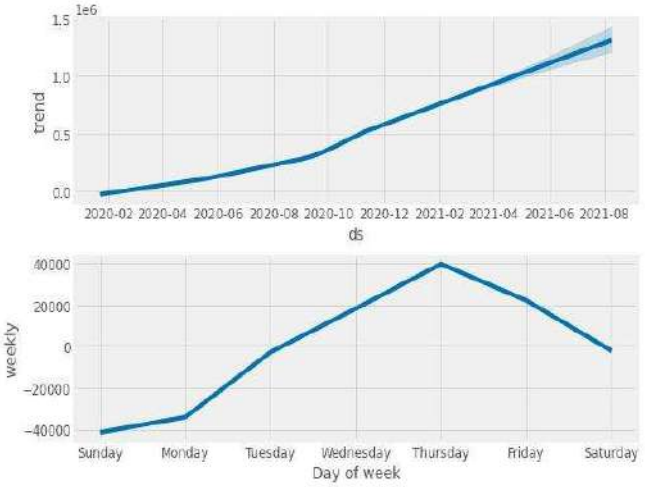


Fig 9. The monthly and weekly predictions of confirmed cases.

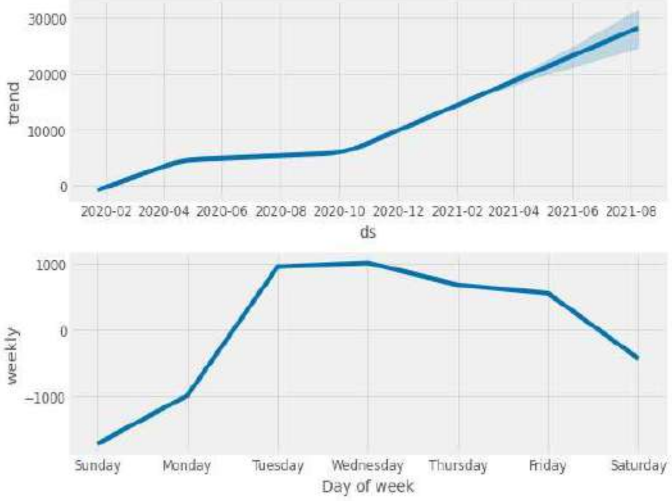


Fig 10. The monthly and weekly prediction of deaths globally.

In the case of ARIMA we can infer that even though the total number of cases in the next 200 days would increase but the number of deaths will probably see a gradual decrease mostly attributed to herd immunity

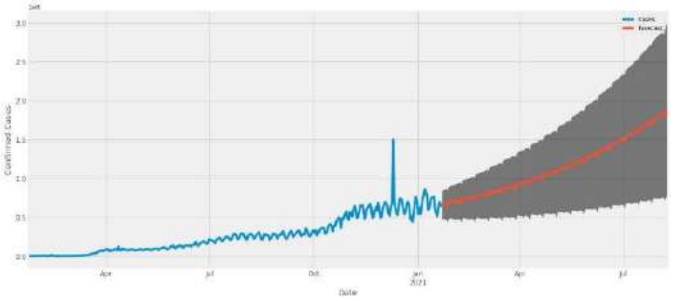


Fig 11. ARIMA prediction of total confirmed cases globally for next 200 days (Shown in red).

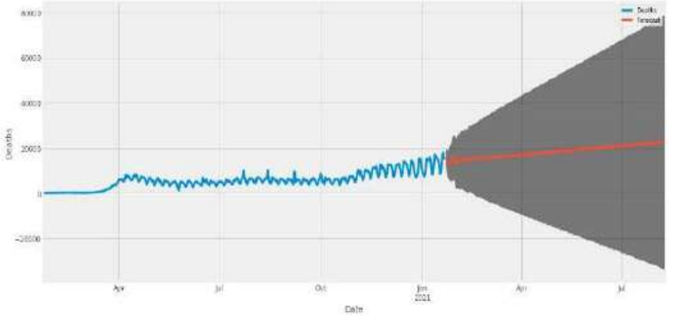


Fig 12. ARIMA prediction of total deaths globally for next 200 days (Shown in red).

The topmost impacting factors and their correlation to CFR can be seen in the below table

TABLE 1: TOP FEATURES HAVING GOOD CORRELATION WITH CFR

Attributes	Correlation	Sign
Stimulants	0.205982	-
Urban Population	0.200971	+
Miscellaneous	0.179596	-
Cereals - Excluding Beer	0.177802	+
Total Population	0.151331	+
Hospital Beds	0.146754	-
GNI/CAPITA	0.146533	-
Starchy Roots	0.140317	-
Fruits - Excluding Wine	0.116855	-
Spices	0.116002	-
HCI	0.114813	-
GDP	0.113702	+
Fish, Seafood	0.113460	-
CHE/CAPITA	0.111517	-
Milk - Excluding Butter	0.110685	-
Doctors	0.109849	-

The '+' correlation of CHE with confirmed, recovered and deaths shows that the countries with higher medical expenditure have covered more tests thus, the higher numbers. The countries with higher urban population reemphasize the above fact.

The regression analysis was conducted along with sequential feature selection with backward and forward selection. The

Linear Regression model has yielded a 0.277 R2 score which indicates the variability of the data around its mean. Random Forest Regressor was able to achieve -0.8489 R2. The table 1 and 2 represents the attributes and their influence on the Target which is the CFR. Of which It is observed that Doctors, HCI, Life Expectancy, CHE/GDP, CHE/CAPITA, Alcoholic Beverages are significant factors.

TABLE 2: FEATURES SELECTED BY BACKWARD SELECTION FOR LINEAR REGRESSION MODEL AND THEIR RESPECTIVE COEFFICIENTS.

Attributes	Coefficients	Sign
Cereals - Excluding Beer	0.012295	+
Animal Products	0.010490	+
HCI	0.008315	-
Obesity	0.008146	+
Life Expectancy	0.005329	+
DOCTORS	0.004749	-
Vegetable Oils	0.003736	+
CHE/GDP	0.003354	+
Alcoholic Beverages	0.002875	+
URBAN Population	0.002606	+

TABLE 3: FEATURES SELECTED BY BACKWARD SELECTION FOR RANDOM FOREST REGRESSOR AND THEIR FEATURE IMPORTANCE.

Attributes	Feature Importance
DOCTORS	0.140603
Spices	0.135485
Alcoholic Beverages	0.134923
CHE/GDP	0.100956
Milk - Excluding Butter	0.087006
CHE/CAPITA	0.079828
Vegetable Oils	0.078902
GNI/CAPITA	0.059623
Undernourished	0.057868
Vegetal Products	0.052789
Sugar Crops	0.040464
Animal Products	0.031553

Doctors, HCI CHE/GDP, CHE/CAPITA have Inverse effect on CFR, which implies higher values of these factors results in lower CFR and vice-versa. Life Expectancy and Alcoholic Beverages have a direct effect on CFR. The early trends of COVID19 outbreak in Italy, established that the aged population was worst affected when it comes to the infection/mortality.

To Understand the Data more and to identify underlying patterns, the classification analysis was conducted and the data was grouped by CFR <0.02.

TABLE 4: FEATURES SELECTED BY BACKWARD SELECTION FOR LOGISTIC REGRESSION AND THEIR COEFFICIENTS.

Attributes	Coefficients	Sign
DOCTORS	0.968987	-
CHE/CAPITA	0.687164	-
Starchy Roots	0.681179	-
CHE/GDP	0.552446	+
Miscellaneous	0.492156	-
Hospital Beds	0.393863	-
Meat	0.301703	+
Cereals - Excluding Beer	0.254172	-
Obesity	0.226581	+
Pulses	0.212649	-
Alcoholic Beverages	0.176067	+
Spices	0.174774	-
Undernourished	0.080433	-
GDP	0.055724	+
Animal Products	0.003265	-

The features that were selected for regression models along with hospital beds had significant effect on the classification of CFR classes. Logistic Regression was able to classify countries into low and high CFR classes with Cross validate mean accuracy of 65% and variance of 0.7%. Whereas the Random Forest Classifier had mean accuracy of 58% and variance of 3.5%.

TABLE 5: FEATURES SELECTED BY BACKWARD SELECTION FOR RANDOM FOREST CLASSIFIER AND THEIR FEATURE IMPORTANCE.

Attributes	Importance
URBAN Population	0.096616
Rural Population	0.090795
DOCTORS	0.085755
Spices	0.085121
Obesity	0.083540
Vegetal Products	0.073592
Starchy Roots	0.070095
Hospital Beds	0.069928
Cereals - Excluding Beer	0.064682
Animal fats	0.063623
Offal	0.057456
Eggs	0.055637
Vegetables	0.053458
Undernourished	0.038416
Sugar Crops	0.011286

V. DISCUSSION AND CONCLUSION

Almost a year ago, not everyone in this world was aware of the severity related to Novel Coronavirus, which has now transpired into a global pandemic and the world is still trying solutions to deal with it. This study certainly hopes to help and clarify the impact of this pandemic. There were questions related to data that has been recorded as the data only reflects based on the number of tests carried out by respective governments. This issue of low number of tests being carried out can be attributed to the lack of medical professionals and infrastructure in some countries, while other countries intentionally carried out low miscellaneous political reasons. This does to an extent hinder research in drawing out a clear picture in dealing with such pandemic.

Our research involves validating the data available to perform modelling and predictions, we have done Exploratory Data Analysis (EDA) to make sure that there are no null/NAN values in the data used, also ensuring that such values are treated appropriately(imputation), The various data types of available columns have been changed accordingly so that the predictive modelling is done without any bias or misappropriation.

With both the algorithms one thing that can certainly be predicted that there will be a gradual increase in the total number of cases globally even though the number of deaths won’t see a huge spike. During the regression and classification study, it was found that there are common diet and socio-economic features that affect CFR. Countries that are predicted to have higher CFR could contain the increasing mortality by investing in healthcare infrastructure. The use of certain ingredients in the diet such as spices can contribute to reducing mortality. The diet factors can explain the low CFR for India and countries that use such ingredients in their cuisines. The timing of this study has not taken the exception of Covid-19 vaccine now available from 5 vaccine makers and vaccination drives set to begin across 150 countries, we can expect to see a sharp drop in the number of cases and deaths reported given the effectiveness of the vaccine. However, there are various mutations of this virus that have developed into regional strands showing divergent characteristics moving ahead, we have to wait and see how mankind would learn and respond to the pandemic that has already affected lives in a big way.

VI. References

[1] Estimating mortality from COVID-19. (n.d.). Retrieved February 1, 2021, from Who.int website: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19

[2] Sarkar, K., Khajanchi, S., & Nieto, J. J. (2020). Modeling and forecasting the COVID-19 pandemic in India. *Chaos, Solitons, and Fractals*, *139*(110049), 110049.

[3] Coronavirus update (live): 103,557,049 cases and 2,238,513 deaths from COVID-19 virus pandemic - worldometer. (n.d.). Retrieved February 1, 2021, from Worldometers.info website: https://www.worldometers.info/coronavirus/

[4] Perc, M., Gorišek Miksić, N., Slavinec, M., & Stožer, A. (2020). Forecasting COVID-19. *Frontiers in Physics*, *8*. doi:10.3389/fphy.2020.00127

[5] Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons, and Fractals*, *135*(109850), 109850.

[6] (N.d.). Retrieved February 1, 2021, from Researchgate.net website: https://www.researchgate.net/publication/344360045_Using_Machine_Learning_to_Develop_a_Novel_COVID-19_Vulnerability_Index_C19VI

[7] Asfahan, S., Shahul, A., Chawla, G., Dutt, N., Niwas, R., & Gupta, N. (2020). Early trends of socio-economic and health indicators influencing case fatality rate of COVID-19 pandemic. *Monaldi Archives for Chest Disease*, *90*(3), 451–457.

[8] Yadaw, A. S., Li, Y.-C., Bose, S., Iyengar, R., Bunyavanich, S., & Pandey, G. (2020). Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *The Lancet. Digital Health*, *2*(10), e516–e525.

[9] Khafaie, M. A., & Rahim, F. (2020). Cross-country comparison of case fatality rates of COVID-19/SARS-COV-2. *Osong Public Health and Research Perspectives*, *11*(2), 74–80.

[10] Malki, Z., Atlam, E.-S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons, and Fractals*, *138*(110137), 110137.

[11] Eltoukhy, A. E. E., Shaban, I. A., Chan, F. T. S., & Abdel-Aal, M. A. M. (2020). Data analytics for predicting COVID-19 cases in top affected countries: Observations and recommendations. *International Journal of Environmental Research and Public Health*, *17*(19), 7080.

[12] Brownlee, J. (2017, January 8). How to create an ARIMA model for time series forecasting in python. Retrieved February 1, 2021, from Machinelearningmastery.com website: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/.

[13] Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. Retrieved from http://arxiv.org/abs/2004.07859

[14] fbprophet. (n.d.). Retrieved February 1, 2021, from Pypi.org website: https://pypi.org/project/fbprophet/

[15] freespirit. (2018, August 24). Time series for beginners with ARIMA. Retrieved February 1, 2021, from Kaggle.com website: https://www.kaggle.com/freespirit08/time-series-for-beginners-with-arima

[16] Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values. (n.d.). Retrieved February 1, 2021, from Psu.edu website: https://online.stat.psu.edu/stat510/book/export/html/665.

[17] Where, Z. X. (n.d.). A generalization of ARMA models which incorporates a wide class of nonstation- ary TS is obtained by introducing the differencing into the model. The simplest example of a nonstationary process which reduces to a stationary one after dif- ferencing is Random Walk. As we have seen in Section 4.5.2 Random Walk is a nonstationary AR(1) process with the value of the parameter φ equal to 1, that is the model is given by. Retrieved February 1, 2021, from Qmul.ac.uk website: http://www.maths.qmul.ac.uk/~bb/TimeSeries/TS_Chapter7.pdf

[18] COVID19-India API. (n.d.). Retrieved February 1, 2021, from Covid19india.org website: https://api.covid19india.org/documentation/csv/

[19] Prophet. (n.d.). Retrieved February 1, 2021, from Github.io website: https://facebook.github.io/prophet/

[20] Home - johns Hopkins Coronavirus resource center. (n.d.). Retrieved February 1, 2021, from Jhu.edu website: https://coronavirus.jhu.edu/

[21] N. Darapaneni, D. Reddy, A. R. Paduri, P. Acharya, and H. S. Nithin, “Forecasting of COVID-19 in India using ARIMA model,” in 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0894–0899.

[22] N. Darapaneni et al., “COVID-19 Infection Dynamics for India-Forecasting the Disease using SIR models,” in 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 387–392.

Steer-by-Wire Control System Based on Carsim and Simulink

1st Haiyuan Wei
Hubei Key Laboratory of Advanced Technology for Autom otive Components, Wuhan University of Technology
Hubei Collaborative Innovation Center for Automotive Com ponents Technology, Wuhan University of Technology
Hubei Research Center for New Energy & Intelligent Conne cted Vehicle, Wuhan University of Technology
China Automotive Technology Research Center Co., Ltd.
Wuhan 430070, China
why_whut@163.com

2nd Jingjing Wang
China Automotive Data Co., Ltd.
China Automotive Technology Research Center Co., Ltd.
Tianjin 300393, China
wangjingjing@catarc.ac.cn

3rd Meng Jian
China Automotive Data Co., Ltd.
China Automotive Technology Research Center Co., Ltd.
Tianjin 300393, China
mengjian@catarc.ac.cn

4th Shengming Mei
Hubei Key Laboratory of Advanced Technology for Autom otive Components, Wuhan University of Technology
Hubei Collaborative Innovation Center for Automotive Com ponents Technology, Wuhan University of Technology
Hubei Research Center for New Energy & Intelligent Conne cted Vehicle, Wuhan University of Technology
Wuhan 430070, China
1041079221@qq.com

Miaohua Huang*
Hubei Key Laboratory of Advanced Technology for Autom otive Components, Wuhan University of Technology
Hubei Collaborative Innovation Center for Automotive Com ponents Technology, Wuhan University of Technology
Hubei Research Center for New Energy & Intelligent Conne cted Vehicle, Wuhan University of Technology
Wuhan 430070, China
mh_huang@163.com
*Corresponding Author

***Abstract*—With the development of artificial intelligence and driverless technology, the steer-by-wire system has become the mainstream direction of future development. Its structure is simple, lightweight, and real-time optimized steering gear ratio according to road conditions, steering road feel, and improved vehicle handling stability and active safety. Therefore, the research on the wire-controlled steering control system has great practical value. It is applied to platforms such as electric vehicles and intelligent networked vehicles to realize unmanned driving and optimize intelligent transportation, which has good application prospects. Based on the analysis of the structure of each component of the steer-by-wire system, this paper established the dynamic models of the steering execution part, the rack and pinion part and the steering wheel part, and completed the simulation model of each part and the three closed-loop control of the motor through Simulink. This paper, finally built a Carsim/Simulink co-simulation platform, integrated the motor control strategy into the vehicle model in Carsim, and verified the following characteristics and fast response characteristics of the steer-by-wire model.**

Keywords—Control, Steer-by-wire, Three closed loop, Carsim, Simulink

I. Introduction

Realizing the real-time control of the steer-by-wire system is crucial to the development and popularization of intelligent vehicles in the future. For this purpose, many scholars have conducted a lot of research on the control methods and strategies of the steer-by-wire system.

Zhao, Huiyong et al. proposed an energy-saving strategy for SBW vehicle steering motors by analyzing the relationship between steering wheel angle and road curvature, and proved the feasibility and effectiveness of the energy-saving strategy through closed-loop simulation[1]; Dankert, Jens etc. through an overview of the system prototype, focusing on the technical and electrical settings, the main hardware and software components, including a brief introduction to the safety concept[2]; Lee, Jaepoong and others have developed a haptic control of the steer-by-wire system. After experimental verification, the haptic controller proposed successfully tracks the steering feedback torque of the steer-by-wire system[3]; Scicluna, Kris et al. proposed a sensorless position control of a permanent magnet synchronous motor on the steering wheel side of a steer-by-wire system of an automobile, and tested the performance of the sensorless drive under forward and reverse load torque conditions[4]; Wang, Zezheng proposed a compound control scheme including proportional integral derivative controller and fuzzy logic system, and proved the rationality and superiority of the method through numerical simulation and hardware-in-the-loop experiment [5]; Scicluna, Kris et al. proposed a sensorless current control for the permanent magnet synchronous motor on the hand wheel side of the “steering-by-wire” system in automobiles, and compared the performance of vehicles with traditional steering devices through experiments[6]; Hua, Min et al. proposed the dual-motor synchronous control using differential negative feedback method for the non-synchronization of the double-steering actuator motor to reduce the adverse effects of the "servo fight" non-synchronization problem. The simulation results and hardware-in-the-loop experiments proved that the proposed

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

control strategy was effective and feasible[7]; Luo, Yutao and others proposed an adaptive neural network sliding mode control method that considers system disturbances. Research showed that RBFSMC has better robustness and stability than SMC[8]; Wu, Xiaodong and others designed the variable steering ratio of the steer-by-wire system. Experimental results showed that the system can not only improve the steering agility at low speeds and the steering stability at high speeds, but also reduce driving workload under critical driving conditions[9]; Ye, Mao et al. proposed a new type of robust adaptive overall terminal sliding mode control strategy based on extreme learning machine, and proved the excellent control performance of the proposed control through simulation [10]; Ma, Bingxin et al. proposed an event-triggered adaptive sliding mode control method for SBW systems, and the effectiveness of the method was demonstrated through numerical simulation and experimental results[11]; Huang, Chao et al. proposed a fault-tolerant model predictive control with fault compensation for an SBW system with actuator failure based on an incremental operator. The simulation results showed that the proposed fault-tolerant control strategy can cope with various types of actuator failures[12]; Zou, Songchun et al. proposed a dual-motor steer-by-wire system to improve the fault tolerance and reliability of the system, and verified the effectiveness of the control strategy proposed in this paper through simulation and experimental results[13]; Huang, Chao et al. proposed a fault-tolerant control method for actuators based on fault detection and isolation based on model predictive control. Simulation and experimental results showed that compared with traditional model predictive controllers, this method has no fault detection and isolation. It can achieve better steering performance and stabilize the entire SBW system in the event of actuator failure[14]; Deng, Bin et al. proposed a recursive sliding mode structure, which can effectively reduce shake without reducing the control accuracy, and experimentally proved the fast convergence speed and smaller shake of the proposed controller[15]; Jin, Zhilin and others proposed the multi-gain ratio of the steer-by-wire system, and studied the anti-rollover control strategy of the vehicle with the steer-by-wire system. The experimental results showed that the proposed anti-rollover strategy can effectively prevent the vehicle from tripping. A rollover occurred both in the case of a rollover without tripping, and the intervention of the driver's steering intention is reduced[16]; Zhang, Jie et al. proposed a robust adaptive sliding mode controller to improve the maneuverability and lateral stability of steer-by-wire vehicles. The results of hardware-in-the-loop simulations prove the proposed control excellent stability control performance of the system under different steering maneuvers[17]; Yu, Shu-You et al. proposed a tube model predictive control scheme for a four-wheel steer-by-wire vehicles. The scheme adopts an event trigger strategy to prevent potential adversary attacks and ensure information security[18].

The control of the steer-by-wire system based on the permanent magnet synchronous motor adopts the closed-loop control technology of position loop, speed loop, current loop and SVPWM. In order to study the steering control strategy of the SBW system, it is necessary to establish SBW vehicle dynamics based on Matlab/Simulink Models, including steering execution model, rack and pinion model, and steering wheel model. The SBW system is a complex mechanical structure. It is difficult to establish an accurate mathematical model. This paper made some necessary simplifications and assumptions when establishing the SBW system dynamics model, and adopted a reduced-order modeling method to establish the SBW system module and the front wheel steering module, and finally combined the vehicle model in the Carsim software to form the SBW vehicle dynamics simulation platform, and verified the accuracy of the co-simulation model established based on the Carsim-Simulink software. In order to verify the SBW vehicle simulation model and the traditional vehicle model with the same vehicle parameters in the Carsim software, chose the double line change test conditions for simulation. The process is shown as Figure 1.

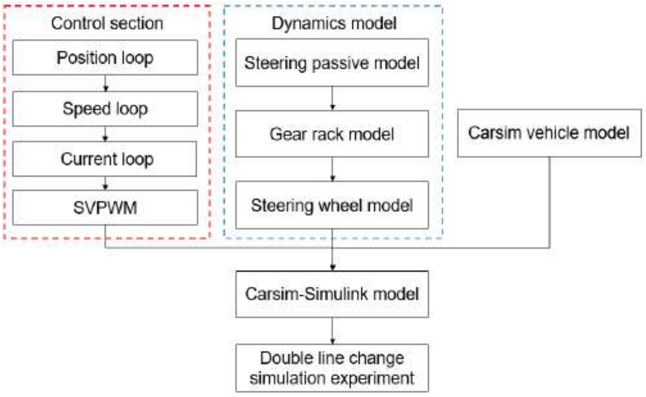


Figure 1 The process of steer-by-wire system

II. Models And Data

A. Dynamic model of steer-by-wire system

Since the driverless vehicle cancels the steering wheel module, the SBW system can be divided into steering actuator components, rack and pinion components, and left and right steering wheel components. As shown in Figure 2, the steering actuator components consists of the steering motor and its reducer. The rack and pinion assembly refers to the rack and pinion steering gear. And left and right steering wheels components refer to the left and right steering knuckles and their steering wheels. In this modeling process, the transmission gap between the steering motor and the steering wheels is ignored.

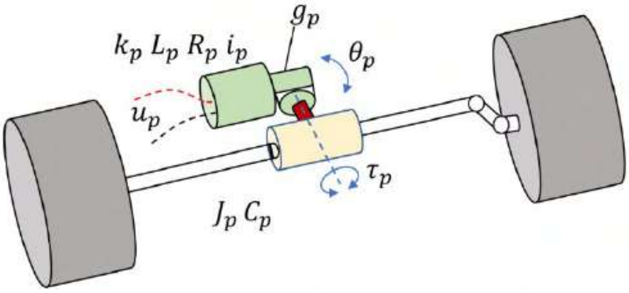


Figure 2 The front wheel steering model structure diagram

The steering actuator components is composed of a steering motor and a reducer. The steering motor is generally a permanent magnet synchronous motor. This motor can be well used in the steer-by-wire system to accurately adjust the steering angle of the vehicle, and then adjust the steering of the vehicle; most of the matched reducers are worm gear reducers, which can change the direction of the steering execution motor while ensuring reduce speed and increase torque, so that the axis of the steering motor and the rack axis are vertically distributed, which can greatly reduce the layout space of the steer-by-wire system along the longitudinal direction of the vehicle, and eliminate the interference of the steer-by-wire system and the suspension, which is called a cross steering gear. The dynamic equation of the steer-by-wire actuator components composed of a permanent magnet synchronous motor and a worm gear reducer can be expressed as formula (1).

$$T_s = J_s\ddot{\theta}_s + B_s\dot{\theta}_s + T_r \tag{1}$$

Where T_s is the electromagnetic torque of the steering motor, J_s is the rotor moment of inertia of the steering motor, B_s is the viscous damping coefficient of the steering motor, θ_s is the rotor angle of the steering motor, T_r is the torque acting on the rack by the steering motor, and shown as formula (2). K_s is the torsional stiffness of the steering motor, X_r is the rack displacement, g_s is transmission ratio of the rotation angle of the motor and the displacement of the rack is executed for the steering, and r_p is the radius of the steering pinion.

$$T_r = K_s(\theta_s - X_r g_s / r_p\ ) \tag{2}$$

The rack and pinion assembly consists of a steering pinion and a rack, as shown in Figure 3. And the dynamic equation can be expressed as formula (3).

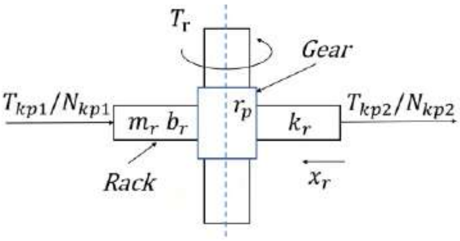


Figure 3 The structure diagram of rack and pinion

$$m_r\ddot{x}_r + B_r\dot{x}_r + k_r x_r = \frac{T_r}{r_p}\eta_p - \frac{T_{kp1}}{N_{kp1}}\eta_n - \frac{T_{kp2}}{N_{kp2}}\eta_n \tag{3}$$

Among them, m_r is the rack mass. B_r is the rack viscous damping coefficient. The k_r is the rack stiffness. The x_r is the rack displacement. $N_{kp1}$and N_{kp2} are the rack to left and right steering knuckle arm transmission ratios. The $\eta_p$and η_n are the steering forward transmission efficiency and reverse transmission efficiency respectively. K_{kp1} and K_{kp2} are the torsional stiffness of the left and right steering wheels around the kingpin respectively. θ_{FW1} and θ_{FW2} are the rotation angles of the left and right steering wheels. T_{kp1} and T_{kp2} are the moments of the left and right steering wheels acting on the kingpin, shown as formula (4) and (5).

$$T_{kp1} = K_{kp1}\left(\frac{x_r}{N_{kp1}} - \theta_{FW1}\right) \tag{4}$$

$$T_{kp2} = K_{kp2}\left(\frac{x_r}{N_{kp2}} - \theta_{FW2}\right) \tag{5}$$

The left and right steering front wheel components are composed of left and right steering knuckles and steering wheels, as shown in Figure 4. The dynamic equation can be expressed as formula (6) and (7).

$$J_{FW1}\ddot{\theta}_{FW1} + B_{kp1}\dot{\theta}_{FW1} = T_{kp1} - T_1 \tag{6}$$

$$J_{FW2}\ddot{\theta}_{FW2} + B_{kp2}\dot{\theta}_{FW2} = T_{kp2} - T_2 \tag{7}$$

Among them, J_{FW1} and J_{FW2} are the moment of inertia of

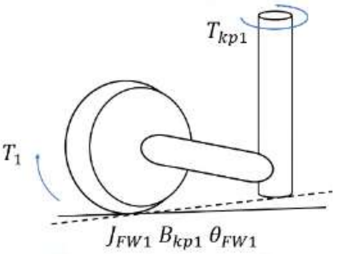


Figure 4 The steering wheel structure diagram

the left and right steering wheels around the kingpin respectively; B_{kp1} and B_{kp2} are the viscous damping coefficients of the left and right steering wheels around the kingpin respectively; T_1 and T_2 are resistance moment, which are respectively acted on the left and right steering wheels by the ground.

B. Simulink model of steer-by-wire system

According to the dynamic equation of the rack and pinion assembly, the basic module provided by Simulink is used to establish the simulation model of the rack and pinion assembly as shown in Figure 5.

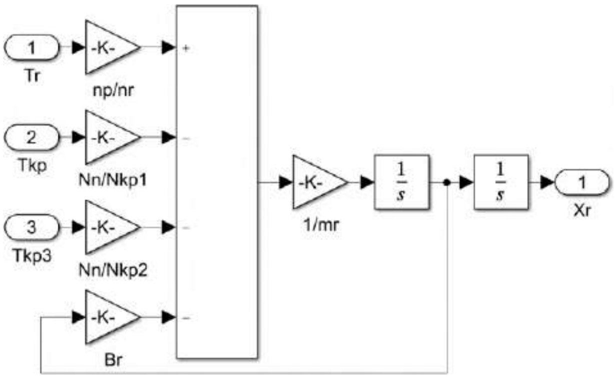


Figure 5 Simulation model of rack and pinion

According to the left and right steering front wheel assembly dynamics equation, the basic module provided by Simulink is used to establish the left and right steering wheel assembly simulation model as shown in Figure 6 and 7.

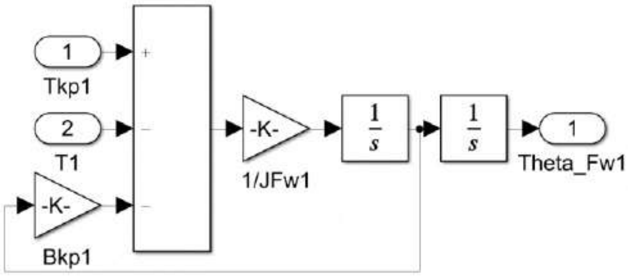


Figure 6 Left steering wheel simulation model

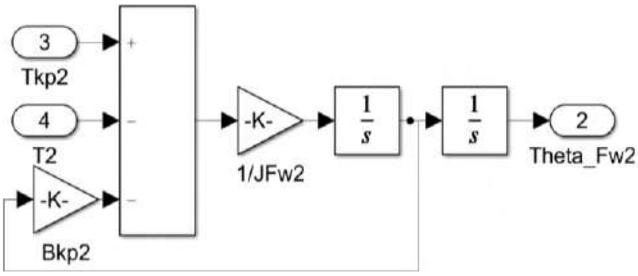


Figure 7 Right steering wheel simulation model

The parameters required for the rack and pinion simulation model and the steering wheel simulation model found above are shown in Table I below.

TABLE I THE STEERING MECHANISM PARAMETER TABLE

Parameter	value	Parameter	value
Steering pinion radius $r_p(m)$	0.0088	Rack quality $m_r(kg)$	5
Steering shaft stiffness $K_c(N\cdot m/rad)$	98	Reverse efficiency $\eta_n(\%)$	65
Forward efficiency $\eta_p(\%)$	75	Rack viscous damping coefficient $B_r(N\cdot s\cdot m^{-1})$	312
Steering kingpin torsion stiffness $K_{kp1}(N\cdot m/rad)$	39951	Rack stiffness $K_r(N/m)$	2.588 $\times 10^8$

According to the dynamic model of the steering motor, the machine module that has been packaged by Matlab/Simulink is used to establish the steering motor simulation model. The relevant parameter settings are shown in Table II below.

TABLE II THE STEERING MOTOR PARAMETER TABLE

Parameter	value	Parameter	value
Rotor moment of inertia $J_s(kg\cdot m^2)$	0.006	Armature inductance $L_s(mH)$	0.00029
Rotor damping $B_s(N\cdot s/m)$	0.0017	Back electromotive force constant $K_{se}(V\cdot s/rad)$	0.06
Torsional stiffness of motor shaft $K_s(N\cdot m/rad)$	121	Torque coefficient $K_{sT}(N\cdot m/A)$	0.086
Armature resistance $R_s(\Omega)$	0.068	Coefficient of viscous friction of motor $B_m(N\cdot s/rad)$	0.01

C. Control model of steer-by-wire system

The motor adopts a three-closed-loop PID control strategy, and the inner loop is current loop tracking control, the middle is speed loop tracking control, and the outer loop is angle tracking control. The motor three-closed-loop control system is actually based on the basic double-closed-loop control, adding a position loop. The double-closed loop can be regarded as a speed control module, and the position regulator will determine the difference between the given position signal and the actual position detection signal. The setting is the speed setting of this control module, so that the motor accurately follows the position setting. The greater position deviation, the greater corresponding speed setting.

Because the PID regulators of the position loop, speed loop and current loop are involved in the paper, in order to reduce the system response overshoot and system response time, the parameters of the three PID regulators are adjusted. This paper sets the input signal to two steps signal, first set the regulator to pure P adjustment. Since most position control does not allow overshoot, adjust the initial parameters of the position P regulator from 500. When K_p=500 is set, a certain amount of overshoot appears in the system response, indicating that the value of K_p is too large, and then the value of K_p is reduced to make the system response result meet the requirements. The value of the position I regulator is the same. The final adjustment result is shown in Figure 8, and the stator three-phase output current is shown in Figure 9. The position loop K_p=250, K_i=0.

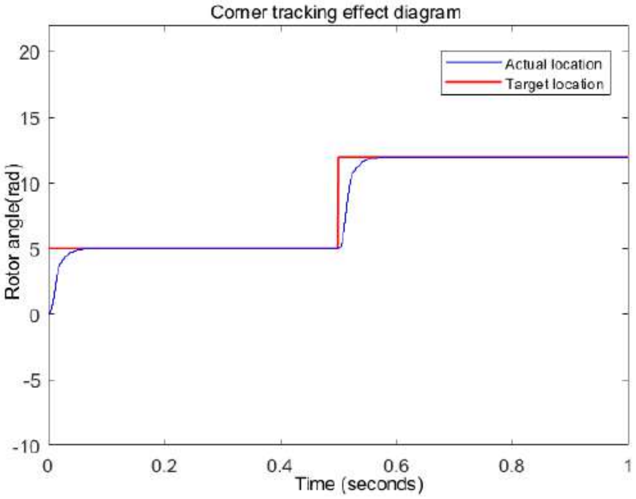


Figure 8 Location tracking result diagram

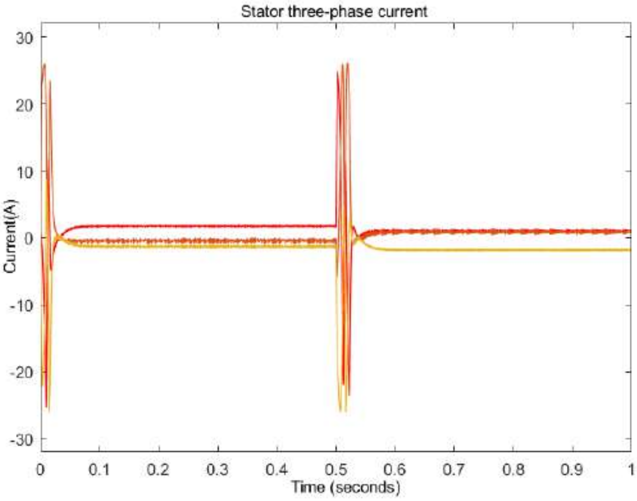


Figure 9 Stator three-phase current output diagram

III. SIMULATION AND RESULTS

The Matlab/Simulink simulation model of the front wheel steering module is embedded into the vehicle model of the Carsim software, instead of the traditional steering system, to obtain the SBW vehicle co-simulation model, as shown in Figure 10. The input of Carsim software are the steering angle of the left and right steering wheels, and the output are the steering resistance torque of the steering wheels and the vehicle state parameters. The vehicle model in Carsim software selects its own B-class vehicle, and some vehicle parameters are shown in Table III.

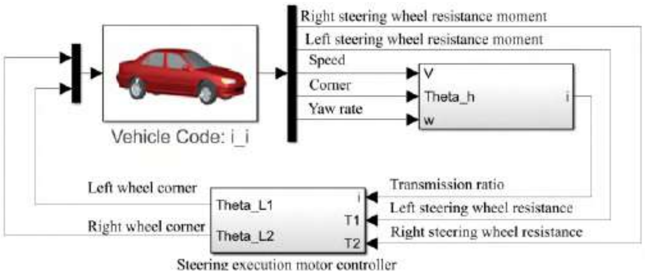

Right steering wheel resistance moment
Left steering wheel resistance moment
Speed
Corner
Yaw rate
V
Theta_h
w
i
Vehicle Code: i_i
Left wheel corner
Right wheel corner
Theta_L1
Theta_L2
i
T1
T2
Transmission ratio
Left steering wheel resistance
Right steering wheel resistance
Steering execution motor controller

Figure 10 Matlab/Simulink simulation model

In order to verify the accuracy of the co-simulation model based on the Carsim-Simulink software, the double-line change condition that best reflects the vehicle path tracking

TABLE III THE VEHICLE PARAMETERS TABLE

Parameter	value	Parameter	value
Vehicle length(mm)	4323	Moment of inertia around Z axis $I_z(kg \cdot m^2)$	1523
Vehicle width (mm)	1765	Front wheel track B_f(mm)	1480
Vehicle height (mm)	1660	Rear wheel track B_r(mm)	1475
Vehicle quality m(kg)	1274	Front axle load $F_{Zf}(N)$	7600
Distance from center of mass to front axle l_f(mm)	1010	Distance from center of mass to rear axle l_r(mm)	1550
Rear axle load $F_{Zr}(N)$	4950	The height of the center of mass of the vehicle h(mm)	650
Wheelbase L(mm)	2560		

characteristics is chose for the simulation test to verify the SBW car simulation model and the traditional car with the same vehicle parameters in the Carsim software model. The angular transmission ratio of the SBW system adopts the transmission ratio of the traditional car model in the Carsim software and is set to a fixed value of 20. The simulation results are shown in Figures 11 and 12.

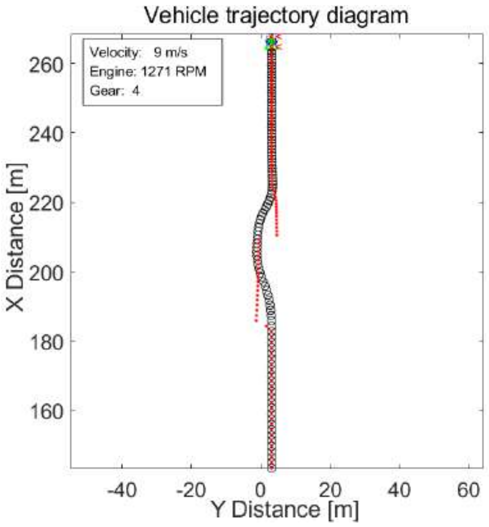


Figure 11 The vehicle trajectory of steering diagram

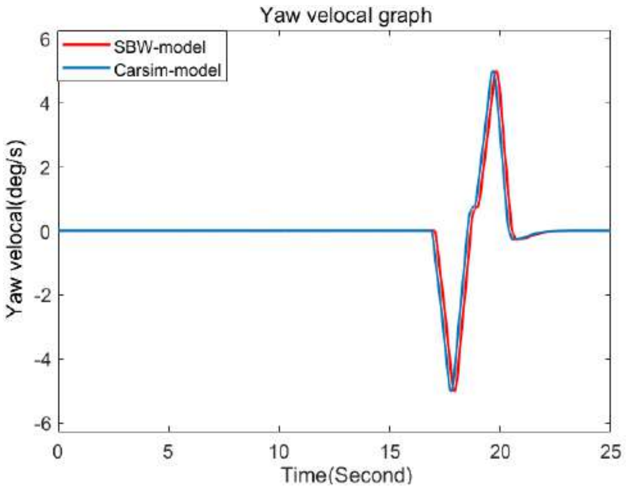


Figure 12 The yaw rate diagram

IV. CONCLUSION

By establishing the dynamic model, simulation model and three closed-loop control strategies of the motor for the steer-by-wire system, the Carsim and Simulink co-simulation platform is finally built. It can be seen from Figure 12 that the yaw rate response of the SBW car model under the double-line-change test conditions lags behind the front wheel angle of the Carsim car model by about 0.15s, but the peak values of the two are not much different, and the curve change trend is the same. The angle tracking error is no more than 8%, which shows that the SBW vehicle model and the three-closed-loop motor control strategy established in this paper also have good angle following characteristics under the two-line change test conditions, and can accurately reflect the dynamic characteristics of the vehicle.

REFERENCES

[1] Zhao, H., et al., An energy-saving strategy for steering-motors of steer-by-wire vehicles. 2020. 6(2): p. 234-262.

[2] Dankert, J., S. Dreyer and L. Eckstein. Intelligent and Safe Components for Steer-by-wire Control Sidesticks as a Smart Actor. 2017. Dortmund, Germany: VDE Verlag GmbH.

[3] Lee, J., et al., Haptic control of steer-by-wire systems for tracking of target steering feedback torque. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2020. 234(5): p. 1389-1401.

[4] Scicluna, K., C.S. Staines and R. Raute. Sensorless Position Control at the Steered Wheel in Steer-by-Wire : using High-Frequency Injection with Search-Based Observer. 2020. Palermo, Italy: Institute of Electrical and Electronics Engineers Inc.

[5] Wang, Z., Adaptive fuzzy system compensation based model-free control for steer-by-wire systems with uncertainty. International Journal of Innovative Computing, Information and Control, 2021. 17(1): p. 141-152.

[6] Scicluna, K., C.S. Staines and R. Raute. Sensorless Current Control at the Handwheel in Steer-by-Wire : using High-Frequency Injection with Search-Based Observer. 2020. Palermo, Italy: Institute of Electrical and Electronics Engineers Inc.

[7] Hua, M., et al., Research on synchronous control strategy of steer-by-wire system with dual steering actuator motors. International Journal of Vehicle Autonomous Systems, 2020. 15(1): p. 50-76..

[8] Luo, Y. and H. Guo, Adaptive Neural Network Sliding Mode Control for Steer-by-Wire System. Huanan Ligong Daxue Xuebao/Journal of South China University of Technology (Natural Science), 2021. 49(1): p. 65-73.

[9] Wu, X. and W. Li, Variable steering ratio control of steer-by-wire vehicle to improve handling performance. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2020. 234(2-3): p. 774-782.

[10] Ye, M. and H. Wang, Robust adaptive integral terminal sliding mode control for steer-by-wire systems based on extreme learning machine. Computers and Electrical Engineering, 2020. 86.

[11] Ma, B. and Y. Wang, Adaptive type-2 fuzzy sliding mode control of steer-by-wire systems with event-triggered communication. 2021.

[12] Huang, C., F. Naghdy and H. Du, Delta Operator-Based Model Predictive Control with Fault Compensation for Steer-by-Wire Systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020. 50(6): p. 2257-2272.

[13] Zou, S. and W. Zhao, Synchronization and stability control of dual-motor intelligent steer-by-wire vehicle. Mechanical Systems and Signal Processing, 2020. 145.

[14] Huang, C., et al., Actuator fault tolerant control for steer-by-wire systems. 2020..

[15] Deng, B., K. Shao and H. Zhao, Adaptive Second Order Recursive Terminal Sliding Mode Control for a Four-Wheel Independent Steer-by-Wire System. IEEE Access, 2020. 8: p. 75936-75945.

[16] Jin, Z., W. Liang and W. Zhao, Multi-gains Ratio of Steer-by-wire System and Anti-rollover Control for Vehicle. Jixie Gongcheng Xuebao/Journal of Mechanical Engineering, 2020. 56(10): p. 172-180.

[17] Zhang, J., et al., Adaptive Sliding Mode-Based Lateral Stability Control of Steer-by-Wire Vehicles with Experimental Validations. IEEE Transactions on Vehicular Technology, 2020. 69(9): p. 9589-9600.

[18] Yu, S., et al., Active information security oriented steering control of steer-by-wire vehicles. Kongzhi yu Juece/Control and Decision, 2019. 34(11): p. 2414-2420.

Predicting the Impact of Covid-19 Pandemic in India

Narayana Darapaneni
Director – AIML
Great Learning/Northwestern
University Illinois, USA
darapaneni@gmail.com

Suma Maram
Student – AIML
Great Learning
Hyderabad, India
srisumarya7@gmail.com

Mandeep Kour
Student – AIML
Great Learning
Hyderabad, India
highmandeep1990@gmail.com

Harpreet Singh
Student – AIML
Great Learning
Hyderabad, India
harpreeth21@gmail.com

Sathish Nagam
Student – AIML
Great Learning
Hyderabad, India
Sathish.Nagam@Brambles.com

Anwesh Reddy Paduri
Data Scientist - AIML
Great Learning
Hyderabad, India
anwesh@greatlearning.in

***Abstract* -- COVID-19 is spreading at an unprecedented pace around the world. It has now spread to over 200 countries around the world. CoVID-19 mathematical modelling is often useful for strategic decision making in highly populated countries like India to gain some understanding of the epidemic's future. The prediction of an outbreak has become more difficult as the pandemic scenario of COVID-19 cases has grown. We hope to forecast the effects of COVID-19 in India by gaining a better understanding of its lifecycle in various Indian states. From historical data of verified COVID-19 cases, we are attempting to forecast potential COVID-19 cases and active cases.. For the prediction of COVID-19 we are implementing Susceptible-Infected-Recovered (SIR) model and FB-Prophet model for time series analysis. SIR modelling is more intuitive and explainable, but requires a lot of trial and error and assumptions. The FB-Prophet prediction process is simple and accuracy is also better compared to SIR modelling. In this model we are trying to understand the spread of COVID-19 in the ten most affected states of India (as on 9th December 2020) using publicly available state-wise time series data of COVID-19 patients. In this paper, we discuss how such continuous and unparalleled factors lead us to design intricate models, as It's time to use data-driven, mathematically proven models with the ability to tune parameters dynamically and automatically over time.**

Keywords— COVID-19, prediction, SIR model, Facebook Prophet.

I. Introduction

The CoVID-19 coronavirus outbreak started in Wuhan, China, and has quickly spread around the world. The outbreak of this virus has been deemed so dangerous by the World Health Organization (WHO) that it was declared a pandemic disease on March 11, 2020 [1, 2]. Because of its rapid spread, this pandemic disease is posing a serious threat to people's health and safety all over the world. Regardless of how advanced the healthcare system is, every nation has limited medical services. The most commonly used method for containing the spread of the novel corona virus has been social distancing. The first COVID-19 positive case in India was discovered on January 30, 2020, in Kerala. The Indian government also imposed a full nationwide lockdown on the 25th of March 2020, which lasted for 21 days and was extended until the 31st of May 2020, in order to minimize the number of people infected and slow the virus's spread. Fever and cough are the most common infection symptoms. Other signs and symptoms include chest pain, sputum production, and a sore throat. There were 10,689,527 reported cases of COVID-19 in India from January 3 to 4:35 p.m. CET on January 27, 2021, with 153,724 deaths. [3] Lockdown 2.0 was implemented from April 15th to May 3rd, 2020. During Lockdown 3.0 and 4.0, which lasted until May 30, 2020, the government divided all districts into three zones based on virus spread - green, red, and orange - with applied relaxations [4]. The number of people visiting hospitals declined over time as a result of the national lockout, which included quarantine, social distancing, hand washing, school closures, curfews, mask wearing, and other measures.The government declared that the lockdown would be lifted on May 30, 2020, with the exception of the containment areas, where the lockdown would be extended until June 30, 2020. Every month from June 1, 2020, to January 31, 2021, there were eight unlock steps imposed, each focusing on how restrictions could be relaxed in a staggered manner while keeping an eye on the pace of progress in reported cases. This model calculates the parameter values for India, and then extracts the trend and uncertainty from the available data to make a forecast. Predictive models can help to do this by predicting the number of COVID-19 cases in the future.

II. Materials

As listed in our References column, we gathered data from regular bulletins that are published online in various websites. The details we gathered included the number of regular positive cases, total cumulative cases, recovered cases, deaths, and active cases, among other things. We used the Kaggle [5] dataset for our research and review, which contains data derived from various official sources.

III. ANALYSIS AND FORECASTING

A. Covid Cases:

1. State wise Confirmed Cases:

Fig.1 shows the confirmed cases between 30th January, 2020 to 9th December, 2020 (a total of 315 days). Here we can see

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

that maximum confirmed cases are noted in Maharashtra followed by Karnataka, Andhra Pradesh, Tamilnadu then Kerala.

.

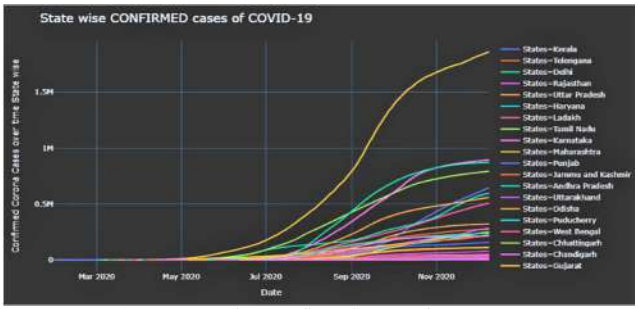


Fig. 1. State wise Confirmed cases of COVID-19

2. State wise Recovered cases:

Fig.2 shows the State wise Recovered cases. Here we can see that maximum Recovered cases are noted in Maharashtra followed by Andhra Pradesh, Karnataka, Tamilnadu then Kerala.

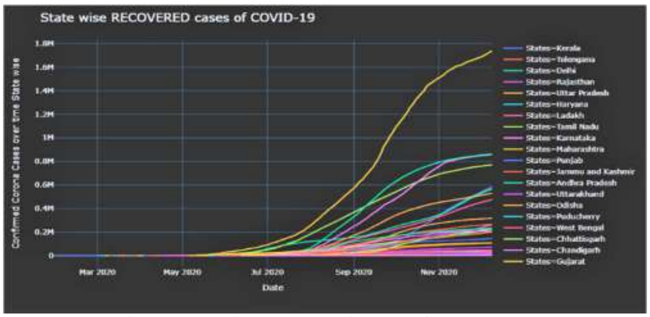


Fig.2. State wise Recovered cases of COVID-19

3. State wise Deceased cases:

Fig.3 shows the State wise Death cases. Here we can see that maximum Recovered cases are noted in Maharashtra followed by Karnataka, Tamilnadu then Delhi.

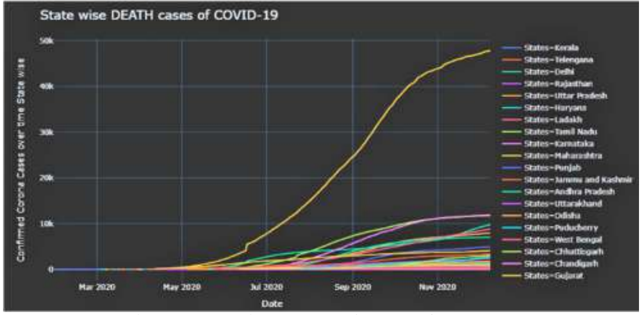


Fig. 3. State wise Deceased cases of COVID-19

4. Reported Cases in India over time

Fig.4 shows the details of cases related to COVID-19 in India. We can observe that the percentage of deceased cases are very less, compared to the total Confirmed cases were around 9.7 million and total recovered cases were around 9.2 million on the 9th of December 2020.

There were a total 141.3k deaths during the same time. The total active cases reached around 383.86k during first week of December 2020 and was maximum on 18th of September 2020 were around 1.01 million.

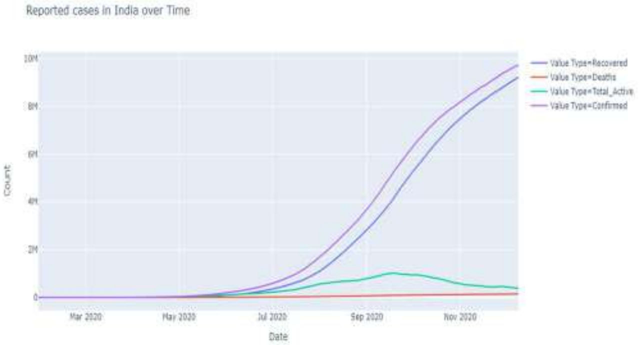


Fig. 4. Reported cases in India

5. State of Patients in India

Fig.5a Shows the percentage of Confirmed, Recovered and Deaths related to COVID-19. We can see that total Confirmed cases are around 51.2%, Recovered cases are around 47.9% and 0.88% Deaths. State wise data can be seen in Fig. 7, Fig. 8, Fig 9.

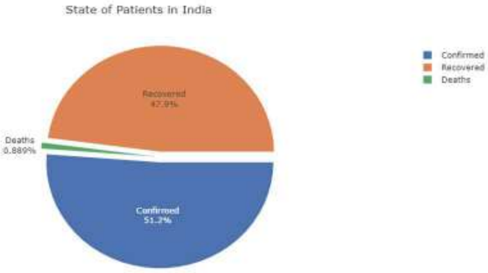


Fig. 5a State of Patients in India

6. COVID-19 CASES IN PERCENTAGES

FIG. 5B, FIG.5C, FIG.5D SHOWS STATES THAT HAD MAXIMUM IMPACT OF COVID-19.

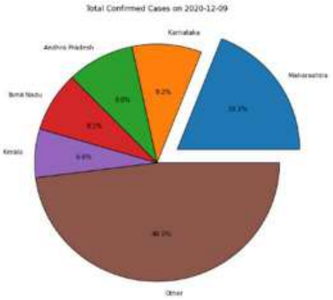


Fig. 5b. Percentage of Confirmed cases

From the above figure we can see that out of Total number of confirmed cases across India, Maharashtra has 19.1% followed by Karnataka with 9.2%, Andhra Pradesh with 9.0%, Tamil Nadu with 8.1%, Kerala with 6.6%.

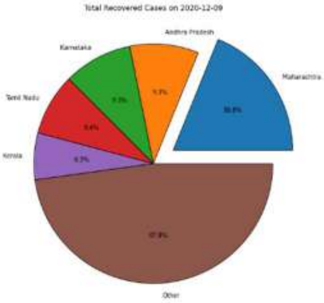


Fig. 5c. Percentage of Recovered cases

From the above figure we can see that out of Total number of Recovered cases across India, Maharashtra has 18.8% followed by Andhra Pradesh with 9.3%, Karnataka with 9.3%, Tamil Nadu with 8.4%, Kerala with 6.3%.

From fig.8 we can see that out of Total number of Deceased cases across India, Maharashtra has 33.8% followed by Karnataka with 8.4%, Tamil Nadu with 8.4%, Delhi with 6.9% and West Bengal with 6.2%.

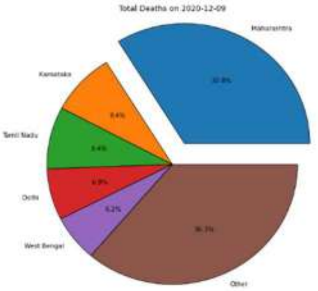


Fig. 5d. Percentage of Deaths

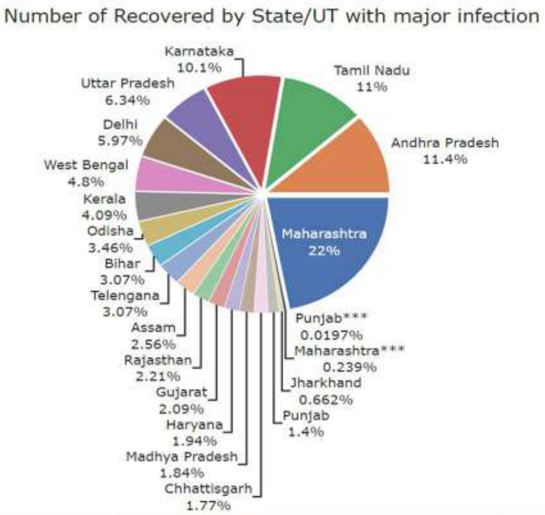


Fig. 6. Recovery cases Percentage for each state

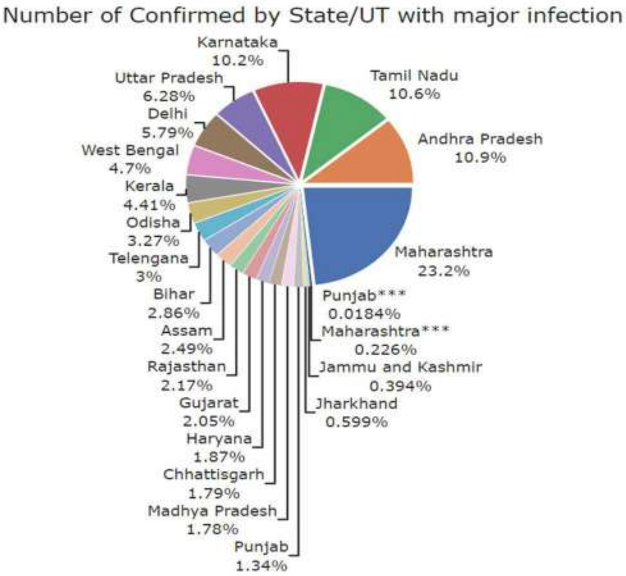


Fig. 7. Confirmed cases Percentage for each state

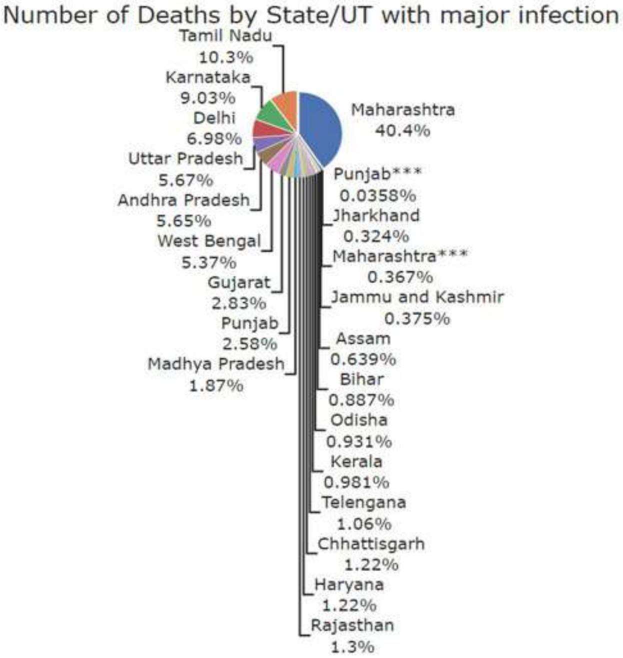


Fig. 8. Death cases Percentage for each state

7. Effect of Lockdown

Fig.9 shows During lockdown, Covid-19 has an effect. As seen in the graph below, there was an increase in reported cases during the Lockdown time, but the number was kept under control. There were 218 confirmed cases on March 25th, and 21.1k confirmed cases by May 31st, the end of the lockdown. However, we can see that once the lockdown in India was lifted, the number of confirmed cases increased exponentially. We reached the 1 million mark in just a few days, on September 12th.

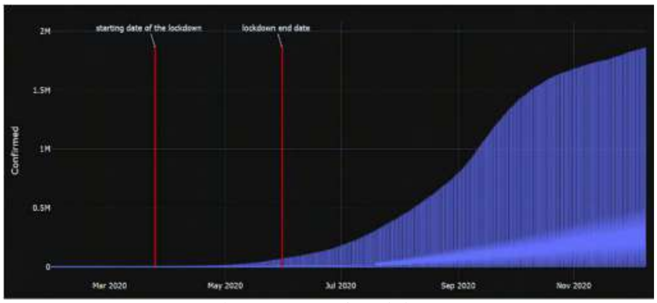
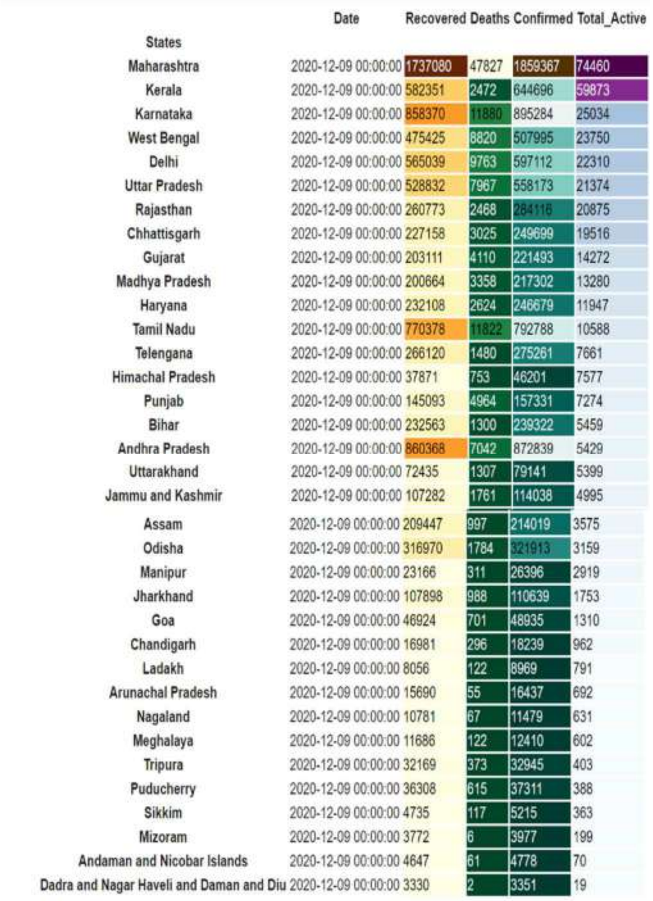


Fig.9. Impact of Lockdown

8. Total Active Cases

Fig.10 Shows state wise data of covid-19. We have calculated total active cases for each state on the latest data.

Total Active cases = Total Confirmed – (Total Deaths + Total Recovered)



States	Date	Recovered	Deaths	Confirmed	Total_Active
Maharashtra	2020-12-09 00:00:00	1737080	47827	1859367	74460
Kerala	2020-12-09 00:00:00	582351	2472	644696	59873
Karnataka	2020-12-09 00:00:00	858370	11880	895284	25034
West Bengal	2020-12-09 00:00:00	475425	8820	507995	23750
Delhi	2020-12-09 00:00:00	565039	9763	597112	22310
Uttar Pradesh	2020-12-09 00:00:00	528832	7967	558173	21374
Rajasthan	2020-12-09 00:00:00	260773	2468	284116	20875
Chhattisgarh	2020-12-09 00:00:00	227158	3025	249699	19516
Gujarat	2020-12-09 00:00:00	203111	4110	221493	14272
Madhya Pradesh	2020-12-09 00:00:00	200664	3358	217302	13280
Haryana	2020-12-09 00:00:00	232108	2624	246679	11947
Tamil Nadu	2020-12-09 00:00:00	770378	11822	792788	10588
Telengana	2020-12-09 00:00:00	266120	1480	275261	7661
Himachal Pradesh	2020-12-09 00:00:00	37871	753	46201	7577
Punjab	2020-12-09 00:00:00	145093	4964	157331	7274
Bihar	2020-12-09 00:00:00	232563	1300	239322	5459
Andhra Pradesh	2020-12-09 00:00:00	860368	7042	872839	5429
Uttarakhand	2020-12-09 00:00:00	72435	1307	79141	5399
Jammu and Kashmir	2020-12-09 00:00:00	107282	1761	114038	4995
Assam	2020-12-09 00:00:00	209447	997	214019	3575
Odisha	2020-12-09 00:00:00	316970	1784	321913	3159
Manipur	2020-12-09 00:00:00	23166	311	26396	2919
Jharkhand	2020-12-09 00:00:00	107898	988	110639	1753
Goa	2020-12-09 00:00:00	46924	701	48935	1310
Chandigarh	2020-12-09 00:00:00	16981	296	18239	962
Ladakh	2020-12-09 00:00:00	8056	122	8969	791
Arunachal Pradesh	2020-12-09 00:00:00	15690	55	16437	692
Nagaland	2020-12-09 00:00:00	10781	67	11479	631
Meghalaya	2020-12-09 00:00:00	11686	122	12410	602
Tripura	2020-12-09 00:00:00	32169	373	32945	403
Puducherry	2020-12-09 00:00:00	36308	615	37311	388
Sikkim	2020-12-09 00:00:00	4735	117	5215	363
Mizoram	2020-12-09 00:00:00	3772	6	3977	199
Andaman and Nicobar Islands	2020-12-09 00:00:00	4647	61	4778	70
Dadra and Nagar Haveli and Daman and Diu	2020-12-09 00:00:00	3330	2	3351	19

Fig. 10. Total Active Cases

Fig.11 helps us to depict the total number of Confirmed cases recorded in all states and union territories till 9th December 2020

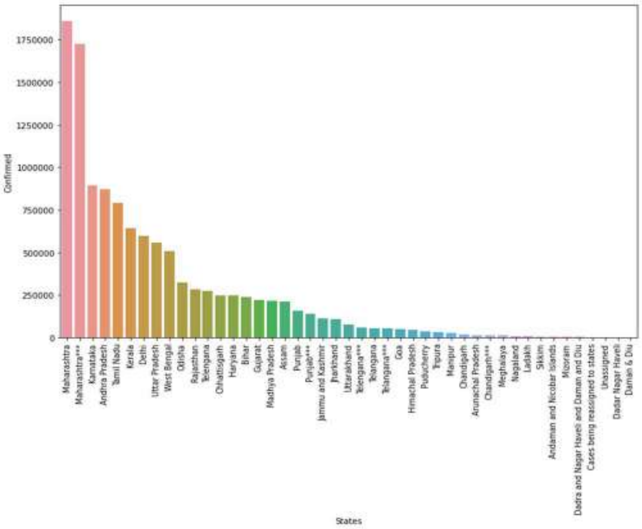


Fig.11. State wise Confirmed Cases

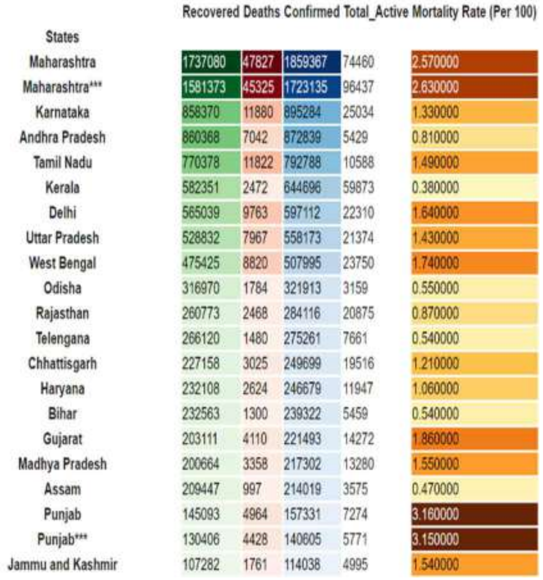
9. Mortality Rate

After its discovery, the COVID-19 virus has spread rapidly throughout the world, with cases registered in 210 countries (till 10:39 GMT on April 26, 2020). The data obtained showed over 2.4 million confirmed cases of CoViD-19.[13] Higher mortality rate (15%) was found in Algeria, Belgium (13.95), Italy and United Kingdom (13%) and Netherland (11.35%). Lower mortality rate was found in countries Qatar 0.17%, Singapore 0.2%, United Arab Emirate 0.6%, and Australia 0.97. The WHO updates and shares these statistics on a regular basis, and as of April 28th, it had issued ninety-seven reports giving country wise details of number of cases.

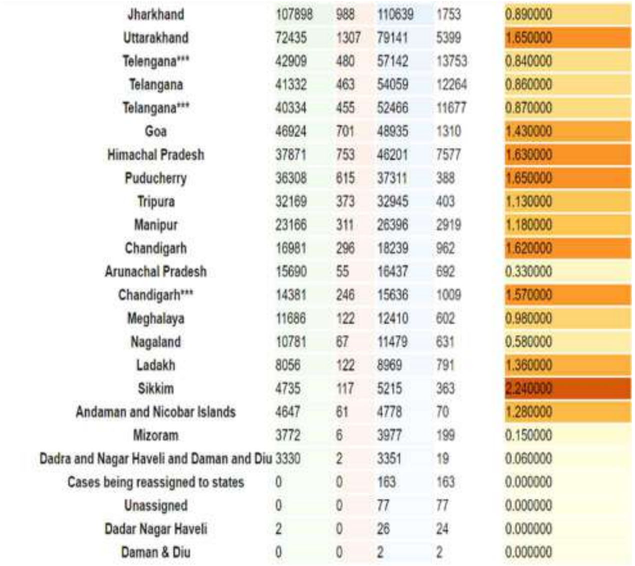
A strong positive correlation r=0.9, n=56 was observed between reported cases and deaths, indicating that disease spread raises the risk of death due to overcrowded hospitals, limited medical facilities, and other environmental factors. COViD-19 had already begun to spread before preventive steps were implemented. [14] Countries that responded quickly suffered less than countries that were unconcerned in the early stages of the pandemic. Another explanation for the pandemic was that since 80 percent of CoViD-19 cases are mild or asymptomatic, symptom-based disease management is difficult and ineffective.

The recovery versus deceased ratio was estimated, and the data revealed that in Singapore, Qatar, and Thailand, recovery was 68, 59, and 35 times higher than death.The ratio of deceased to recovered was found to be lower in the United Kingdom (0.03), the Netherlands (0.08), Ireland (0.16), and Norway (0.21). [13]In contrast to CoViD-19 prevalence, a previous study found that community acquired pneumonia cases were prevalent in males from lower socioeconomic groups who were illiterate and lived in rural areas.[14] Patients recoveries are significantly correlated with the number of cases (r = 0.63, n = 56), indicating that recoveries increase as the number of cases increases. The potential factors involved in recovery could be a strong immune system among the population, good dietary habits, and early treatment, as well as a Bacillus Calmette-Guerin (BCG) vaccination policy in some countries, which showed fewer cases than non-BCG vaccinated nations. [16]

To measure the Mortality Rate, divide the number of people who have died by the total number of active cases in the population, then multiply by 100.
In the fig.12 we can see the calculated Mortality rate on cases related to COVID-19 in India.



States	Recovered	Deaths	Confirmed	Total_Active	Mortality Rate (Per 100)
Maharashtra	1737080	47827	1859367	74460	2.570000
Maharashtra***	1581373	45325	1723135	96437	2.630000
Karnataka	858370	11880	895284	25034	1.330000
Andhra Pradesh	860368	7042	872839	5429	0.810000
Tamil Nadu	770378	11822	792788	10588	1.490000
Kerala	582351	2472	644696	59873	0.380000
Delhi	565039	9763	597112	22310	1.640000
Uttar Pradesh	528832	7967	558173	21374	1.430000
West Bengal	475425	8820	507995	23750	1.740000
Odisha	316970	1784	321913	3159	0.550000
Rajasthan	260773	2468	284116	20875	0.870000
Telengana	266120	1480	275261	7661	0.540000
Chhattisgarh	227158	3025	249699	19516	1.210000
Haryana	232108	2624	246679	11947	1.060000
Bihar	232563	1300	239322	5459	0.540000
Gujarat	203111	4110	221493	14272	1.860000
Madhya Pradesh	200664	3358	217302	13280	1.550000
Assam	209447	997	214019	3575	0.470000
Punjab	145093	4964	157331	7274	3.160000
Punjab***	130406	4428	140605	5771	3.150000
Jammu and Kashmir	107282	1761	114038	4995	1.540000



Jharkhand	107898	988	110639	1753	0.890000
Uttarakhand	72435	1307	79141	5399	1.650000
Telengana***	42909	480	57142	13753	0.840000
Telangana	41332	463	54059	12264	0.860000
Telangana***	40334	455	52466	11677	0.870000
Goa	46924	701	48935	1310	1.430000
Himachal Pradesh	37871	753	46201	7577	1.630000
Puducherry	36308	615	37311	388	1.650000
Tripura	32169	373	32945	403	1.130000
Manipur	23166	311	26396	2919	1.180000
Chandigarh	16981	296	18239	962	1.620000
Arunachal Pradesh	15690	55	16437	692	0.330000
Chandigarh***	14381	246	15636	1009	1.570000
Meghalaya	11686	122	12410	602	0.980000
Nagaland	10781	67	11479	631	0.580000
Ladakh	8056	122	8969	791	1.360000
Sikkim	4735	117	5215	363	2.240000
Andaman and Nicobar Islands	4647	61	4778	70	1.280000
Mizoram	3772	6	3977	199	0.150000
Dadra and Nagar Haveli and Daman and Diu	3330	2	3351	19	0.060000
Cases being reassigned to states	0	0	163	163	0.000000
Unassigned	0	0	77	77	0.000000
Dadar Nagar Haveli	2	0	26	24	0.000000
Daman & Diu	0	0	2	2	0.000000

Fig.12. Mortality rate

IV. METHODOLOGY AND RESULTS

From the above analysis, we can conclude that if we want to know the impact of COVID-19 in India, it is essential for us to understand Mortality rate. Mortality rate is an important parameter for the fight against the Global Pandemic.

We devised the following method for estimating the number of positive cases, recovered cases, and active cases in order to estimate and forecast the recovery rate.
1) Examine the data and determine the pattern.
2) We'll try to explain the Susceptible, Recovered, Infected Individuals phenomenon using the SIR model.
3) Using Prophet to fills the existing gaps in generating fast reliable forecasts.

In our case, we used the SIR model to figure out the pattern of positive cases and the number of recovered cases. We also experimented with Time Series analysis and prediction model Facebook Prophet.

a. SIR modelling

The standard SIR model divides the entire population into three compartments, namely,

- *Susceptible individuals,* S: These are those individuals who are not infected but may become infected in the future. A person who is susceptible to infection may become infected or remain susceptible. If the virus spreads from its origins or new sources emerge, more people will become infected, increasing the susceptible population for a time (surge period).
- *Infected individuals,* I: These are those individuals who have already been infected with the virus and can now pass it on to those who are susceptible. A person who has been infected will stay infected and be removed from the infected population to recover or die.
- *Removed/Recovered individuals* R: These are those individuals who have recovered from the virus and are thought to be resistant, or who have died as a result of the virus.

The disease dynamics are then solved and the model is propagated using a series of ordinary differential equations. The equations for the corresponding equations are as follows

$$dS/dt = -\beta SI \tag{1}$$

$$dI/dt = \beta SI - \gamma I \tag{2}$$

$$dR/dt = \gamma I \tag{3}$$

N is the total number of people in the population.

The rate of change of infected population (dIdt) (dIdt) depends on two primary factors:

1. The number of people falling ill, and

2. The number of people recovering.

The number of people who become ill is determined by the degree of contact between the infected and susceptible populations, and is related by the constant β, which stands for the transmission/infection rate and βSI represents the number of susceptible individuals that become infected per day. γ is the rate of recovery. γI is the number of infected individuals that recover per day; 1/ γ is the infectious period i.e., the average duration of time an individual remains infected. the average amount of time that a person is infected. The reproductive number R0, which represents the average number of secondary infections produced by one infectious person in a fully susceptible population, is an important quantity in any disease model. The total (constant) population size is denoted by the letter N. Regarding the SIR model,

$$N = S + I + R \tag{4}$$

$$R0 = \beta N / \gamma \tag{5}$$

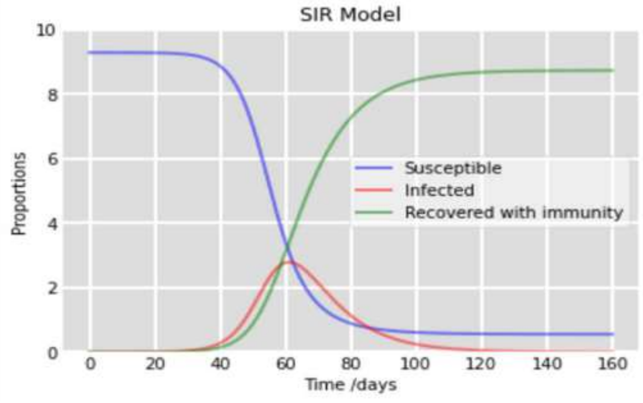


Fig.13. SIR modelling

We used trial and error and visual analysis to try to match the data to the SIR model. The parameters: initial susceptible population, infection rate (beta), and recovery rate (gamma) are manually calibrated through trial and error to match the data as closely as possible, resulting in the graph shown below in. Fig.12. The SIR model is used to simulate the susceptible, infection and removed cases from 9th December, 2020 for next 160 days until 18th may, 2021. From Fig.12 we can observe that the infection cases will reduce to zero by June- July, 2021, if there are no second wave of infections.

b. Facebook Prophet - Time series analysis:

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data [7].

Advantages of prophet are:

-Prophet makes it much more straightforward to create a reasonable, accurate forecast.

-Prophet forecasts are customizable in ways that are intuitive to non-experts.

Prophet is open-source software released by Facebook’s Core Data Science team. It is available for download on CRAN and PyPI. We use Prophet, a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

By building a base model both with and without changing the seasonality-related parameters and additional regressors, we can create a week-ahead forecast of verified COVID-19 cases using Prophet, with specific prediction intervals.

Predicting daily Death cases using FB-Prophet (Base Model):

FB-prophet prediction of death/ Morality cases and their trend.

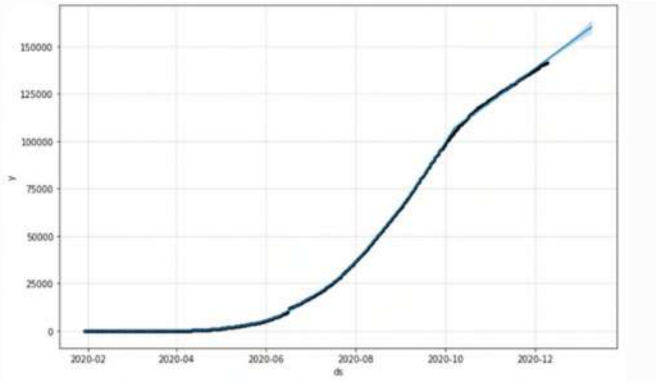


Fig.14. Prediction plot

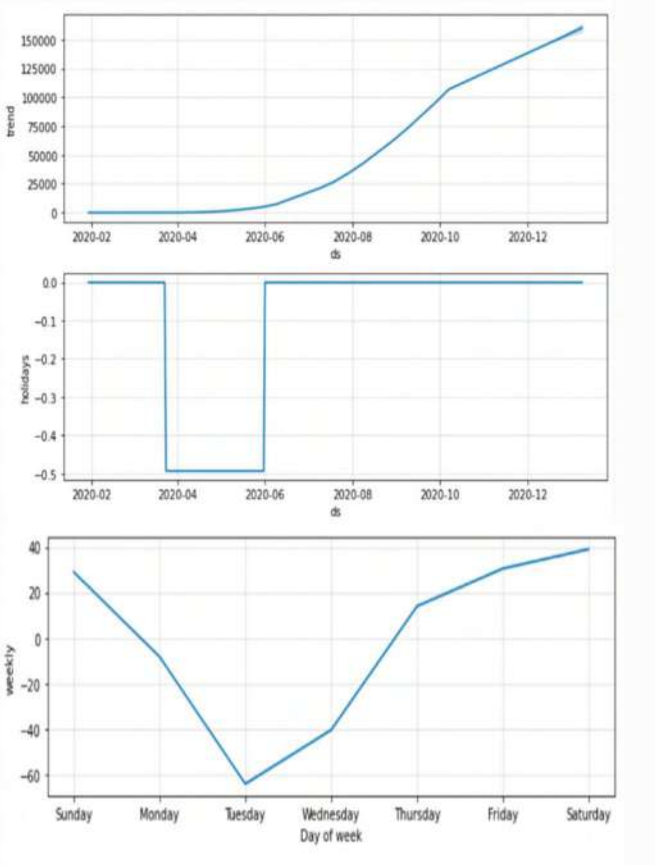


Fig.15. Plot for Predicted value /week/trend

We cross validated data from 23th April, 2020 to 23th may, 2020. Plot in Fig.16 shows the RMSE curve for all the predicted values

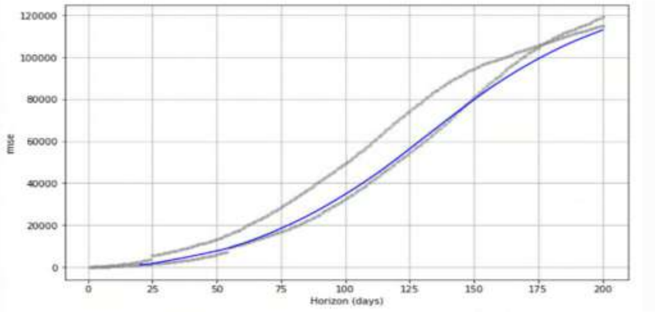


Fig.16a. RMSE plot for FB-Prophet(death)

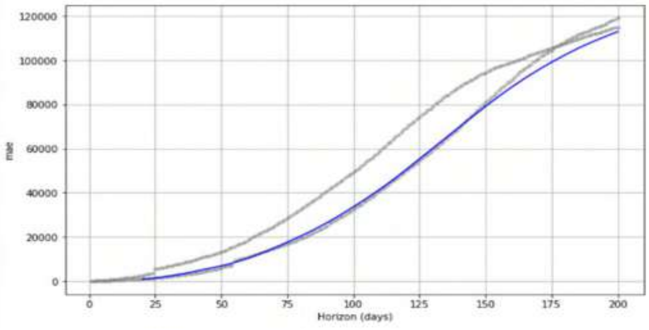


Fig.16b. MAE plot for FB-Prophet(death)

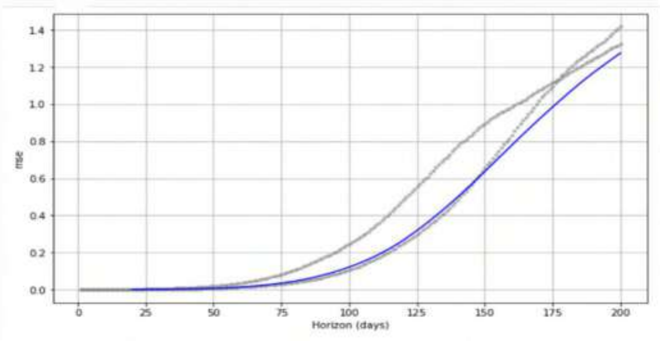


Fig.16c. MSE plot for FB-Prophet(death)

V. DISCUSSION AND CONCLUSION

With data analytics and data mining, information and communication technology aid in decision-making based on historical data. The amount of data available is massive, and extracting information and creating an interesting pattern from the accumulated data is a difficult task. With the current data on confirmed, recovered, and death across India for a long period of time, it is possible to predict and forecast the near future.

The country's epidemic was primarily caused by the movement of people from various foreign countries to India. With the current data on confirmed, recovered, and death across India for a long period of time, it is possible to predict and forecast the near future.

The country's epidemic was primarily caused by the movement of people from various foreign countries to India. The rate of recovery increased while the rate of transmission decreased. As a result, the value of the basic reproduction number decreased, flattening the COVID-19 epidemic spread curve.

Similarly, the percentage of people who remained susceptible after the infection had passed has decreased, and this should be checked in the near future. Furthermore, the peak percentage value of infectious people decreased and showed better results.

In the case of India, additional strict governmental interventions should be carried out, and, of course, pandemic cases can be drastically reduced by raising awareness among the local population. SIR modelling is more intuitive and explainable, but it requires a great deal of trial and error as well as assumptions.

The FB-Prophet prediction process is simple that is why FB-Prophet predictions can be used for quick and accurate prediction of daily COVID-19 positive and active cases. Investigations have been made on the issue of COVID-19 pandemic spread in India in the current challenging scenario. Trend knowledge has been observed with the help of recovery rate and case load rate obtained for the data available.

The various strategies implemented as lockdown; quarantine of population have played a significant role in reducing the risk of spread of epidemic.

This study predicts that when the case load rate gets lesser than recovery rate, there after COVID-19 patients would start to decline.

VI. REFERENCES

[1] "Coronavirus Disease (COVID-19) Situation Reports," *Who.int*. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. [Accessed: 15-Mar-2021].

[2] E. 402 562., "Coronavirus Disease 2019 (COVID-19)," *Ecdhd.ne.gov*. [Online]. Available: https://ecdhd.ne.gov/file_download/inline/e549726d-52ce-4e0e-9fc7-fe237fd971b0. [Accessed: 15-Mar-2021].

[3] "India: WHO Coronavirus disease (COVID-19) dashboard," *Who.int*. [Online]. Available: https://covid19.who.int/region/searo/country/in. [Accessed: 15-Mar-2021].

[4] SRK, "COVID-19 in India." .

[5] "COVID-19 Confirmed Cases india," *Uri.sh*. [Online]. Available: https://flo.uri.sh/visualisation/1977187/embed. [Accessed: 15-Mar-2021].

[6] "fbprophet," *Pypi.org*. [Online]. Available: https://pypi.org/project/fbprophet/. [Accessed: 15-Mar-2021].

[7] "The SIR epidemic model," *Scipython.com*. [Online]. Available: https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/. [Accessed: 15-Mar-2021].

[8] K. Ghosh, N. Sengupta, D. Manna, and S. K. De, "Inter-state transmission potential and vulnerability of COVID-19 in India," *Progress in Disaster Science*, vol. 7, no. 100114, p. 100114, 2020.

[9] R. Gupta and S. K. Pal, "Trend Analysis and Forecasting of COVID-19 outbreak in India," *bioRxiv*, 2020.

[10] "Media Bulletins," *Gov.in*. [Online]. Available: https://covid19.telangana.gov.in/announcements/media-bulletins/. [Accessed: 15-Mar-2021].

[11] "Coronavirus update (live): 120,476,225 cases and 2,666,439 deaths from COVID-19 virus pandemic - worldometer," *Worldometers.info*. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 15-Mar-2021].

[12] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?," *Lancet*, vol. 395, no. 10228, pp. 931–934, 2020.

[13] *Academicjournals.org*. [Online]. Available: https://academicjournals.org/journal/IJMMS/article-abstract/8D3E82155907. [Accessed: 15-Mar-2021].

[14] A. Miller, M. J. Reandelar, K. Fasciglione, V. Roumenova, Y. Li, and G. H. Otazu, "Correlation between universal BCG vaccination policy and reduced mortality for COVID-19," *bioRxiv*, p. 2020.03.24.20042937, 2020.

[15] The Lancet, "India under COVID-19 lockdown," *Lancet*, vol. 395, no. 10233, p. 1315, 2020.

[16] N. Darapaneni, D. Reddy, A. R. Paduri, P. Acharya, and H. S. Nithin, "Forecasting of COVID-19 in India using ARIMA model," in 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0894–0899.

NARROWBAND HAIRPIN BANDPASS FILTER FOR 4G LTE APPLICATIONS

Vinod Babu Pusuluri*
Assistant Professor, Dept. of ECE
RGUKT APIIIT Nuzvid
Nuzvid, India
vinod@rguktn.ac.in*

Varun Mannam
mannamvarun@gmail.com

V A Sankar Ponnapalli
sankar.p@sreyas.ac.in

A Mallikarjuna Prasad
Professor,Dept. of ECE
JNTU, Kakinada, India
a_malli65@jntucek.ac.in

***Abstract*— Narrowband bandpass Filter plays an important role in modern wireless communication systems specially in the sub 6GHz space. In many cases, transmitted and received signals must be filtered at a certain center frequency with specific bandwidth limitations asper specifications. This research article represents the principle of microwave filter design, fabrication, and performance evaluation of bandpass filter design using the fundamental hairpin resonator structure. It is proposed to design a fifth-order Chebyshev bandpass filter with center frequency at 2.01 GHz using two different FR4 and Arlon 25N substrates separately. The simulated designs are implemented using the full-wave electromagnetic solver Genesys and fabricated on both substrate materials. We observed the accurate and reasonable filter parameters like return loss and insertion loss S11, S21 in both simulation and measurements. The design specifications bandwidth of400MHz, (-30dB, -1.4dB) and (-20dB, -2.5dB) of return, insertion losses were obtained respectively on Arlon 25N, FR-4 materials respectively. The designed filter has multiple potential applications for 4G LTE bands.**

Keywords*— *Bandpass filter, Microwave Filter, Genesys Simulator, return loss, insertion loss, Hair pin resonator.

I. Introduction

Microwave filters are unavoidable components in various electronic systems at various frequencies across the electromagnetic spectrum, like mobile radio, satellite communication transceivers, and Radar. Due to advancements in the technological world of electronics, mainly in Communications, fully integrated filters for high-frequency applications are now receiving great interest worldwide [1]. The usage of microstrip devices in the design of microwave systems and integrated circuits have gained huge popularity for the last two decade due to its compactness, cost-effective and ease of design, and further microstrip can operate in a wide range of frequencies.

Many researchers have presented numerous equations for the analysis and synthesis of microstrip devices, but with the advent of various full-wave electromagnetic simulators like high-frequency structural simulator (HFSS), Zeland IE3D, CST Microwave Office, the design of microwave structures has become simple and challenging. So far, scientific society has resulted in many advanced designs in microwave filters and antennas using microstrip structures [2]-[4].[12]. The purpose of this article is to design and analysis the simulated and fabricated bandpass filters as per specifications needed for various applications like LTE, GSM, S-band mobile transceivers, etc., with minimum return loss and attenuation of 15 dB and 3 dB respectively. Here the design of the fifth-order Chebyshev bandpass filter prototype microwave filters is carried by using the Hair Pin Bend microstrip technique. The synthesis is more convenient and smoothly done by the hairpin bends. The circuit performance is simulated and optimized using the full-wave electromagnetic simulator Genesys developed by Agilent technologies for the intended circuit fabrication. The designed filters were fabricated on the two different types of substrate material, FR4 and Arlon 25N material, with a dielectric constant of 4.4 and 3.38, respectively which are readily available in the market and a high tolerance for electromagnetic noises.

The performance of filters by hairpin bend structures on both substrates is compared in term of scattering parameters Sll and S22, the Arlon 25N microstrip filter design results in low loss condition compared with FR4, and it shows that Arlon 25N has produced a better performance for the microstrip filter design. The filter design presented in this paper can be extended to the filter bank design in the future. The following chapter presents the design and analysis procedure for the proposed bandpass filters. Here, the individual filter's design is done with the futuristic aim to filter banks such that a single device is applicable for multiple advanced applications [6]-[7].

II. Filter Design

There are various procedures in designing the microwave filters, like the image parameter method and insertion loss method. The insertion loss method with coupled line microstrip, stripline filter design is easy for the design and fabrication; further, the parallel-coupled lines give a bandpass response. The selected parallel-coupled transmission-line filter in microstrip, stripline technologies provides bandpass and bandstop filters with required bandwidth up to 20% of centre

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

frequency [8]. Due to their relatively weak coupling, this type of filter has narrow fractional bandwidth. Instead, it has desired advantages such as low-cost fabrication, easy integration, and simple designing procedures. Design equations for the coupled line parameters such as space-gap between lines and line widths and lengths are readily available in the literature. The model prototype of the width and lengths of the parallel-coupled line is demonstrated in figure 1. We could calculate the line's length and widths to design the filter from the existing literature [9]-[10]. The characteristic impedance based on the calculated width and height for the given material is provided in Equation.1.

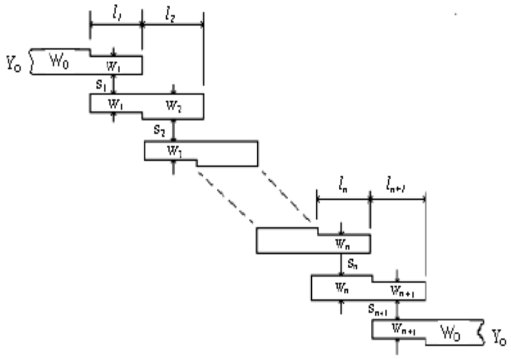


Figure 1: Parallel coupled line filter with parameters.

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\epsilon}} \ln\left(\dfrac{8d}{W} + \dfrac{W}{4d}\right) & \dfrac{W}{d} < 1 \\ \dfrac{120*\pi}{\sqrt{\epsilon}\,(\frac{W}{d}+1.393+0.667\ln(\frac{W}{d}+1.414))} & \dfrac{W}{d} \geq 1 \end{cases} \qquad (1)$$

III. HAIRPIN STRUCTURE

Among various techniques available for bandpass microstrip filters in parallel-coupled lines, the Hairpin structural filter design is preferred. The concept of the hairpin filter is the same as parallel-coupled half-wavelength resonator filters [11]. The advantage of hairpin filter over end-coupled and parallel-coupled microstrip realizations is the optimal space utilization. This space utilization is achieved by folding the structure as the half-wavelength long resonators. The design of the simple half-wavelength resonator is shown in figure .2, and the corresponding bandpass filter layout design using multiple hair-pin resonator structures is in figure.3, where α gives the values of the slid angle, and the small slid angle leads to greater coupling between the resonators in figure .3.

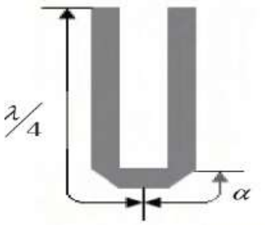


Figure 2: Illustration of a single Hairpin resonator structure.

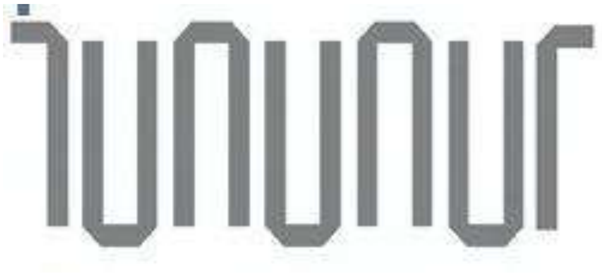


Figure 3: Hairpin filter layout.

IV. DESIGN SIMULATION

Many full-wave and half-wave Electromagnetic simulators are available in the market like Zeland IE3D, High-Frequency Structural Simulator (HFSS), CST Microwave office, etc. Here in this research article, the design of the bandpass filter is done in the Genesys full-wave simulator from the Agilent technologies. The Genesys simulator enables to design of circuits composed of schematics and electromagnetic (EM) structures from an extensive electrical model database and then generates layout presentations for the designs. The design's output will be in a wide variety of graphical forms based on the analysis needed. The advanced tunable tool in this simulator enables us to tune or optimize the designs, and the changes reflect immediately in the layout according to the given filter specifications. The designers can observe the effects of the test signals before investigating in hardware prototypes. The Genesys simulation design procedure consists of the following tasks in the design environment:

- Creating projects to organize and save the designs.
- Creating system diagrams, circuit schematics, and EM structures.
- Placing circuit elements into schematics.
- Placing system blocks into system diagrams.
- Creating and displaying output graphs.
- Running simulations for schematics and system diagrams.
- Tuning simulations to optimize the design in terms of the space.

- Creating layouts.

The final design in the simulation layout for the proposed filters is shown in figure.4, here extreme ends in both filters are input and output ports and it was designed for port characteristic impedance of 50 Ω which is common standard value in any electromagnetic designs. There are **five half wave resonators** in each design which indicates the fifth-order filter. The fifth-order Chebyshev proposed filter gives more precise cutoffs frequencies (sharp transitions between passband and stopband frequencies) than the popular and simple Butterworth filters.

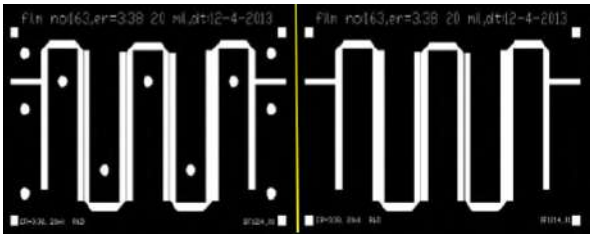

Fln no163,er=3.38 20 mil,dt:12-4-2013
Fln no163,er=3.38 20 mil,dt:12-4-2013

Figure 4: Simulator filter layout for two-different fifth-order bandpass filters.

V. FILTER FABRICATION

In modern technology there are advanced semiconductors machines for the fabrication of many electronic devices. This section gives detail explanation about the processes involved in the fabrication of microstrip circuits using printed circuit board technologies. The following are the basic steps involved in actual fabrication process of microstrip line bandpass filter fabrication.

- Preparing the mask
- Creating a photo resist pattern
- Etching the unwanted copper material
- Housing for the microstrip
- Soldering the edges for the input and output terminals.

The fabrication is done in this research article on both substrates Arlon 25N and FR-4 substrate materials. The substrate characteristics were mentioned in the Table.1. The width and lengths were calculated from the literature in such a way that it matches for the 50 Ω characteristic impedance of the input and output ports of the 2-port network and the corresponding values are represented in the Table 2. The proposed filters were designed assuming the center frequency of the bandpass filter is at 2010MHz which is the 4G LTE band (example: TDD, band 34, FDD, band 1 etc.).

Material Parameter	**FR4**	**Arlon 25N**
Dielectric Constant	4.60	3.38
Loss Tangent	0.011	0.0022
Metal Thickness	1.42 mil	1.6 mil
Substrate Height	59 mil	20 mil

Table.1: Substrate material characteristics.

S.No	**Parameter**	**Arlon 25N**	**FR-4**
1	Length (mm)	23.1	20.5
2	Width (mm)	1	2.56

Table.2: Hairpin structure length and width values of Arlon 25N & FR-4 substrates.

The sample design of the filter after the due fabrication is as shown in figure.5 & figure.6 with Arlon 25N and FR-4 substrate materials, respectively. Here the input and output ports are matched to the standard characteristic impedance of 50Ω. The real-time fabricated filter parameters are measured and obtained by using the 3.6 GHz R&S Network Analyzer, and the analysis of the simulated and measured parameters are explained in the next section.

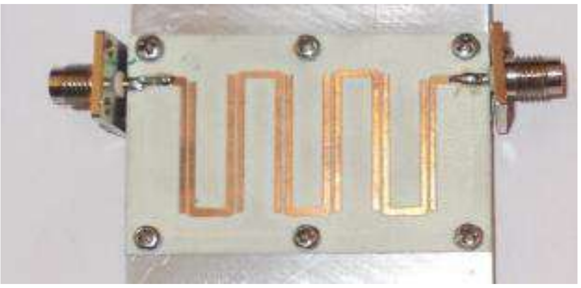


Figure 5: Fabricated fifth-order bandpass filter with Arlon N25 substrate.

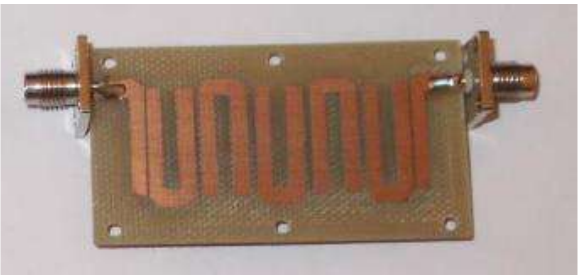


Figure 6: Fabricated fifth-order bandpass filter with FR-4 substrate.

VI. RESULT ANALYSIS

The fabricated filters were tested using the R&S Network Analyzer of 3.6 GHz capacity. The Insertion loss (S21) and Return losses (S11) of each fabricated filter has been observed and recorded in the Table 3. The filter parameters are observed that these filter characteristics are of required standards like low insertion loss in passband, high attenuation in stopband whereas it has high return loss in passband. The simulated filter response for the Arlon 25N and FR-4 substrates are mentioned in figure.7 and figure.8, respectively. The measured filter responses are mentioned in the figure.9 and figure.10, respectively. Clearly, the results from simulations and measurements indicate a low insertion loss in pass band. Losses in case of bandpass filters with Arlon 25N substrate material and FR4 substrate material are compared in Table.3.

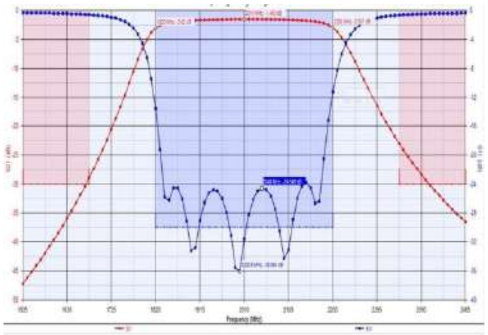


Figure 7: Simulated bandpass filter Response with Arlon 25N substrate.

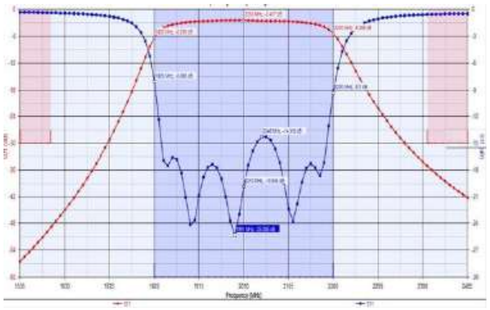


Figure 8: Simulated bandpass filter Response with FR-4 substrate.

S. No	Parameter	FR-4 Filter	Arlon 25N Filter
1	S11 (dB)	-14 to -25	-24 to -36
2	S21 (dB)	-2.5	-1.4

Table 3: Fabricated filter measurements (return loss and insertion loss) for each substrate.

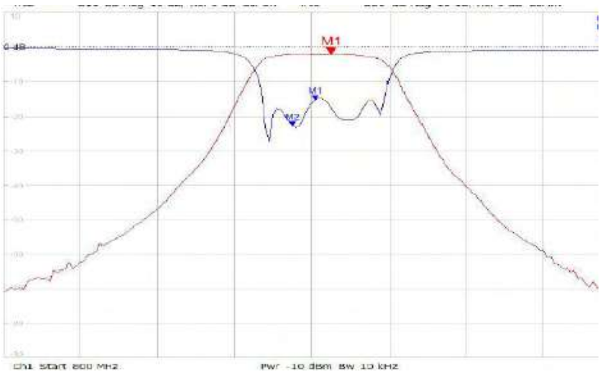


Figure 9: Measured bandpass filter characteristics with the Arlon 25N substrate.

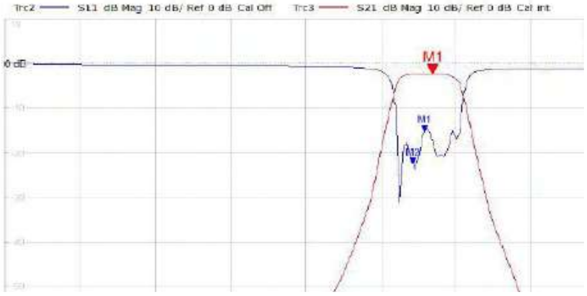


Figure 10: Measured bandpass filter characteristics with the FR-4 substrate.

VII. CONCLUSIONS

The successful simulated and fabricated bandpass filters with Arlon 25N and FR4 substrates are comparable and are found to be close enough (98%) with the provided specifications, further they exhibit the nearly 90% elimination of unwanted signals in the respective bands. Use of hairpin structure results to obtain the desired characteristics for a filter with line coupling and reduced size. Finally, Arlon 25N microstrip filter design shows low insertion loss and high return loss in the passband compared with FR4 filter design for the proposed specifications interims of insertion loss and return loss. The designed bandpass filter has the potential applications in 4G LTE band application as filters specially for the central frequency of 2.01GHz like TDD band 34 and FDD band 1. The design of multiple filters with different center frequencies and specifications (like bandwidth, insertion loss, return loss in the passband and attenuation in the stopband) can be implemented on a single substrate either on Arlon 25N or on FR-4 substrate such that it can be a filter bank which helps the multiband radio wave applications in future.

References

[1] Mannam, A. V. V., & Veeranki, B. Y. R. (2016). Design of narrowband bandpass filter using open-loop square resonators with loading element. Indian Journal Science Technology, 9(47), 1-9.

[2] Ponnapalli, V. S., & Babu, V. (2020, September). A Study on Scilab Free and Open Source Programming for Antenna Array Design. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-3). IEEE.

[3] Hong, J. S., & Lancaster, M. J. (1995). Canonical microstrip filter using square open-loop resonators. *Electronics Letters*, *31*(23), 2020-2022.

[4] Vasa, S. K., Vinod, P., Singuluri, P., & Babu, D. R. (2016, March). Design of cylindrical annular dielectric resonator antenna (CDRA). In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 2092-2094). IEEE.

[5] Harshasri, K., Babu, P. V., & Rao, P. N. (2018, April). Design of Compact C-Band Concave Patch Antenna for Radar Altimeter Applications. In 2018 International Conference on Communication and Signal Processing (ICCSP) (pp. 0542-0546). IEEE.

[6] LEVY, R. (1981). Filters with single transmission zeros at real or imaginary frequencies. *IEEE Trans. Microwave Theory and Tech.*, *29*(3), 215-222.

[7] Hong, J. S., & Lancaster, M. J. (1996). Couplings of microstrip square open-loop resonators for cross-coupled planar microwave filters. IEEE Transactions on Microwave theory and Techniques, 44(11), 2099-2109.

[8] Hong, J. S., & Lancaster, M. J. (1997). Theory and experiment of novel microstrip slow-wave open-loop resonator filters. IEEE Transactions on Microwave theory and Techniques, 45(12), 2358-2365.

[9] Wang, N. F., Tang, I. T., Hung, C. I., & Houng, M. P. (2004). Investigation of novel microwave surface-acoustic-wave filter on different piezoelectric substrates. *Japanese journal of applied physics*, *43*(12R), 8139.

[10] Brady, D. (2002). The design, fabrication and measurement of microstrip filter and coupler circuits. *High Freq. Electronics*, *1*(1), 22-30.

[11] Pozar, D. M. (2011). *Microwave engineering*. John wiley & sons.

[12] Vye, David, John Dunn, Dan Swanson, Jim Assurian, Ray Hashemi, and Philip Jobson. "Designing a Narrowband 28 GHz Bandpass Filter for 5G Applications." *Microwave Journal* 62, no. 4 (2019).

Designing a Teaching Aid for Microprocessor Class: Case Study Microprocessor Interconnection with Memory

1st Gunna Cahya Wardiyani
School of Electrical Engineering
Telkom University
Bandung, Indonesia
gunnacahyawardiyani@student.telkomuniversity.ac.id

2nd Nyoman Karna
School of Electrical Engineering
Telkom University
Bandung, Indonesia
aditya@telkomuniversity.ac.id

3rd Istikmal
School of Electrical Engineering
Telkom University
Bandung, Indonesia
istikmal@telkomuniversity.ac.id

***Abstract*— The learning process requires specific methods to achieve effective and efficient learning goals. Learning methodology is a way of carrying out activities between educators and students when interacting in the learning process. One of the learning methodology methods is the demonstration method, which using objects or other teaching materials at the time of teaching. Students expected can understand the correlation between microprocessors and memory practically is one of the main goals in the microprocessor lectures. This research aims to help the learning process in the microprocessors lecture by designing and implementing the teaching aid of the "Microprocessor Interconnection with Memory" module. This aid demonstrates an input and processes it into output in the form of lit LED using the 80C88 Microprocessor. The test results of this aid obtain an average of 75.34% by using the Mean Opinion Score (MOS) method.**

Keywords—* *Teaching aid, microprocessor, memory read, memory write, qualitative analysis.

I. Introduction

The media of learning can be in the form of trainers or teaching aids. A trainer is a learning medium that forms kit equipment to simulate material in teaching and learning activities so students can understand [1]. Conventional ways to teach microprocessor and microcontroller are commonly based on simulators. In practical such approach will not always effective due to the complexity of the microprocessor components [2]. One of the courses in the Telecommunication Engineering Program at Telkom University is the Microprocessor course. This course talks about microprocessors, microcontrollers, and microcomputers.

The microprocessor is a computer electronic central processing unit (CPU) made from mini transistors and other circuits on top of a semiconductor integrated circuit. The course typically comprises of lectures and laboratory work. The laboratory gives students the opportunity to use software for the programming part of the course and linking the output to the hardware. It is essential that the students know how to operate the functionality of both part [3]. In this research, the authors made a learning media in the form of trainers or teaching aids using an Intel 80C88 microprocessor, Dual In-line Package (DIP) Switch, and several Light Emitting Diode (LED) that help students to understand the material about microprocessor interconnection with memory.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. Basic Theory

A. Numeral System

The numeral system is a method that represents a quantity with a symbol. In the digital world, there are only two voltage levels that indicate a value. The two voltages are 0 Volt and +5 Volt. Therefore, the base-2 numeral system or binary numeral system can express those two voltage levels, which use only two symbols. These two symbols are 0 represents a 0 Volt and 1 represents +5 Volt [4].

B. 80C88 Microprocessor

A microprocessor is an integrated circuit or IC in the form of a Very Large-Scale Integration (VLSI) chip component that can execute commands sequentially in the form of a program and carry out counting, reasoning, and digital control operations. An 80C88 Microprocessor has 8 bits of the data path and 20 bits of address path [5].

The microprocessor course in Telkom University still uses 8088 as the main reference for CISC (Complex Instruction Set Computer) with comparison to DLX processor for RISC architecture [6] and this teaching aid is focused on 8088 microprocessors.

Fig. 1 is the pin configuration of the 80C88 microprocessor [5].

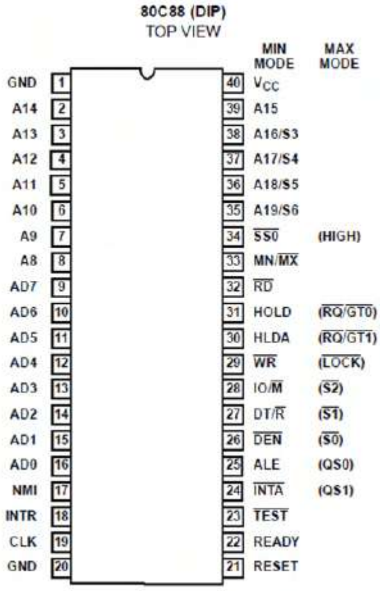


Fig. 1. Pinout of microprocessor 80C88 [5]

The Work Process of The 80C88 Microprocessor

While carrying out an instruction, the 80C88 microprocessor performs 3 stages of processing as follows [7]:

1. Instruction Fetch (IF)

IR ← [CS+IP]. The process begins with the pickup of new instruction from memory to IR.

- Control Unit (CU) translates the register contents of CS and IP, then determine the location of a new instruction in the memory.
- CU sends the translated CS and IP contents to memory via the address bus.
- CU sends a Mem-READ signal to indicate the memory that the CU wants to read the data.
- In response to the Mem-READ signal, the memory returns the data stored at the cell memory specified by the CU on the data bus (1 byte wide)
- CU reads the contents of the data bus and places it on the IR.

2. Instruction Decode/ID.

CU translates the new content from the IR to find out the instructions wanted. It uses the help of the instruction table in the Instruction Decoder.

3. Execution/EX

CU will carry out one of three functions depending on the results of the translation of the instructions, including:

- Arithmetic or Logical Operations
- Data transfer
- Control

C. Memory

Memory is a tool or medium that is used as a storage medium for data information or instructions and can be issued again. There are two types of memory, including [8]:

1. Volatile, is a memory that stores information as long as there is a power supply source (uninterrupted). Example: Random Access Memory (RAM).
2. Non-Volatile, a memory that can store memory for a long time even when the power supply source is cut off. Example: Read Only Memory (ROM), magnetic disk, magnetic core, magnetic tapes, hard disk, etc.

Interconnecting microprocessor with memory

The work process of reading data by the microprocessor from memory including:

1. The microprocessor prepares the address of the data to be read.
2. The microprocessor sends a read signal to memory.
3. After receiving the read signal, the memory looks for the desired data by the microprocessor according to the address given.
4. The data is sent by memory to the microprocessor.

The work process of writing data by the microprocessor to memory including:

1. Microprocessor prepares the address where data will be written.
2. The microprocessor prepares the data to be written.
3. The microprocessor sends a write signal to memory.
4. After receiving the write signal, the memory will read the data that has been given by the microprocessor and put it according to the predetermined address.

D. Mean Opinion Score

Mean Opinion Score (MOS) is an assessment obtained from listening to a voice directly, this scoring system is subjective [9]. The following is a table of MOS values used in this research.

TABLE I. MOS VALUES

No	Quality	Value
1	Excellent	5
2	Good	4
3	Fair	3
4	Poor	2
5	Bad	1

III. SYSTEM PLANNING

A. Block Diagram

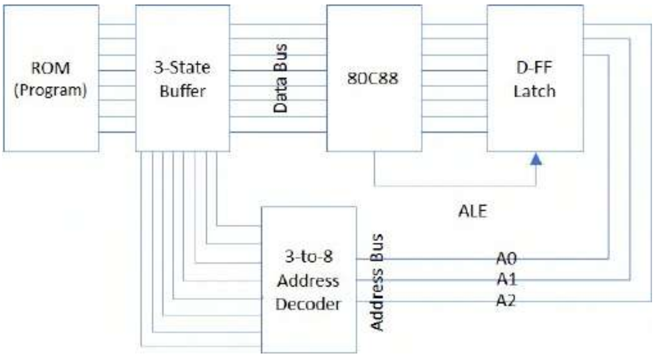


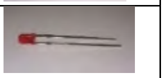
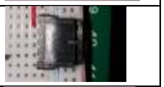
Fig. 2. Block diagram for the teaching aid

Fig. 2 shows that the 80C88 Microprocessor is connected to the D-FF Latch. D-FF is activated using the ALE pin. This D-FF functions to separate the Address bus from the data bus of the 80C88 microprocessor. If the ALE pin is active (condition 1), then the address bus line is active, while the ALE pin is not active (condition 0) then the data bus is active. The data bus line is connected to a 3-state buffer output. In this trainer, the input of the 3-state buffer is the program the DIP Switch. In this DIP Switch, the program inputted to or from the memory register. DIP Switch is used to replace ROM. This makes it easier for users to use this teaching aid, because if you use DIP Switch, the user will immediately program it. If using ROM, then the ROM must be removed to enter the program / instruction to be executed. To activate the 3-state buffer, use a 3 to 8 decoder. The input of this decoder is A0 - A2 and the output of this decoder is connected to 3 state buffer pins. the 80C88 microprocessor only uses a clock generator in the form of a push button because there is no minimum clock frequency.

B. Component Specification

TABLE II. COMPONENT SPECIFICATION

No	Component Name	Qty	Description	Figure
1	Microprocessor 80C88	1	8-bit CMOS Microprocessor	

2	74LS373N	1	8-bit D-Flip flop	
3	74HC138N	1	Decoder 3 to 8	
4	DIP Switch 8 bit	4	8 SPST	
5	74HC541N	8	Octal buffer	
6	LED Lamp	16	3mm	
7	Resistor 330 Ohm	16	¼ W	
8	Resistor 10k Ohm	32	¼ W	
9	Push button	1		
10	USB Socket	1	Female, type A	
11	Single copper cable		Multi colour cable to represent each functionality	
12	USB to USB Cable	1	Male – Female	

C. *System Design*

1. Separating address bus with data bus using D-FF

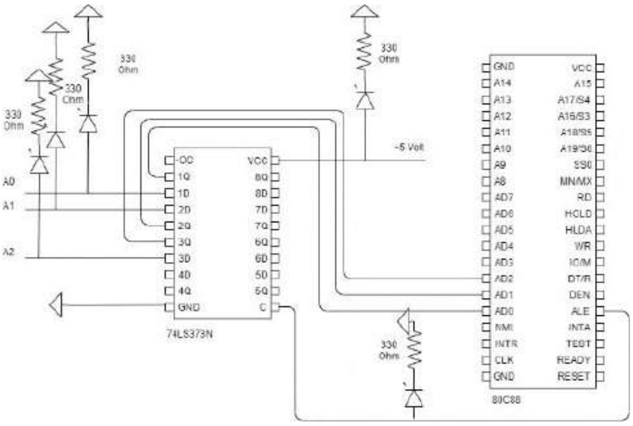


Fig. 3. Schematic diagram to separate data bus and address bus from 80C88 using 74LS373N

Fig. 3 shows interconnection of 80C88 microprocessor with IC 74LS373N. This schematic only uses 3 address line, which is A0, A1, and A2 since we only use 4 DIP Switch to replace the ROM (while A2=0) and to access 4 other bytes from Static RAM (while A2=1). Thus, we connect pins AD0 - AD2 to IC 74LS373N, which will provide the address bus output (A0 to A2). Microprocessor connection with the 74LS373N IC using the ALE pin. If the ALE pin is active (condition 1), then the address bus line is active, while the ALE pin is not active (condition 0) then the data bus is active. Connect the address bus with the LED and each led is connected to a 330 Ohm resistor. The ALE pin is given an LED indicator so that the user knows when the ALE pin is active or inactive. The VCC pin is also given an LED indicator so that the user knows the IC has been given a voltage.

2. Connecting DIP Switch to data bus

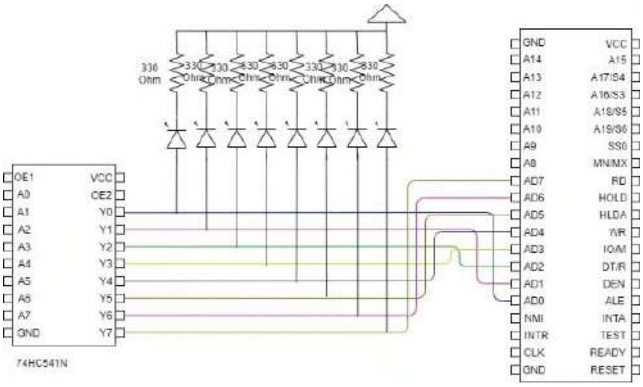


Fig. 4. Connecting DIP Switch to data bus using 74HC541N

Fig. 4 shows interconnection of 80C88 microprocessor with IC 74HC541N. The output 3-state buffer (IC 74HC541N) is connected to the microprocessor data bus (AD0 - AD7). In this schematic we use four IC 74HC541N, each for one DIP Switch. Each component is connected to the data bus. Connect the data bus to the LED. The LED is also connected to a 330 Ohm resistor which is connected to GND.

3. Control bus of 80C88

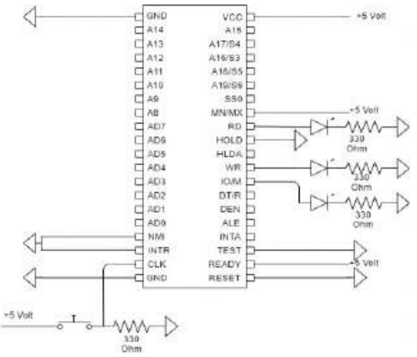


Fig. 5. Control bus used from 80C88

Fig. 5 shows the control signal on the 80C88 microprocessor. On the RD pin, an LED is connected to determine whether the RD signal is active or not (RD signal is active low). The WR pin is also given an LED indicator to find out the status of WR signal, whether it is active or not (WR signal is active low). Because the RD pin and WR pin are active low, if they are active it provides 0 Volt, while inactive it provides +5 Volt. The IO/M pin is also given an LED indicator to provide the status of access, whether 80C88 is accessing memory or I/O. The CLK pin is given a push button to replace the clock generator. The MN/MX pin and READY pin are connected to +5 Volt, while the HOLD pin, TEST pin, RESET pin, NMI pin, INTR pin, and GND pin are connected to ground. VCC pin is connected to +5 Volt to provide power to 80C88.

4. Replacing ROM with DIP Switch to store program

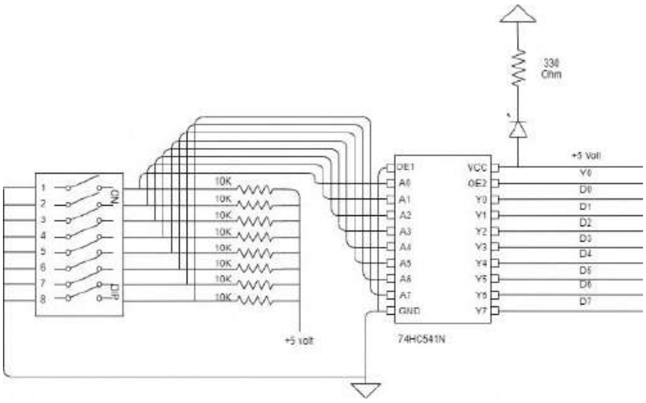


Fig. 6. Schematic for DIP Switch to store program with 74HC541N to connect with data bus

Fig. 6 shows the DIP Switch interconnection with the IC 74HC541N. One side of the DIP Switch is connected to the ground, while the other side is connected to 74HC541N and thus connected to data bus with 10 kOhm pull-up resistor.

5. Address decoder for DIP Switch using 3-to-8 decoder 74HC138

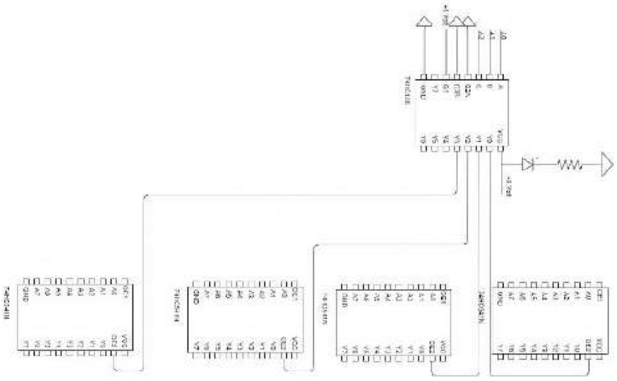


Fig. 7 Schematic of address decoder using3-to-8 decoder 74HC138

Fig. 7 shows interconnection between the 74HC138 and 74HC541N decoders. The input of this decoder is A0 - A2. With the output to be connected to the 74HC541N. This creates a binary combination. Each binary combination will activate each of the 75HC541N to connect one DIP Switch at a time.

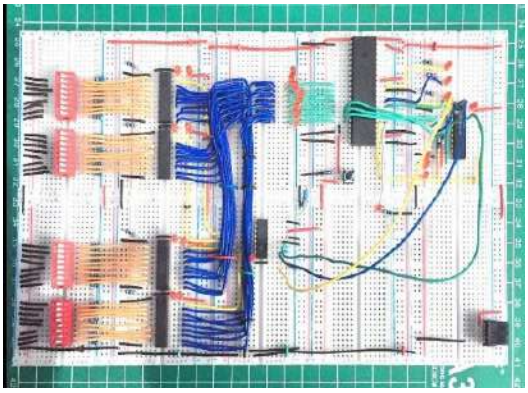


Fig. 8. Layout of teaching aid for microprocessor

Fig. 8 shows that the 80C88 microprocessor, along with all components in the teaching aid, is activated using a USB power supply. LEDs are installed in the address bus line, data bus, read signal, write signal and IO / M signal. The LED serves as an indicator of which path is active. From this path, it can be seen which memory address the information and data entered is pointing to. When inputted information, the information will be read by the 80C88 microprocessor. The information will be processed by the 80C88 microprocessor and will be output in the form of an LED. The illuminated LEDs are analyzed by the user from various aspects, namely the data bus LED, the address bus LED, the write signal LED or the read signal.

IV. ANALYSIS AND RESULT RESEARCH

This research conducted tests on several respondents who had different backgrounds. These criteria include:

- First Respondent: Senior Undergraduate Student of Telecommunications Engineering, Telkom University.
- Second Respondent: Junior Undergraduate Student of Computer Engineering, Telkom University.
- Third Respondent: Senior Undergraduate Student of Informatic Engineering, Telkom University.
- Fourth Respondent: Senior Undergraduate Student of Telecommunications Engineering, Telkom University.
- Fifth Respondent: Magister Undergraduate Student of Electro Engineering, Telkom University.

A. Testing The Microprocessor Trainer

In this test, 2 operations were performed, namely the write to memory operation and the reading from memory operation.

- The reading from memory operation

TABLE III. RESULT OF TESTING READING FROM MEMORY OPERATION RESPONDENT 1 USING MOV AL,[BX]

Binary	**Result of LED**						
	Clock	***Address Bus***	***Data Bus***	***RD***	***WR***	***IO/ M***	***Decoder***
1000 1010	9	100	0000 0111	Off	On	Off	Y1
0000 0111	17	000	1000 1010	Off	On	Off	Y0

TABLE IV. RESULT OF TESTING READING FROM MEMORY OPERATION RESPONDENT 2 USING MOV AL,[BX]

Binary	**Result of LED**						
	Clock	***Address Bus***	***Data Bus***	***RD***	***WR***	***IO/M***	***Decoder***
1000 1010	5	000	1000 1010	Off	On	Off	Y0
0000 0111	10	100	0000 0111	Off	On	Off	Y1

TABLE V. RESULT OF TESTING READING FROM MEMORY OPERATION RESPONDENT 3 USING MOV AL,[BX]

Binary	**Result of LED**						
	Clock	***Address Bus***	***Data Bus***	***RD***	***WR***	***IO/M***	***Decoder***
1000 1010	4	100	0000 0111	Off	On	Off	Y1
0000 0111	6	000	1000 1010	Off	On	Off	Y0

TABLE VI. RESULT OF TESTING READING FROM MEMORY OPERATION RESPONDENT 4 USING MOV AL,[BX]

Binary	**Result of LED**						
	Clock	***Address Bus***	***Data Bus***	**RD**	**WR**	**IO/ M**	***Decoder***

1000 1010 0000 0111	4	100	0000 0111	Off	On	Off	Y1
	6	000	1000 1010	Off	On	Off	Y0

TABLE VII. RESULT OF TESTING READING FROM MEMORY OPERATION RESPONDENT 5 USING MOV AL,[BX]

Binary	Result of LED						
	Clock	*Address Bus*	*Data Bus*	*RD*	*WR*	*IO/M*	*Decoder*
1000 1010 0000 0111	11	000	1000 1010	Off	On	Off	Y0
	12	100	0000 0111	Off	On	Off	Y1

Table III - Table VII show that the test results of 5 respondents have different numbers of clocks. The results of the read signal or write signal show the same results. The props also when displaying the programs inputted on the data bus LED are also not sequential and are not the same for each respondent.

- The write to memory operation

TABLE VIII. RESULT OF TESTING WRITING TO MEMORY OPERATION RESPONDENT 1 USING MOV [BX],AL

Binary	Result of LED						
	Clock	*Address Bus*	*Data Bus*	*RD*	*WR*	*IO/M*	*Decoder*
1000 1000 0000 0111	5	100	0000 0111	On	Off	Off	Y1
	11	000	1000 1000	On	Off	Off	Y0

TABLE IX. RESULT OF TESTING WRITING TO MEMORY OPERATION RESPONDENT 2 USING MOV [BX],AL

Binary	Result of LED						
	Clock	*Address Bus*	*Data Bus*	*RD*	*WR*	*IO/M*	*Decoder*
1000 1000 0000 0111	10	100	0000 0111	On	Off	Off	Y1
	18	000	1000 1000	On	Off	Off	Y0

TABLE X. RESULT OF TESTING WRITING TO MEMORY OPERATION RESPONDENT 3 USING MOV [BX],AL

Binary	Result of LED						
	Clock	*Address Bus*	*Data Bus*	RD	WR	IO/M	*Decoder*
1000 1000 0000 0111	44	000	1000 1000	On	Off	Off	Y0
	48	100	0000 0111	On	Off	Off	Y1

TABLE XI. RESULT OF TESTING WRITING TO MEMORY OPERATION RESPONDENT 4 USING MOV [BX],AL

Binary	Result of LED						
	Clock	*Address Bus*	*Data Bus*	*RD*	*WR*	*IO/M*	*Decoder*
1000 1000 0000 0111	44	000	1000 1000	On	Off	Off	Y0
	48	100	0000 0111	On	Off	Off	Y1

TABLE XII. RESULT OF TESTING WRITING TO MEMORY OPERATION RESPONDENT 5 USING MOV [BX],AL

Binary	Result of LED						
	Clock	*Address Bus*	*Data Bus*	*RD*	*WR*	*IO/M*	*Decoder*
1000 1000 0000 0111	14	100	0000 0111	On	Off	Off	Y1
	22	000	1000 1000	On	Off	Off	Y0

Table VIII - Table XII show that the test results of 5 respondents have a different number of clocks. The results of the read signal or write signal show the same results. The visual aid also when displaying the program inputted on the bus data LED is also not sequential and not the same for each respondent.

- Subjective Testing

This research uses three MOS parameters to conclude that the teaching aid can help students to have more understanding on how the microprocessor 80C88 works. The three parameters are:

1. aesthetic based on the shapes, colors, and neatly assembled
2. easy to use based on the worksheet provided
3. completeness based on the compliance with the syllabus

The formula to identify whether the teaching aid is beneficial for students is: if more than 51% respondents state that the teaching aid is easy to use, and if more than 51% respondents state that the worksheets were comply with the syllabus, and if more than 51% respondents state that the teaching aid has simple shapes, colors, and are neatly assembled.

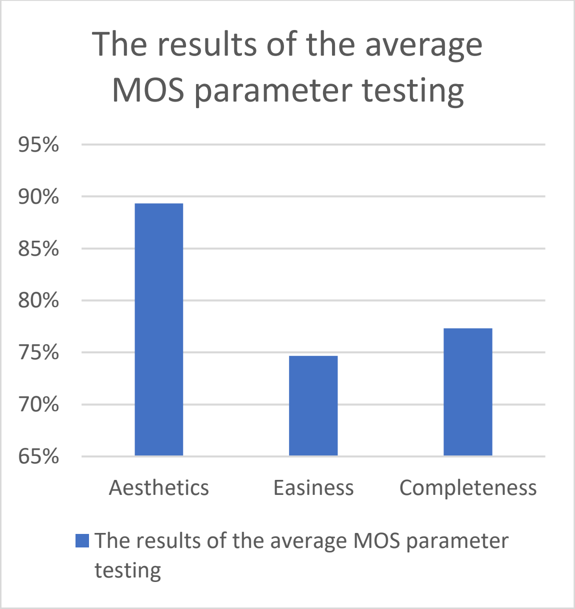


Fig. 9. Control bus used from 80C88

Fig. 9 is the result of the quantitative tests using three MOS parameters. Fig. 9 shows that the MOS test results are based on the assessment aspects that have been submitted to respondents. The average MOS test results were obtained based on the assessment aspect, namely the aesthetic aspect of 89.33%, the easy-to-use aspect of 74.67%, the completeness aspect of 77.33%.

V. CONCLUSION

The conclusion of this research is the trainers are functioning correctly and effectively to improve students' understanding of the "Microprocessor Interconnection with Memory" module. The test results of the props are in accordance with the expected parameters. The test results with the MOS parameter get an average value of more than 51% for each respondent. The average MOS test results were obtained based on the assessment aspect, namely the aesthetic aspect of 89.33%, the easiness aspect of 74.67%, the completeness aspect of 77.33%.

It is recommended that this trainer use RAM memory so that data can be stored in memory. The trainer is still in breadboard form, so the component cannot hold it in the socket. Therefore, the author suggests designing PCBs for microprocessor trainer.

ACKNOWLEDGMENT

Authors are grateful to Renesas Electronics Corporation (https://www.renesas.com) for providing the microprocessor 80C88, giving the opportunity to implement this teaching aid to support microprocessor course at Telkom University.

REFERENCES

[1] A. Arsyad, “Media Pembelajaran,” Jakarta: Raja Grafindo Persada, 2013.

[2] Karyono and A. Wicaksana, "Teaching microprocessor and microcontroller fundamental using FPGA," 2013 Conference on New Media Studies (CoNMedia), Tangerang, Indonesia, 2013, pp. 1-5, doi: 10.1109/CoNMedia.2013.6708541.

[3] M. A. H. A. Ghani, N. I. M. Enzai and N. Ahmad, "Effect of using simulator as teaching aid on microprocessor course performance," 2017 IEEE 15th Student Conference on Research and Development (SCOReD), Wilayah Persekutuan Putrajaya, Malaysia, 2017, pp. 215-219, doi: 10.1109/SCORED.2017.8305391.

[4] N. Karna, "Memahami Cara Kerja Mikroprosesor 8088 dan Antarmukanya," OpenLibrary Telkom University, 2021 [online: https://openlibrary.telkomuniversity.ac.id/pustaka/166942/memahami-cara-kerja-mikroprosesor-8088-dan-antarmukanya.html].

[5] Intersil, "80C88 CMOS 8/16-bit Microprocessor," Intersil, 1997.

[6] N. Karna, N. Fatihah and D. -S. Kim, "Evaluation of DLX Microprocessor Instructions Efficiency for Image Compression," 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2019, pp. 612-616, doi: 10.1109/ICTC46691.2019.8939709.

[7] R. Nugroho, "Modul Praktikum Mikroprosesor," Bandung: Telkom University, 2019.

[8] Y. Somantri, "Macam-macam Memori," 2017.

[9] L. A. A. Ramadhina Fitriyani, "Analisis Perbandingan Mean Opinion Score Alikasi VoIP Facebook Messenger dan Google Hangouts menggunakan Metode E-Model pada Jaringan LTE," vol. 6, no. 3, pp. 379 - 392, 2018.

Short-term Prediction of Lightning in Southern Africa using Autoregressive Machine Learning Techniques

Yaseen Essa
School of Computer Science and Applied Mathematics,
The University of the Witwaterstand,
Johannesburg, South Africa
YaseenEssa@essamail.co.za

Hugh G.P. Hunt
The Johannesburg Lightning Research Laboratory,
School of Electrical and Information Engineering,
The University of the Witwaterstand,
Johannesburg, South Africa
Hugh.Hunt@wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics,
The University of the Witwaterstand,
Johannesburg, South Africa
Ritesh.Ajoodha@wits.ac.za

***Abstract*—Lightning is responsible for both human and economic loss but its prediction remains challenging. We seek to find a lightning prediction model in South Africa that uses historical lightning-flash data only. This type of prediction model is cost-effective, easy to interpret and may be used for real-time forecasting. We evaluated and compared three popular time-series machine learning techniques on their ability to predict the number of Cloud-to-ground lightning flashes in South Africa for three-hours ahead. These models are the Auto Regressive (AR), Auto Regressive Integrated Moving Average (ARIMA) and the Long-Short-Term-Memory Recurrent Neural Network (LSTM) models. We used historical lightning data from the South African Lightning Detection Network during 2018. Our prediction model parameters were AR(lag=8), ARIMA (AR lag=8, integrate=0, MA lag=2) and LSTM (2x50 layers, activation=ReLU, optimizer=adam) and models were minimized for Root Mean Square Error but evaluated based on Mean Absolute Percentage Error (MAPE). We used a 70%/30% Train-test split. The AR and ARIMA models performed comparably with a MAPE of 15312 and 15080 respectively. The LSTM Model outperformed considerably with a MAPE of 3705. Although the LSTM model outperformed, predictions errors in absolute terms were still high. This paper highlights the usefulness of non-parametric predictions models for lightning prediction.**

***Index Terms*—lightning forecast, univariate, Long-Short-Term-Memory Recurrent Neural Network, weather forecasting, ARIMA, autoregressive**

I. Introduction

Lightning is responsible for both human and economic loss. It is estimated that 264-people on average die from lighting strikes in South Africa every year [4]. Further, lightning is responsible for about twenty-percent of all outages of electrical distribution in South Africa [5]. Knowing when lightning will occur, will help reduce human loss and assist in planning for expected lightning damage.

Lightning is an electrostatic discharge that results in a spectacular display of electromagnetic radiation and a pressure-wave called thunder [1]. There are between 30-to-100 lightning strokes every second on earth [2]. Seventy-five-percent of these lightning strikes originates and ends in the clouds; this is termed cloud-to-cloud lightning [3]. The remaining twenty-five-percent of lightning originates from the cloud and discharges to the ground; this is called Cloud-to-Ground (CG) lightning.

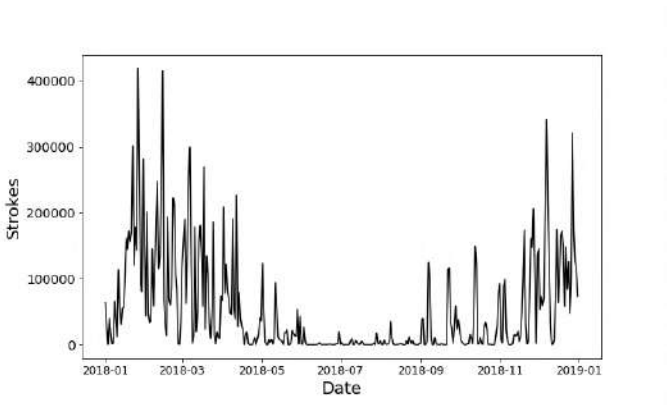


Fig. 1. Number of Lightning Strokes per day in South Africa during 2018.

Although lightning is familiar and well-researched, its prediction remains challenging. Recent academic studies have focused on using Numerical Weather Prediction (NWP) data to build a weather model for lightning forecasting. NWP models attempts to use mathematical models of the atmosphere and oceans to predict the weather based on current weather conditions. NWP model data is the culmination of sophisticated and complicated weather modeling often using supercomputers that relies on weather stations for data.

A lightning prediction model based only on actual historical lightning would be advantageous. A model based only on historical time-series would be cost-effective and can be used for real-time forecasting. Three common machine learning models for univariate time-series models are the Auto Regressive (AR), Auto Regressive Integrated Moving Average (ARIMA) and Long-Short-Term-Memory Recurrent-Neural-Network (LSTM) models.

The AR model is a linear predictive modeling technique. The model predicts future values based on past values of the same series by using the AR parameters as coefficients [14]. The ARIMA model also uses this concept but builds on it by including moving average error and integrated components.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

The moving average component indicates that the regression error is a linear combination of error terms whose values occurred contemporaneously and at various times in the past [10]. Moving average component removes non-determinism or random movements from a time series. The integrated component assists to make the data stationary if required.

The LSTM RNN Model is a type of artificial recurrent neural network (RNN). Unlike standard feedforward neural networks, RNN has feedback connections. LSTM models have a unit called a cell that is composed on an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data since there can be lags of unknown duration between important events in a time series.

Our aim is to compare and evaluate the short-term (three-hours) lightning predictive ability (number of flashes) of the AutoRegressive model, ARIMA model and the LSTM Recurrent Neural Network model using historical CG lightning flash data only. Doing this will lower the cost of lightning prediction and allow for real-time lightning prediction. Time-series univariate analysis have been applied in various datasets; although their accrual is limited, they are easy to implement and often provide a decent approximation for predictions.

Our study will contribute to current literature. This is the first time a LSTM RNN model has been applied to historical lightning data events from the SALDN dataset to evaluate its effectiveness against other common autoregressive techniques. This study will set the foundations of a more accurate lightning forecast model that will also incorporate weather data variables [15].

II. Methodology

Dataset Historical Cloud-to-ground Lightning Data in year 2018. The South African Weather Service established and maintains the South African Lightning Detection Network that records mainly CG lightning strikes [11]. The network currently consists of 24 Vaisala lightning sensors. The SALDN can detect lightning with a location accuracy of approximately 0.5km and an estimated detection efficiency of 90% over most of South Africa [12]. Our dataset had just under 20 million lightning observations that was grouped for every thee-hourly. Data was scaled using with a feature rand between 0 and 1.

Train/test Split. We trained the models using 70% of data and predicted the remaining 30%. For the year, this corresponds to training dates between 1 Jan 2018 to 12 Sep 2018, and testing dates between 13 Sep 2018 and 31 Dec 2018. All negative test predictions were considered as zero, as it is impossible to have a negative number of lightning strikes.

Test for stationery. The AR and ARIMA models require the historical time-series lightning data to be stationary in nature. We used the Augmented Dickey-Fuller test to test for stationary data. The test results indicate a test statistic of t = 5.96 with a p-value of 2.03×10^-7 (p <0.05). This indicates that we can reject the null hypothesis that the data in non-stationary.

A. Machine Learning Models

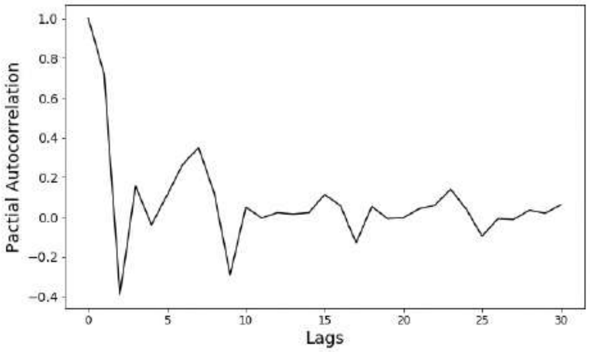


Fig. 2. Partical AutoCorrelation Function with SALDN Historical Lightning Data. The graph indicates a lag value of 8 is inclusive.

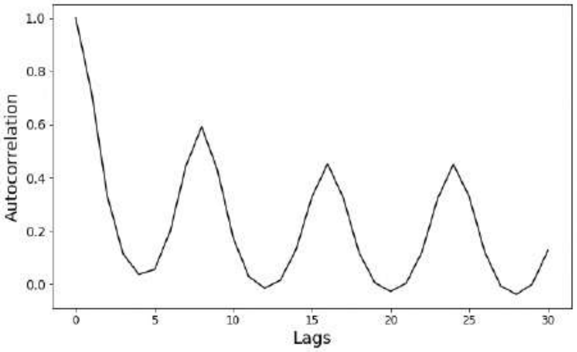


Fig. 3. AutoCorrelation Function with SALDN Historical Lightning Data. The graph indicates a lag value of 1 is optimal.

AutoRegressive Model. The general equation for the AR model is as follows:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-1} + \epsilon_t,$$

where p indicates number of lag steps, ϕ_i are the parameters of the model, c is a constant and ε_t is white noise. We used a lag value based on the Partial-Auto-Correlation-Function 2, which corresponds to a lag value of eight. The AR model from the Scikit-learn library of Anaconda Python v2019.10 was used. The Ordinary Least Square (OLS)/MSE error was minimized.

ARIMA Model. ARIMA is a popular class of autoregressive models that builds on the AR. model. In addition to AR component, the ARIMA model makes data stationary and also incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Given a time series data X_t where t is an integer index and the X_t are real numbers, This defines an ARIMA(p,d,q) process with drift $\frac{\delta}{1-\sum \phi_i}$,

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t,$$

where L is the lag operator, the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average part and the ε_t are error terms.

Our ARIMA model is optimized with an AR value of 8, Integrated value of 0, and MA value of 2. The AR parameter is taken from Fig. 2. The MA value of 1 is taken from the Auto-Correlation-Function of Fig. 3. An integrated value of 0 is used as the data is stationary. The ARIMA model from the Scikit-learn library of Anaconda Python v2019.10 was used. The Ordinary Least Square (OLS)/MSE error was minimized.

LSTM Recurrent Neural Network Model. Long Short-Term Memory Recurrent Neural Network. As mentioned earlier, LSTM is a special kind of RNN with additional features to filter recurrent data that is well-suited for time-series data forecasting. Our network has two dense network layer of fifty units followed by one dense layer with an activation function of rectified Linear Unit (ReLU) to reduce negative forecast values. We ran the model for 200 epochs; Fig 4 indicates that 200 epochs are sufficient minimize loss. The LSTM Keras module version 2.3.1 was used for LSTM RNN Network.

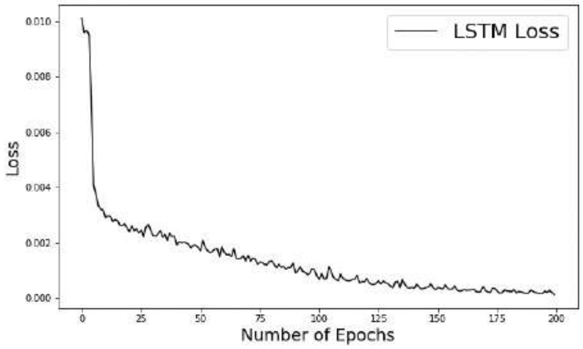


Fig. 4. Loss Function of LSTM RNN Model.

Limitations. Our data only processed one year of historical lighting data; this data should have seasonality which will improve prediction power.

III. Results

A graphical summary of the predicted values versus the actual data is shown in Figures 5, 6, and 7. The AR and ARIMA models performed comparably similar with MAPE values of 15312 and 15080 respectively. The RMSE values were 8579 and 8301 respectively I. The LSTM model had MAPE and RMSE values of 3705 and 9426, indicating it had outperformed AR and ARIMA considerably.

TABLE I
Prediction Model Error Values

Model	MAPE Value	RMSE Value
AR(p)	15312	8579
ARIMA	15080	8301
LSTM	3705	9426

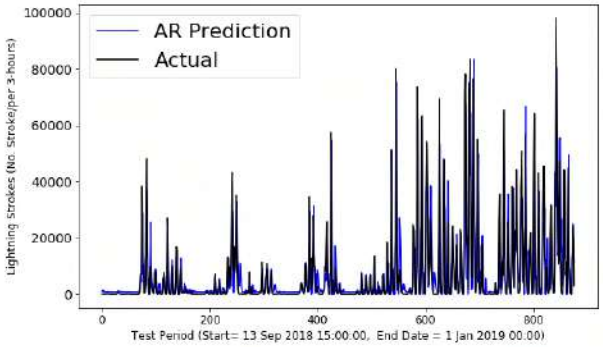


Fig. 5. AR Prediction Model Results vs. Actual Flashes.

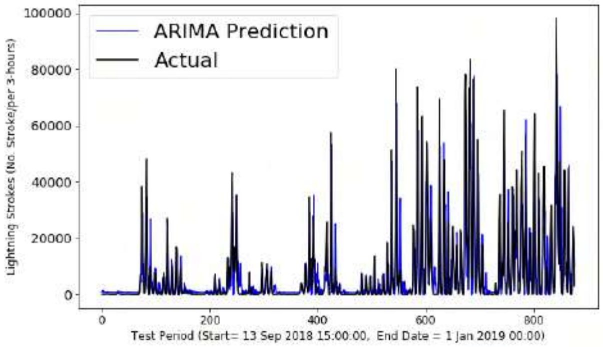


Fig. 6. ARIMA Prediction Results vs. Actual Flashes.

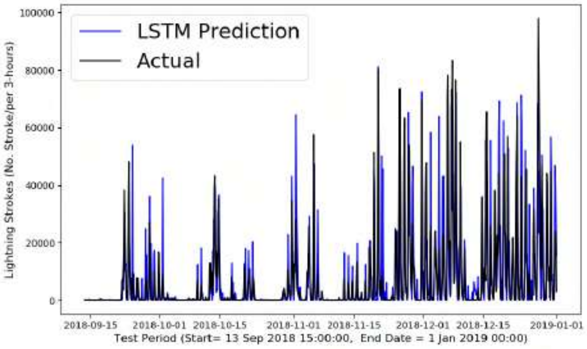


Fig. 7. LSTM RNN Prediction Results vs. Actual Flashes.

A graphical summary of the predicted values versus the actual data is shown in Figures 5, 6, and 7. The AR and ARIMA models performed comparably similar with MAPE values of 15312 and 15080 respectively. The RMSE values were 8579 and 8301 respectively I. The LSTM model had MAPE and RMSE values of 3705 and 9426, indicating it had outperformed AR and ARIMA considerably.

TABLE II
RESULTS OF AR COEFFICIENT AND INTERCEPT

Lag	AR Coefficients
AR Lag 1	0.9479
AR Lag 2	-0.5274
AR Lag 3	0.2505
AR Lag 4	-0.1218
AR Lag 5	0.0615
AR Lag 6	-0.127
AR Lag 7	0.3206
AR Lag 8	0.0702
Intercept	0.0054

TABLE III
RESULTS OF ARIMA COEFFICIENT AND INTERCEPT

Lag	ARIMA Coefficients
AR Lag 1	0.2063
AR Lag 2	-0.0816
AR Lag 3	0.1331
AR Lag 4	-0.0736
AR Lag 5	-0.0054
AR Lag 6	-0.019
AR Lag 7	0.0936
AR Lag 8	0.4647
MA Lag 1	0.7489
MA Lag 2	0.3323
Intercept	0.0456

Table II shows that the first lag value has the highest correlation compared to other lag steps for the AR model. Table III indicate the Moving Average Lag 1 values have the highest correlations actual data.

IV. DISCUSSION

In this study, we evaluated the AutoRegressive, ARIMA and LSTM models to forecast the number of lightning strikes for a period of 3-hours into the future. The AR and ARIMA performed comparably, and the LSTM considerably outperformed all models based on MAPE values (Table I).

Our study is not directly comparable with recent lightning prediction studies. The most recent study to investigate same-day lightning prediction in South Africa is by [11]. In this study, the authors found an AUC ratio of 90% using a stepwise logistic regression mode. But the study predicted the occurrence of a lightning strike occurring rather than the quantitative number of lightning strikes.

Numerous studies have found that LSTM models outperform AR and ARIMA models in forecasting using univariate time-series data [6], [7]. Even with weather series data [8], [9]. Reference [6] found the average reduction in error rates obtained by LSTM was between 84 - 87% when compared to ARIMA indicating the superiority of LSTM to ARIMA for various time-series data. Reference [8] found LSTM models performed about 20% better than ARIMA to predict wind-speed based on MAPE.

LSTM RNN models have several advantages over linear regression models but does not provide explanatory ability. Firstly, LSTM models are able to perform more complex functions than regression models. Secondly, they are able to analyses data with less restrictions such as the stationary requirement of data in regression models. The fact that LSTM models outperform autoregressive models indicates that lightning is non-parametric in nature and involves numerous dependencies. A limitation of neural network models is that they do not provide an understanding of how the results arises.

Recommendation for future studies. We believe that LSTM prediction accuracy can improve if we used a larger dataset to incorporate seasonal trends. This should increase predictive accuracy as lightning is seasonal. If we want an more accurate but data-intensive model, Numerical Weather Prediction parameters can also incorporated into the LSTM model. To gain a better understanding on the factors that influence lightning density, we suggest using a non-parametric machine learning techniques,

V. CONCLUSION

In this study, we predicted the number of lightning strikes for a three-hour period within South Africa. We found that the LSTM RNN model significantly outperforms AR and ARIMA models based on MAPE values. But all models still have relatively high error rates. This study indicates that lightning is better predicted using non-parametric techniques to due its nature. Future models may focus on using non-parametric modelling to better understand and predict lighting.

ACKNOWLEDGMENT

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835 and 98244). The authors would also like to thank The South African Weather Service, specifically Morné Gijben and Andrew van der Merwe, for the use of the SALDN data.

The support of the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NEPTTP

REFERENCES

[1] M. A. Uman, “The Lightning Discharge,“ ISSN. Academic Press, ISBN: 9780080959818, 1987, pp. 10—18.

[2] J. R. Dwyer, and M. A. Uman, “The physics of lightning,“ In: Physics Reports 534.4, ISSN: 0370-1573, pp. 147—241, 2014.

[3] V.A. Rakov and M.A. Uman, “Lightning: Physics and Effects,“ Cambridge University Press, ISBN: 9780521583275, pp. 15—20, 2003.

[4] R. L. Holle, and M. A. Cooper, “Lightning Fatalities in Africa From 2010-2017,“ In: 2018 34th International Conference on Lightning Protection (ICLP), pp. 1—4, 2018.

[5] T. B, Andersen, and C. J. Dalgaard, “Power outages and economic growth in Africa,“ In: Energy Economics 38, ISSN: 0140-9883, pp. 19—23, 2013.

[6] S. Siami-Namini, N. Tavakoli, and A. Siami-Namin, “A comparison of ARIMA and LSTM in forecasting time series,“ 2018 17th IEEE International Conference on Machine Learning and Applications, 2018.

[7] E. S. Karakoyun, and A. O. Cibikdiken, “Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting,“ The 13th multidisciplinary academic conference in Prague, 2018.

[8] Q. Cao, B. T. Ewing, and M. A. Thompson, “Forecasting wind speed with recurrent neural networks,“ European Journal of Operational Research, Volume 221, Issue 1, pp 148—154, 2012.

[9] Z. Pala, and R. Atici, “Forecasting Sunspot Time Series Using Deep Learning Methods,“ Solar Physics, 294(5), 50, 2019.

[10] G. E. P Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, “Time Series Analysis: Forecasting and Control,“ 5th Edition, John Wiley & Sons, pp 55—69, 2016

[11] M. Gijben, L. Dyson, and M. Loots, “A statistical scheme to forecast the daily lightning threat over southern Africa using the Unified Model,“ Atmospheric Research, vol. 194, pp 78—88, 2017.

[12] M. Gijben, “The lightning climatology of South Africa,“ South African Journal of Science, vol. 108, pp 44—53, 2012.

[13] V. Gandhi, “Brain-Computer Interfacing for Assistive Robotics Electroencephalograms,“ ISBN-13: 978-0128015438, Elsevier, pp 7—63, 2015.

[14] V. Gandhi, “Brain-Computer Interfacing for Assistive Robotics Electroencephalograms,“ ISBN-13: 978-0128015438, Elsevier, pp 7—63, 2015.

[15] Y. Yaseen, R. Ajoodha, and H. G. P. Hunt, ”A LSTM Recurrent Neural Network for Lightning Flash Prediction within Southern Africa using Historical Time-series Data,” 6th IEEE International Conference on Sustainable Technology and Engineering, 2020.

Two-stream Emotion-embedded Autoencoder for Speech Emotion Recognition

Chenghao Zhang
School of Communication and Information Engineering
Shanghai University
Shanghai, China
zhangch@shu.edu.cn

Lei Xue
School of Communication and Information Engineering
Shanghai University
Shanghai, China
16301098@163.com

Abstract*—Speech emotion recognition is an important part of the human-computer interaction process, which has been receiving more attention in recent years. However, although a wide diversity of methods had been proposed in decades, these approaches still cannot improve the performance. The main reason for the low accuracy of emotion recognition system is how to effectively extract emotion-oriented features. In this paper, we propose a novel autoencoder architecture, two-stream emotion-embedded autoencoder, to extract deep emotion feature. The input is projected to two latent representations in our method. One of them is meant to learn the best representation of the input which contains all information of speech; whereas the other is used to capture emotion-independent information. Next, the difference between two latent representations is considered as the deep emotion feature. Furthermore, the deep emotion feature is concatenated with global acoustic features obtained by openSMILE toolkit. Finally, based on the concatenated feature vector, fully connected network is adopted to conduct emotion classification. Besides, to improve generalization of our method, a simple data augmentation approach is applied. IEMOCAP that is a publicly available and highly popular databases is chosen to evaluate our method. Experimental results demonstrate that the proposed model achieves significant performance improvement compared to other speech emotion recognition systems.

Keywords—speech emotion recognition, two-stream autoencoder, emotion embedding

I. Introduction

In human speech interaction, people convey the underlying intent of speech through paralinguistic characteristics such as emotions, intonations and styles. Since human emotions help us to understand each other better, speech emotion recognition (SER) has gradually become a significant research interest. This technology has promising prospects and plays an important role in natural language understanding. For example, robots, mobile services and call centers. Recognizing these paralinguistic characteristics can help intelligent systems understand user intention and further improve the user experience. In this paper, an algorithm that analyzes the human emotions in speech with deep learning algorithm is proposed.

Recognizing human emotions from speech are a complex task. The main challenges are as follows: 1) due to their abstraction, human emotions may be treated as noise and discarded in many current speech recognition methods. 2) generally speaking, human emotion in a long utterance can only be detected in some specific moments [1]. Many previous works mainly focused on selecting speech acoustic features that can distinguish different emotion such as statistical features and prosodic features. Finally, basic machine learning algorithm (e.g., hidden Markov model (HMM) [2], Gaussian mixture model (GMM) [3], support vector machine (SVM) [4]) are utilized for SER. Recently, deep learning (DL) algorithm have made noteworthy progress in image processing field. Therefore, DL algorithms, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are also introduced into speech signal processing. The automatic extraction of useful features from speech signals by Deep Neural Networks (DNNs) has become a very powerful technique. Prior researchers used DNNs has demonstrated that DL has the most promising results compared with traditional algorithms.

Encouraged by the recent success of autoencoder structure [5] with deep unsupervised learning, and the idea of word embedding [6] in natural language processing (NLP), a novel autoencoder architecture, two-stream emotion-embedded autoencoder, is proposed to improve both learning and generalization capacities of SER system. Our modified autoencoder method projects the input to two hidden spaces. One of them is meant to learn the best representation of the input which contains all information of speech; whereas the other is used to capture information that is related to human emotions. Emotion embedding layers in our method can allow the model to efficiently learn a priori information from the ground truth. Besides, instance normalization (IN) [7] is introduced into autoencoder. In emotion classification stage, the deep emotion feature extracted by modified autoencoder and IS10 [8] feature set obtained by openSMILE toolkit [9] are fused for SER task. Finally, we evaluate our suggested SER model on IEMOCAP dataset and it achieves 71.86% recognition results. In the comparative analysis, our system shows outperformed recognition performance.

The rest of this paper is organized as follows. Section II describes the related algorithm of the autoencoder. Section III presents our proposed novel algorithm in detail. Section IV shows the experimental details and database. Section V demonstrates the experimental results.

II. Related Work

SER is considered a challenging task in human-computer interaction domain. In the early stage, several SER approaches attempted to use handcrafted speech features and low-level descriptors (e.g., fundamental frequencies, pitch, prosody, voice quality) to train classic machine learning models. Recently, increasing attention has been drawn to the study of the DNNs. However, there are two major issues observed in SER domain: (1) insufficient amount of labelled speech data, (2) the difficult of extracting emotion-oriented features from audio.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

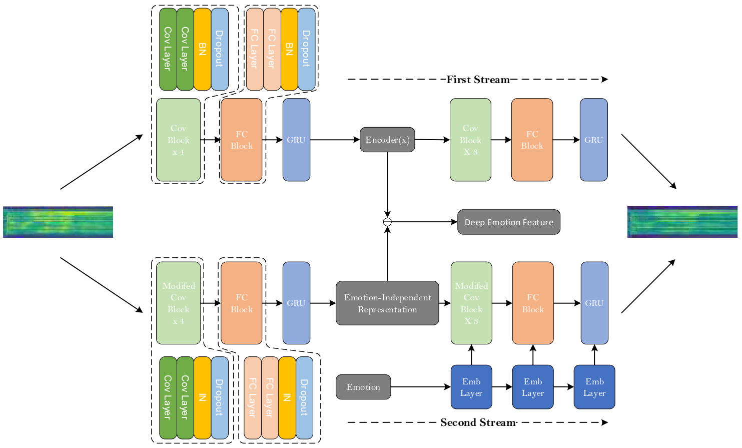


Fig. 1. The framework of two-stream emotion-embedded autoencoder.

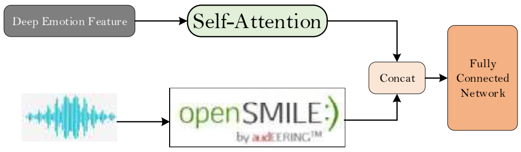


Fig. 2. The framework of emotion classification network.

To address the scarcity of training data, multiple methodologies were implemented in many works. Generally speaking, there are three approaches to solve this obstacle. (1) Collecting and annotating new data. However, it is expensive and time-consuming to create a big enough dataset. (2) Data augmentation. It is a most common method that has been widely used in DL field. (3) Transfer learning. This method is a popular research problem in DL that focuses on storing knowledge gained while training one model and applying it to another task. It has been successfully applied in various domains [10][11]. A major assumption in transfer learning is that the training and future data must be in the same feature space and have the same distribution. However, the mismatch between the dataset in SER field is a common problem. This is the reason why transfer learning has not further improved the accuracy of SER system. In this paper, a simply data augmentation method is applied for proposed SER algorithm.

In recent years, to strengthen the capability of feature extraction, autoencoder structure with unsupervised learning have been proposed. With the successful application of autoencoder, there is an increasing trend in most work to use it. Autoencoder is an unsupervised learning model used to reconstruct the input with minimum reconstruction error. Basic autoencoder has one input layer, one hidden layer and one output layer. Autoencoder first maps the input vector to the best latent representation through a non-linear layer and then this representation is mapped back to output layer, which purpose is to reconstruct input vector. If the number of hidden layers is greater than one, the network is considered to be deep. Many previous works directly utilized the latent representation learned by basic autoencoder for SER tasks. For instance, [12] proposed a deep autoencoder based on a multilayer perceptron for SER. Arghya Pal et al. [13] proposed deep dropout autoencoder based multilayer perceptron. Moreover, autoencoder was also applied to extract the bottleneck features for dimensionality reduction in [14] and [15]. Finally, the features extracted by autoencoder were utilized to train some machine learning algorithms, such as SVM and long short-term memory (LSTM).

Furthermore, denoising autoencoder (DAE) are also investigated in SER field to extract more robust features. The major difference between DAE and traditional autoencoders is that DAE is trained to recover from corrupted inputs. Inspired by the motivation behind this, Sayan Ghosh et al. [16] explored stacked DAEs, the deep structure of DAE, for representation learning. In addition, Zixing Zhang et al. [17] proposed a memory-enhanced recurrent denoising autoencoder (rDA). Experiment results have shown that this method can achieve significantly performance improvement.

In aforementioned methods, the purpose of autoencoder is to learn a lower-dimensional distributed representation of the input data directly. However, many researchers have also explored many modified autoencoder network. To reduce the discrepancy between the training and test set, shared hidden-layer autoencoder (SHLA) was proposed to learn common feature representations shared across them in [18]. Besides, Zefang Zong et al. [19] proposed a novel framework named multi-channel auto-encoder (MTC-AE) on emotion recognition. MTC-AE contains multiple local DNNs based on different low-level descriptors with different statistics functions that are partly concatenated together. It allows the model to consider both local and global features simultaneously. Pengcheng Wei et al. [20] proposed an algorithm based on autoencoder, denoising autoencoder, and sparse autoencoder. The first layer of the structure is based on a denoising autoencoder which purpose is to learn a hidden feature with a larger dimension, and the second layer employs a sparse autoencoder to learn sparse features.

Obviously, even if such methods can further improve the performance of SER, the high-level features learned by reconstructing input mainly contain the content information rather than emotion-oriented feature. Moreover, these above-mentioned works do not consider the significance of a priori knowledge. To address this problem, we design a new autoencoder architecture, two-stream emotion-embedded autoencoder, to increase the modeling capacity. In our model, we impose a constraint, emotion embedding, on decoder stage. Emotion embedding layers in our method lead the model to efficiently learn a priori emotion information from the label, which allows autoencoder focus more on deep emotion feature during reconstruction process.

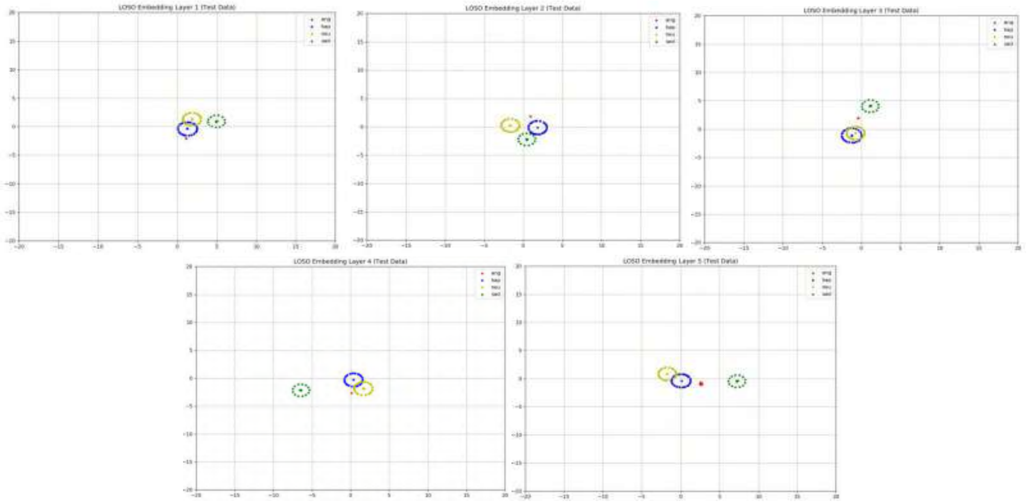


Fig. 3. T-SNE visualization of emotion embedding on IEMOCAP (LOSO).

III. Proposed Method

In this section, we describe our proposed method. There are three parts including input speech feature, two-stream autoencoder with emotion embedding and emotion classification network. Fig. 1 depicts the model framework, which includes two autoencoder paths, an emotion embedding path, and an emotion classification net. Let us consider that a dataset with N labelled samples $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$ and M unlabeled samples $\{(x_{N+1}),(x_{N+2}),\ldots,(x_{N+M})\}$, where x_i is denoted as the i-th acoustic feature sequence of speech sample and y_i is the emotion label corresponding to x_i. $y_i \in \{1,2,3,\ldots,K\}$, and K is the number of emotion categories.

A. Input Speech Feature

1) Log magnitude spectrogram: Spectrogram is a useful feature for analysis of speech and audio signals. Many previous researches are performed in the spectral domain rather than in the original time domain. The reason is that the magnitude spectrograms of audio signals tend to be highly structured in terms of both spectral and temporal regularities [21]. It is easier to deal with many problems by processing magnitude spectrograms than directly processing time-domain signals. In fact, magnitude spectrograms have been employed in many speech processing fields including audio separation and speech synthesis systems [22][23]. In this paper, log magnitude spectrogram is utilized to input our model, and the detailed spectral analysis was the same as the previous work [24].

2) IS 10 feature set: We utilize openSMILE [9] toolkit to extract statistics feature which was used in the INTERSPEECH 2010 Paralinguistic Challenge [8]. The open-source media interpretation by large feature-space extraction (openSMILE) toolkit is a modular tool for signal processing and machine learning applications. It can flexibly extract the features of signals and is mainly used for audio signal feature extraction. 1582-demensional feature vector is generated by extracting 38 kinds of LLDs and applying 21 statistic functions in this work. Details about these features can be found in [8]. In the emotion classification stage, it is concatenated with the deep emotion feature obtained by autoencoder.

B. Two-stream Autoencoder with Emotion Embedding

In this part, we interpret the complete scheme of our modified autoencoder in detail, as shown in Fig. 1. In this letter, due to the 2D representation of spectrogram, our proposed autoencoder is mainly based on a CNN. To further remove redundant information that is not related to human emotions and obtain robust deep emotion feature, the input is projected to two hidden representations. The first path in our model is a basic autoencoder network which consists of three blocks: convolution blocks, fully connected blocks and gated recurrent unit (GRU) [25]. In the second path, emotion embedding is introduced into the autoencoder. Additionally, in our method, we replace batch normalization (BN) [26] with IN in our network. BN is one of the most common components in many CNNs, and it can reduce the internal covariate shift during training process. The key difference between BN and IN is that the latter applies the normalization to an individual sample instead of a whole batch of samples. Generally, IN is mainly used in the style transfer field, for instance, image style transfer [27]. Many existing works disclose that IN learns features that are invariant to appearance changes, such as colors, styles, and virtuality, while BN is essential for preserving content related information [28]. We consider that human emotion is a kind of style features. To this end, we introduce IN into our autoencoder network, which purpose is to lead model to attract more attention to features related to emotion.

In the encoder process, the network is trained to map an input x to two latent representations: $encoder(x)$ (the first stream), $encoder_{EI}(x)$ (the second stream). EI donates emotion-independent. In the decoder process, the decoder network in the first path is trained to reconstruct input directly, as shown in (1). However, the decoder in the second path is equipped with the emotion embedding path, leading the model to efficiently learn a priori emotion information from the label. The main contribution of emotion embedding is that we expect that the projection can capture the emotion-independent information of the utterance. In this path, the decoder is trained to generate x' which is a reconstruction of x from $encoder_{EI}(x)$ given the emotion label y, as shown in (2).

$$x' = decoder(encoder(x)) \tag{1}$$

$$x' = decoder(encoder(x), y) \tag{2}$$

The mean absolute error (MAE) is used as the reconstruction loss since it generates a sharper output than the mean square error [29], as shown in (3).

$$L_{AE}(\theta_{Enc}, \theta_{Dec}) = \sum_{(x,y)\in D} \|x' - x\|_1 \tag{3}$$

Where θ_{Enc} and θ_{Dec} are the parameters of the encoder and decoder respectively.

Finally, the deep emotion feature is considered as:

$$emotion\_feature = encoder(x) - encoder_{EI}(x) \tag{4}$$

C. Emotion Classification with Feature Fusion

In the classify process, classification network takes the output of two-stream autoencoder and learns the links between it and the emotion label, as shown in Fig. 2. The output is fed into the self-attention [30] layer firstly, and the details of attention layer is same as the previous work [30]. With the attention mechanism, the network can focus more on emotion-oriented information of speech utterance.

Moreover, while the progressive downsampling of CNNs provides strong capability in local context modeling, Runnan Li et al. [31] believed that the temporal structure of speech that is highly related to emotions will gradually be lost in the downsampling process [32]. To overcome this problem, we concatenate the deep emotion feature extracted from attention layer and acoustic features obtained by openSMILE toolkit. These features contain global information of speech. Finally, the concatenated feature vector is fed into the fully connected network for emotion classification.

The emotion classification network takes $emotion\_feature$ as input and outputs the predicted emotion class. The classifier is trained to minimize the negative log-probability, as shown in (5).

$$L_{EC}(\theta_{Att}, \theta_{Cla}) = \sum_{(x,y)\in D} -\log P_{EC}(y \mid emotion\_feature) \tag{5}$$

Where θ_{Att} and θ_{Cla} are the parameters of the attention layer and classification network respectively.

During the training process, the object function of our network is a joint function decided by both reconstruction error and the negative log-probability:

$$\begin{aligned} L_{Total}(\theta_{Total}) = L_{AE1}(\theta_{Enc1}, \theta_{Dec1}) \\ +\lambda_1 L_{AE1}(\theta_{Enc1}, \theta_{Dec1}) + \lambda_2 L_{EC}(\theta_{Att}, \theta_{Cla}) \end{aligned} \tag{6}$$

where λ_1 and λ_2 are constant controlling the weighting between encoder path and classify path.

IV. Experiment

A. Data augmentation

Currently, there are two common problems about dataset in SER field: 1) the typical inherent mismatch between the dataset, 2) the difficulty in creating corpora. Generally, due to different emotion annotation schemes, the distribution of data between different datasets is often mismatch. In addition, high data collection often come with high annotation costs. Therefore, data augmentation has been proposed as a method to generate additional training data. In [33], Navdeep Jaitly et al. proposed a method named Vocal Tract Length Normalization for data augmentation. In [34], authors superimposed clean audio with a noisy audio signal. In LVSCR tasks [35], they have applied speed perturbation in their work. Besides, the use of an acoustic room simulator [36] and generative adversarial networks (GANs) [37] have been also proposed for data augmentation. However, aforementioned approaches all operated on the raw audio itself, rather than the spectrogram. In [38], D. S. Park et al. proposed a simple and computationally cheap method for data augmentation, which directly acted on the log mel spectrogram and did not require any additional data. Three deformations of spectrogram were chosen in their work: time warping, frequency masking and time masking. More generally, many works have demonstrated that data augmentation techniques have achieved state-of-the-art performance in ASR. In this paper, to not lose local context information of the speech signal, we randomly sampled 128 frames of log magnitude spectrogram with overlap. It means that 128 consecutive time steps $[t, t+128)$ is termed as a training sample, where t is chosen from a uniform distribution $[0, T-128)$, and T is the length of log magnitude spectrogram.

B. Dataset

To investigate the performance of the proposed method, a publicly available and highly popular database, namely the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [39] is chosen as source set. IEMOCAP was collected by SAIL lab at USC, USA, and it consists of 5 sessions. It has 10 professional actors (5 male and 5 female) acting in two different scenarios: scripted play and spontaneous dialog. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions. Each interaction has been segmented into sentences that are labeled by at least 3 annotators. In this paper, we used four emotion categories: angry, happy, sad and neutral. Note that, like many previous works, Happy and Excited in the original annotation were merged into one class: happy. Only the audio signals were used in the experiments.

C. Experiment Setup

Since there are 10 speakers in IEMOCAP and each session consists of 2 speakers, leave-one-speaker-out (LOSO) cross-validation were applied in our experiments, so that there is no speaker overlap between the training and test data. Moreover, 10-fold cross-validation strategies were also used to evaluate the proposed method.

For performance comparison, we utilize unweighted accuracy (UA) [40], which have been used in several previous emotion challenges. Weighted accuracy is the accuracy over all testing utterances in the dataset. It is quite a good measurement in this case since the class distribution is imbalanced.

We used sampled log magnitude spectrogram as the inputs. We trained the network using Adam optimizer with $lr = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$. The model was trained for 40 epochs on the dataset. All the experiments were performed using an Nvidia GTX 1080Ti with 11 GB memory.

V. Results and Analysis

A. *Impact of emotion embedding*

In this part, the IS10 feature set is combined with the traditional machine learning algorithm SVM (IS10+SVM) to serve as a comparison baseline. Moreover, to verify that our modified autoencoder (TSAE+EE) can efficiently improve SER performance, contrast experiments are performed on two different models (OSAE, TSAE).

TABLE I. The Impact of Emotion Embedding

Method	**Two Stream**	**Emotion Embedding**	**UA**	
			10-fold	**LOSO**
IS10+SVM	--	--	--	58.6
OSAE	×	×	69.98	65.46
TSAE	✓	×	69.43	65.18
TSAE+EE	✓	✓	**71.86**	**66.23**

Table I shows the performance of different classifiers on the IEMOCAP speech database. For 10-fold cross-validation, the UA obtained with the proposed method is improved by 1.88% and 2.43% compared with OSAE and TSAE, respectively. For LOSO cross-validation, the UA obtained with TSAE+EE is improved by 7.63%, 0.77% and 1.05% compared with IS10+SVM, OSAE and TSAE, respectively. Moreover, we can find that OSAE and TSAE have similar performance. This is because two streams in TSAE are mutually independent. In summary, the performance of SER is further improved by introducing emotion embedding.

B. *Visualizing the Emotion Embedding Using T-Distributed Stochastic Neighbor Embedding (t-SNE)*

T-SNE is an algorithm developed for visualizing multidimensional data, based on the idea of dimensionality reduction. We visualize the emotion embedding of our modified autoencoder model by t-SNE. There are five emotion embedding layers in decoder network, and they are trained to map an emotion to a 512-dimensional representation. T-SNE was then used to reduce the dimensions to only two for a 2D plot, as shown in Fig. 3. From Fig. 3, we can clearly see the separation between " ang" and "sad". Such result is expected since there are obviously different characteristics between them. However, we can also see that it is not clearly separated between "neu" and the other three emotions. One possible explanation is that low energy state of emotion "neu" do not have salient characteristics compared with the other emotions. Meanwhile, we can find the fact that "ang" is easily confused with "hap" due to the reason that the anger and happiness emotions correspond to high activation.

In summary, experimental results demonstrate that our proposed autoencoder naturally learn useful emotion representations from the label, and the training process discovers the intrinsic attributes that are necessary to solve the emotion recognition.

C. *Proposed method on IEMOCAP*

Table II clearly shows that the proposed method, two-stream emotion-embedded autoencoder, offers an improved performance in SER compared to previous studies. From the results of Table II, we can see that the highest accuracies obtained by the proposed method are 71.86% and 66.23% on IEMOCAP.

TABLE II. Performance Comparisions on IEMOCAP

Method	**Validation Setting**	**UA**
[41]	10-fold LOSO	64.2
[42]	10-fold LOSO	62.8
[43]	10-fold LOSO	59.54
[44]	10-fold	68.8
Proposed Method	10-fold LOSO	**66.23**
Proposed Method	10-fold	**71.86**

VI. Conclusion

In this paper, we designed a novel autoencoder network, two-stream emotion-embedded autoencoder. In modified autoencoder, the first stream is a basic autoencoder which purpose is to learn the best representation of speech. In second path, we combine both autoencoder and emotion embedding and replace BN with IN. The emotion embedding path focus on learning strong emotionally information from label. In emotion classification process, IS10 feature set was fused with the deep emotion feature from autoencoder. Experimental results with one publicly available corpora show that the proposed algorithm further enhances the classification accuracy.

References

[1] B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in 2019 IEEE International Conference on Signals and Systems (ICSigSys), pp. 40-44, 2019.

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603-623, 2003.

[3] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 4, pp. IV-957-IV-960, 2007.

[4] A. K. Samantaray, K. Mahapatra, B. Kabi, and A. Routray, "A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages," in 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 372-377, 2015.

[5] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 490-497, 2014.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, No. ICS-8506, 1985.

[7] D. Ulyanov, A. Vedaldi, and V. J. a. p. a. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016,

arXiv: 1607.08022. [Online]. Available: https://arxiv.org/abs/1607.08022.

[8] B. Schuller et al., "The INTERSPEECH 2010 paralinguistic challenge," in Eleventh Annual Conference of the International Speech Communication Association, 2010.

[9] [38] Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor," Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp. 1459-1462.

[10] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," IEEE Transactions on Affective Computing, vol. 10, no. 2, pp. 265–275, 2019.

[11] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification- an effective transfer learning technique," 2018, arXiv:1801.06353. [Online]. Available: https://arxiv.org/abs/1801.06353.

[12] Cibau, Neri E., Enrique M. Albornoz, and Hugo L. Rufiner, "Speech emotion recognition using a deep autoencoder," Anales de la XV Reunion de Procesamiento de la Informacion y Control, vol. 16, pp. 934-939, 2013.

[13] A. Pal and S. Baskar, "Speech emotion recognition using deep dropout autoencoders," in 2015 IEEE International Conference on Engineering and Technology (ICETECH), pp. 1-6, 2015.

[14] K. Huang, C. Wu, T. Yang, M. Su, and J. Chou, "Speech emotion recognition using autoencoder bottleneck features and LSTM," in Proc. International Conference on Orange Technologies (ICOT), 2016.

[15] W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang, and S. Shang, "Research on speech emotion recognition based on deep auto-encoder," in 2016 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 308-312, 2016.

[16] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation Learning for Speech Emotion Recognition," in Interspeech 2016, pp. 3603-3607, 2016.

[17] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," 2016.

[18] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4818-4822, 2014.

[19] Z. Zong, H. Li, and Q. J. a. p. a. Wang, "Multi-Channel Auto-Encoder for Speech Emotion Recognition," 2018, arXiv: 1810.10662. [Online]. Available: https://arxiv.org/abs/1810.10662.

[20] P. Wei, Y. J. P. Zhao, and U. Computing, "A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep auto-encoder model," Personal and Ubiquitous Computing, vol. 23, no. 3-4, pp. 521-529, 2019.

[21] K. Oyamada, H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and H. Ando, "Generative adversarial network-based approach to signal reconstruction from magnitude spectrogram," in 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2514-2518, 2018.

[22] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. J. I. S. P. M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 66-75, 2014.

[23] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct Modeling of Frequency Spectra and Waveform Generation Based on Phase Recovery for DNN-Based Speech Synthesis," in INTERSPEECH 2017, pp. 1128-1132, 2017.

[24] Y. Wang, RJ. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," 2017, arXiv:1703.10135. [Online]. Available: https://arxiv.org/abs/1703.10135.

[25] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, arXiv: 1406.1078. [Online]. Available: https://arxiv.org/abs/1406.1078.

[26] S. Ioffe and C. J. a. p. a. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, arXiv: 1502.03167. [Online]. Available: https://arxiv.org/abs/1502.03167.

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6924-6932, 2017.

[28] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 464-479, 2018.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125-1134, 2017.

[30] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, pp. 5998-6008, 2017.

[31] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated Residual Network with Multi-head Self-attention for Speech Emotion Recognition," in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6675-6679, 2019.

[32] S. J. Mozziconacci and D. J. Hermes, "Expression of emotion and attitude through temporal speech variations," in Sixth International Conference on Spoken Language Processing, vol.2, pp. 373-378, 2000.

[33] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, vol. 117, 2013.

[34] A. Hannun et al., "Deep Speech: Scaling up end-to-end speech recognition," 2014, arXiv: 1412.5567. [Online]. Available: https://arxiv.org/abs/1412.5567.

[35] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in INTERSPEECH 2015, 2015.

[36] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in INTERSPEECH 2017, 2017.

[37] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition gan for visual paragraph generation," in Proceedings of the IEEE international conference on computer vision, pp. 3362-3371, 2017.

[38] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," 2019, arXiv: 1904.08779. [Online]. Available: https://arxiv.org/abs/1904.08779.

[39] C. Busso et al., "IEMOCAP: interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, p. 335, 2008/11/05 2008.

[40] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in Tenth Annual Conference of the International Speech Communication Association, pp. 312-315, 2009.

[41] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms," in INTERSPEECH 2018, pp. 3683-3687, 2018.

[42] S. Latif, R. Rana, J. Qadir, and J. J. a. p. a. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," 2017, arXiv: 1712.08708. [Online]. Available: https://arxiv.org/abs/1712.08708.

[43] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7390-7394, 2019.

[44] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. J. I. T. o. A. C. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," 2020, arXiv: 1907.06078. [Online]. Available: https://arxiv.org/abs/1907.06078.

Energy-Efficient Vertical Handover in Heterogeneous Networks

Pratyashi Satapathy^1 and Judhistir Mahapatro, Member, IEEE^2
Department of Computer Science and Engineering
National Institute of Technology, Rourkela
{518cs1010^1, mahapatroj^2}@nitrkl.ac.in

***Abstract*—In heterogeneous networks, vertical handover plays a major role in providing seamless connectivity to the mobile users that are passing through different network access technologies and connecting to different point of network attachments. A mobile node scans all of its interfaces to find a suitable network with which it can connect. However, scanning of a large number of networks requires a huge amount of energy thereby the mobile node drains the battery very fast. The Media Independent Handover (MIH), which was introduced by IEEE, to facilitate seamless and energy efficient handover of mobile nodes across heterogeneous networks. The MIH Information Service (MIIS) offers a variety of criteria and services that can be used to avoid network scanning. But sometimes scanning avoidance leads to inconsistent handover, that is, increased handover failure rates. Thus, an optimal network scanning procedure, which can maintain a consistent handover and should consume the energy as least as possible is required. Our proposed work incorporates two additional functional units into MIH – one is responsible for making an optimal network scanning decision (it could be either full scanning, partial scanning, or avoid scanning) and other one is responsible for computing handover decisions by taking both network conditions and user preferences into account. In the second functional unit, we have introduced a utility function based TOPSIS algorithm that computes the handover decisions. Whereas, in principle, MIH users make the handover decision. Almost all of these existing methods suffer from severe ping-pong effect, unnecessary handovers, handover failures, and excess energy consumption due to inaccurate scanning method and ineffective network choice, whereas the performance analysis of our proposed work indicates that the suggested scheme performs better and consumes less amount of energy than existing works.**

***Index Terms*—Vertical Handover, Heterogeneous network, IEEE 802.21, MIH, Energy, TOPSIS, Utility Function**

I. Introduction

Heterogeneous networks [1] have been designed to include several access technologies of diverse forms aimed at providing better network coverage and capacity to mobile users. A mobile node should use an enhanced mobility management technique to make use of the access technologies in best way for meeting it's requirements [2]. Due to better handover management, a mobile user can retain its connectivity while switching from one point of network attachments to other [3]. As the handover takes between different access technologies, it is called as Vertical Handover. From now on, handover refers to Vertical Handover. Since handover consumes a lot of resources for managing it, hence it can greatly impact the overall efficacy of the networks unless managed aptly.

In order to serve multiple web applications in heterogeneous systems, next-generation communication systems must integrate multiple network interfaces (NICs) [4]. The usage of multiple interfaces opens up various ways to overcome many of the limits of transmission networks and provide numerous exciting new opportunities like Resource Sharing, Bandwidth Aggregation, Mobility Support etc. Using multiple interfaces at the same time a mobile node can extend the communication across the interfaces that reduce the risk of an interruption of communication. The increase in numbers of network interfaces and emergence of various standards, leads to compatibility issues and consumes excess energy during handover [5, 6]. IEEE 802.21 [7], a standard for Media Independent Handover (MIH), was developed to address these challenges. The handover protocol proposed in this standard is capable of communicating with all types of IEEE 802.x networks and other mobile networks (non-IEEE) like LTE, GSM etc. [8]. It fits in between Layer 2 and Layer 3. The central and logical entity of MIH known as, MIH Function (MIHF), operates as an intermediary layer between top and bottom layers, whose primary purpose is to manage the transfer of commands and data between various devices participate in the decision-making and the execution of handover process. MIHF offers three types of services to handle a seamless handover between various technologies. They are Media Independent Event services (MIES), Media Independent Command services (MICS), Media Independent Information services (MIIS). These services enable users of MIHF to obtain information relating to handover process, and to send commands to L2 Layer or to the network. The events created in L2 layers are transferred and delivered to MIHF, asynchronously, whereas the commands and data produced by request/response method are delivered synchronously. The MIIS is used to enable the mobile node to carry out a handover procedure without any network scanning [9]. As the MIIS includes the Information Elements (IE) that incorporates the network attributes, in compliance with the norm, the mobile node can conduct a handover instead of using any scanning procedure. In real scenarios, the updation of IE happens after a complete scanning procedure. If a mobile node is carried out a handover process by avoiding any scanning process then it fails to deliver the IE update notification to the Information Server (IS). This suggests that the information about the network and the IS are not consistent, which could lead to handover failures. Consequently, scanning is required or not required can be determined. During the network discovery process of handover, the mobile terminals conduct the total network scanning procedure by continuously activating all network interfaces. The energy usage of a given mobile node directly depends on the number of network interfaces it uses [10]. The existing network scanning methods are of the following types: Always active scheme (conventional method), periodic scanning, and adaptive scanning [11]. Always Active scheme performs better in terms of handover delay, but it fails to address the issue of energy efficiency. Periodic scanning [9] is another network scanning approach, where a mobile node scans the neighbouring network periodically over a time frame. In order to conserve energy, a mobile node enters into an idle slate when scanning process is not carried out which prevents vertical handover decisions from being taken in real time. Therefore, the existing approaches may not be suitable for heterogeneous networks. Different algorithms based on the IEEE 802.21 architecture have been reported by various researchers in order to solve and optimise the problem of vertical handover. Such algorithms are categorized into many groups

978-1-6654-4067-7/21/$31.00 © 2021 IEEE

[12] such as: RSS based, distance based, cost function based, utility function based, multi attribute decision making (MADM) based etc. Utility function-based algorithms are widely used to define the mobile node's degree of satisfaction with the various functionalities provided by network technology [13]. The contributions of the paper are outlined as follows:

1) An efficient vertical handover is proposed to minimize the energy consumption of a mobile node by improving the network scanning process of MIH protocol.
2) Although the complete scanning procedure for the heterogeneous network is very heavy, the objective of our approach is to limit or avoid the number of scans on each NIC with a purpose of ensuring lower energy usage. Here we have considered a heterogeneous network consisting of numerous access points and base stations, and suggested an optimal network scanning approach intended to make the network very energy-efficient.
3) This work also employees a partial scanning method which is based on application's priority level to speed up the scanning process and reduced energy consumption by minimizing unwanted scans.
4) In our proposed handover technique, the handover decisions are computed by taking both network conditions and user preferences into account. To take care of user demands along with handover consistency, a hybrid approach of utility function and TOPSIS is considered in this work.

The remaining portion of the paper is organized as follows: Section II includes Related Work. Section III describes the Proposed Methodology. In Section IV, the Performance Analysis of our proposed handover technique is discussed using the simulation results along with simulation settings. Section V lastly concludes this paper.

II. Related Works

A) *Energy Efficient Handover Techniques*

In heterogeneous network the transmission ranges of most of the point of attachments are overlapped due to their random deployments which leads to higher interference and packet loss rate. The co-existence of point of attachments often lead to unnecessary energy consumption even at the time of low traffic condition as the communication systems does not have an uniformity in terms of energy usage relative to traffic loads. For example, even though these nodes are in idle condition, they continue to consume considerable amount of energy [14]. Here some earlier study has been carried out on the channel scanning method. Moon et al. [15] suggested a scanning method, focused on Received Signal Strength (RSS) and dwell time. This method extends the scanning period when the dwell-time time is sufficiently long and the RSS is enough for communication. Many studies mentioned MIH for energy consumption purposes [16, 17] but most solution approaches only referred to the life-span of mobile node batteries while the actual energy usage of the system was ignored. The authors of [16] suggest a fuzzy logic-based algorithm that involves multiple criteria for network selection. This work gives a good QoS and encourages the mobile nodes to utilize low energy power. The authors in [17], proposed an algorithm in which network scanning is carried out by considering channel coherence time to conserve MN battery life. In Ref. [18], the MIIS is used to handle the scanning process. The goal of the suggested approach is to reduce energy consumption by minimizing the number of network scan. Xenakis et al. [19] proposed a context aware vertical handover framework to reduce energy consumption in MN. Liu et al. [20], suggested an energy-efficient handover method in which they used IEEE 802.21 MIIS for information gathering. In their work, handover is triggered only when power consumption level, RSS, and cost exceeds a predefined threshold. Although conventional network scanning techniques are improving the accuracy of the scanning performance they consume a lot of energy. In this context, the periodic [21, 22] and adaptive scanning mechanisms [23] were proposed.

B) *Network Selection using Hybrid Algorithms*

Current research [24, 25] suggests a number of vertical handovers schemes focused on MADM approaches for selection of the best target network. The hybrid approach of Simple Additive Weighting (SAW) and Analytic Hierarchy Process (AHP) was used widely in order to make network selection judgments [26]. In [26, 27], the authors formulated the problem of network selection using two approaches named as AHP and Multiplicative Exponential Weighting (MEW). The TOPSIS is extensively used by many researchers to rank the available networks [28, 29]. Even though AHP approach is generally used to assign weights to decision criteria, limitations remain in this method. For this purpose, the authors in [30] have suggested an improved TOPSIS algorithm to rank the alternatives using ANP to assign weights to criteria. As a result, it gives better performance than conventional approaches. Authors in [31, 32], have proposed a hybrid algorithm that uses MADM approaches and utility function to solve the network selection issue where the utility functions is useful in describing the required application specifications and evaluate the state of network resources. Another smart network selection technique based on utility function and MADM approaches was introduced in Ref. [33]. The suggested approaches helped the mobile nodes to choose appropriately a network to connect and reduced the number of ping-pong effects significantly.

III. Proposed Methodology

A. *Problem Formulation*

The network consisting of N base stations, S mobile nodes, each mobile node is assumed to have K number of interfaces and running M different applications on each interface. Before triggering a handover, the network scanning process go on to search for a network which suits well to the requirements of the application which is currently running on the mobile node.

The objective of our work is to minimize the energy consumption of the mobile node due to unnecessary scanning and maximize user preferences during network selection. Mathematically, the objective function is explained as follows:

$$Minimize \quad P_{total} = \sum_{i=1}^{S} (P_{HO})_i \tag{1}$$

Where, P_{total}: Total power consumption due to network scanning, $(P_{HO})_i$: Power consumption of the mobile node due to i^{th} handover, S: The number of handovers.

B. *Solution Approach*

We have introduced two functional units in the existing MIH architecture which will minimize total energy consumption and maximize user satisfactions of the mobile nodes after a handover takes place. Our proposed energy efficient vertical handover framework is illustrated in Figure 1. The functional units are Energy Consumption Unit (ECU) and Handover Decision-making Unit (HDU).

The purpose of ECU unit is to reduce the unnecessary network scanning during a handover process. Sometimes scanning avoidance leads to inconsistency in handover process and increases handover failure rates whereas excess network scanning increases energy consumption in the system. So, an optimal network scanning procedure is introduced here to minimize overall energy consumption.

HDU unit is responsible for taking handover decisions. This unit not only takes handover decisions but also maintains user satisfactions and helps in meeting requirements of the mobile nodes. For this purpose, we have used the idea of utility functions. Finally, a hybrid approach of TOPSIS and utility function is proposed to take handover decisions.

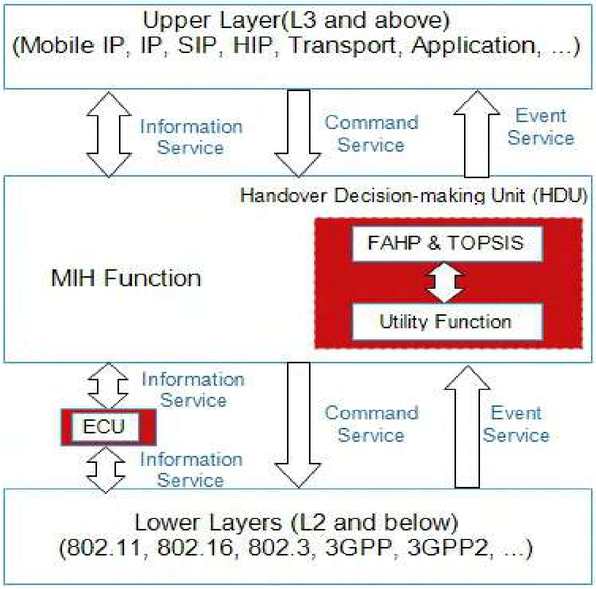

Upper Layer(L3 and above)
(Mobile IP, IP, SIP, HIP, Transport, Application, ...)
Information Service
Command Service
Event Service
Handover Decision-making Unit (HDU)
MIH Function
FAHP & TOPSIS
Utility Function
Information Service
ECU
Information Service
Command Service
Event Service
Lower Layers (L2 and below)
(802.11, 802.16, 802.3, 3GPP, 3GPP2, ...)

Fig. 1. Energy efficient vertical handover framework

C. Energy Efficient Vertical Handover in Heterogeneous Networks

1) Energy Consumption Unit (ECU)

The proposed solution approach is relying on IEEE 802.21, a latest IEEE standard which allows the continuity of service between heterogeneous networks like IEEE 802.x, 3GPP and 3GPP2. In order to minimize the energy consumption of mobile nodes, here we have used the Media Independent Information Service (MIIS) to handle the network scanning procedure. MIH offers the MIIS to enable the mobile node to conduct a handover process by avoiding network scanning [34]. Though the MIIS gives the IE, which includes relevant information related to network, as per the norm, the mobile node can conduct a handover even without performing any scanning process. Initially, the most important concept of the MIH is the reduction of network scanning using MIIS. The data provided by the MIIS are not always relevant because the updation period of the IE is not consistent and there are also not any representatives appointed to perform the duties of updating the IEs. In the real scenarios, the values of IE are updated after a complete scanning phase. When a mobile node conducts a handover by avoiding the scanning phase, then it is unable to transmit the updated value of IE to the Information Server (IS). So, the system information's are not accurate and valid. As a result, there appears a consistency issues between the system and IS which leads to handover failure.

The working procedure of ECU is shown in Figure 2. It's role is to decide the necessities of scanning process in the network. That's why the mobile node always conducts the complete scan process of the network and as a result it works against MIH's main objective. Thus, the criteria for evaluating whether to scan or skip needs to be established. Therefore, in this context, a new functional entity called the Energy Consumption Unit (ECU) has been introduced. The objective of the ECU is to limit the amount of network scans, which is the initial objective of the MIH. The ECU assists the overall minimizing process of the network scanning towards the chosen network.

Once the mobile node is alerted by the Connection Going Down (CGD) signal, it realizes that the handover mechanism needs to be carried out soon, so, the mobile node requests Information Elements (IE) to the information server (IS) of MIH (All networks are expected to have an access connection to the IS). After receiving the request message, IS then delivers the appropriate IE to the mobile node. When the mobile node obtains the IE, the ECU of the node retrieves the message and produces the membership degree. To verify the validity of the values of IEs, we provided two Time fields to the IE which helps to validate the IE data. It is restored with its initial values, whenever the data is modified. The MIHF conducts the task of network selection using the IE obtained and informs to the ECU. The ECU then measures the consistency of the preferred network. By using the outcome of the consistency checking process, it will determine whether to perform or skip the network scanning phase to trigger the handover process.

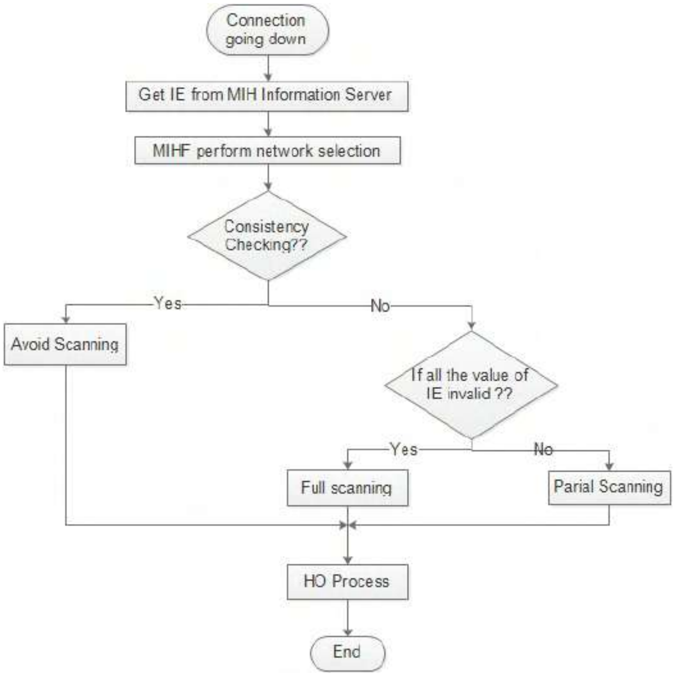


Fig. 2. The Working Procedure of ECU

- Consistency Checking Unit:

It checks whether the IE values are valid to use or not. It uses a fuzzy theory based mechanism [35] to measure consistency level of IE values. The scheme uses a Normal Distribution Function with mean 1.0 and standard deviation 0.4 as per Ref. [35]. With the help of input value and membership function we get the membership degree. If the value of membership degree is higher than the threshold value then the consistency of the value is still preserved.

- Network Scanning:

In this section, various types of scanning mechanism are explained that we have considered throughout the handover process by taking consistency level into account. After going through validity checking process, if all the IE values are consistent then the mobile node ultimately avoid the scanning process. Secondly, If some fields of IE values are valid i.e., partially valid, having some basic network related parameters, e.g., Service Set Identifier (SSID), location etc., then the mobile node goes for partial or fast scanning method, where, network scanning is performed only for the chosen network (not for all enabled networks).

a) Application's Priority-based Scanning Method

It actually falls under the category of partial scanning method. It is expected to have various applications operating on a single interface. During this, for each interface the mobile node is allocated with various time slots for scanning, which is based on the importance of applications that are running on the mobile node at that time. Usually, a higher prioritized application requires a lot of energy relative to a lower one. A mobile device typically has various applications including conversational (e.g., VoIP, video telephony, video game), streaming (e.g., watching multimedia), interactive (e.g., web surfing) and background class (e.g., WWW, emails) [36]. While travelling through heterogeneous wireless networks, each interface of a mobile device uses various types of applications. When an interface uses a prioritized application, then that will become the first choice for scanning. The application's importance level in terms of their priority is described in Table 1. Algorithm 1 describes the partial scanning method. According to algorithm, when a mobile device detects a high prioritized application then the node proceeds to scan that interface. Before the maximum time limit expires if that interface identifies a suitable target network then that time mobile node proceeds to scan a different interface having second prioritized application. Likewise, every interface executes their scanning processes. Finally total power consumption during partial scanning is calculated by adding each interface's power consumption in one scan time together.

TABLE I
APPLICATION'S PRIORITY VALUES

Various Applications	Priority Level
Conversational Class (e.g., VoIP, video telephony, video game)	High=1
Streaming Class (e.g., watching multimedia)	
Interactive Class (e.g., web-surfing)	Low=0
Background Class (e.g., news, WWW, Emails, file transfer)	

Algorithm 1: Partial Scanning Method

```
Input: M, K, $t_{inf}$=0, $t_{max}$=0, $p_{scan}$=0, $p_{inf}$
//M:Number of applications, K:Number of interfaces, $t_{inf}$:Interface scanning time, $t_{max}$:Maximum scanning time, $p_{scan}$:Power consumption during total interface scanning, $p_{inf}$:Power consumed by an interface in one scan time
Output: Scanned networks
// P:Priority, // inf:Interface
 $high \leftarrow inf1$;
 for $inf = 1$ to $K$ do
     if $(P[inf] > P[inf-1])$ then
         $high \leftarrow P[inf]$;
     end
     $inf$++;
end
while $(t_{inf} < t_{max})$ do
    $Scan(high)$; // Scan the interface having highest prioritized application.
    $p_{scan} = p_{scan} + p_{inf}$;
    if $(Scan(high) == finish)$ then
        Scan the interface having $2^{nd}$ highest prioritized application
    else
        $t_{inf}$++;
    end
end
```

b) Full Scanning Method

Lastly, if all IE values are inconsistent/invalid then complete network scanning process is conducted. Here complete scanning means all the enabled target networks are scanned. The total power consumption due to network scanning throughout a simulation having multiple handover process is calculated as follows:

$$P_{Total} = \sum_{i=1}^{H} P_{FS} \cdot (N_{FS})_i + \sum_{i=1}^{H} P_{PS} \cdot (N_{PS})_i \tag{2}$$

$$P_{HO} = P_{FS} \cdot N_{FS} + P_{PS} \cdot N_{PS}$$

$$P_{FS} = P_{AP} \cdot t_{real\_scan} \cdot N$$

$$P_{PS} = \sum_{inf=1}^{K} p_{inf}$$

Where, P_{Total}: The total power consumption due to network scanning, P_{HO}: Power consumption due to a single handover, P_{FS}: Power consumption due to full scanning, P_{PS}: Power consumption due to partial scanning, N_{FS}: Number of full scanning, N_{PS}: Number of partial scanning, P_{AP}: Average received power of different access technologies involved, t_{real_scan}: The actual scanning time per scanning period, K: Number of interfaces, H: Number of handovers, p_{inf}: Power consumed by an interface in one scan time.

2) Handover Decision-making Unit (HDU)

HDU is divided into three phases as shown in Figure 3.

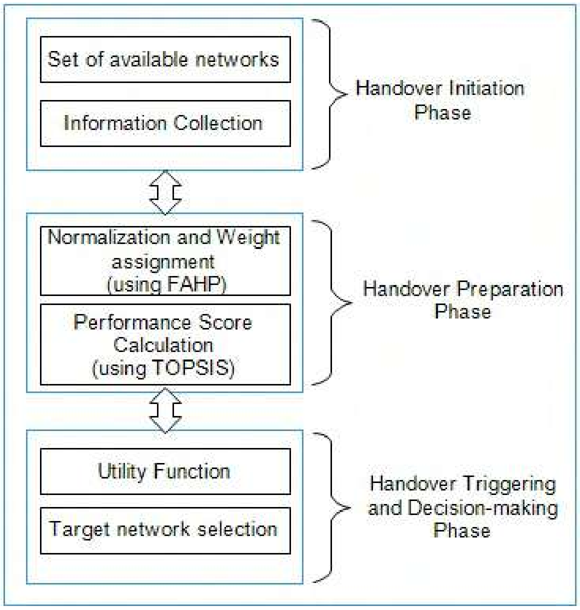

Set of available networks
Information Collection
Handover Initiation Phase
Normalization and Weight assignment (using FAHP)
Performance Score Calculation (using TOPSIS)
Handover Preparation Phase
Utility Function
Target network selection
Handover Triggering and Decision-making Phase

Fig. 3. Handover Decision-making Unit (HDU)

- Phase1: This phase is known as handover initiation phase. After the completion of network scanning phase, the mobile node gathered necessary information (e.g., network, terminal, and service related information) about the networks. Handover decisions are taken using these information.

- Phase2: This phase is called preparation phase of handover where all necessary steps are carried before triggering of actual HO process. Here, we have considered multiple decision criteria to make handover decision seamless and consistent. Fuzzy AHP (FAHP) is used to find the relative weights of decision criteria. Then the performance score of the networks is evaluated using TOPSIS algorithm.
- Phase3: Lastly this is called the decision-making phase of handover process. With the help of utility function and network performance score (which is calculated in Phase 2), the appropriate target network is decided for a seamless handover process.

The core element of this unit is an integrated module consisting of TOPSIS and utility function. TOPSIS usually picks the BS having highest solution score as the target node, irrespective of whether the user or the system requirements are fulfilled or not. Here the user or system demands and requirements are clearly neglected. To avoid these shortcomings, utility function is used along with TOPSIS to avoid the abnormalities created during the network ranking process. TOPSIS is used to aggregate multiple parameters and finds the relative closeness to the ideal solution or performance score. Then the utility function is used to rank the networks by using the above performance score. Here utility function measures the degree of satisfaction when MN travels from one Point of Attachment (PoA) to other.

The complete procedure of Utility Functions based TOPSIS method is described as follows:

- The conventional TOPSIS algorithm is used to distinguish the alternatives by ranking them. Then as a result, the highest ranked alternative is known as the best option for handover. But here we have not considered the highest ranked network as the target one. We only use TOPSIS for calculating the performance score of all the networks. We avoid the last step i.e. the ranking step of TOPSIS [37]. First the decision matrix with dimension m×n is constructed to compare available networks using different parameters. Then, the value in each cell of decision matrix is normalized. After that a weighted normalized decision matrix is created. The positive as well as the negative ideal solutions (optimal best and worst values) are evaluated. Then the deviation from the optimal solutions for each BS is determined. Finally, the performance score or the relative closeness to the ideal solution is measured.
- After measuring the performance score of the networks using TOPSIS then Utility function is used to rank the networks. The utility functions is useful in describing the required application specifications and evaluate the state of network resources. Here we have used exponential utility function to identify the appropriate target network. This is described below [38]:

$$\begin{cases} u(x) = \alpha * [\frac{1}{1+e^{-a(x-b)}} - \beta] \\ \alpha = (1+e^{ab})/e^{ab} \\ \beta = \frac{1}{(1+e^{ab})} \end{cases} \tag{3}$$

Where, x: Relative closeness to the ideal solution (computed in TOPSIS algorithm), a: value of anti-ideal solution, b: value of ideal solution.

IV. Performance Analysis

A. Simulation Settings

To facilitate our illustration, we consider a heterogeneous network scenario consisting of twenty BSs, three access technologies: LTE, WLAN, and WiMax and five MN. Table II contains all the simulation parameters. This framework is implemented in MATLAB. The decision criteria used in this work are grouped as, Network related: RSS, bandwidth, security, network condition, network performance, time to trigger, delay, QoS, velocity; Terminal related: power; Service related: cost, quality factor. FAHP is used to assign the relative weights to these criteria. We have used Empirical Hata propagation model since it is a widely used pathloss model along with all four traffic classes: Conversational, Streaming, interactive and Background traffic class. The Random waypoint mobility model is used in the proposed work to represent the behaviour of mobile nodes in the heterogeneous network. It is a widely used network mobility model [39]. In this model a node is arbitrarily selected at any point to reside or move with a certain probability. When it is decided to move, the direction and speed are randomly chosen.

TABLE II
Simulation Parameters

Parameters	Values
Area of interest(A)	80 x 80 m^2
Total no of nodes(N)	20
Transmitter Power (PT)	43dBm
Transmitter gain (GT)	18dB
Receiver gain (GR)	-1dB
Loss due to Transmitter (LT)	3dB
Loss due to Receiver (LR)	8dB
Pathloss Type	Urban/Sub-urban Hata model
Number of MS	5
Simulation Time	100s
MS Height	1m
BS Height	40m/50m
MS Speed	5-10(m/s)
Threshold value of Bandwidth	1800MHz

B. Result Analysis

In order to make the simulation more accurate, we ran the simulation 10 times and averaged the results. We didn't concentrate on the energy usage of a single node, but rather the entire network's energy consumption. An Optimum energy-efficient vertical handover is being achieved by employing an energy efficient scanning process and user requirements-based handover decision mechanism. Here the mobile node scans the neighbouring networks using the ECU employed in MIH. ECU, as described in Figure 2, decides whether to scan or skip. The suggested scanning method is compared with conventional, periodic and adaptive scanning schemes. Figure 4 shows the total energy consumption comparison between proposed and

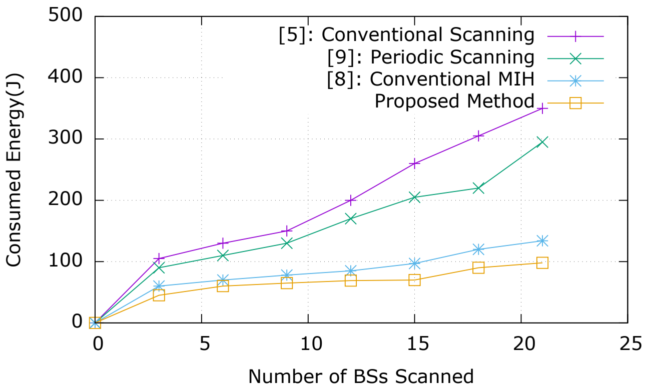


Fig. 4. Comparison of Consumed Energy versus Number of Base Stations

other existing scanning techniques. As a result, the proposed scheme considerably decreases the total energy consumption, occurred due to network scanning process. Figure 5 shows the comparison of network scanning rate between conventional MIH and our proposed method. Simulation results indicate that the proposed scheme lowers network scanning rate profoundly. Due to inadequate scanning technique, the existing schemes suffers from increased unnecessary scanning rate. The conventional technologies scan all the neighbouring networks at the time of handover which ultimately requires a huge amount of energy. There are certain applications which demands continuous connections across heterogeneous networks. So, the proposed technique gives more attention to the highest priority applications as compared to others. This often eliminates repeated handovers, which ultimately saves substantial amount of energy. Figure 6 shows the comparison of unnecessary handovers, where the proposed scheme performs very well as compared to the conventional MIH standard because our work considered a hybrid handover decision algorithm which avoids the occurrences of unnecessary HOs. The proposed work maintains a good balance between energy consumption and user requirements. Figure 7 shows the comparison of handover failure rate by considering all traffic classes. As a result, our suggested scheme is superior than TOPSIS and utility function algorithm. Figure 8 displays the average ping-pong rate versus all traffic classes. We observed that the hybrid approach of TOPSIS and utility function based algorithm can effectively minimizes the ping pong effect than other algorithms. This figure illustrates that the utility function-based TOPSIS outperforms the traditional TOPSIS. As a result, we conclude that introducing the utility function with TOPSIS will improve efficiency for various types of services.

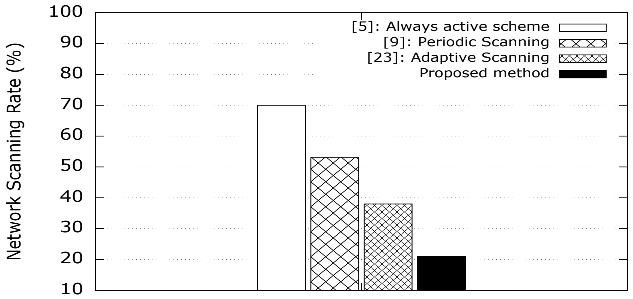


Fig. 5. Comparison of Network Scanning Rate

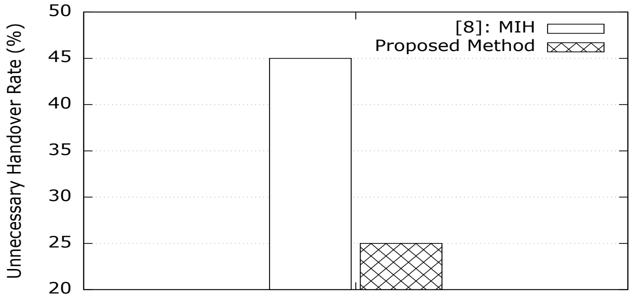


Fig. 6. Comparison of Unnecessary Handover Rate

V. Conclusion

To ensure smooth and seamless connectivity and to allow the optimal use of available network resources, it is essential to improve the performance of vertical handover in heterogeneous network in terms of unnecessary handovers and handover failure rates. With this in mind, an energy-efficient improved vertical handover technique is proposed in this work. This work introduces an energy efficient scanning scheme for IEEE 802.21 protocol that considers two functional units ECU and HDU to minimize network scanning process as well as meeting user requirements. MIH Information Server offers quick and energy-efficient channel scanning outcomes to the mobile nodes. Now a days, multiple applications are running on a mobile device at a given point of time with different priorities level. The proposed scheme scans the interfaces according to their priority of applications thus avoiding the scanning of entire interfaces which ultimately reduced the energy consumption. From the simulation result, it is confirmed that the suggested TOPSIS and utility function-based hybrid approach lowers the handover failure rate and ping-pong effect for all types of traffic classes. Thus the proposed work reduced the energy consumption while maintaining user preferences and their demands. This work may be further enhanced by proposing a prediction and forecasting based handover scheme for better quality of experience of users.

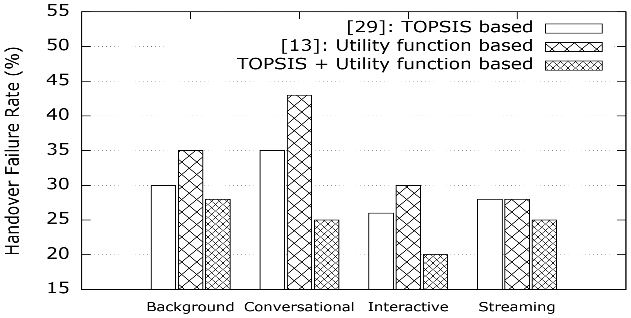


Fig. 7. Comparison of HO Failure Rate versus all Traffic Classes

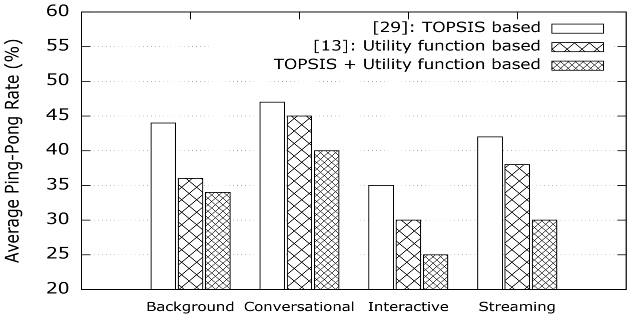


Fig. 8. Comparision of Average Ping-Pong Rate versus all Traffic Classes

References

[1] R. Almosbahi and M. Elalem, “Optimization Of Coverage and Handover for Heterogeneous Networks,” International Journal of Advanced Research and Publications, vol. 3, no. 3, pp. 213–219, 2019.

[2] S. Pahal, B. Singh and A. Arora, “Performance Evaluation of Signal Strength and Residual Time based Vertical Handover in Heterogeneous Wireless Networks,” International Journal of Comput. Netw. Technol., vol. 02, no. 01, pp. 25–31, 2014.

[3] E. M. Malathy and Vijayalakshmi Muthuswamy, “State of Art: Vertical Handover Decision Schemes in Next-Generation Wireless Network,” Journal of Communications and Information Networks, vol. 3, pp. 43–52, 2018.

[4] J. Amaro de Sarges Cardoso, F. Pereira Ferreira da Silva, T. Costa de Carvalho, J. Jailton Henrique Ferreira, NL. Vijaykumar and CR. Lisboa Francês, “Heterogeneous wireless networks with mobile devices of multiple interfaces for simultaneous connections using Fuzzy System,” PLoS ONE, vol. 16, no. 2, 2021.

[5] B. Narwal and A K Mohapatra, "Energy efficient vertical handover algorithm for heterogeneous wireless networks," International Journal of Control Theory and Applications, vol. 9, no. 19, pp. 9221-9225, 2016.

[6] GHS. Carvalho, I. Woungang, A. Anpalagan and SK. Dhurandher, "Energy-efficient radio resource management scheme for heterogeneous wireless networks: a queueing theory perspective," Journal of Convergence, vol. 3, no. 4, pp. 15–22, 2012.

[7] C. Imane, B. Jamila, K. Azeddine and K. Mohamed, "Overview on technology of vertical handover and MIH architecture," in Proceedings of the 4^{th} IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, Morocco 2016.

[8] M. Naresh, D. Venkat Reddy and K. Ramalinga Reddy, "A Comprehensive study on Vertical Handover for IEEE 802.21 Wireless Networks," Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020.

[9] D. Triantafyllopoulou, T. Guo and K. Moessner, "Energy-Efficient WLAN Offloading through Network Discovery Period Optimization," IEEE Transactions on Vehicular Technology, no. 99, pp. 1-11, 2014.

[10] StemmM, Gauthier P, Harada D, Katz RH, "Reducing power consumption of network interfaces in hand-held devices," in Proc. of 3^{rd} workshop on mobile multimedia communications (MoMuC-3), 1996

[11] M. Khan, J. Kim, J. Yun, K. Cho and K. Han, "An application dependent and sequential scanning scheme for vertical handover management in heterogeneous wireless networks," in Proceedings of the International Conference on Soft-Computing and Networks Security (ICSNS), Coimbatore, India, 2015.

[12] X. Yan, A. Şekercioğlu and S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks," Computer networks, vol. 54, no 11, pp. 1848-1863, 2010.

[13] P. Kosmides, A. Rouskas and M. Anagnostou, "Utility-based RAT selection optimization in heterogeneous wireless networks," Pervasive and Mobile Computing, vol. 12, pp. 92-111, 2014.

[14] A. Bianzino, C. Chaudet, D. Rossi and J. Rougier, "A survey of green networking research," IEEE Communications Surveys Tutorials, vol. 14, no. 1, pp. 3–20, 2012.

[15] CL. Moon, SH. Yang and IJ. Yeom, "Performance analysis of decentralized RAN (Radio Access Networks) discovery schemes," in Proceedings of the IEEE VTC,Baltimore, MD, USA, pp. 41–4, 2007.

[16] A. Calhan and C. Ceken, "Speed sensitive-energy aware adaptive fuzzy logic based vertical handoff decision algorithm," Proceedings of the 18^{th} International Conference on Systems, Signals and Image Processing (IWSSIP), Sarajevo, Bosnia and Herzegovina, pp. 1–4, 2011.

[17] C. Desset, N. Ahmed and A. Dejonghe, "Energy Savings for Wireless Terminals through Smart Vertical Handover," IEEE International Conference on Communications, pp. 1–5, 2009.

[18] W. Lee, W. Kim and I. Joe, "A power-efficient vertical handover with MIH-based network scanning through consistency check," Journal of Supercomputing, vol. 69, no. 3, pp. 1027–1038, 2013.

[19] D. Xenakis, N. Passas, L. Di Gregorio and C. Verikoukis, "A context-aware vertical handover framework towards energyefficiency," Vehicular Technology Conference (VTC Spring), pp. 1–5, 2011.

[20]] H. Liu, C. Maciocco, V. Kesavan and A.L.Y. Low, "Energy efficient network selection and seamless handovers in mixed networks," in Proce. of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops, Kos, Greece, pp. 1–9, 2009.

[21] C.-S. Wu, Y.-S. Chu, and C.-H. Fang, "The periodic scan and velocity decision handover scheme for next generation femtocell/macrocell overlay networks," in Proceedings of the International Conf. on ICT Conv. (ICTC), Jeju, Korea (South), 2013, pp. 201-206.

[22] D. Triantafyllopoulou, T. Guo, and K. Moessner, "Optimal Network Discovery Period for Energy-Efficient WLAN Offloading," in Proceedings of the IEEE 78^{th} Vehicular Technology Conference, Las Vegas, NV, 2013, pp. 1-5.

[23] G. Castignani, N. Montavont, A. Arcia-Moret, M. Oularbi and S. Houcke, "Cross-layer adaptive scanning algorithms for IEEE 802.11 networks," in Proceedings of the IEEE Wireless Comm. and Networking Conf., Cancun, Mexico, pp. 327-332, 2011.

[24] S. Baghla and S. Bansal, "VIKOR MADM based optimization method for vertical handover in heterogeneous networks," Advances in Systems Science and Applications, vol. 18, no. 3, pp. 90-110, 2018.

[25] G. A. Preethi, P. Gauthamarayathirumal and C. Chandrasekar, "Vertical Handover Analysis Using Modified MADM Method in LTE," Mobile Networks and Applications volume, vol. 24, pp. 1139–1151, 2019.

[26] L. Sheng-Mei, P. Su, M. Zheng-kun, M. Qing-min, and X. Minghai, "A simple additive weighting vertical handoff algorithm based on SINR and AHP for heterogeneous wireless networks," in Proceedings of IEEE International Conference on Intelligent Computation Technology and Automation (ICICTA), vol. 1, pp. 347-350, 2010.

[27] A. Davalos, L. Escobar, A. Navarro, A. Arteaga, F. Guerrero and C. Salazar, "Vertical handoff algorithms a new approach for performance evaluation," IEEE Globecom Worshop on Ubiquitous Computing and Networks, pp. 1724-1728, 2010.

[28] A. Sgora, D. Vergados, P. Chatzimisios, "An access network selection algorithm for heterogeneous wireless environments," in Proceedings of IEEE symposium on Computers and Communications, Riccione, Italy, pp. 890-892, 2010.

[29] L. Sheng-mei, P. Su and X. Ming-hai, "An improved topsis vertical handoff algorithm for heterogeneous wireless networks," in the Proceedings of the 12^{th} IEEE International Conference on Communication Technology, pp. 750-754, 2010.

[30] M. Lahby, C. Leghris and A. Adib, "An Enhanced-TOPSIS Based Network Selection Technique for Next Generation Wireless Networks," in the Proceedings of the 20th International Conference on Telecommunications (ICT 2013), pp. 1-5, 2013.

[31] M. Lahby, A. Attioui and A. Sekkaki, "An improved policy for network selection decision based on enhanced-topsis and utility function," in Proceedings of the 13^{th} International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, 2017.

[32] I. Chamodrakas, and D. Martakos, "A utility-based fuzzy TOPSIS method for energy efficient network selection in heterogeneous wireless networks," Applied Soft Computing, vol. 12, no 7, pp. 1929-1938, 2012.

[33] Y. Yu, B. Yong and C. Lan, "Utility-dependent network selection using MADM in heterogeneous wireless networks," in Proceedings of the IEEE 18^{th} International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, Greece, pp. 1-5, 2007.

[34] IEEE standard for local and metropolitan area networks: Media Independent Handover (2008).

[35] LA. Zadeh, "Fuzzy sets," Information and Control, vol. 8, no. 3, pp. 338–353, 1965.

[36] M. A. Senouci, M. S. Mushtaq, S. Hoceini, and A. Mellouk, "TOPSIS based dynamic approach for mobile network interface selection," Computer Networks, vol. 107, pp. 304–314, 2016.

[37] V. G. Vassilakis, G. A. Kallos, I. D. Moscholios, and M. D. Logothetis, "On the handoff-call blocking probability calculation in W-CDMA cellular networks," 4^{th} Adv. Int. Conf. Telecommun. AICT 2008, no. July 2020, pp. 173–179, 2008.

[38] M. Lahby and A. Sekkaki, "Optimal vertical handover based on TOPSIS algorithm and utility function in heterogeneous wireless networks," in Proceedings of the International Symposium on Networks, Computers and Communications (ISNCC), Marrakech, Morocco, 2017.

[39] A. Nayebi, M. R. Rahimi and H. S. Azad, "Analysis of Time-Based Random waypoint Mobility Model for Wireless Mobile Networks," in Proceedings of the 4^{th} International Conference on Information Technology, pp. 42-47, 2007.

Proactive Measures to Mitigate Cyber Security Challenges in IoT based Smart Healthcare Networks

Marshal R*, Gobinath K &, V Venkateswara Rao$
Indian Computer Emergency Response Team (CERT-In)
Ministry of Electronics and Information Technology
New Delhi, India
marshalrgj@gmail.com*, gobinathk@cert-in.org.in &, venkat4u.rao@gmail.com$

***Abstract*—The wide range of applications and advantages offered by the Internet of Things has attracted every sector to deploy it in their environment to exploit its advantages. The deployment of devices manufactured by different manufacturers in an Internet of Things environment has also opened the doors for threat actors to launch variety of attacks by exploiting the vulnerabilities present in these devices. The miniaturized size of most of the devices offers a very less space to incorporate the security elements. Healthcare sector has greatly benefited from the growth of Internet of Things. The COVID-19 pandemic has further highlighted the role of Smart Healthcare applications in the future. Deployment of life saving devices with limited security features make them one of the critical sector that needs more attention. The impact of a vulnerability exploited in smart healthcare devices can even be life damaging. In this work, the security challenges faced by smart health devices are analyzed and the necessary measures to be taken to improve the security are suggested.**

Keywords— Attack, Healthcare, Internet of Things (IoT), Security, Vulnerability

I. Introduction

Internet of Things (IoT) is a network of different physical devices that are connected to each other through different communication mechanisms. The surge in growth of IoT is evident in almost every sector and it has minimized the gap between the Information technology (IT) and Operation Technology (OT). The ability to operate with varied type of devices with varied operational features has made them an indispensable part of every sector. IoT has also taken the healthcare sector and the number of Smart Healthcare applications are on the rise [1, 2]. The growth in wearable technology coupled with the increase in device connectivity technologies like RFID, 5G, WBAN, Bluetooth, Wi-Fi and other evolving standards have greatly contributed to the progress of Smart Healthcare [3, 4]. It has reduced the complexity involved in the collection and analysis of data from patients located in remote areas. It has become a boon to both the medical practitioners as well as patients who need continuous monitoring. The beneficiaries list of smart healthcare applications even extend to insurance companies as the entire patient history can be accessed with ease [5].

The COVID-19 pandemic had an impact on the economy of many sectors but IoT has helped many sectors to survive the pandemic period and Smart Healthcare is one such critical sector that has grown strong during the pandemic period [6]. As in the case of all IoT devices, smart healthcare application devices are also vulnerable to attacks. The size of the devices and their mode of operation has limited the ability to add more security features to the device. The critical and sensitive nature of the data used in smart healthcare make them a favourite attacking ground for attackers. As these devices are involved in life saving applications, security is an indispensable ingredient in such applications [7] In this work, the impact of IoT in smart healthcare and the major devices that are used in healthcare applications are analysed in detail. It also explores the major cyber security challenges faced in the deployment of such lifesaving networks. Proactive measures are also suggested to ensure the security of smart healthcare networks.

The reminder of this work is structured as follows. Section II presents the various elements involved in Smart Healthcare. Section III gives a brief overview on the stakeholders of Smart Healthcare. Section IV presents the widely used smart healthcare devices. Section V discusses the cyber security challenges present in smart healthcare. Section VI presents the best practices and proactive measures that can be taken to secure a smart healthcare network. Finally, conclusion is provided in Section VII.

II. Smart Healthcare Elements

There are three main elements present in IoT based smart healthcare networks. They are Data collection and Pre-processing, Data storage and Processing and Data analysis. An overview of the elements involved in smart healthcare is presented in Fig.1.

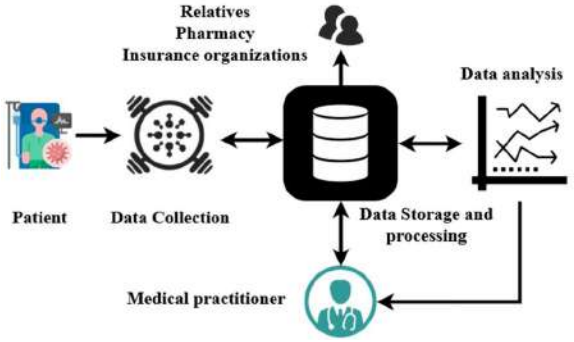


Fig. 1. Smart Healthcare elements

A. Data Collection and Pre-processing

Sensors, detectors, or monitors collect the information from the patient. Sensors will be present in the wearable devices at the patient end, in case if the patient is located in a remote place. If the device is not wearable, the patient may be present at locations where the devices are available to transmit the information to the network. Data collection transceivers present at the patient end may either transmit the data immediately or it may process the data and send the processed data to the data storage centre. However, the data will be converted into a form suitable for transmission. For security reasons, in most cases, the information will be pre-processed for encrypting the data before transmission [8].

B. Data Storage and processing

Information collected from the patient will be received by the data storage centre. In most cases, it could be a cloud storage medium. The data will be then be processed and it will

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

be transmitted in accordance to the user who needs access to the data. The format of the data will be varied in accordance to the accessing person. Data sent to the medical practitioner may not be the same data that will be sent to the relatives. The data storage medium will receive the feedback of the medical practitioner and it will send the necessary data that has to be sent to the concerned people. Depending on the information type, it may send the data to the patient, hospital, relatives, insurance agencies or any other relevant person defined in the network.

C. *Data analysis*

The data analysis will be done by the medical practitioner or by the research team that will analyse the response of the patient to the treatment provided to him/her. The data can also be used for behavioural analysis and for different research purposes. It can also be used to forecast the health condition of the patient.

III. Stakeholders of Smart Healthcare

Smart Healthcare is beneficial not only to the patient and the medical practitioner. But it is also useful to different stakeholders of Smart Healthcare [9]. In this section, the stakeholders of Smart Healthcare presented in Fig. 2 are discussed in brief.

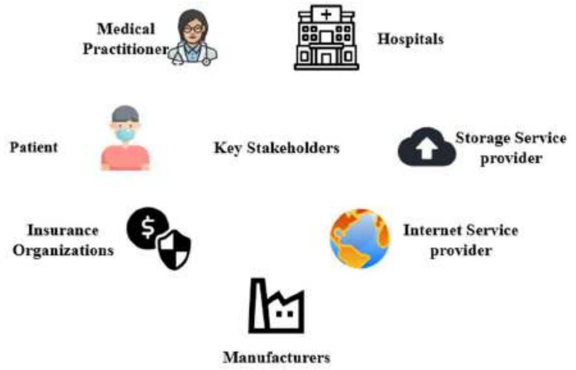


Fig. 2. Stakeholders in Smart Healthcare

A. *Patients*

The rise in wearable technologies and miniaturization of critical monitoring devices with high speed connectivity features have increased the rise of Smart Healthcare applications used by patients. Constant monitoring is made possible with these applications and the deployed smart devices are useful to alert the doctors, patients and their relatives, in case of any fluctuation in their regular body functioning. It has helped to predict and prevent health problems. Even in cases of emergency cases, the time gap between any health incident and the initial treatment has got dropped.

B. *Medical practitioners*

The constant supply of patient data from the smart devices helps the practitioners to constantly monitor the progress and response to the treatment provided to the patient. It offers the flexibility to the practitioners to change their treatment in accordance to the progress made by the patient. Practitioners also have better access to patients' history as they are readily available anytime and anywhere.

C. *Hospitals*

Hospitals provide the platform to most of the critical healthcare monitoring cases. Hospitals play a key role in monitoring the location of patients. They also monitor the performance and operation of critical devices. In addition to that, pharmacy inventory and the need for medicines can be generated based on the information received from the practitioners and patients. It will also help to acquire critical medicines and equipment well in advance before an emergency arises.

D. *Storage service provider*

Data storage is one of the important elements in Smart Healthcare system. Transmission, storage and retrieval of data have to be done effectively. As most healthcare applications require continuous monitoring, the volume of the data will be large and it requires a proper storage and retrieval mechanism. In addition to that, to perform a complete analysis on the data of the patient, the data must be available anytime and anywhere for analysis to the authorized users.

E. *Insurance organizations*

Insurance claims can be made more transparent as the entire data of the patient will be available for analysis and documentation. It can also help them to make the risks associated with the health of the patients. Insurance organizations also can track the progress made by the patient and adherence to treatment guidelines.

F. *Manufacturers*

The important stakeholder in Smart Healthcare system is the manufacturer. The smart health device performance is crucial to determine the efficiency and security of the network. The device must have secured connectivity features and also it should have better performance, in terms of power consumption and information processing capabilities. The manufacturers must comply with the standards setup by the regulatory bodies and legal agencies.

G. *Internet Service Providers*

Internet Service providers determine the communication platform for communication between the devices and the storage medium. Integration of dedicated IoT communication protocols along with the standard communication channels are on the rise. This will increase the security and connectivity of IoT devices.

IV. Smart Healthcare devices

COVID-19 pandemic has increased the deployment of Smart Healthcare devices and it is likely to develop in the post-pandemic era. The Smart Healthcare devices can be classified as wearable devices, implantable devices and clinical devices. Wearable devices can be worn by the patient and it can be removed even by the patient. Implantable devices are placed inside the body of the patient and it requires the help of medical practitioner to place or remove it from the body. Clinical devices will be present only at places where assistance will be provided by medical experts, such as MRI and X-Ray machines. The devices must comply with the CIA triad. A few of the major devices used in Smart Healthcare are presented in Fig. 3.

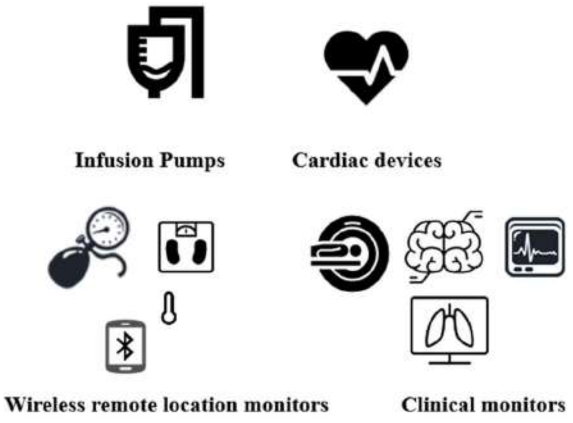


Fig. 3. Smart Healthcare devices

A. *Infusion pumps*

Infusion of medical fluids and monitoring the progress of the fluids can be done through the infusion pumps. Infusion pumps can be controlled remotely and it can be used to infuse multiple doasages at multiple places inside the body. This controlled infusion without any large scale surgery has reduced the complexity and expenses of several critical treatments. Insulin can also be infused through such pumps.

B. *Cardiac devices*

Programmable implantable cardiac devices can be used to monitor the performance of the heart. These devices will be continuously monitored and any change in the performance of the heart will be immediately intimated to the physician and the concerned authorities.

C. *Wireless monitors*

Devices such as blood pressure monitors, thermometers and heart rate detectors monitor the patient performance at periodic intervals and the progress of the patient will be monitored on a regular basis. These devices are generally connected to the mobile phone of the patient mostly through Bluetooth communication and by using dedicated apps. In some cases, cameras are also used to monitor the health condition of the patient.

D. *Clinical monitors*

Devices such as MRI, CT scanners, X-ray, lasers, surgical devices, ventilators, dialysis machines, ECG, EEG and other such similar devices which are deployed in hospitals or clinics are also connected to the IoT environment to track the continuous progress of the patients [10]. They are also used to analyze, process and interpret the data for better diagnosis.

V. Cyber Security Challenges in Smart Healthcare

Health information of a person is a very sensitive information and attacks on gathering such data have taken a steep rise after the COVID-19 pandemic. Attacks are launched by threat actors through multiple means to gather the patient related information. Since the smart devices carry vital data about the health, the attackers target such devices as they are more vulnerable.

Attackers can exploit the vulnerabilities that can be present in any layer, such as application layer, network layer or perception layer [11]. If the attacker gains access to control these live saving devices, such as infusion pumps or cardiac devices, the impact can be even life damaging. Even a small change in the dosage of drugs, or a Denial of Service (DoS) attack can result in death of patients with critical diseases.

A. *Device size and Power limitations*

The devices in most of these applications are too small and they have very few features for security. Incorporation of security features increases the size of the device and it limits the addition of security features to the smart healthcare devices. In addition to that, these devices are battery powered and loading them with too many features can affect the battery life of the device. The size also limits the use of stronger encryption mechanisms to transmit and receive the data.

B. *Software Update*

The devices may have vulnerabilities and if the devices are not updated with the patches, the device will remain vulnerable to be exploited by an attacker. Software update of Smart Healthcare devices are complicated since some critical devices cannot stop its operation even for few seconds. In case of Zero day attacks or if no patches are available for the vulnerability, it is difficult to replace the implanted devices. The attacker can also exploit any device that is connected in the network. In a hospital environment, multiple devices can be connected and they can be outdated. An attacker can exploit such systems and intrude into the network.

C. *IT-OT gap*

The healthcare industry has less IT experts to handle IT related incidents compared to other sectors. Hence, the sector has limited ability to handle incidents or attacks. The limited resources available with them increases the risks associated with the device. The networks maintained at the hospitals lack proper cyber security experts in most cases and hence, there is always a possibility of cyber threat in such vulnerable networks. The famous Wannacry ransomware attack was one such attack that exploited systems located at hospitals [12].

D. *Network connectivity failure*

The availability of Internet is the backbone for any IoT system. Patients with critical diseases need to stay in Internet connected areas for continuous monitoring. A small lapse in the connection, can turn crucial in some cases.

VI. Proactive Measures

As Cyber security is a key issue in Smart Healthcare, the following proactive measures can be taken to improve the security level of the Smart Healthcare networks.

A. *Deployment of cyber security experts*

Sufficient number of cyber security experts must be deployed at hospital and Smart Healthcare networks to regularly monitor, update and protect the connected devices in the network. This will try to reduce the IT-OT gap in a Smart Healthcare environment. The hospitals must be equipped with expert teams and also with proper incident response procedure to handle incidents.

B. *Inventory maintanence*

The IoT network must have an inventory of all the devices connected to the network. Their behaviour must be continuously monitored including the traffic and connections made by them. A Vulnerability database must be developed and it must be updated with the latest vulnerability reports related to the products used in the network and the corresponding devices must be updated with the suitable patches released by its vendors.

C. Compliance

The devices must be tested before deployment for compliance with the security standards such as ISO/IEC 82304, ISO/IEC 62304 and other relevant health products standards [13]. The devices must be tested for the ability to do security update. The data collection and distribution must comply with the legislative terms and conditions.

D. Secure update

Update should be allowed from authorized IPs and any connection made with unauthorized IPs must be considered as a malicious traffic. Connection attempts for security update must be made only through specific ports and the ports must be closed once the update is completed. Communication must be done only with whitelisted IPs.

E. Product security

The network must be formed as much as possible by using devices manufactured by the same manufacturer and the manufacturer must try to reduce the use of third party products in their design to reduce the risk of supply chain attacks. The default password of the devices must be changed before deployment in to the network. It is always preferable to deploy specific application based devices in a healthcare environment instead of third party general purpose IoT devices.

F. Network segmentation

Microsegmentation can be done to lock the critical devices from any unauthorized access outside the network. It has to be ensured that the devices in the network are loosely coupled and hence, failure of one device may not affect the overall performance of the network.

G. Data Integrity

The data that is stored in the storage medium must be made accessible only to the authorized user and all the information must not be made available to the users. The device must gather only the required information and it must be ensured that it do not gather any unnecessary information [14]. Periodical backup of data must be made to mitigate any unprecedented attack.

H. Security audit

Third party audits must be conducted periodically on the network and the vulnerabilities present in the networks must be assessed. Devices which cannot be updated with patches must be replaced with new devices.

VII. Conclusion

Smart Healthcare offers many advantages such as rapid diagnosis, improved decision making and proactive treatment. Cyber security is a key ingredient in such networks. But most of the Smart Healthcare networks are vulnerable to attacks due to several factors. In this work, the elements involved in Smart HealthCare networks and the cyber security challenges associated with them are analysed. The proactive measures that can be taken to improve the security of Smart Healthcare networks is also presented. The provided proactive measures can be used as best practice guidelines for developing a secure Smart Healthcare environment.

References

[1] L. Catarinucci et al., "An IoT-Aware Architecture for Smart Healthcare Systems," in *IEEE Internet of Things Journal*, vol. 2, no. 6, pp. 515-526, Dec. 2015, doi: 10.1109/JIOT.2015.2417684.

[2] S. U. Amin and M. S. Hossain, "Edge Intelligence and Internet of Things in Healthcare: A Survey," in *IEEE Access*, vol. 9, pp. 45-59, 2021, doi: 10.1109/ACCESS.2020.3045115.

[3] S. C. Mukhopadhyay, "Wearable Sensors for Human Activity Monitoring: A Review," in *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1321-1330, March 2015, doi: 10.1109/JSEN.2014.2370945.

[4] S. Amendola, R. Lodato, S. Manzari, C. Occhiuzzi and G. Marrocco, "RFID Technology for IoT-Based Personal Healthcare in Smart Spaces," in *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 144-152, April 2014, doi: 10.1109/JIOT.2014.2313981.

[5] H. Zhu et al., "Smart Healthcare in the Era of Internet-of-Things," in *IEEE Consumer Electronics Magazine*, vol. 8, no. 5, pp. 26-30, 1 Sept. 2019, doi: 10.1109/MCE.2019.2923929.

[6] O. Taiwo, and A.E Ezugwu. “Smart healthcare support for remote patient monitoring during covid-19 quarantine.” *Informatics in medicine unlocked*, vol. 20, Article ID: 100428, 2020. doi:10.1016/j.imu.2020.100428

[7] A. Alabdulatif, I. Khalil, X. Yi and M. Guizani, "Secure Edge of Things for Smart Healthcare Surveillance Framework," in *IEEE Access*, vol. 7, pp. 31010-31021, 2019, doi: 10.1109/ACCESS.2019.2899323.

[8] S. Tian, W. Yang, J. M. Le Grange, P. Wang, W. Huang, and Z. Ye, "Smart healthcare: making medical care more intelligent", *Global Health Journal*, vol. 3, no. 3, 2019, pp. 62-65. doi: 10.1016/j.glohj.2019.07.001.

[9] P. Sundaravadivel, E. Kougianos, S. P. Mohanty and M. K. Ganapathiraju, "Everything You Wanted to Know about Smart Health Care: Evaluating the Different Technologies and Components of the Internet of Things for Better Health," in *IEEE Consumer Electronics Magazine*, vol. 7, no. 1, pp. 18-28, Jan. 2018, doi: 10.1109/MCE.2017.2755378.

[10] I. Masood, Y. Wang, A. Daud, N. R. Aljohani, and H. Dawood, "Towards Smart Healthcare: Patient Data Privacy and Security in Sensor-Cloud Infrastructure,” *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 2143897, 23 pages, 2018. Doi:10.1155/2018/2143897

[11] A. Djenna and D. Eddine Saïdouni, "Cyber Attacks Classification in IoT-Based-Healthcare Infrastructure," 2018 2nd Cyber Security in Networking Conference (CSNet), Paris, France, 2018, pp. 1-4, doi: 10.1109/CSNET.2018.8602974.

[12] Wannacry Attack [Online] https://www.bbc.com/news/technology-39901382 (Accessed on 14th March 2021).

[13] S. Zeadally, F. Siddiqui, Z. Baig, and A. Ibrahim, "Smart healthcare: Challenges and potential solutions using internet of things (IoT) and big data analytics,” *PSU Research Review*, vol. 4 No. 2, pp. 149-168, 2019. Doi: 10.1108/PRR-08-2019-0027

[14] S. M. Karunarathne, N. Saxena and M. Khurram Khan, "Security and Privacy in IoT Smart Healthcare," in *IEEE Internet Computing* (Early access), 2021. doi: 10.1109/MIC.2021.3051675.

Success Criteria and Factor for IT Project Application Implementation in Digital Transformation Era: A Case Study Financial Sector Industry

Ni Wayan Trisnawaty
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
ni.wayan05@ui.ac.id

Teguh Raharjo
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
teguhr2000@gmail.com

Bob Hardian
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
hardian@cs.ui.ac.id

Adi Prasetyo
Faculty of Computer Science
Universitas Indonesia
Jakarta, Indonesia
adip12@ui.ac.id

***Abstract*— One of Indonesia's companies engaged in investment management is PT PNM Investment Management (PNMIM). Entering the digitalization era, PNMIM intends to use an information system in web and mobile-based applications, currently known as fintech, making it easier for investors to conduct mutual fund transactions online. This application is known as SiJAGO. However, in the first quarter of 2020, it did not achieve the launch of SiJAGO to the public. The impact of not achieving the launch of SiJAGO to the public will be affected the adaptive competition in technological development. Besides, the investment that PNMIM has spent for constructing the SiJAGO development project will be hampered by meeting several targets from the top management of PNMIM. This research is to identify the criteria and success factors for Information Technology (IT) projects in application development in the financial sector industry (FSI) in the era of digital transformation. The SLR analysis and expert judgment results show ten criteria and 18 factors (grouped into five categories) of IT project success, using the waterfall life cycle, using PMBOK® guidelines in analyzing each process. In the subsequent development of SiJAGO, PNMIM, and IT vendors, SiJAGO developers can focus on the criteria and factors that can support IT project development's success.**

Keywords*— *IT projects, information technology, application implementation, fintech, CSF, Critical Success Factors, Systematic Literature Review, SLR, expert judgment

I. Introduction

Investment is a common thing at this time, the commitment to several funds or other resources carried out at this time, intending to obtain several benefits in the future [1]. One of the fastest-growing investment products today is mutual funds. Mutual funds have turned into an investment product of choice for people who want to invest and cannot separate the increasing facilities and ease of investing. One of Indonesia's companies engaged in investment management is PT PNM Investment Management (PNMIM). PNMIM is a group of PT Permodalan Nasional Madani (Persero). PNMIM runs its main business in investment management, especially mutual funds and other managed funds, in the form of discretionary funds, business advisory, and corporate finance, both private and state-owned companies.

Entering the digitalization era, PNMIM intends to use an information system in web and mobile-based applications that can make it easier for investors to conduct mutual fund transactions online, known as financial technology (fintech). The implementation of an online mutual fund application development project, starting now referred to as SiJAGO, begins with submitting a request for proposal (RFP) document to an Information Technology (IT) consulting vendor. The RFP lists PNMIM's expectations regarding the SiJAGO development project, among them: (1) launching SiJAGO to the public in the 1st quarter of 2020, (2) 40,000 SiJAGO users when the application launched to the public in the first year. The SiJAGO development project began with a kick-off meeting between PNMIM and IT vendors. The SiJAGO development project's life cycle uses the waterfall life cycle and application development using the most common life cycle in IT, namely the Systems Development Life Cycle (SDLC). Planning, analysis, design, implementation, maintenance, and support are the five primary SDLC phases [2].

The implementation of SiJAGO implementation was three months late from the set schedule. After the implementation phase is carried out or known as go live, public users cannot directly use SiJAGO, the process related to licensing and SOP (Standard Operating Procedure) requirements issued by the Financial Services Authority (OJK) [3]. The condition impacts the delays of PNMIM's plan to launch SiJAGO to the public in the first quarter of 2020. Furthermore, PNMIM proposed additional features for SiJAGO development to the IT vendor. In the Software Maintenance Life Cycle, this process is called a modification request [4]. The addition of features to this application is considered a new project for application development and also carried out the System Integration Test (SIT) and User Acceptance Test (UAT) phases. In both stages, the application's issue findings into the ticket system of JIRA, the recapitulation of JIRA data in the SIT and UAT phases showed 49% of application bugs and 51% of change requests from users.

The fishbone diagram in Fig 1 illustrates the causes of the problems on the previously described exposure. Describe the causal relationship of various aspects and root causes that affect the failure to reach the public release of SiJAGO in the 1st quarter of 2020. It can be concluded that this condition is

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

a risky thing, especially since SiJAGO is a transaction tool that will be accessed at any time by the user. The company engaged in the financial sector industry (FSI) and transforming into the digital era cannot tolerate transaction errors due to technological factors. The condition will have an impact on customer perceptions and also on the increasingly competitive FSI competition. Therefore, it encourages analyzing the factors that can affect IT projects on application development at FSI.

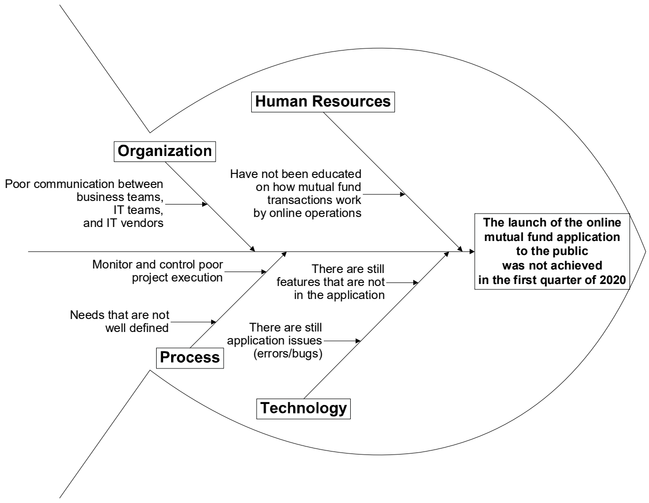

Human Resources
Organization
Poor communication between business teams, IT teams, and IT vendors
Have not been educated on how mutual fund transactions work by online operations
The launch of the online mutual fund application to the public was not achieved in the first quarter of 2020
Monitor and control poor project execution
There are still features that are not in the application
Needs that are not well defined
There are still application issues (errors/bugs)
Process
Technology

Fig. 1. *Fishbone Diagram of Root Problems Identification*

A survey was conducted by The Standish Group from 2012 to 2016 for all software projects around the world. The Standish Group also identifies the success rate of IT projects on that period [5] as follows:

- 36% of projects are successful (the project was delivered on time, on budget, and will all features).
- 27% of projects are challenged (the project was eventually delivered but either over budget, not on time, or not fully completed).
- 17% of projects are unsuccessful (nothing was delivered).

Furthermore, the survey results based on project resolution based on the industry for projects in the financial sector showed that projects in the project criteria were victorious by 29%, on challenging project criteria by 55%, and on project criteria for failing by 14% [5].

In previous research literature studies, not many focuses on IT projects' success factors on application development at FSI in the era of digital transformation or those that focus on implementing fintech applications. Furthermore, this research question is, what are the criteria and factors for IT projects' success in application development in the FSI in the era of digital transformation. Furthermore, the research is structured as follows: Part II describes a literature review related to the topic. Section III describes the methodology used for this research. Section IV reports the results and analysis of findings. Section V concludes and provides recommendations for further research on this research area.

II. LITERATUR REVIEW

A. IT Project

The IT projects are activities to build or implement IT-based service products, an organization's investment. Will be expecting to return the spent value of the time, money, and resource [2]. Information systems projects require resources with different expertise to work together to create software products, including systems analysts, software programmers or developers, testers, system installers, trainers, and other specialized skills [6].

B. IT Project Success Criteria

The success criteria for IT projects have different variations; the criteria for the success of IT projects that are used as a reference source in this research are [2], [6], [7], [8], [9], and [10]. The results of the analysis of the six references are described in Table I.

TABLE I. IT PROJECT SUCCESS CRITERIA

Criteria	Reference					
	[2]	[6]	[7]	[8]	[9]	[10]
Cost	√		√	√		√
Communication						√
Quality			√		√	√
Process		√				
Risk			√	√		√
Scope	√	√	√	√		√
Stakeholder		√			√	√
Resource		√		√	√	√
Time	√		√	√		√

The results of the analysis of six references obtained nine success criteria for IT projects. Furthermore, interviews were conducted with the teams involved in the SiJAGO development project, both from PNMIM and from IT vendors, to obtain validation and insight for research. The interview results showed that they agreed, and there was no revision of the nine success criteria for the IT project. The PNMIM team and IT vendors' interviews indicated that these nine factors were essential criteria for project success. The interviewee also conveyed that these nine factors are essential for the SiJAGO project, related to the collaboration between PNMIM and IT vendors. The communication factor becomes the determinant of the project's success. The Stakeholder factor is also important because any decision regarding the SiJAGO project requires stakeholder approval.

C. Critical Success Factor (CSF)

Since 1960 the success factor concept has been studied; D. Ronald Daniel introduced the origin of the success factors' concept [11]; first developed and refined into CSF in 1961 [12]. Daniel highlighted the importance of the factors that determine organizational success in addressing the risk of data overload. CSF is the key to information system design [13]. Several essential factors in management information systems have a positive and sustainable influence on a company's success; using these factors, competitive advantage can be realized [12].

CSF could prioritize an organization's strategic planning to support the CSF to achieve a competitive advantage with concentrate the resources. There are usually three to ten influential CSFs to determine the success of a project [14]. The CSFs are considered a powerful and applicable method used to address many of the challenges of implementing IT in several domains [15]. There is no differentiation of CSF for business models in general and fintech specifically. Moreover, the fast-paced and positively changing environment naturally demands a continuous observation of the critical factors. The dynamic development of the fintech field creates a need for future investigations [16].

D. Systematic Literature Review (SLR) for CSF in IT Project

This section is carried out at the literature research stage to find out the literature on the research results that have been done regarding CSF application implementation, especially in the FSI. The literature research is carried out a systematic literature review (SLR) using the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) model approach. The PRISMA model approach was used to eliminate excess factors in the literature [17]. The protocol used in the PRISMA model approach in this research is as follows:

1) *Identification*, literature searches were conducted on online databases with large repositories for IEEE Xplore, Elsevier (SCOPUS), ScienceDirect, and Springer Link. First, inclusion and exclusion criteria are determined to filter the search results. Inclusion criteria: literature published between 2015 and 2020, written in English, published in international journals and conferences, which conduct evaluation or analysis of CSF IT projects for implementation or application development in the FSI or digital transformation. Exclusion criteria: literature not related to presentations/opinions/papers, not focused on IT, related to IT projects but did not describe CSF, not published as proceedings of international conferences or international journals.

2) *Keyword*, literature search using keywords: ("project *" OR "implementation *" OR "information technology project *" OR "project management" OR "IT project implementation") AND ("financial technology" OR "financial sector" OR "fintech "OR" digital transformation "OR" case study") AND ("critical success factor * "OR" success factor * "OR" CSF"). The search results obtained are as follows: IEEE Xplore found 98 kinds of literature, Elsevier (SCOPUS): found 84 literature, ScienceDirect: found 53 literature, Springer Link: found 69 pieces of literature.

3) *Screening*, at this stage, a selection is made from the search results by looking at the title's suitability, abstract, and keywords previously identified. The total number of articles obtained in this section is 20 articles.

4) *Eligibility*, this section is done by reading in full to determine the article will be included in the next research that matches the eligibility criteria. The total number of articles obtained in this section is seven articles.

5) *Included*, in this section, five articles are selected that best fit the predetermined criteria.

E. CSF in IT Project Application Implementation

This section describes the CSF mapping of five previous types of research, which was obtained from the literature analysis in Part D, namely: [18], [14], [19], [20], [21]. In the five previous research literature, the research was conducted using the same framework, namely The Project Management Body of Knowledge (PMBOK®) [10], to analyze CSF in this research.

CSF is grouped based on factors and sub-factors, according to [18] and [14]. There are six-factor groupings and thirty-nine sub-factors of project success. Thirty-nine sub-criteria for project success were analyzed with the same characteristics, impacts, and risks on the development of SiJAGO. The results of these analysts obtained nineteen subfactors.

TABLE II. MAPPING OF PREVIOUS RESEARCH CSFS

Factor	Subfactor
Organizational	Communication
	Evaluation Progress
	Management Commitment and Support
	Management Leadership
	Relationship with Third Party
	Strategic Planning
Project	Project Control and Monitor
	Project Management
	Project Performance
	Project Planning and Requirements
	Project Risk
	Project Duration
Project Manager	Project Manager Experience
	Project Manager Formal Power
	Project Manager Skill
Project Team	Team Commitment and Participation
External Environment	Rules and Regulation
	Subject Matter Expert (SME)
	Tren

III. RESEARCH METHOD

This research uses qualitative and quantitative research methods (mixed methods) to draw more robust conclusions, provide a greater diversity of different views, and enable researchers to simultaneously answer confirmatory and exploratory questions, verify, and generate theories simultaneously [22]. In this research, the research strategy was carried out using a case study using a single case. Case study research is a form of qualitative method that has a robust approach to information systems research. Case studies are used both for confirmatory purposes (theory testing) and for exploratory purposes (theory building) [22]. The case study uses a single case because it is unique and essential, fulfilling all the conditions for testing the existing theories. Furthermore, single-case case studies can be generalized to various other conditions [23]. Research with the case study method seeks to examine the subject under research as much data as possible.

A. Instruments and Data Sources

In this research, the research instrument was related to qualitative research's basic principles, namely, data triangulation. A more nuanced picture of the situation can be obtained through data triangulation and increases the findings' reliability and validity [22]. Triangulation involves using more than one data source and collection methods to confirm the research data's authenticity, analysis, and interpretation [24]. Data triangulation in this research was carried out by interview, observation, and literature research. There are two types of data sources taken, namely primary data and secondary data. Primary data from this research were obtained from interviews obtained directly from informants or research objects. Secondary data is obtained from written documents and other things related to this research.

B. Data Collection Technique

Data collection was carried out using a technical triangulation approach. Triangulation is defined as a data collection technique that combines various data collection techniques and existing data sources [25]. Technique triangulation itself means that researchers use different data collection techniques to get data from the same source. Researchers used participatory observation, interviews, and documentation for the same data source.

- Observation, participant observation is carried out because researchers participate in doing what data sources do [25]; observation is also carried out directly at the PNMIM organization.
- Interviews were conducted with PNMIM's VP Retail Business Development & Customer Relations and PNMIM Information Technology & General Affairs Operations, two PNMIM business user teams, two IT vendor project managers, and two IT vendor application maintenance teams.
- The documentation source is from conducting studies on documents related to SiJAGO development projects. The source is from (1) kick-off meetings, Functional Specifications Document (FSD) containing SiJAGO functional specifications, (2) Technical Specifications Documents (TSD) containing SiJAGO system technical specifications, (3) Minutes of Meeting (MoM) documents, (4) JIRA ticket data, and (5) sign-off documents.
- Expert judgment, as part of an interview using a Likert scale, to seven application development project management experts.

IV. RESULT AND ANALYSIS

This section describes the results of semi-structured interviews with seven IT project management experts for application development. The interview is related to the SiJAGO development project. It is associated with assessing the criteria and success factors of the IT project in application development, described in sections B and E.

A. Expert Profile

The seven experts' selection was based on experience criteria in application development project management, especially in the FSI, and had attended IT project management training. Six experts work at the IT vendor company that does the development of SiJAGO, one person outside the IT vendor organization. Two of the seven experts have the Project Management Professional (PMP) certification. Table III describes the profiles of the seven experts.

TABLE III. EXPERT PROFILE

Expert Code	**Position**	**Company**	**Project Management Experience (years)**
P1	Chief Executive Officer (CEO)	IT Consultant	>15
P2	Head of Professional Delivery Services	IT Consultant	>10
P3	Head of Project Management Officer (PMO)	IT Consultant	>8
P4	Head of Wealth Management	IT Consultant	>6
P5	Project Manager (have PMP certification)	IT Consultant	>5
P6	Project Manager	IT Consultant	>5
P7	Head of Financial Solutions Group (have PMP certification)	Telecommuni cation Consultant	>10

B. Expert Judgement Interviews

Interviews with experts were conducted for two days, 07 - 08 January 2021, through virtual meetings media. At the time of the interview, the experts were presented with a list of criteria and success factors in IT projects in application development. The assessment was carried out using the survey.ui.ac.id site, with the link https://survey.ui.ac.id/955715. The scale used to rate responses to the criteria and success factors is a five-Likert scale from strongly disagree. Besides, each expert is also given the option to add criteria and success factors for IT projects that are not yet on the selection list.

C. Expert Judgment Results for IT Project Success Criteria

The results of the assessment carried out by experts are described in Table IV. The criteria proposed will be used in this research if the experts' average scale value is more significant than three. Next, a discussion was held regarding the assessment given by these experts.

TABLE IV. EXPERT JUDGMENT RESULTS FOR IT PROJECT SUCCESS CRITERIA

Criteria	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**Average**
Biaya	2	1	4	3	5	4	4	3,29
Communication	4	4	5	5	5	5	5	4,71
Quality	5	4	5	4	5	4	5	4,57
Process	5	4	5	4	5	5	5	4,71
Risk	4	2	4	3	5	5	4	3,86
Scope	5	4	4	3	5	5	4	4,29
Stakeholder	5	4	4	5	5	4	5	4,57
Resource	5	1	4	4	5	4	5	4,00
Time	4	1	4	4	5	4	5	3,86

Four experts provide additional criteria from the proposed. P1 provides additional commitment from the entire project team to support the IT project's success; these related discussion results are acceptable to add to the IT project success criteria. P3 adds lessons learned and expert judgment for the success criteria of IT projects. However, after further discussion, these criteria had the same meaning as the Project

Performance factor, so the proposed criteria from P3 were not included.

P6 provides handover to maintenance entry as criteria. However, from the discussion results, these criteria are more appropriately included in one of the sub-factors of IT project success, the Project factor. P7 adds business direction and government regulations as the success criteria for IT projects. The proposed business direction criteria have the same meaning as the Strategic Planning sub-factors on Organizational factors from the discussion results. Table IV shows that the Communication and Process criteria have the highest value in the Average column, followed by Quality and Stakeholders, and Scope and Resources in the next sequence.

None of the criteria has a mean value below three, although P2 assesses one for the criteria Cost, Resources, and Time. From the interview results with P2, it was revealed that the criteria for Communication, Process, Quality, and Stakeholders have the most crucial role in the success of the project, while other criteria are supporting factors. From the interviews with other experts, on average, they have the same opinion that these nine criteria are critical in IT projects' success.

D. *Expert Judgment Results for IT Project Success Factor*

Similar to Part B, this section describes the assessment results conducted by experts for IT projects' success factors. The criteria proposed will be used in this research if the experts' average scale value is more significant than three. Next, a discussion was held regarding the assessment given by these experts.

TABLE V. EXPERT JUDGMENT RESULTS FOR IT PROJECT SUCCESS FACTOR

Factor	Sub Factor	P1	P2	P3	P4	P5	P6	P7	Ave-rage
Organiza-tional	Communication	4	4	5	5	5	5	4	4,6
	Evaluation Progress	3	4	4	5	5	4	5	4,3
	Management Commitment and Support	4	3	5	4	5	5	5	4,4
	Management Leadership	4	3	5	5	5	4	5	4,4
	Relationship with Third Party	3	4	4	5	5	5	4	4,3
	Strategic Planning	4	4	4	5	5	4	5	4,4
Project	Project Control and Monitor	5	2	4	5	4	4	5	4,1
	Project Management	5	2	5	5	4	4	5	4,3
	Project Performance	4	2	4	4	4	4	5	3,9
	Project Planning and Requirements	5	2	4	5	5	5	5	4,4
	Project Risk	4	2	3	4	5	5	4	3,9
	Project Duration	5	2	3	4	5	4	3	3,7
Project Manager	Project Manager Experience	4	4	3	4	5	4	5	4,1
	Project Manager Formal Power	3	5	3	4	5	3	4	3,9
	Project Manager Skill	4	4	3	5	5	3	5	4,1
Project Team	Team Commitment and Participation	5	4	3	5	5	5	5	4,6
External Environment	Rules and Regulation	4	4	3	4	5	3	5	4,0
	Subject Matter Expert (SME)	4	4	4	4	5	4	5	4,3
	Tren	3	4	3	3	5	2	4	3,4

Four experts add to the success factors of an IT project. P1 proposes the Project Management Maturity (PMM) level as an additional success factor for IT projects; this is in line with the literature, which explains that the higher the PMM level, the higher the success rate [26]. The type of industry is a supporting factor related to PMM and organizational performance; it also explained that the service sector, especially the IT industry, has benefited the most from high PMM rates [27]. Furthermore, P1 describes the project team's roles and responsibilities in understanding its goals and objectives. The discussion results revealed the Project Team factor could add to the team's roles and responsibilities. P3 provides additional factors related to team knowledge; the discussion results revealed that these factors could be added to the Project Team factor. Furthermore, P7 provides additional factors related to the technology used, but these factors have the same meaning as the discussion's knowledge team factor. Organizational factors in communication and project team commitment and participation are the factors that have the highest average. On the other hand, the Project factor in duration and the External environment factor in the trend are the factors that have the lowest average value.

E. *Analysis*

From the results of interviews with project teams from the PNMIM organization as well as from IT vendors, document observations, literature studies using SLR, as well as the results of interviews and assessments conducted by experts, then an analysis regarding the criteria and success factors of IT projects in application development in the most suitable FSI in the digital transformation era.

Fig 2 outlines the criteria and success factors for an IT project. In the IT project success criteria, eight criteria are obtained from [10]; the PMBOK® guide defines the project success criteria: cost, communications, quality, risk, scope, stakeholder, resource, and time. Process criteria are obtained from [8], which explains that process performance is related to project completion on time and within budget. Commitment criteria are criteria added by expert judgment; project success should not be separated from all parties' commitment and activities in the project.

The factors and sub-factors of IT project success were obtained from SLR and expert judgment. There are five factors and eighteen subfactors, three of which are included in the expert judgment. The following describes an explanation for each of the success factors of an IT project.

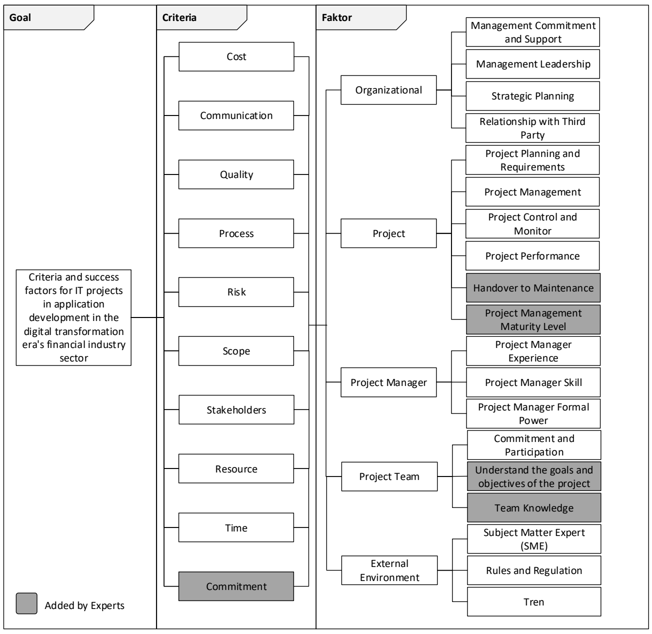

Goal
Criteria
Faktor
Criteria and success factors for IT projects in application development in the digital transformation era's financial industry sector
Cost
Communication
Quality
Process
Risk
Scope
Stakeholders
Resource
Time
Commitment
Organizational
Project
Project Manager
Project Team
External Environment
Management Commitment and Support
Management Leadership
Strategic Planning
Relationship with Third Party
Project Planning and Requirements
Project Management
Project Control and Monitor
Project Performance
Handover to Maintenance
Project Management Maturity Level
Project Manager Experience
Project Manager Skill
Project Manager Formal Power
Commitment and Participation
Understand the goals and objectives of the project
Team Knowledge
Subject Matter Expert (SME)
Rules and Regulation
Tren
Added by Experts

Fig. 2. *Structure of Success Criteria and Factor for IT Project Application Implementation in Digital Transformation Era*

1) Organizational: this factor is used by literature [14] and [20], projects are the primary way to create value and benefits in organizations, organizational leaders must be able to manage with tighter budgets, shorter schedules, scarcity of resources, and changing technology with fast [10], also added by experts that the role of the organization for IT projects in the FSI has an important role related to cooperation and communication.

2) Project: this factor is found in all SLRs used in this research; the experts added the handover to the maintenance sub-factor, with the reason that if the user has used the IT project, minimal bug maintenance is one of the success factors of IT projects.

3) Project Manager: three SLRs use this factor as the success of IT projects, [14], [19], [21], and also explained in discussions with experts that the ability of project managers to handle a project at FSI in the era of digital transformation has challenges separately to be able to communicate with all parties to align business needs with the technology used.

4) Project Team: this factor is found in all SLRs used in this research; the experts add sub-factors Understand the goals and objectives of the project and Team Knowledge; the reason for adding these two factors is because the team's understanding and ability in IT project are one of the factors that can support success IT project.

5) External Environment: two SLRs use this factor as the success of IT projects, [14] and [19], which were also described in discussions with experts; in Indonesia, the regulatory authority at the FSI is regulated by the government, in this case, the OJK.

V. CONCLUSION

This research is to identify the criteria and success factors for IT projects in application development at FSI in the digital transformation era, with a case study of SiJAGO development in PNMIM organizations. It is hoped that in the subsequent development of SiJAGO, PNMIM, and IT vendors as SiJAGO developers can focus on the criteria and factors that can support IT project development's success. The results of discussions with experts at FSI, the Subject Matter Expert (SME) factor has an essential role in the success of the SiJAGO project. Organizations that will transform a business that was initially done manually to digital need an SME regarding the processes and rules that must be carried out.

The organization is a factor that is no less important in IT projects' success, in communicating and collaborating with third parties, such as banks or payment gateway vendors. The rules and regulation factor are a factor that cannot be ignored in IT projects at FSI because the development of financial technology, otherwise known as fintech, must follow the rules and regulations of the Indonesian government, in this case, the OJK.

This research is limited to IT projects' criteria and success factors in application development at FSI with the SLR method and expert judgment. As a criterion and success factor for IT projects at FSI in Indonesia, it cannot be used as a reference because it only represents the fintech field. Expanding the scope of IT project research in other fields and CSF ranking can improve further research quality. The parties involved in project work can prioritize strategies in project management.

REFERENCES

[1] E. Tandelilin, *Pasar modal : manajemen portofolio & investasi*. 2017.

[2] J. T. Marchewka, *Information technology project management: Providing measurable organizational value*. 2015.

[3] OJK, "Peraturan Otoritas Jasa Keuangan Nomor 77/POJK.01/2016 tentang Layanan Pinjam Meminjam Uang Berbasis Teknologi Informasi," *Otoritas Jasa Keuangan*. 2016.

[4] S. Mohapatra, "Best Practices in Software Maintenance Projects," *Int. J. It/bus. Alignment Gov.*, 2013, doi: 10.4018/jitbag.2013010102.

[5] "CHAOS Reports - The Standish Group." [Online]. Available: https://www.standishgroup.com/chaosReport/index. [Accessed: 01-Mar-2021].

[6] D. Olson, *Information Systems Project Management*. Business Expert Press, 2014.

[7] Axelos and S. Office, *Managing Successful Projects with PRINCE2*. Stationery Office, 2017.

[8] D. Pimchangthong and V. Boonjing, "Effects of Risk Management Practice on the Success of IT Project," *Procedia Eng.*, vol. 182, pp. 579–586, 2017, doi: https://doi.org/10.1016/j.proeng.2017.03.158.

[9] A. Ć. Hasibović and A. Tanović, "PRINCE2 vs Scrum in digital business transformation," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2019, pp. 1514–1518, doi: 10.23919/MIPRO.2019.8756716.

[10] Project Management Institute, *A guide to the Project Management Body of Knowledge (PMBOK guide)*, vol. 45. Newton Square, PA: Project Management Institute, 2017.

[11] R. A. Dickinson, C. R. Ferguson, and S. Sircar, "Critical Success Factors and Small Business," *Am. J. Small Bus.*, vol. 8, no. 3, pp. 49–57, Jan. 1984, doi: 10.1177/104225878400800309.

[12] D. R. Daniel, "Management Information Crisis," *Harvard Business Review*, vol. 39, no. 5. Harvard Business School Publ. Corp., Boston, Mass., pp. 111–121, 1961.

[13] J. F. Rockart, "Chief executives define their own data needs.," *Harv.*

Bus. Rev., vol. 57, no. 2, pp. 81–93, 1979.

[14] A. Priambodo, P. W. Handayani, and A. A. Pinem, "Success Factor for IT Project Implementation in Banking Industry: A Case Study," in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, 2019, pp. 1–5, doi: 10.1109/ICICoS48119.2019.8982404.

[15] Z. A. Al-Sai, R. Abdullah, and M. H. Husin, "Critical Success Factors for Big Data: A Systematic Literature Review," *IEEE Access*, vol. 8, pp. 118940–118956, 2020, doi: 10.1109/ACCESS.2020.3005461.

[16] O. Werth, D. R. Cardona, J. Nowatschin, M. Werner, N. Guhr, and M. H. Breitner, "Challenges of the financial industry - An analysis of critical success factors for FinTechs," in *25th Americas Conference on Information Systems, AMCIS 2019*, 2019.

[17] M. Jelassi, S. Ben Miled, N. B. Ben Saoud, and J. Demongeot, "Obesity determinants: A systematic review," in *2015 Third World Conference on Complex Systems (WCCS)*, 2015, pp. 1–6, doi: 10.1109/ICoCS.2015.7483277.

[18] R. Hassani and Y. El Bouzekri El Idrissi, "A framework to succeed IT project management in an era of digital transformation," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 630–636, 2020, doi: 10.30534/ijatcse/2020/88912020.

[19] R. Octavianus and P. Mursanto, "The analysis of critical success factor ranking for software development and implementation project using AHP," in *2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*, 2019, pp. 313–318, doi: 10.1109/ICACSIS.2018.8618147.

[20] R. N. Kasayu, A. N. Hidayanto, and P. I. Sandhyaduhita, "Critical success factors of software development projects using analytic hierarchy process: A case of Indonesia," *Int. J. Innov. Learn.*, vol. 22, no. 1, pp. 1–22, Jan. 2017, doi: 10.1504/IJIL.2017.085245.

[21] O. P. Sanchez, M. A. Terlizzi, and H. R. de O. C. de Moraes, "Cost and time project management success factors for information systems development projects," *Int. J. Proj. Manag.*, vol. 35, no. 8, pp. 1608–1626, 2017, doi: 10.1016/j.ijproman.2017.09.007.

[22] J. Recker, *Scientific Research in Information Systems: A Beginner's Guide*. Springer Publishing Company, Incorporated, 2012.

[23] R. K. Yin, *Case Study Research and Applications: Design and Methods*. SAGE Publications, 2017.

[24] U. Sekaran and R. Bougie, *Research methods for business a skill-building approach*. 2016.

[25] Sugiyono, "Metode Penelitian Kuantitatif,Kualitatif dan R&D," in *ke-26*, 2018.

[26] M. E. Mullaly and J. Thomas, "Re-thinking project management maturity: perspectives gained from explorations of fit and value," in *PMI Research and Education Conference*, 2010.

[27] R. Busse, H. Zafer, and M. Warner, "Rethinking the roles of project management maturity and organisational culture for perceived performance: an empirical study based on German evidence," *Eur. J. Int. Manag.*, vol. 14, p. 730, Jan. 2020, doi: 10.1504/EJIM.2020.107605.

Simulation of the Impact of a Submerged Jet with a Shield During Metal Welding for Use in the Arctic

Alexey Lagunov
Fundamental and Applied Physics Department
Northern (Arctic) Federal University
named after M.V. Lomonosov
Arkhangelsk, Russian Federation
a.lagunov@narfu.ru

Anton Losunov
Radio monitoring center
Northern (Arctic) Federal University
named after M.V. Lomonosov
Arkhangelsk, Russian Federation
a.losunov@narfu.ru

***Abstract*— The Arctic has a unique resource potential. Many countries around the world are ready to explore this territory for mineral exploration. This circumpolar region has been covered with many summer ice for a long time. Recently, due to global warming, the amount of ice in the Arctic has decreased significantly, but still, it is pretty enough to make it difficult for navigation. Delivery of extracted minerals in this region is possible only by sea transport. For navigation in the Arctic, icebreakers or ice-class vessels are required. In the manufacture of such ships, welds are most often used. For the seams not to oxidize during welding, it is necessary to carry out this process by blowing the seam with inert gas, for example, carbon dioxide. The method of selecting the parameters of a gas jet is very complex. Our work's theme is the impact of a submerged jet with a shield during metal welding for Arctic use. We have created a mathematical model that simulates various operation modes and chooses the optimal one.**

Keywords—computer, modeling, welding, gas, shield, Arctic

I. Introduction

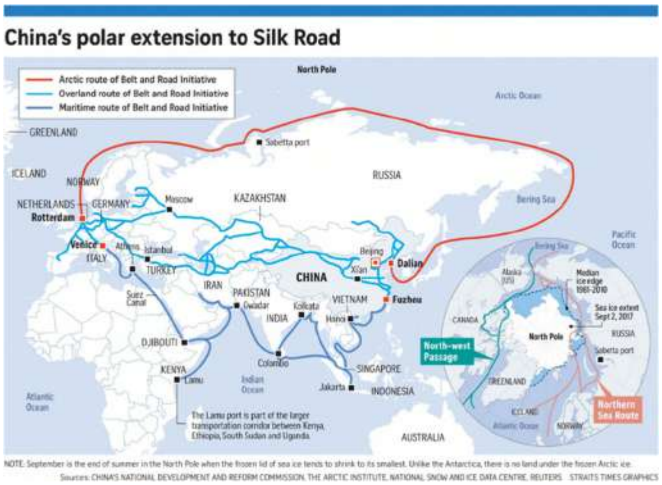
The territories adjacent to the Arctic zone occupy a special place in the system of strategic national interests of circumpolar countries in economy and transport, environmental protection, innovation, defense, and geopolitics. This region's unique resource potential allows, subject to the formation of a particular system of state regulation of the economy, to ensure dynamic, sustainable development of both the Arctic regions and the country.

Minerals produced in the Arctic, their reserves and forecast reserves explored in the 20th century, constitute the main part of the circumpolar region countries' mineral resource base. According to calculations of the U.S. Geological Survey, the last pantry of the Earth contains 30 % (47 trillion cubic meters) of the world's natural gas reserves, 13 % (90 billion barrels) of oil, 9 % of coal, as well as significant volumes of metals (uranium, copper, titanium, silver, gold) and other rare minerals (diamonds, graphite). In addition to its resource potential, which has yet to be clarified, the Arctic opens up new opportunities in the field of transportation [1]. For example, the sea route from Asia to Europe through the Arctic waters is average two times shorter than the route currently used through the Suez Canal.

Long-term forecasts of the circumpolar region countries' economic development show that the total forecast estimate of recoverable hydrocarbon resources of the continental margin of the Arctic Ocean made by Academician I.S. Gramberg is about 110 billion tons of fuel equivalent. This significantly exceeds the reserves of the continental margins of each of the Earth's oceans.

The mined minerals can only be delivered by sea transport. Because the Arctic Ocean is almost constantly covered by ice fields [2], shipping in the Arctic is risky. Over the past decades, the summer ice cover in the Arctic has significantly decreased. Some scientists make predictions about global warming, which may substantially reduce ice fields and ice thickness so much that navigation in this region will be possible for non-ice class ships.

Simultaneously, facts and more pessimistic forecasts suggest some reduction in ice thickness in the Arctic and the possibility of organizing navigation. For example, the government of the People's Republic of China has adopted a document "China's Arctic policy" [3], which proposes to add a "blue economic corridor" linking China and Europe through the Arctic Ocean to the 21st century Maritime Silk Road (Fig. 1).


China's polar extension to Silk Road

NOTE: September is the end of summer in the North Pole when the frozen lid of sea ice tends to shrink to its smallest. Unlike the Antarctica, there is no land under the frozen Arctic ice.

Sources: CHINA'S NATIONAL DEVELOPMENT AND REFORM COMMISSION, THE ARCTIC INSTITUTE, NATIONAL SNOW AND ICE DATA CENTRE, REUTERS   STRAITS TIMES GRAPHICS

Fig. 1. The maritime silk road of the 21st century.

The Suez and Panama canals' capacity, which currently carry the main flow of maritime traffic, is limited. This fact requires a

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

search for other short and safe ways of cargo transportation. In addition to the best-known shipping routes, there are two different routes in the Arctic: the "Northwest Passage" (Fig.2, red line), the "Middle (Polar) Passage" (Fig. 2, green line), and the "Northeast Passage," which has a safer component, the "Northern Sea Route" (Fig. 2, turquoise). (Fig. 2, turquoise line). These transport corridors, which allow the transport of goods from Asia to Europe and back, and the delivery of energy resources, are shorter than traditional transport corridors. The route from China to Western Europe along the NSR is about 8,100 nautical miles long. The way through the Suez Canal is 2,400 miles longer. If you go around Africa, you need to add more than 4 thousand miles.

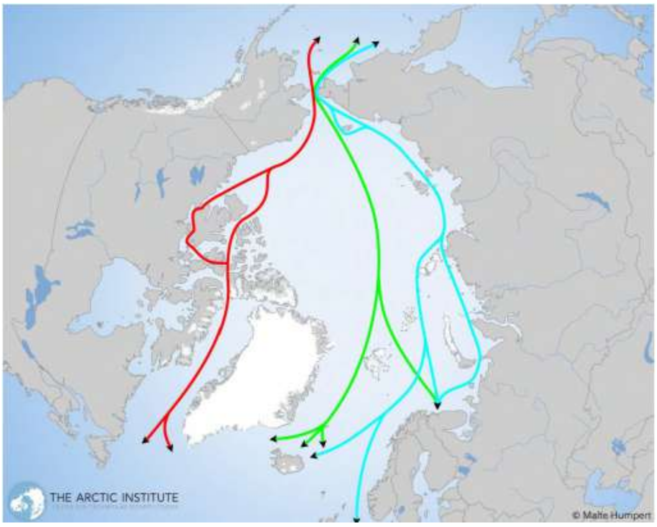


Fig. 2. Arctic shipping routes.

However, the use of these transport corridors affects states' geopolitical security, some of whose zones of influence extend along these transport corridors (Fig.3). Geopolitical security describes the extent to which international relations (both pan-Arctic and global) are cooperative or contradictory; and the level to which political tensions can affect demand for and access to Arctic resources, militarization efforts, and, ultimately, maritime operations. Instability in some regions affects fuel prices and has a direct impact on naval logistics and planning. Also, control of sea lanes (i.e., protecting shipping lanes worldwide is a costly undertaking) affects maritime transport in the Arctic region.

This key factor involves the complex role that energy prices play in increasing maritime operations' profitability in the Arctic. On the one hand, the entire naval industry is strongly affected by high fuel prices and the resulting increase in operating costs. On the other hand, the fossil fuel sectors are driven by increased profitability as the oil per barrel price rises. While high bunker fuel prices encourage the use of trans-Arctic routes, the extent to which Arctic routes are used in practice depends on inevitable tradeoffs in transit times: the need to reduce speed in ice-covered areas may tip the scales in favor of alternative, traditional maritime routes. While this factor is related to the critical factor "Global Economic Trends," it highlights the importance of energy price trends in creating economic incentives to increase or decrease activity in the Arctic region.

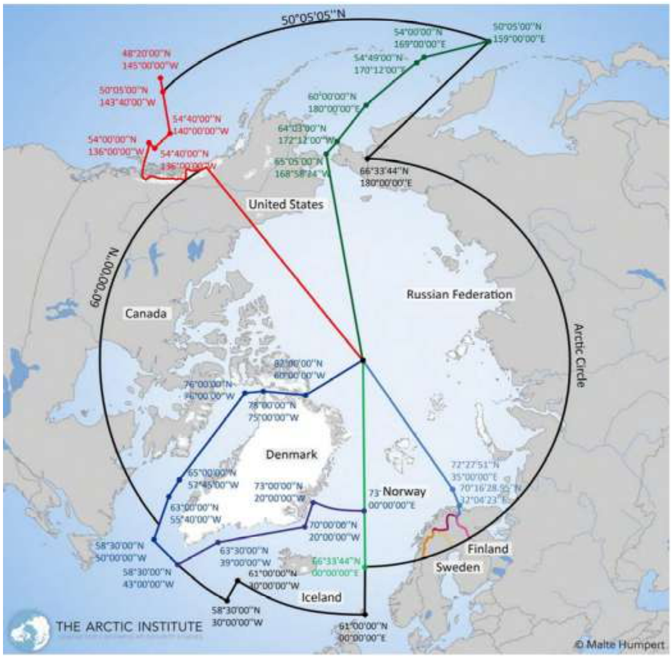


Fig. 3. Distribution of zones of influence among countries.

Increased activity in the Arctic region concerning navigational safety is impossible without the countries that will use the Northwest and Northeast Passages with an icebreaker fleet. For polar nations, icebreakers are strategic to national sovereignty and crucial in opening up safe waterways for other shipping to supply northern indigenous communities, explore resources and study climate change.

For the construction of icebreakers, it is necessary to have high-quality metal welding. In arc welding under the circumpolar region's harsh conditions, the weld metal is formed from molten metal. Despite the short duration of the metal in the liquid state, it has a detrimental effect on the weld's quality. The contact of the molten metal with air leads to the oxidation of the metal and nitrogen and hydrogen dissolution. This contributes to the appearance of pores and cracks in the welded seam, reduces the seams' plasticity, and reduces the structure's efficiency.

Protection of the welding zone from atmospheric air is a condition for ensuring high-quality welds in all arc welding methods. Gas-shielded welding is one of the most common types of welding. It is widely used in all industrial countries globally since it makes it relatively easy to mechanize technological processes. We have already considered this issue in the works [4,5]. In this paper, we will consider modeling the collision of a flooded nozzle scene.

II. THEORETICAL STUDY

When the welding torch is positioned at an angle to the surface to be welded, the gas shielding zone configuration undergoes some changes: the burner angle α plate and the burner symmetry axis. According to [6], in the range of angles,

$\alpha < 25^0$ the protection zone's size practically does not change. At angles, $\alpha > 25^0$ $\alpha \approx 55^0 - 65^0$. At curves, the $\alpha > 70^0$ effectiveness of the gas protection is completely compromised. If the length of the protection zone is denoted by b, then the function b/DZ within $15^0 < \alpha < 65^0$ can be approximated by the following expression:

$$\frac{b}{D_3} = 1 + 0{,}2 \cdot \left(\frac{\alpha}{65}\right)^2 \tag{1}$$

Following [6], let us outline an approximate solution for the impact of a flooded turbulent jet with a boundless plane surface. Schematically, the jet flow is divided into three regions (Fig.): 1) a free jet, 2) a jet turning zone, and 3) a jet flowing over a solid surface.

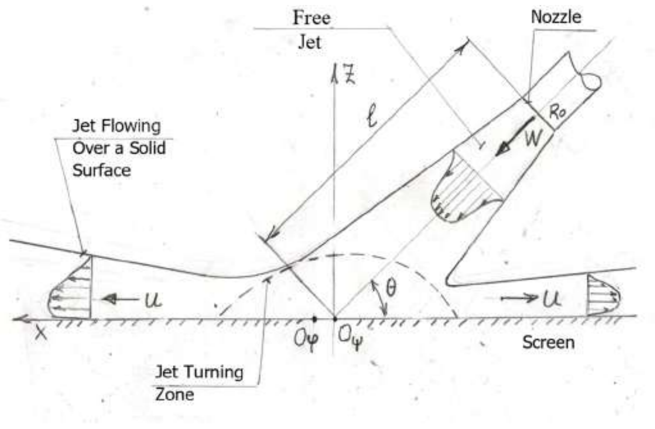


Fig. 4. Distribution of zones of influence among countries.

The calculation scheme is based on experimental data characterizing the flow in each of these areas. Experiments have shown that the static pressure on the jet axis up to the turning zone practically does not differ from the pressure on the free flooded jet's axis, i.e., up to the turning zone, the flow parameters obey the laws of the free plane. The turning zone's flow is characterized by a significant change in pressure and a large curvature of the current lines. This zone has a transverse size on the order of the free jet's diameter before it comes into contact with the screen. Measurements of static pressure on the screen at different angles of impact of the plane with the screen θ showed that outside the turning zone, it is close to atmospheric, and at the exit from this zone does not exceed 3% of the velocity head in the free jet at the entrance to the turning zone. The jet's flow area, flowing on the screen surface, is characterized by constant pressure, almost equal to atmospheric pressure. The movement of liquid in this zone has a radial character, and the center of flow O_Ψ is the point of intersection of the axis of the free jet with the plane of the screen.

The line of maximum velocity in the jet above the screen surface corresponds to the exit from the turning zone and is close to a circle. Thus, the flowing jet flow is realized as if it spreads from a cylindrical circular source of variable height b*. The center of this source O_φ is displaced relative to the point O_Ψ by a distance (Fig.5). The results of measuring the vertical velocity distribution at different screen points showed that the relative velocity profiles u^0 in the close point coordinates z^0 are similar. The velocity profile is well described by Schlichting's formulas [7].

$$u^0 = \left(1 - \xi^{1,5}\right)^2 \text{ωηερε}\ \xi = \frac{z - \delta}{b\ - \delta} \quad \text{ατ}\ z \geq \delta$$

$$u^0 = \left(z / \delta\right)^{0,1} \ \text{ατ}\ z \leq \delta \tag{2}$$

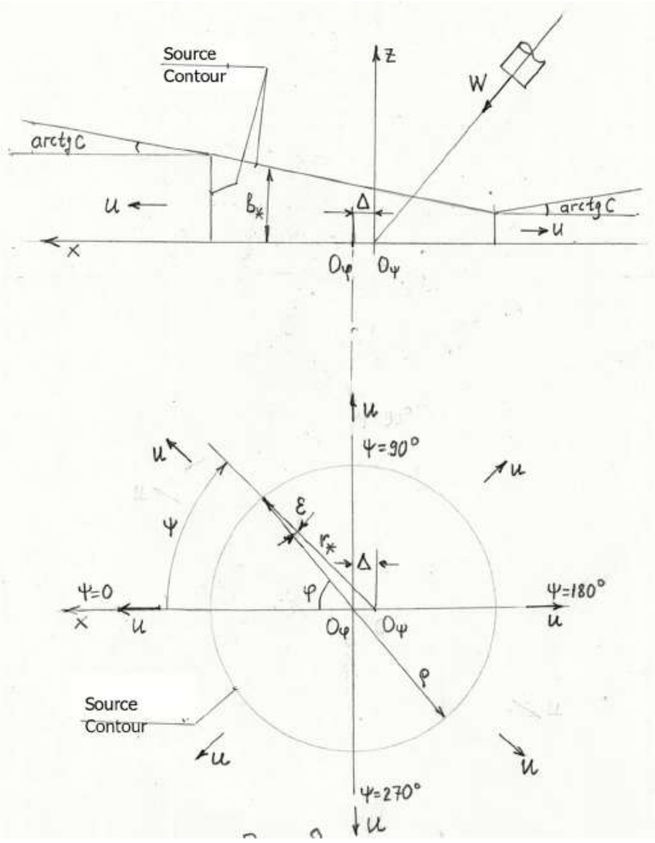


Fig. 5. Jet spread pattern.

Here δ is the thickness of the boundary layer; b is the thickness of the jet. Let's denote by $l^0 = l / R_0$, where l is the distance from the exit section of the nozzle to the screen, measured along the jet's axis, R0 is the radius of the nozzle. Experiments have shown that for all values of parameters and θ l^0 the thickness of the boundary layer along the screen is approximately 5 to 10% of the flowing jet's thickness. With distance from the center of propagation, $O_\Psi\ \psi$), increases linearly:

$$\frac{b - b_*(\psi)}{r\ - r_*(\psi)} = C \tag{3}$$

Where are $b_*(\psi), r_*(\psi)$ some initial values of the jet thickness and radius. It follows from the experimental data that the angle of inclination of the boundaries of the flowing jet concerning the screen's plane is equal to arctg C and does not depend on the impact angle θ and the angle ψ of. The velocity at the source outlet is constant and equal to u_{m*} . The outflow occurs in the direction of the radii drawn from the point O_Ψ . According to the experimental data, it is possible to assume that the circular source's upper base is a plane that makes some angle with the plane of the screen. In this case, the height of the formations of the circular source, depending on the direction of propagation, will be expressed by the relation:

$$b_* = A + B \cdot \cos\varphi \qquad (4)$$

The angle φ differs from the angle by the value $\varepsilon\ \psi$, depending on ψ . The radius of the cylindrical source ρ R* when approaching the screen. Thus, to determine the flow at the exit from the turning zone, it is necessary to find three geometrical parameters A, B, Δ , which depend on the angle of impact of the jet with the screen plane θ and the shape of the longitudinal velocity profile in the free plane before the turning zone, and one kinematic parameter u_{m*} , which depends on the value and distribution of velocity W in the free jet at the entrance to the circular source. The laws for the free plane determine the jet parameters before the turning zone. This is justified because the screen's presence does not significantly affect the flow in the jet up to the turning zone. Assume that the flow rate and kinetic energy of the flow are conserved during the turning process since the turning zone is short and there are no significant energy losses. With this in mind, we can write that

$$\int_{t^*} \int_0^{b^*} u \cos\varepsilon \cdot dz dt = \int_{s_0} W ds$$

$$\int_{t^*} \int_0^{b^*} u^3 \cos\varepsilon \cdot dz dt = \int_{s_0} W^3\ ds \qquad (5)$$

where

ε - is the angle between the radius vectors and $r_*\ \rho$, drawn in the plane of the screen from points and O_φ , respectively O_Ψ, t_* - the boundary of the circular source, b_* - the height of the formations of the cylinder, z - coordinate axis normal to the plane of the screen, s_0 - cross-sectional area of the free jet at the entrance to the turning area. The left parts of the above equations represent the flow rate and kinetic energy at the liquid exit from the turning zone, and the right parts represent the same values in the jet when approaching the turning zone. Assume that the velocity profiles do not depend on the angle φ , that is, $u_*(z^0) = u_{m*} \cdot f(z^0)$ where $z^0 = z / b_*$, $W_*(R^0) = W_{m*} \cdot f_0(R^0)$ where $R^0 = R / R_*$.

Then the flow and energy equations can be converted to the following form:

$$\int_0^1 f(z^0) dz^0 \int_{t^*} b_* \, u_{m*} \cos\varepsilon \cdot dt = 2\pi \cdot R_*^2 \int_0^1 f_0(R^0) R^0 \cdot W_{m*} dR^0$$

$$\int_0^1 f^3(z^0) dz^0 \int_{t^*} b_* \, u^3{}_{m*} \cos\varepsilon \cdot dt = 2\pi \cdot R_*^2 \int_0^1 f_0{}^3(R^0) R^0 \cdot W^3{}_{m*} dR^0 \qquad (6)$$

where

R is the radial distance from the axis of the free jet to an arbitrary point, R_* - boundary radius of the free plane before the turning zone, W_{m*} - maximum speed (on the axis) in the aircraft before the turning zone, u_{m*} - is the maximum velocity at the outlet of the annular source, which does not depend on the angle of φ . If we take out and u_{m*} W_{m*} flow and energy, we obtain:

$$u_{m*} = \alpha \cdot W_{m*}$$

$$\alpha = \{ \int_0^1 f_0{}^3(R^0) R^0 dR^0 \cdot \int_0^1 f(z^0) dz^0 \}^{0,5} \cdot (\int_0^1 f_0 (R^0) R^0 dR^0 \cdot \int_0^1 f^3(z^0) dz^0)^{-0,5} \qquad (7)$$

To find the parameters A and B, characterizing the height of the cylindrical source, we will use the conditions of conservation of flow and momentum projection on the screen's plane. Let us write down the equations of conservation of flow and momentum projection, taking into account the similarity of velocity profiles:

$$\int_0^1 f(z^0) dz^0 \int_{t^*} b_* \cdot u_{m*} \cos\varepsilon \cdot dt = \int_0^1 f_0(R^0) R^0 dR^0 \cdot \pi R_*^2 \cdot W_{m*} \quad ,$$

$$\int_0^1 f^2(z^0) dz^0 \int_{t^*} b_* \, u^2{}_{m*} \cos\phi dt = \cos\theta \cdot (\int_0^1 f^2{}_0(R^0) R^0 dR^0 \cdot \pi R_*^2 \cdot W^2{}_{m*}) \cdot$$

Converting these ratios, we obtain

$$2\rho \int_0^\pi (A + B\cos\varphi)\cos\varepsilon \cdot d\varphi = \frac{\beta}{\alpha} s_0 \qquad ;$$

$$2\rho \int_0^\pi (A + B\cos\varphi)\cos\varepsilon \cdot \cos\phi \cdot d\varphi = s_0 \cdot \cos\theta \cdot \frac{\gamma}{\alpha^2} \quad .$$

Here

$$\beta = 2(\int_0^1 f_0 (R^0) R^0 dR^0) \cdot (\int_0^1 f(z^0) dz^0)^{-1}$$

$$\gamma = 2(\int_0^1 f_0^2(R^0) R^0 dR^0) \cdot (\int_0^1 f^2(z^0) dz^0)^{-1} \qquad (8)$$

Let us express $\cos\psi$ as a function of the angle and φ Δ . From the geometric diagram of the reversal (Fig.5), we can derive the relation $\cos\varepsilon = \dfrac{\rho + \Delta\cos\varphi}{\sqrt{\rho^2 + 2\rho\Delta\cos\varphi + \Delta^2}}$.

Using these relations and neglecting the change compare $\cos\varepsilon$ to $\cos\psi$, we obtain the flow and momentum equations, respectively, in the following form:

$$A\left(l_0 + \Delta^0 l_1\right) + B\left(l_1 + \Delta^0 l_2\right) = \frac{s_0\beta}{2\rho\alpha}$$

$$A\left(l_1 + \Delta^0 l_0\right) + B\left(l_2 + \Delta^0 l_1\right) = s_0 \cdot \cos\theta \cdot \gamma \cdot \frac{1}{2\rho\alpha^2}$$

Here $\Delta^0 = \dfrac{\Delta}{\rho}$, $l_j\left(\Delta^0\right) = \displaystyle\int_0^{\pi} \frac{\cos^j\varphi \cdot d\varphi}{\sqrt{1 + 2\Delta^0\cos\varphi + {\Delta^0}^2}}$, where j = 0, 1, 2. (9)

Dependencies ca $l_j(\ \Delta^0\ )$ be obtained by numerical integration. Calculations have shown that at $A \approx \dfrac{s_0\beta}{2\rho\alpha\left(l_0 - \Delta^0 l_1\right)}$ $0 \le \Delta^0\ \le 0{,}85$. This formula is simplified:

$$A \approx \frac{s_0\beta}{2\rho\alpha\pi} \qquad (10)$$

This result means that the turning area's average height is almost independent of the angle θ at which the jet collides with the screen. Let us determine the fourth unknown - the value of Δ^0 . By analogy with an ideal fluid, let us assume the integral form about conservation of the flow rate in the angle element before and after the turn. Shows the jet section before the turning zone and the section at the base of the annular source. Drawing in the bottom of the cylinder's plane through the point O_Ψ the line a-a, perpendicular to the axis X, we divide the liquid flow into two parts. Since the issue O_Ψ is the quasi-source of the jet, all of the fluid to the right of a-a will move to the right, and the liquid to the left of this line will move to the left. In the jet's cross-section s_0 at a distance Δ_* from the plane's axis, it is also possible to draw a line $a_* - a_*$ perpendicular to the axis X, which similarly divides the flow in the jet. Let us assume by analogy with the theory of an ideal fluid that the distances Δ and are Δ_* proportional to the radii of the corresponding circles $\dfrac{\Delta}{\rho} = \dfrac{\Delta_*}{R_*} = \Delta^0$. Based on the adopted flow diagram, write the flow rate balance. The flow rate of gas flowing from the turning area is equal to $G_* = \displaystyle\int_\lambda^\pi \rho \int_0^{b*} u(z)dzd\omega$, where $\omega = \pi - \varphi$, $\lambda = \arccos\Delta^0$. Since $b_* = A - B\cos\omega$ it makes sense to represent G_* as a sum of $G_* = G_{*1} + G_{*2}$. Here $G_{*_1} = A\displaystyle\int_0^\rho u(\omega)dr\int_\lambda^\pi d\omega$, $G_{*_2} = -\displaystyle\int_0^\rho u(\omega)dr\int_\lambda^\pi B\cos\omega d\omega$.

The flow rate of gas flowing into the turning area is equal to $G = G_1 + G_2$, where $\sigma = \Delta^0 / \cos\omega$ $G_1 = \displaystyle\int_\lambda^\pi\int_0^{R_*} u(R)\cdot R \cdot dR \cdot d\omega$ $G_2 = \displaystyle\int_0^\lambda\int_0^\sigma u(R)\cdot R \cdot dR \cdot d\omega$

From the comparison of these relations, it follows that $G_1 = G_{*_1}$, and since $G_* = G$, and $G_2 = G_{*2}$ or else: $\displaystyle\int_0^\lambda\int_0^\sigma u(R)\cdot R \cdot dR \cdot d\omega = -\int_0^\rho u(\omega)dr\int_\sigma^\pi B\cos\omega \cdot d\omega$.

Converting this expression and neglecting the change in $\cos\varepsilon$, we obtain equality of costs through the unshaded part of the sections: $B\sqrt{1 - {\Delta^0}^2} = \dfrac{R_*^2 \cdot W_{m*}}{\rho \cdot u_{m*}}\left(\displaystyle\int_0^1 f\left(z^0\right) \cdot dz^0\right)^{-1} \cdot \int_0^\lambda\int_0^\sigma f_0\left(R^0\right)R^0 dR^0$.

Because of the difficulties in calculating the integral in the right-hand side of this equality, we considered two limiting cases corresponding to triangular and rectangular velocity profiles at the turning area entrance. It turned out that in a wide range of values of the jet's angle of impact with the screen, the following dependence is well performed $\Delta^0 = \cos\theta \cdot (\ 2\displaystyle\int_0^1 f_0\ \left(R^0\right)R^0 dR^0\ )^{0{,}5}$. This formula can be used in the field of $0 \le \cos\theta \le 0{,}7$. The valuecan be obtained either by numerical integration or by extrapolation using the condition for large values. Using the results obtained above and introducing the notation

$$A^0 = \frac{k \cdot A}{R_*},\ ,B^0 = \frac{k \cdot B}{R_*}\quad k = \frac{\rho}{R_*} \qquad (11)$$

Let us write out the final formulas for calculating the parameters of the ring source:

$$A^0 = \frac{\beta\pi}{2\alpha\left(l_0 - \Delta^0 l_1\right)}\ \text{ορ}\ A^0 \approx \frac{\beta}{2\alpha}\ \text{ιφ}\ \Delta^0 < 0{,}6$$

$$B^0 = \frac{-A^0\left(\Delta^0 l_0 + l_1\right) + 0{,}5\pi\cos\theta \cdot \gamma\alpha^{-2}}{\left(l_2 + \Delta^0 l_1\right)} \qquad (12)$$

The value of the coefficient k is an experimental constant and is equal to 1.5. To calculate the propagation of the jet over the screen, we will use the empirical fact of its boundaries' straightness at a constant angle of inclination to the screen surface for propagation directions. Consider the motion of a fluid element in cylindrical coordinates O_Ψ , z, r, ψ . Let us write down for the selected elementary volume of fluid the equation of change of momentum: $\int_0^{b_1} u_1^2 dz_1\, dL_1 = \int_0^{b_2} u_2^2 dz_2 dL_2 + (\ T_1 - T_2 + T_3)$, $b_1 = (r - r_*)C + b_*$, $b_2 = (r + dr - r_*)C + b_*$ $\quad dL_1 = rd\psi$, $dL_2 = (r + dr)d\psi$ where b is the height of the element, S1, S2 are the areas of the first and second faces of the component, respectively, T1 is the friction force on the bottom edge adjacent to the screen, T2, T3 - friction forces on the side faces of the element.

The action of the pressure forces is mutually balanced. After estimating the magnitude of the equation terms, we conclude that friction forces in total give no more than 5% of the importance of the change in momentum. This makes it possible to neglect the friction forces. Given the similarity of the velocity profiles and discarding the terms of the second-order of smallness, we finally obtain the differential equation $du_m^2 \cdot r[(r - r_*)C + b_*] = -u_m^2[(2r - r_*)C + b_*]dr$. Integration of the equation is performed under the boundary condition: $u_m = u_{m*} = \alpha \cdot W_{m*}$ $r = r_*$.

The result is that

$$u_m^2 = \alpha^2 \cdot W_m^2 \cdot \frac{r_*}{r} \cdot \left[1 + \frac{Ck}{A^0 + B^0 \cos\varphi} \cdot \frac{r_*}{R_*} \cdot \left(\frac{r}{r_*} - 1\right)\right]^{-1} \qquad (13)$$

Values and $\cos\varphi$ ρ are determined from geometric considerations

$$r_* = \rho \cdot \left(\sqrt{1 + \Delta^{0^2} \sin^2\psi} + \Delta^0 \cos\psi\right)$$

$$\cos\varphi = -\Delta^0 \sin^2\psi + \cos\psi \sqrt{1 + \Delta^0 \sin^2\psi} \qquad (14)$$

It follows from these formulas that at sufficiently large r, the value of the maximum velocity at an arbitrary point of the screen weakly depends on l and is determined only by the angle of impact of the jet with the screen. If the screen is located within the initial section of the plane ($l^0 = l / R_0 \leq 9$), then

$$W_{m*} / W_{m0} = 1 ; R_* / R_0 = 1 + k_1 l^0 \qquad (15)$$

If the screen is placed on the central part of the jet ($l^0 \geq 9$), then

$$W_{m*} / W_{m0} = k_3 l^0 ;\ R_* / R_0 = k_2 l^0 \qquad (16)$$

Here $k_1 = 0{,}14$; $k_2 = 0{,}22$; $k_3 = 12{,}4$. When determining the velocity profiles at the entrance to the turning zone $l^0 \leq 9$, we need to know the ordinate of the inner boundary of the mixing zone R_1 , which is found from the ratio

$$(R_0 - R_1) / R_0 = 0{,}08 \cdot l^0 \qquad (17)$$

When calculating the values of α , β , γ the importance of integrals $\int_0^1 f_0^{\,n}(R^0) R^0 dR^0$ for the initial section of the jet is calculated by the formula:

$$\int_0^1 f_0^{\,n}(R^0) R^0 dR^0 = 0{,}5(R_1^0)^2 + (1 - R_1^0)^2 \cdot \int_0^1 f_0^n(\eta)\eta d\eta + (1 - R_1^0) R_1^0 \int_0^1 f_0^n(\eta) d\eta \qquad (18)$$

Here $R_1^0 = R_1 / R_*$, $\eta = (R\ - R_1)/(\ R_* - R_1\ )$. The Schlichting formula describes the velocity profile at the source outlet

$$f_0(\eta) = (1 - \eta^{1,5})^2 \qquad (19)$$

To describe the velocity profile for further propagation of the jet through the screen, the Schlichting profile with boundary layer in mind is used

$$u^0 = \frac{u}{u_m} = (1 - \xi^{1,5})^2 \ \text{ατ}\ b \geq z \geq \delta ;$$

$$\text{ατ}\, u^0 = (z / \delta)^{0,1} \ z \leq \delta \qquad (20)$$

where b is the thickness of the jet flowing on the screen,

δ - is the thickness of the boundary layer in the jet flowing over the screen $\xi = \frac{z - \delta}{b\ - \delta}$.

III. RESULTS OF COMPUTER MODELING

The results of calculations using the obtained formulas are shown in Fig.6-7. Unfortunately, the solution given here is only marginally applicable to protective jets since it satisfactorily describes the flow only in the jet's flowing part. The size of the protection zone DZ of interest to us in this theory can be estimated very approximately by the value of the conditional source's radius ρ . The solution consists of the following main stages:

1) Selecting the computational domain

2) Setting the equations of the mathematical model

3) Setting the initial and boundary conditions

4) Setting the computational grid

5) Carrying out a numerical calculation

6) Viewing and analyzing the results obtained

The calculation was carried out using the FlowVision-HPC software package [8], which is designed for numerical modeling of three-dimensional laminar and turbulent, stationary, and unsteady liquid and gas flows.

The software package is based on the finite volume method, high-precision difference schemes, effective numerical methods, and reliable mathematical models of physical processes.

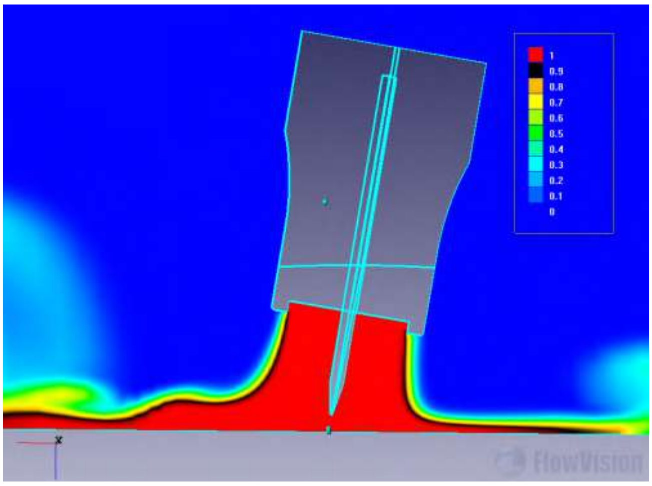


Fig. 6. Carbon dioxide concentration in the section plane.

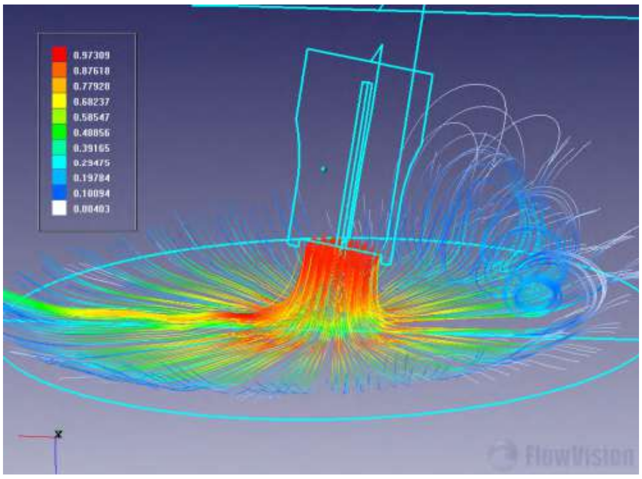


Fig. 7. Streamlines.

An experiment at the Joint Stock Company "Production Association 'Northern machine-building enterprise" has shown the proposed model's relatively high efficiency.

IV. CONCLUSION

The wide range of shielding gases used makes this method widespread concerning metals to be welded and their thickness. The main advantages of the considered welding method are as follows:

- high quality of welded joints on a variety of metals and their alloys of different thicknesses, especially when welding in inert gases due to low waste of alloying elements;
- the possibility of welding in various spatial positions;
- no operations for filling and cleaning flux and removing slag;
- the ability to observe the formation of a seam, which is especially important in mechanized welding;
- high productivity and ease of mechanization and automation of the process;
- low cost when using active shielding gases;
- the ability to weld at different ambient temperatures, which is very important in the Arctic.

This welding method's disadvantage is the complex process of adjusting the welding location, pressure, and direction of shielding gas supply. To eliminate this drawback, we created a mathematical model and implemented it in the FlowVision program. An experiment at a shipbuilding plant showed reasonably high efficiency of the proposed model.

REFERENCES

[1] T.S.Ahlbrandt, "Future petroleum energy resources of the world," *International Geology Review*, 44 (12), pp. 1092-1104, 2002, DOI: 10.2747/0020-6814.44.12.1092.

[2] B.Blair and M.Mueller-Stoffels, "Maritime futures 2035: the arctic region workshop report & technical documentation", 2019, 10.13140/RG.2.2.16234.26561.

[3] China's Arctic Policy, Full text. Available from http://english.www.gov.cn/archive/white_paper/2018/01/26/content_281476026660336.htm

[4] A.Lagunov, A.Fofanov, A.Losunov, "Peculiarities of metal welding process modeling for the Arctic," *AIP Conference Proceedings*, 1886, статья № 020096, 2017, DOI: 10.1063/1.5002993.

[5] A.Lagunov, A.Losunov, "Simulation of gas flow for welding process control in the arctic environment," *Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019*, 1, 8924233, pp. 362-367, 2019, DOI: 10.1109/IDAACS.2019.8924233.

[6] G. A. Fedorenko, "Theory of gas protection while arc welding in protecting gases," *Internet Engineering*, 224p., 2012.

[7] H.Schlichting, K.Gersten, "Boundary-Layer Theory," *Boundary-Layer Theory*, pp. 1-799. 2016.

[8] FlowVision, Official site, Available from: https://fv-tech.com/index.php/en/

[9] Joint Stock Company "Production Association 'Northern machine-building enterprise'," Official site, Available from: http://www.sevmash.ru/eng/

Smart Food Service System For Future Restaurant Using Overhead Crane

Kazi Rabiul Alam
Electrical and Electronic Engineering
International Islamic University
Chittagong,Bangladesh
kazirabiulalameee@gmail.com

Farhadul Islam
Electrical and Electronic Engineering
Chittagong University of Engineering and Technology
Chittagong,Bangladesh
farhademon77@gmail.com

Sk. Md. Golam Mostafa
Electrical and Electronic Engineering
International Islamic University
Chittagong,Bangladesh
mostafa_93eee_iiuc@yahoo.com

Md.Askar Nayen
Electrical and Electronic Engineering
International Islamic University
Chittagong,Bangladesh
askarnayenabir@gmail.com

***Abstract*— The world is globalized day by day so the demand of automated machine has increased rapidly. In restaurant or offices, factories, homes, hospitals, automation robot holds a significant contribution. In this paper we have designed and implemented a smart automated system using over head crane. As the proposed system is overhead type so it needs no extra space in the ground. Most of the smart restaurants in the world are using line following robot to serve food to customer. The installation cost of line following robots is high. For this reason, it's difficult for developing countries to build a smart restaurant. So, we have designed a smart system for serving at restaurant by using overhead crane, which is cost effective. Also due to the ongoing Covid-19 epidemic, the restaurant business is in recession. With this proposed system, it is possible to provide food directly from the kitchen to the customer's table while maintaining social distance and following hygienic rules.**

***Keywords*—PIC16F72 microcontroller; ULN2003 Motor Driver IC; IR LED; LM7805 Voltage Controller; LM2576 DC-DC Converter; Covid-19.**

I. INTRODUCTION

Automation control is the use of different control systems for running devices, such as machines, processes in plants, boilers and heat treating ovens, switching on telecommunication networks, steering and stabilizing ships, aircraft and other applications and vehicles with limited or decreased human interference. A mobile robot is a software-controlled system that uses sensors and other technologies to recognize and move around its surroundings. The use of mobile robots makes the complex and laborious tasks easy, reduces human error, increases efficiency, and enables continuous work without taking rest. We have made a frame work that is an overhead wheel based system which works with relay, IR sensor and microcontroller PIC16F72. The restaurant has been divided into a certain number of row and column sections through IR sensor. The table on which the food will be served in any row and column section of the restaurant will be calculated by the microcontroller PIC16F72. The overhead crane will collect food from the kitchen and serve it at the designated table.

II. LITERATURE RIVIEW

There have been a number of studies on smart restaurant serving in the past. Automated Restaurant Service Program Using a Robot Line Follower has developed a system on automated food serving [1]. This project can be used to maintain a restaurant to take orders and serve the same route the follower robot allows also travels with a particular course. This direction perhaps a black belt on a white frame or a white belt on a black frame, but it may be transparent to the magnetic field. This route utilized for this venture was a black one on a white frame. With the help of IR sensors, this route had tracked, also this sensor readings were utilized to operate the engine operator that was mounted over this robot which was the main reason for the robot motion.

To communicate between the robot and the customer, NRF transceivers were used. A. Jain et. al. developed a system in which robot's first task was to move around the room to find collectible locations, i.e. tables [2]. Once a table has been set up by a computer, the job is to collect the various objects on the table.

The robot was designed to identify and collect the table below. N. Malik et. al. from Moradabad Institute of Technology Moradabad in India designed a robot which were designed in such a way that they can greet the visitor, order and sever the food to the customer [3].

R.Poonguzhali et. al. from Periyar Manniammai Institute of Science and Technology designed a Wheeled Robotic System in which the LCD, Keypad and Bluetooth modules were the basis for the menu bar [4]. By using the automated menu bar to place the order the customer needs. This order was transmitted via the communication network to the kitchen and reception. Then the Waiter robot will transport food to the customer from the kitchen.

Ademola Abdulkareem et. al. from covenant university, Ota ,Nigeria developed smart autonomous mobile robot [5]. When customers arrive, they have to press the button to

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

refresh themselves. The LED at the start point and the LED at the junction from which the robot moves to work towards the table will glow as the button is pressed (switch is on). The robot began after a black line when it got a white light when it turned left or right and served the refreshmentaccordingly. Upon working, the returning black line route will lead again and arrive at the starting position.

M. Asif et. al. from Riphah International University, Islamabad designed a line following waiter robot which provides the client with adequate restaurant service [6]. If a person wants to place an order, he or she can call the robot simply by pressing a switch on his desk. The whole device made use of RF technology.

Afsheen Salmiya et. al. designed a self-sufficient waiter robo in which a waiter based on robotic technique was planned and implemented for a two-room office [7]. The waiter robot was built to accept requests from the kitchen boy via android applications, collect food / drinks (max600grams), travel to the destination (the individual sending the request) and return to the place of origin after delivery of the requested food / drinks.

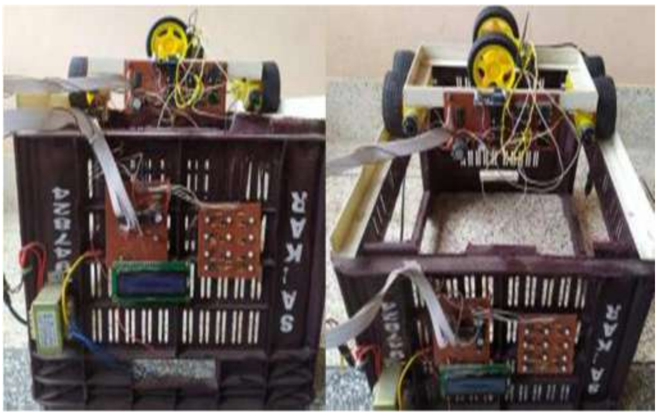
The goal is to develop a small-scale robot called the Serving Robot, which can aid in the development of robotic assistance technologies. A robot that acts as a personal assistant should be able to help in a variety of settings, whether it is a research lab, a clinic, or even at home.

- Effective and effective jobs as we do with robots:
- Reduces customer waiting time.
- One-time system investment.
- Work can be easier and can reduce labor costs.
- When customers place their own orders, the number of waiter staff may be reduced.
- Procedures are carried out with accuracy and high repeatability.
- Serving food to the customer while maintaining social and safe distance due to Covid 19 pandemic.

Considering the previous work, we have seen that each of them has basically used Line following Robot, which is very expensive. Perhaps it is better for developed countries to bear the cost. But it is difficult for developing countries to bear this high cost and also need a certain amount of space allocated for restaurants that are using line following robots. If there is any kind of involuntary obstruction in that space, the robot needs more time than the allotted time, which is undesirable for the restaurant business. In addition, the perimeter of the restaurant needs to be enlarged for the installation of robots and lines. On the other hand, for an overhead system doesn't require extra space to set up, so the restaurant has the potential to be economically viable and make the restaurant environment more attractive by setting up the required number of tables. So, a smart restaurant using overhead crane has been developed at a very low cost to keep pace with the developed world.

III. SYSTEM OVERVIEW

This prototype is microcontroller and relay based. Also IR sensor will play important role in this project.



(a) Front view (b) Top view

Fig 1. Implemented Prototype of Smart Food Service.

In this system we mainly used PIC16F72 microcontroller, ULN 2003 motor driver IC, IR LED, Liquid Crystal Display (LCD),LM7805 voltage regulator,LM2576 DC-DC Converter, Gear Motor, Limit Switch, Resistor, Capacitor, Diode. "Fig.1." shows the implemented prototype of smart food serving using overhead crane.

In this prototype, we have designed a total number of nine tables and one kitchen which is separated by four rows and three columns with the help of IR sensors. "Fig.1." shows the implemented prototype of smart food serving using overhead crane. The block diagram of proposed project is shown in "Fig.2."

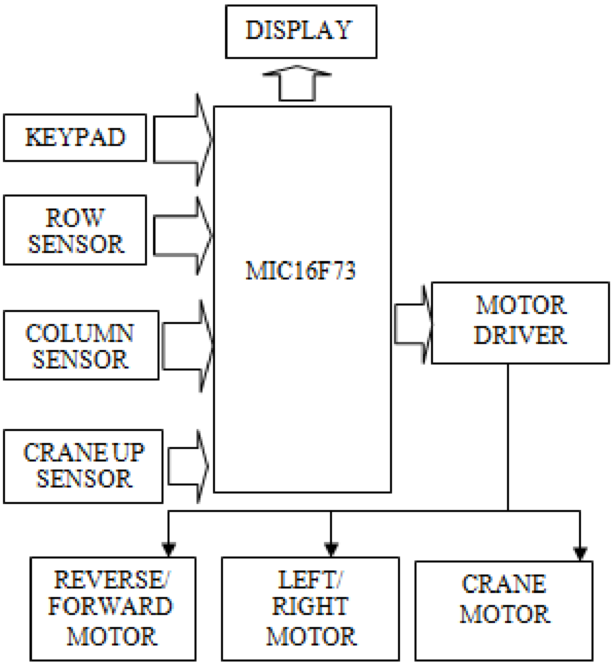


Fig 2. Block diagram.

There are total seven motors which have been used in the overhead crane, of which four motors are used to move the overhead crane to forward /backward direction, two to the right/left direction and one to up/down for collecting food from the kitchen and serving food at the designated table. After activating the system in the initial state, any table will be selected with the help of push button. Different analog voltages are generated due to the use of different resistors with push button. Converting analog voltage to digital voltage through ADC , microcontroller will Calculate the row and column number for the table. “Fig.3.” shows the flow chart of the proposed system.

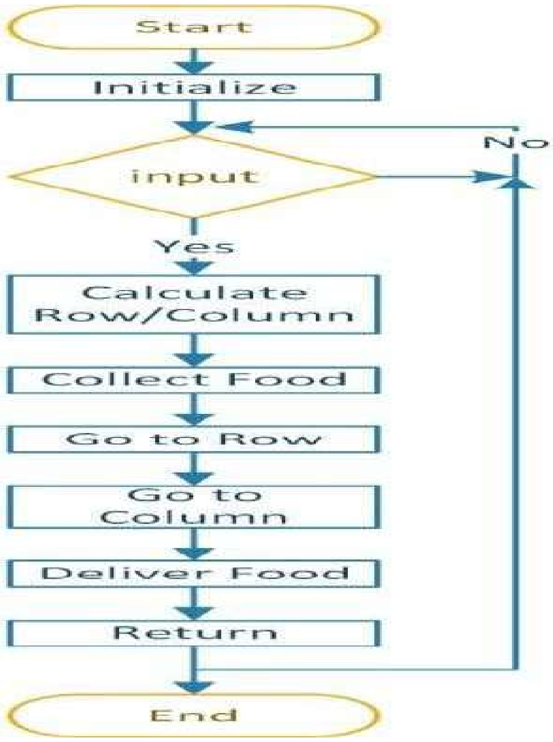


Fig 3. Flow chart.

There are seven motor connected to ULN 2003 motor driver which is connected to port C of microcontroller. Crane up /down motor connected to pin number 13 and 14. Forward/backward motor connected with 15 and 16 number pin and right/left direction motor connected to 17 and18 number pin of PIC16F72 “Fig.4.” shows the circuit diagram.

In order to serve the food, we use row sensors and column sensors in port A, port B and port C. After pressing any key, the microcontroller will measure the row and column number, then the crane will down when pin number 14(RC2) will be high. After collecting the food, the crane will up when pin number 13(RC1) will be high. It will go up until the crane limit or pin number 23 goes high then it will go forward. In this situation pin number 15 (forward direction) will high to go to row with the aid of row sensor Again the crane will go down to deliver the food and after serving it will go up and go back to middle position with the aid of right/left motor. Suppose the crane is in the left position, so it needs to go right direction to go to middle point that means pin number 17 (right direction) will high. After reaching middle point it will go back to the kitchen, for this purpose pin number 16 will high. “Fig.5.” shows the structure of over head crane.

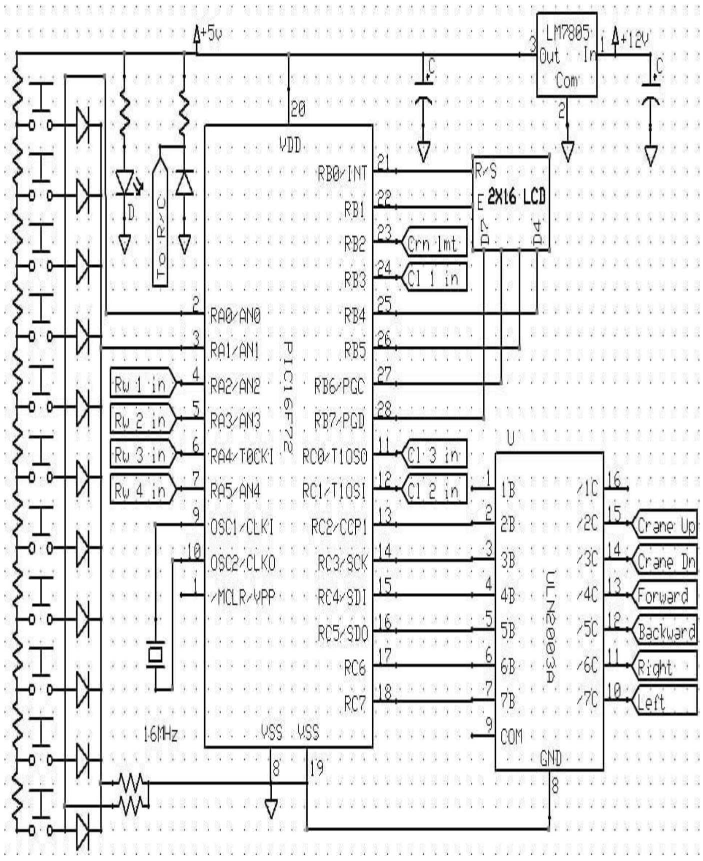


Fig 4. Circuit diagram.

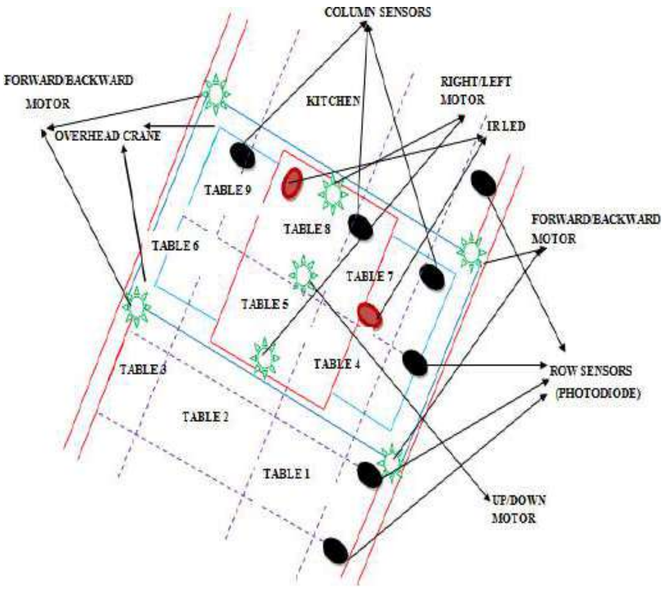


Fig 5. Structure of overhead crane.

IV. WORKING MODE AND RESULT ANALYSIS

The output of the system is shown in "Fig.6." to "Fig.9."

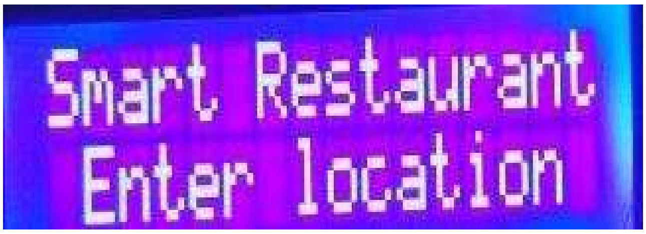

Smart Restaurant
Enter location

Fig 6. When turn on the device then it needs location.

A. *Initial State of working mode*

When we turned on the device then microcontroller send the information in main circuit and main circuit transfer the information to microcontroller circuit. Then it needs location. The program selects the row and column for different tag. Then the information is collected by touch switch then motor driver circuit will be activated. ."Fig.6." shows when we turned on the device, it needs location.

Then location will display either yes or no to confirm the table number which is shown in "Fig.7."

B. *Serving state of working mode*

When the information is sent, microcontroller software selects the row and column then photodiode can detect row and column because the PIN structure provides quick response time. PIN photodiodes are mainly used in high-speed applications and then it will collect food from the kitchen area. After taking food from kitchen, it would reach table number 1 and serving food. In the same way, we will assign a location to the number 2 table and that it is getting ready to serve. Same process it will serve food table number 3, 4, 5, 6, 7, 8 and 9. "Fig.8." shows the collecting food from the kitchen.

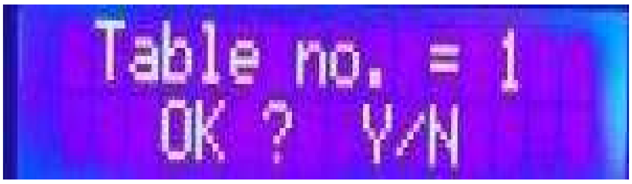


Table no. = 1
OK ? Y/N

Fig 7. Location will display either yes or no to confirm the table number.

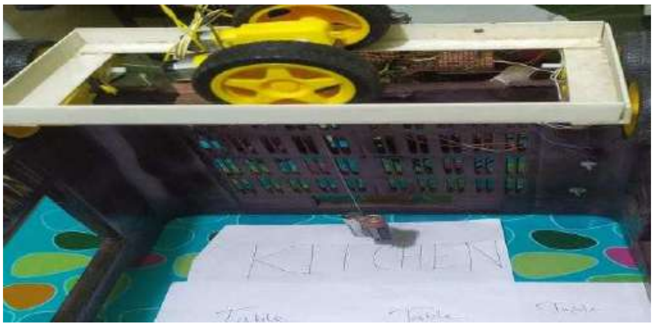


Fig 8. Collecting food from the kitchen.

The crane will go to row and column with the help of IR Sensor. The IR sensor is composed of two parts, the circuit emitter and the circuit receiver. This is known collectively as a Photo-Coupler, or Opto-Coupler. The emitter is an IR LED and the detector is a photodiode of the IR. The IR photodiode is sensitive to the IR light produced by an IR LED. The resistance to the photodiode and the output voltage change in relation to the received IR light. That is the IR detector 's basic operating theory, at last running the gear motor. "Fig.9." shows getting ready to serve food and then it delivers the food to the customers. After delivering the food, it will go back to the kitchen automatically and waiting for the next order.

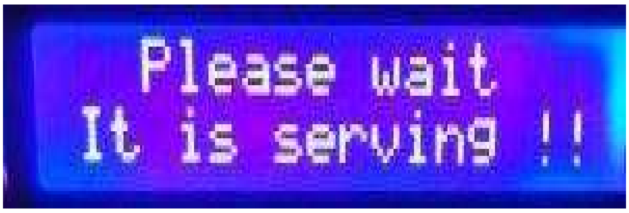

Please wait
It is serving !!

Fig 9. Getting ready to serve.

V. COST ANALYSIS

If we consider the 1st month (when the system is set up): Installment Cost = (1700+920) Tk = 2620 Tk

Variable Cost = (60+60) Tk = 120 Tk Total Cost = Installment Cost + Variable Cost

= (2620+120) Tk

= 2740 Tk = 34.25 USD

There is no installment cost after the 1st month and the overall cost is just the variable cost.

Table 1. COST CALCULATION OF THE PROPOSED SYSTEM

Cost	Equipment Description	Amount (Tk)
Installment Cost	Overhead Crane &Frame Cost	1700/-
	Additional Equipment Cost	920/-
Variable Cost	Motor, Wheel, Wire & LCD Cost	60/-
	Additional Cost	60/-

VI. CONCLUSION

The idea of delivering food using a robot is not old but there are a number of technical challenges to address. It would be the cost involved first. To convince people that this automated food delivery system is workable, the point is that people should actually compare the cost of hiring a worker and purchasing a robot. It is therefore essential to keep costs down. The proposed project is an idea to keep pace with the developed world at low cost. The system can also be used in addition to restaurants, parks, offices, hospitals or market places which will help to continue activities while maintaining social distance in the ongoing Covid-19 epidemic in the current world.

REFERENCES

[1] B. F. Lew, K. L. Tan, K. V. Goh, K. L. Lee, and Z. C. Khor Y. C. Tan, "A new automated food delivery system using autonomous track," in *IEEE Conference on Sustainable Utilization and Developement in Engineering and Technology 2010*, 2010.

[2] S. Chauhan, A. Hirlekar, and S. Sarange A. Jain, "Automated Restaurant Management System," *Int. J. Innov. Res. Electr. Electron,* vol. 4, no. 5,pp. 35-38, 2016.

[3] N. Rani, A. Singh, Pratibha, and S. Pragya N. Malik, "Serving Robot New Generation Electronic Waiter," *Int. J. Innov. Res. Sci. Technol.,* vol. 2,no. 11, pp. 775-777, 2016.

[4] R. POONGUZHALI, "Wheeled Robotic System for Restaurants," *Int. J. TRENDS Eng. & Technol.,* vol. 32, no. 1-August 2018, September 2020.

[5] V. Ogunlesi, A. S. Afolalu, and A. Onyeakagbu A. Abdulkareem, "Development of a smart autonomous mobile robot for cafeteria management," *Int. J. Mech. Eng. Technol.,* vol. 10,no. 1, pp. 1672-1685, 2019.

[6] M. Sabeel, Mujeeb-ur-Rahman, and Z. H. Khan M. Asif, "Waiter Robot-Solution to Restaurant Automation," in *MDSRC-2015*, Wah/Pakistan, 2015.

[7] L. Fathima, Z. Khan, M. Elahi, and S. Afsheen, "A self-sufficient waiter robo for serving in restaurants," *Int. J. Adv. Res. Dev,* vol. 3, no. 5,pp. 57-67, 2018.

[8] C. S. Chang, and Y. F. Chen T. H. Tan "Developing an intelligent e-restaurant with a menu recommender for customer-centric service," *IEEE Trans. Syst. Man Cybern. Part C Appl.Rev,* vol. 42, pp. 775-787, 2012.

[9] S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry K. Dautenhahn, "What is a robot companion-Friend, assistant or butler?," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2005*, 2005.

[10] A. K. M, V. Nanthagopal, and R. Raguram J. S. Shiny, "Automation of Restaurant (Ordering , Serving ,Billing)," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.,* vol. 6, pp. 1532-1539, 2017.

[11] A. Green, and H. Hüttenrauch K. Severinson-Eklundh, "Social and collaborative aspects of interaction with a service robot," *Robotics and Autonomous Systems,* vol. 42, pp. 223-234, March 2020.

[12] P. Jadhav, S. Korade, and P. Teli V. Chavan, "Implementing Customizable Online Food Ordering System Using Web Based Application," *Int. J. Adv. Res. Comput. Eng. Technol,,* vol. 6, pp. 722-727, 2017.

[13] S. Reddy K and C. KGK, "An Online Food Court Ordering System," *J. Inf. Technol. Softw. Eng.,* vol. 6, 2016.

Practical IoT-Enabled Monitoring Platform for Solid Waste Collection

A case study for the municipality of Baba Hassen, Algiers, Algeria

Aouache Mustapha*, Zaid Bousmina†, Ilyes Keskas†, Zakarya Guettatfi‡, Abdelfetah Hentout‡, Ali Mansoul*
*Division Télécom, † Départment DSTSIA, ‡ Division Productique et Robotique
Centre de Développement des Technologies Avancées (CDTA),
Baba Hassen 16303, Algiers, Algeria.
{maouache, zbousmina, ikeskas, zguettatfi, ahentout, amansoul}@cdta.dz

***Abstract*—This work presents a step forward solution for solid waste collection using an adequate IoT-based practical monitoring platform. The system is designed to assist Algerian municipalities in collecting solid waste by monitoring the fullness status of outdoor bins. The solution framework is based on (1) wireless sensor nodes connecting the (2) intelligent bins to the (3) monitoring station by sending the bins filling level status for analysis in order to control the waste in real-time and avoid bins over-filling. This process leads to an improvement in waste management via bins collection-based route optimization process. With the proposed solution, the benefit would be the timely-efficient waste collection, operating costs and gas emissions reduction, and preserve the environment from pollution with future prediction of waste generation.**

***Index Terms*—Municipal solid waste, IoT, Intelligent bin, Gateway, Route optimization, Monitoring platform.**

I. Introduction

A. Motivation

Nowadays, the volume of Municipal Solid Waste (MSW) production is increasing very drastically and become a major challenge for governments and countries due to unprecedented population growth, industrial and urban expansion, and changes in the consumption habits and lifestyle of urban populations [1] [2]. According to worldwide studies, the annual production of solid waste is expected to reach approximately 3.40 billion tonnes by 2050, or an approximate cost of 635.5 billion USD for municipal waste management [3]. Additionally, waste generation in large cities has increased rapidly over the past two decades. Furthermore, 85% of the total MSW management budget is spent on waste collection and transport [3].

Several studies have indicated the direct link between the increase in waste production and the difficulty of its management by traditional solutions, which leads to an inadequate mechanism of collection, disposal, and mismanagement of waste in cities resulting in a deterioration in the citizen' quality of life and economic losses. Thus, solid waste management has become a major issue for the authorities responsible for the lack of satisfactory solutions.

B. Algerian situation regarding waste management

In general, urban population growth and the average standard of living are the main factors contributing to large-scale waste generation [4]. In the Algerian situation, a significant increase in the rate of solid waste production of nearly $0.93\ kg/person/day$ [5] due to several factors, such as the economy has evolved steadily over the past two decades, the regular increase in population (44.18 millions, 2020) [6], improved financial resources, an increased level of consumption of community life and more people are turning to city life.

Currently, the waste collection in Algerian context is based on a static and passive model, carried out once a day by a mechanical pressurized truck with a capacity of 7 to 10 tons which is lifted, transported, and emptied mechanically. There are several fixed waste collection points (bins) in a residential area that make up the truck visit sequence, where they are all collected in one way. In such a model, the capacity of collection trucks is not used at the limit, and several bins are overflowing as in Figure 1. As with other emerging countries, an effective solution for Algerian solid waste management has become the urgent need to replace the conventional way in which municipalities manage waste collection.

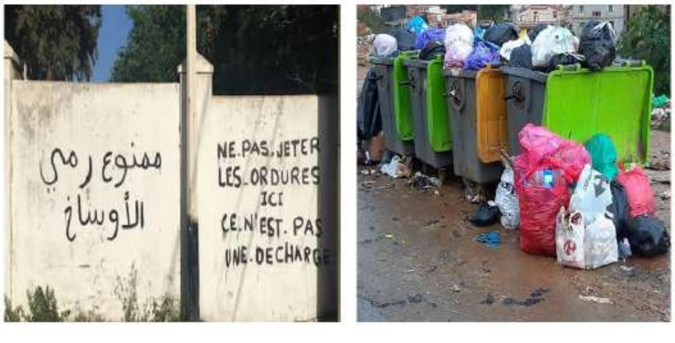


Fig. 1. MSW collection points in a residential area (over-filling).

This work is financially supported in part by the DGRSDT (Direction Générale de la Recherche Scientifique et du Développement Technologique) and CDTA (Centre de Développement des Technologies Avancées), Algeria, as part of Socio-Economic Projects (SEP) 2019-2021. (Project Code: I-Bin, $N°07/CDTA/DGRSDT$).

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

Therefore, a technology-based solution would be the best option and could reduce the effort of the workforce involved in the waste collection by adopting smart bins alerting that send alerts for collection. This can be achieved by using a combination set of modern information technologies that provide necessary data [7] (sensor network, databases) and algorithms for intelligent dynamic routing as shown in Figure 2(b) instead of the traditional method of Figure 2(a). The application of technology in environmental waste management will have major benefits [8] in particular: (1) improve waste collection services, (2) optimize resources and routes for drivers, (3) minimize collection costs, (4) reduce carbon emissions and enhance environmental quality.

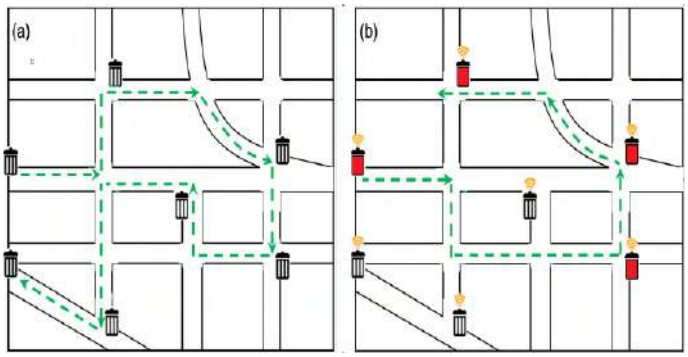


Fig. 2. Waste collection plan (a) static model vs. (b) dynamic model [9].

II. Literature survey

There are many waste management systems proposed to deal with different kinds of challenges existing with current waste management systems in smart cities [10]. In terms of adopting capacity, weight, temperature, humidity, and chemical (C-W-T-H-Ch) sensors, various models have been proposed as in [11]–[14]. The authors in [11] introduced a MSW management platform that exploits information from recycling collection based on IoT technology, and which serves as a model for waste collection introduced in Wuhan. The research findings help city authorities to effectively use the information produced at every step of the waste collection process and ultimately achieve the goal of the smart cycle.

Authors in [12] proposed a dynamic optimization model for solid waste recycling based on material recycling and dynamic optimization. The work represents a significant development of a dynamic decision model that includes state variables and is compatible with the quality of the waste in each bin on a daily basis by (1) controlling the variables, thus determining the quality of collected materials, (2) determining the routes of each waste-truck to reduce the total costs of collection. Furthermore, the decision model is integrated into a decision-support system that is activated with a Geographic Information System (GIS). The demonstration of the concept and its efficiency was carried out as a case study in the municipality of Cogoleto in Italy. A significant advantage has been obtained by improving waste collection 2.5 times more than the current estimated policy.

A monitoring system for solid waste bins and trucks that are activated by Radio Frequency Identification (RFID) and Information Communications Technology (ICT) has been proposed in [13]. The work provides an integrated system for efficient waste collection by adopting Global Positioning Systems (GPS), GIS, and cameras to design an intelligent container and truck surveillance system by introducing an integrated theoretical framework based on hardware engineering and an inference algorithm, in which the model includes a database that stores information about garbage containers and trucks. In addition, data is collected and used via a monitoring system for effective planning and routing using an advanced Graphical User Interface (GUI) which includes real-time images processing, graphical analysis, waste estimation, and container information. Experimental results prove the stability and high performance of the proposed monitoring system.

Proposed in [14], a prototype for monitoring solid bins with an energy-efficient detection model; this research expands [11] by describing a system that responds online when waste is thrown into the bins. The system architecture is based on three levels, (1) the lower one includes the sensors integrated into the bins while the energy-efficient model detects and transmits the measurements to (2) the intermediate level, which includes a gateway that stores and transmits data from the bins to the control station, and (3) the top level stores and analyzes the data for later use. This system is able to reduce truck operating costs and pollutant emissions by providing collected data to the Decision Support System (DSS) which includes dynamic routing of waste collection.

Related works on waste collection containers for content estimation and collection optimization based on capacity, weight, temperature, humidity, and pressure (C-W-T-H-P) sensors are discussed below. Authors in [15] designed and implemented an appropriate urban SWM system based on the amount and variety of waste prediction via adopting measures to correlate waste capacity with residential population and consumption index at different seasons of the year. The system prepared and tested in the Pudong New Area, Shanghai (People's Republic of China) includes an intelligent container and sensors to exploit the data used in further statistical inference processes.

In [16], the authors propose a collection monitoring model for early detection and evaluation of waste through sensitized bins. This research goes from [17] to the description of a new application for monitoring MSW, based on distributed sensor technology and GIS. This system has specific monitoring requirements related to the rapid increase in the rate of waste generation, the model tested and evaluated in Pudong, Shanghai.

In addition, an energy-based model has been proposed to improve solid waste collection in [18], which is applied in a large urban area. The research extends [16] by providing three dynamic scheduling and routing models which have been enhanced by the improvement of paths. A model is proposed in [19] to analyze the impact of solid waste Source Segregation (SS) intensity on fuel consumption as well as collection costs. The work provides an assessment of the fuel consumed and an analysis of the costs of solid waste collection. A simulation model is also integrated to calculate the time spent, the capacity of waste collected, and the fuel consumption for a specific waste collection channel.

Moreover, a Web-GIS system has been proposed in [20] to improve the viability of selective MSW collection in developed and transient economies; issues related to the implementation of the proposed model were also discussed. The model is critically evaluated through two scenarios, (1) two European case studies, where Web-GIS systems in Italy were the best examples of selective collection optimization, and (2) two case studies outside Europe, this scenario achieved 80% efficiency of waste source separation in the recycling purposes.

The Smart-M3 platform was examined and exploited by [21] to model smart planning and monitoring of urban solid waste management by leveraging IoT technology. The authors seek to solve the problem of waste collection by integrating the context of heterogeneous interconnected devices and by sharing data involving a large part of the urban population. They stated the benefits for service providers and users can achieve significant cost savings, while users enjoy Quality of Service (QoS).

Furthermore, many researchers are integrating RFID on a large scale for tagging and identification as part of solid waste collection infrastructure [10]. Authors in [22] proposed RFID and weight-sensor technology as part of the intelligent real-time waste collection system. Weight measurement processes are integrated into the proposed real-time waste collection system; the work ended with an application that leverages RFID data and weight sensors to define an automatic Waste Identity, Weight, and Stolen Bins Identification System (WIWSBIS).

A further effort by [23] who proposed real-time traceability data based on a multi-objective waste collection model supported by various IoT mechanisms to implement an efficient and innovative waste collection routing model. The authors declared that knowing the real-time data related to trucks and the actual level of the bins allows a dynamic routing and scheduling plans. The framework is integrated with an innovative dynamic routing model using real-time traceability data from an Italian city of $100,000$ inhabitants. The model has been tested and validated by simulation; while an economic feasibility study is also being carried out.

In [24], authors proposed an automated bin level detection system by using an advanced image processing approach integrated with ICTnd a camera for bin level detection. The system is evaluated by training and testing features extracted by Gray-Level Cooccurrence Matrix (GLCM) with Multilayer Perception (MLP) and K-Nearest Neighbor (KNN) classifiers. The authors stated the robustness of their system and the ability to apply it at different types of level detection of waste and bins under various conditions.

A. Contribution

This article describes an integrated smart solution developed for solid waste collection to support the current waste collection practices in Algerian municipalities (static model) and responds to the lack of a real-time monitoring system (dynamic model). The solution consists mainly of two modules, (1) monitor system receives the fill level of the outdoor bins measured using sensor node, reported via cellular networks (GPRS) to the platform known as I-Bin Platform, providing the (2) waste collector module with a route map allowing a truck management system and drivers to focus on routes containing most bins with optimum fill levels for collection. Most contributions concerning the proposed solution are: (1) Design and deploy real-time IoT sensor node (ability to measure the bin's capacity). (2) Design a new platform (I-Bin Platform) for data collection and analysis for smart management. (3) Integration and real-time testing of the developed system.

Furthermore, recycling is one of the most efficient ways to manage waste that is environmentally friendly. Among MSW recycling, authors in [25] proposed a real-time robust identification system for polyethylene terephthalate (PET) plastic using a probabilistic approach to model the color, size, and distance of the adjacent tape to differentiate between PET and non-PET plastic materials.

This article is organized as follows. Literature review of relevant works are highlighted in Section II. The examined methodology includes a theoretical framework, system architecture, and physical designs are presented in section III. System setup and implementation are given in section IV. Results and performances with brief limitations are discussed in section V. Conclusions and future work are provided in section VI.

III. Methodology

The effective design methodology for a real-time solid waste management system uses specific hardware, IoT devices, and programming tools, where the methodology is divided into different parts: (1) theoretical framework, (2) system architecture, and (3) physical design based soft-decision algorithms.

A. Theoretical Framework

Figure 3 illustrates the theoretical framework for remote monitoring of bins filling level, using IoT technologies to prevent waste overflow via municipal authorities notification of waste collection manner and times. Main actors involved: smart bin, monitoring system, and collector truck for disposal.

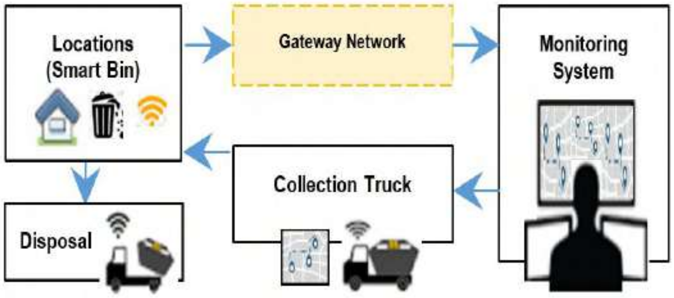


Fig. 3. A theoretical framework to remotely monitor the bins.

The smart bin located in the intended location (house, mosque, hospital, school, etc.) will be registered in the central system. The bin's cover will be embedded with an integrated circuit board for a real-time waste data sensor. The sensors equipped with a cellular network as a gateway to transfer the bins' fill level status to the central database of the system. The monitoring system manager browses and retrieves the information processed for waste collection and transportation; then, it determines the optimal path for truck drivers to collect and dispose of waste.

B. System Architecture

Three main parts from the lower-tier (sensor node) to the upper-tier (control station) through the middle-tier (gateway) represent the system architecture as shown in Figure 4. The sensor node (smart bin) collects the waste status and communicates with the gateway, which is an aggregation layer and message broker providing linking and transmission capabilities to the control station (end-user), which contains a server to store collected data from smart bins located in the city streets, in addition to a processing module for data computing and routing optimization.

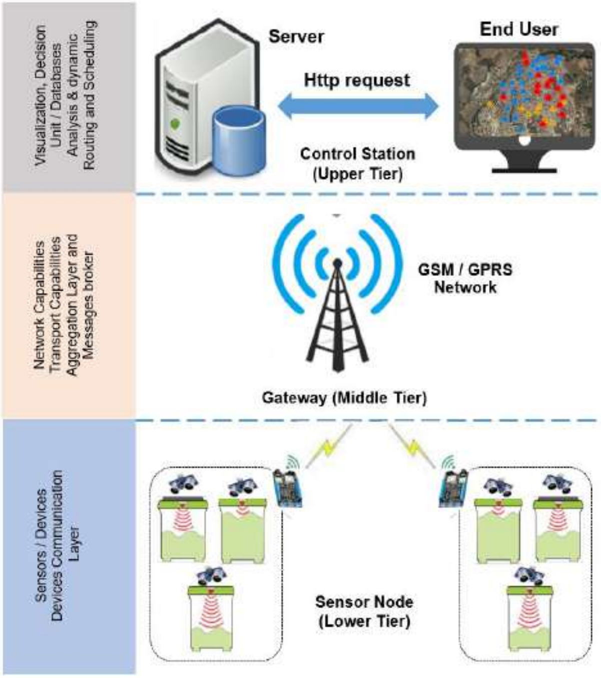

Visualization, Decision Unit / Databases Analysis & dynamic Routing and Scheduling
Server
End User
Http request
Control Station (Upper Tier)
Network Capabilities Transport Capabilities Aggregation Layer and Messages broker
GSM / GPRS Network
Gateway (Middle Tier)
Sensors / Devices Devices Communication Layer
Sensor Node (Lower Tier)

Fig. 4. Main part of the system architecture.

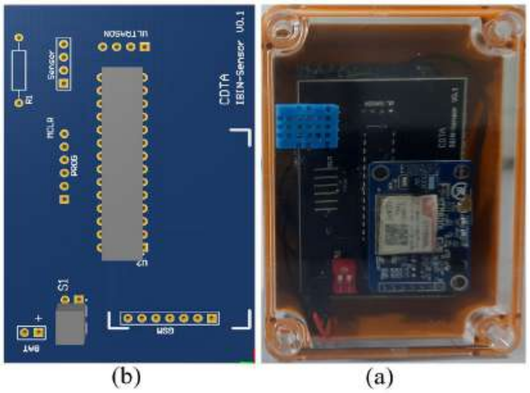
C. Physical Design

Physically, the designed system is based on three main parts:

1) Senor node: consists of sensors, micro-controller, network interface, and power supply attached to the bin cover for data collection and transmission.

- *Sensors*: composed of three types, (1) the HC-SR04 ultrasonic sensor to measures the bins fill-level. (2) the DHT-11 model for temperature and humidity sensing inside and surrounding the bin, helping to predict incoming fires or unwanted events.
- *Micro-controller*: memory and processing requirements, minimum power consumption, and low-cost are factors to consider when selecting a micro-controller. The PIC18F2550 considered as a good enough choice to receive and transmit data from sensors to the server, respectively via a network interface card and Internet services, and is more cost-effective.
- *Network interface*: quad-band GSM/GPRS network expansion module (SIM800L V2.0) with an internal antenna is used to send sensed data over GPRS to the remote web server application.
- *Battery*: Currently, rechargeable batteries ($5V$) are used to power the first prototype. But wireless technology, sensor data, and transmission rates are directly affected by energy consumption, which leads to considering solar energy charging for future use.

Figure 5 depicts the integrated circuit and sensing node that sends and transmits the information to PIC via a specific protocol to coordinate and control signals.



(b) (a)

Fig. 5. Sensing node, (a) PCB model along with (b) physical design.

2) Gateway: receives sensed data from multiple bins, where PIC collects the data from the sensors and sends it to the gateway via TCP/IP using MQTT protocol through the GSM/GPRS module. A broker running on the server, to receive all data and route it to the appropriate NoSQL database for storage. Rule engines will be used for data analysis and display to end-user dashboard.

3) Control station: called I-Bin platform, developed and contains a central server that hosts the database and retrieves the data received from the gateway. The central server hosts a web application-based GUI that monitors bin status and user's interaction with the system. This data can be used by the control station to power programs such as data analysis, optimization engines, and routing and planning applications. The control station basically consists of three components that work together as a processing unit:

- *Storage unit*: all data collected and information received (sensor measurements, location, etc.) from different sensors will be stored in the NoSQL database in an organized manner to be processed for future uses such as predicting filling level and periodic planning of waste collection.

- *Analysis unit*: this is an algorithm-based computation unit that helps to (1) calculate the best bins location with respect to crowd and waste generation ratio in an outdoor area, (2) calculate the distances to optimize the truck's routing paths, (3) scheduling and forecasting tasks for waste collection (4) statistical analysis for further use.
- *Visualisation unit*: provides visual display and tracking services to the end-user, including (1) real-time bins status, (2) optimized routing paths, sub-system, and panels designed for trucks, employees, and tasks management based on updated information.

IV. System setup and implementation

A. Deployment site

Calibration and setup regarding the proposed I-Bin platform for the practical waste collection began with the site selection. Figure 6 illustrates Google map of Baba Hassen city, Algeria, chosen as deployment site for testing and examining the solution. It is a municipality in the capital Algiers (area: $8.75\ km^2$) and a suburb in northern Algeria. The city's population was $23,756$ according to 2008 census [26].

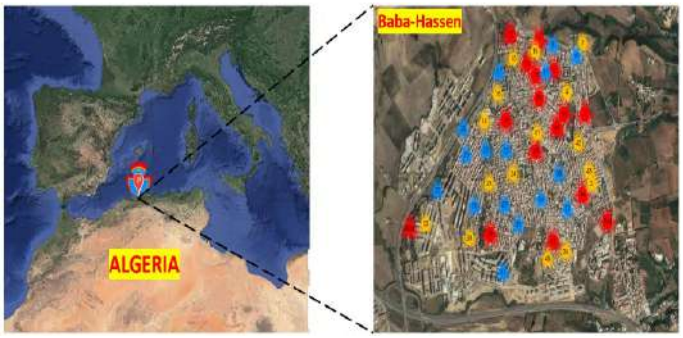


Fig. 6. Deployment site, Baba Hassen city in Algiers, Algeria.

B. Bin layer: Fill levels setting

As shown in Figure 7, the state in the bins was calibrated and monitored according to the three filling levels as follows:

- *Bin-Low Level* (*BLL*): this level defined as an initial step, bins are empty or emptied during the allotted time.
- *Bin-Half Level* (*BHL*): this level defined as a new status of bins to predict their filling time, useful for estimating the waste cost of trucks in terms of time and fuel.
- *Bin-Full Level* (*BFL*): this level reflects the overcapacity of the bins, which requires alerting the authority and truck drivers for the collection time.

The three fill-level for bin monitoring were examined with two different calibration modes:

- *Adjust-Calibration* (*Adju-Calib*): when the user selects this mode on the panel platform, the filling level will be arranged as indicated in Table I (first row).

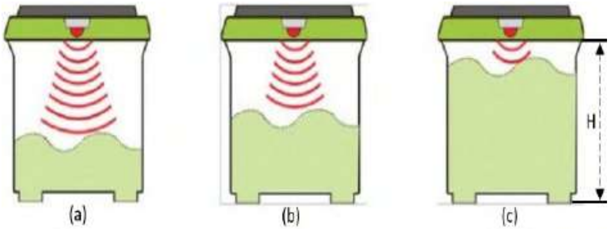


Fig. 7. Ultrasonic fill level sensor, (a) BLL, (b) BHL, and (c) BFL.

- *Adaptive-Calibration* (*Adpt-Calib*): when the user handles this mode, regardless of the status of high-level and low-level bins, the level will be controlled manually via the control panel slider to required values in Table I (second row).

TABLE I
Level bin and calibration setting modes

Mode types	Fill levels setting		
	BLL	***BHL***	***BFL***
Adju-Calib	$BLL \leq \frac{H}{3}$	$\frac{H}{3} < BHL \leq 2 \times \frac{H}{3}$	$2 \times \frac{H}{3} < BFL$
Adpt-Calib	$BLL \leq Th$	$Th < BHL \leq 2 \times Th$	$2 \times Th < BFL$

H=Bin Height. Th=Setting Index (Threshold)

C. Data Layer: Device and gateway work-flow

Figures 8 (a) and (b) are flowcharts and data layer structures used for operating principles and gateway networks and include key elements and logical decision-making processes. Clearly shown the way of real-time detection of bins status, device activity, gateway, and their intra-communication. In short, after turning on the device, the micro-controller will start reading the data from the sensor and send it at regular intervals through a gateway that receives it and transmits it to the server. All data is stored in the database and required data is visualized in real-time.

D. User layer

At this point, end-users manage the waste monitoring and collection process through the I-Bin web platform which includes real-time system monitoring and waste collection modules.

1) System Monitor: in short, the system administrator is the end-user that is able to monitor the current state of the entire system by adding, updating, and removing smart bins from the central system. In addition to providing mapping of the current waste situation, statistical data and various analysis can be performed on historical data provided to the municipality in order to make important decisions regarding updates to the waste management policy.

2) Waste Collector: this unit for users collects targeted bins according to the shortest optimized path. The unit's purpose is to guide and improve the collection path, unlike the static path model that collects every bin, whether full or not. The waste collection can be done quickly, efficiently, and

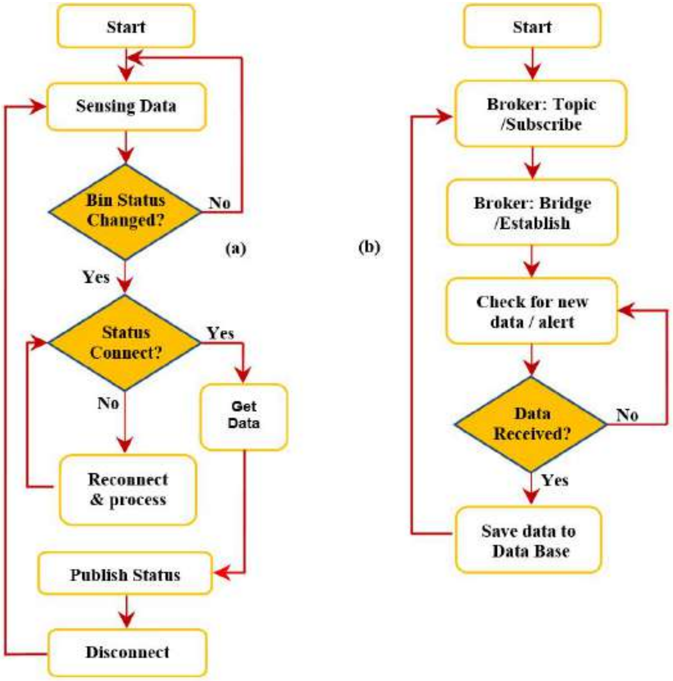


Fig. 8. Workflow of (a) sensing device and (b) gateway.

inexpensively only when the truck is able to perform collection based on the shortest way. In this step, the platform introduces a Dynamic Routing Model (DRM) to the waste collector for waste collection. The DRM carried out in two different stages, (1) DRM-Pattern Mining and (2) DRM-Optimal Routing.

a) DRM-Pattern Mining: for the purpose to extract a pattern function reflecting the distribution of the bins by analyzing the data fill-level and determine when the bins are full or nearly full. This process permits collecting the maximum amount of waste and gives more time to fill the bins. This can be achieved in two different ways: Pattern-A (pick only fully bins), and Pattern-B (pick medium and fully bins) according to Variable Assessment References (*VAR*) computed from equation (1). BFL and BHL reflect number of the daily collected bins with fully and medium filling levels respectively, from $9:00\ p.m.$ to $9:00\ a.m.$ with one-hour interval.

$$\begin{cases} VAR = \frac{\sum(BFL_i)+\sum(BHL_i)}{\sum Bin_i} \times 100 \\ \\ Adopt\ Pattern_{(A)} \quad if\ VAR \geq 50\% \\ Adopt\ Pattern_{(B)} \quad if\ VAR < 50\% \end{cases} \tag{1}$$

b) DRM-Optimal Routing: for the purpose to reduce waste collection path, this step adopted A-Star (A*) search algorithm [27] [28] using either pattern (*A* or *B*) extracted from the previous step to optimize and approximate the shortest path in real-time situations. In addition, two basic aspects can be achieved: (1) determine the optimal path between different sites; (2) estimate time or set of times when a particular route is to be performed.

V. Results and Discussions

A similar case study is shown in Figure 9. Baba Hassen city map was examined to collect the fill level of the 50 bins placed in different places, and the colors indicate the packing capacity of the bins: yellow, blue, and red indicating low, medium, and full levels, respectively.

A. I-Bin platform

This is a web-based monitoring platform provides to the end-user a real-time screening status of bins and fill-level, its sensing data, and amount of waste collected tagged with indicated three different colors. Shown in Figure 9 is the control panel from the I-Bin platform, displaying information such as bin ID, fill-level, date and time, and details including sensor node coordinates, temperature, humidity, and current battery power are retrieved and displayed in the status tool.

The progress bars appear and show the recent fill-level bins status. All this data stored in the database will be further used to optimize the waste collection path. The results returned by the calculation process appear visually in the *Panel Map* with additional settings tool includes VAR index measurement, algorithms for VAR-based routing, and pattern mining (*A* or *B*) for finding the shortest path and alternates routes in real-time. All these data are stored in the database and will be used further to optimize the waste collection route.

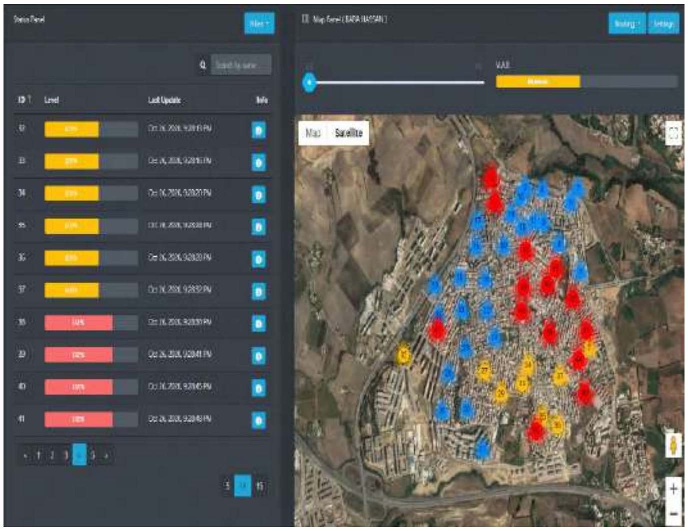


Fig. 9. Control panel from I-Bin platform with bins distribution and their waste fill-levels status, blue (low), yellow (medium), and red (full).

B. Event processing

The waste collection managed dynamically with timely-based updates. The waste truck visits and collects the specified bins only when it receives alerts from the sensors of outdoor bins within a specified period of time or according to the VAR index setting. Figures 10 and 11 represent the results of the optimized waste collection route of the targeted bins when employing the VAR index with both patterns (*A*) and (*B*) respectively. In this instance, the truck speed was set to

20 km/h when is weightless and otherwise at 80 km/h. Due to traffic rules and truck capacity restrictions, each bin pick-up and unload with a specific estimated time.

In short, Figure 10 shows the optimal extraction of the roadmap for waste collection using pattern (A) as in equation (1) with alerts indicator of $VAR \geq 50\%$. In this case, only the filled bins (red color) send an alert for collection and emptying. Waste collection process and truck expenses cost (C_i) was calculated with setting of $C_i = [20\ Km/h,\ 7.08\ Km,\ 1h31m]$, reflects the truck's speed, estimated distance, and additional estimated-time (E-Time++) respectively, and i indicates the number of routes generated for bins collection. On the other hand, Figure 11 illustrates the waste to be collected according to the optimal route extracted using pattern (B), with equation (1), and alerts index $VAR < 50\%$. Both filled and half-filled bins (red and yellow colors) were alerted for collection and emptying. In this case, $C_i = [20\ Km/h,\ 11.55\ Km,\ 1h44m]$ was obtained as the waste-collection cost and expenses.

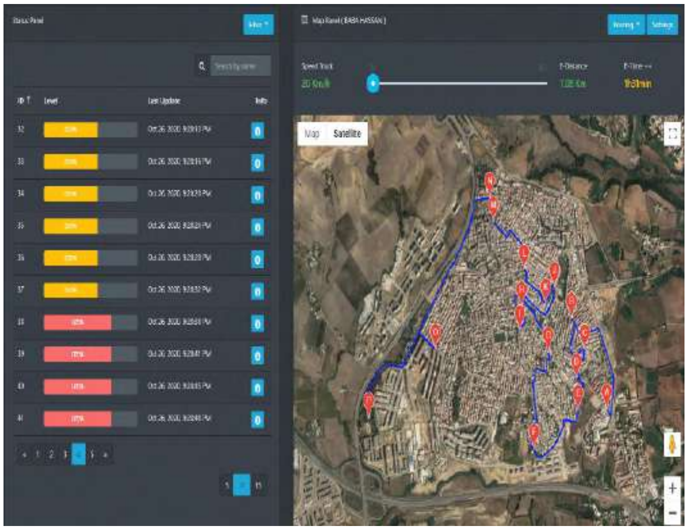


Fig. 10. Control panel from I-Bin platform, DRM-optimal route map through Pattern-A.

From the above, waste collection via DRM roadmap based on pattern (B) required more expense in terms of distance and time consumption compared to pattern (A). Meanwhile, using pattern (B) improves the efficiency of the collection in terms of number of bins collected daily, resulting in further improvement of waste collection and monitoring. In order to effectively monitor and collect accurate data in real-time and help reduce operating costs, the optimal route can be set auto mode to alert for waste collection based on VAR measurement. Moreover, the optimal routing can also be set manually according to the needs of waste management user.

C. Dynamic Scheduling

Besides the above, the added value is also the developed model for dynamic routing and prior planning that helps in dynamic scheduling and better system management. In addition to specifying a time or set of times when a certain route is run, the dynamic planning system provides three subsystems, related to personnel, trucks, and tasks to support collection, management, and reducing waste costs.

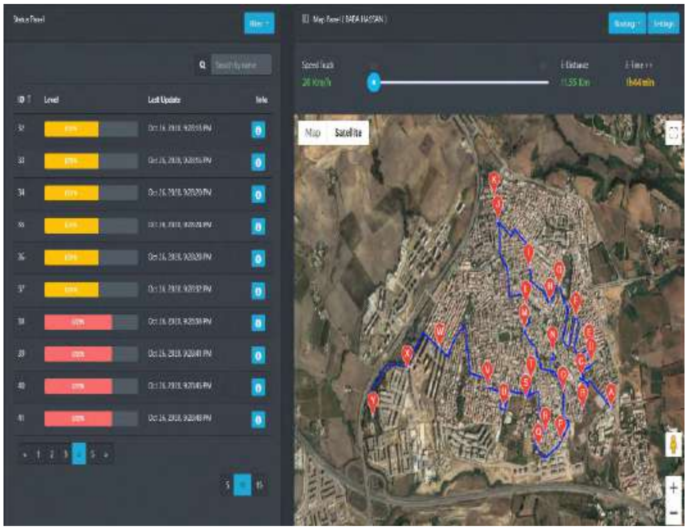


Fig. 11. Control panel from I-Bin platform, DRM-optimal route map through pattern-B.

So far, hardware implementation is the critical challenge with many issues such as (1) scarcity of various high-quality components required, (2) real-time fault reading has emerged with many components of detection, (3) the system tested and limited to only 50 bins, and more challenges will arise when it comes to managing more bins, (4) IoT-enabled technology directly affected by power consumption, prompting consideration of enabled solar charging for future use.

VI. Conclusion

This article discussed practical IoT-enabled monitoring solid waste management system to be implemented in the city of Baba Hassen (Algiers), combines the use of outdoor bins with remote fill level sensor technology for triggering of smart collection when outdoor bins are full or almost. The waste level in each bin sensed and transmitted to a central management platform can be accessed remotely in any web browser to perform the collection process. The system provides route details to the collection truck fleet management system, allowing drivers to focus on routes with most bins of optimum fill levels for collection. Using the I-Bin platform, city authorities can now explore avenues for smart waste management using the power of technology, resulting in cost reduction for waste management, automation, and optimization of daily activities, simultaneously improving services and avoiding bins overfilling on street. Future work includes; (1) seamless integration of smart bins with a cloud-based web application in order to avoid delays in the internet service of mobile operators, and (2) develop mobile apps through which citizens can report incidents and locations of illegal or abandoned waste.

References

[1] B. Mundial, "What a waste 2.0: A global snapshot of solid waste management to 2050," *Washington, EUA*, 2018.

[2] M. Hannan, M. Arebey, R. A. Begum, A. Mustafa, and H. Basri, "An automated solid waste bin level detection system using gabor wavelet filters and multi-layer perception," *Resources, conservation and recycling*, vol. 72, pp. 33–42, 2013.

[3] K. Kansara, V. Zaveri, S. Shah, S. Delwadkar, and K. Jani, "Sensor based automated irrigation system with iot: A technical review," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 5331–5333, 2015.

[4] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, "Towards the implementation of iot for environmental condition monitoring in homes," *IEEE sensors journal*, vol. 13, no. 10, pp. 3846–3853, 2013.

[5] S. Kouloughli and S. Kanfoud, "Municipal solid waste management in constantine, algeria," *Journal of Geoscience and Environment Protection*, vol. 5, no. 1, pp. 85–93, 2017.

[6] (29, Nov. 2020) The current population of algeria as of november, 2020. https://www.worldometers.info/world-population/algeria-population/.

[7] A. Tiberkak, A. Hentout, and A. Belkhir, "Lightweight remote control of distributed web-of-things platforms: First prototype," in *2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2021, pp. 103–108.

[8] T. Ali, M. Irfan, A. S. Alwadie, and A. Glowacz, "Iot-based smart waste bin monitoring and municipal solid waste management system for smart cities."

[9] T. V. Anagnostopoulos and A. Zaslavsky, "Effective waste collection with shortest path semi-static and dynamic routing," in *International Conference on Next Generation Wired/Wireless Networking*. Springer, 2014, pp. 95–105.

[10] T. Anagnostopoulos, A. Zaslavsky, K. Kolomvatsos, A. Medvedev, P. Amirian, J. Morley, and S. Hadjieftymiades, "Challenges and opportunities of waste management in iot-enabled smart cities: A survey," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 3, pp. 275–289, 2017.

[11] C. Tao and L. Xiang, "Municipal solid waste recycle management information platform based on internet of things technology," in *2010 International Conference on Multimedia Information Networking and Security*. IEEE, 2010, pp. 729–732.

[12] D. Anghinolfi, M. Paolucci, M. Robba, and A. C. Taramasso, "A dynamic optimization model for solid waste recycling," *Waste management*, vol. 33, no. 2, pp. 287–296, 2013.

[13] M. Hannan, M. Arebey, R. A. Begum, and H. Basri, "Radio frequency identification (rfid) and communication technologies for solid waste bin and truck monitoring system," *Waste management*, vol. 31, no. 12, pp. 2406–2413, 2011.

[14] M. A. Al Mamun, M. Hannan, A. Hussain, and H. Basri, "Wireless sensor network prototype for solid waste bin monitoring with energy efficient sensing algorithm," in *2013 IEEE 16th International Conference on Computational Science and Engineering*. IEEE, 2013, pp. 382–387.

[15] F. Vicentini, A. Giusti, A. Rovetta, X. Fan, Q. He, M. Zhu, and B. Liu, "Sensorized waste collection container for content estimation and collection optimization," *Waste management*, vol. 29, no. 5, pp. 1467–1472, 2009.

[16] A. Rovetta, F. Xiumin, F. Vicentini, Z. Minghua, A. Giusti, and H. Qichang, "Early detection and evaluation of waste through sensorized containers for a collection monitoring application," *Waste Management*, vol. 29, no. 12, pp. 2939–2949, 2009.

[17] S. Longhi, D. Marzioni, E. Alidori, G. Di Buo, M. Prist, M. Grisostomi, and M. Pirro, "Solid waste management architecture using wireless sensor network technology," in *2012 5th International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2012, pp. 1–5.

[18] X. Fan, M. Zhu, X. Zhang, Q. He, and A. Rovetta, "Solid waste collection optimization considering energy utilization for large city area," in *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, vol. 3. IEEE, 2010, pp. 1905–1909.

[19] F. Di Maria and C. Micale, "Impact of source segregation intensity of solid waste on fuel consumption and collection costs," *Waste Management*, vol. 33, no. 11, pp. 2170–2176, 2013.

[20] E. C. Rada, M. Ragazzi, and P. Fedrizzi, "Web-gis oriented systems viability for municipal solid waste selective collection optimization in developed and transient economies," *Waste management*, vol. 33, no. 4, pp. 785–792, 2013.

[21] V. Catania and D. Ventura, "An approch for monitoring and smart planning of urban solid waste management using smart-m3 platform," in *Proceedings of 15th conference of open innovations association FRUCT*. IEEE, 2014, pp. 24–31.

[22] B. Chowdhury and M. U. Chowdhury, "Rfid-based real-time smart waste management system," in *2007 Australasian Telecommunication Networks and Applications Conference*. IEEE, 2007, pp. 175–180.

[23] M. Faccio, A. Persona, and G. Zanin, "Waste collection multi objective model with real time traceability data," *Waste management*, vol. 31, no. 12, pp. 2391–2405, 2011.

[24] M. Hannan, M. Arebey, R. A. Begum, and H. Basri, "An automated solid waste bin level detection system using a gray level aura matrix," *Waste management*, vol. 32, no. 12, pp. 2229–2238, 2012.

[25] M. A. Zulkifley, M. M. Mustafa, A. Hussain, A. Mustapha, and S. Ramli, "Robust identification of polyethylene terephthalate (pet) plastics through bayesian decision," *PloS one*, vol. 9, no. 12, p. e114518, 2014.

[26] (2008) Baba hassen. https://fr.wikipedia.org/wiki/Baba-Hassen/.

[27] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[28] A. Hentout, A. Maoudj, D. Yahiaoui, and M. Aouache, "Rrt-a*-bt approach for optimal collision-free path planning for mobile robots," *Algerian Journal of Signals and Systems (AJSS)*, vol. 4, no. 2, pp. 39–50, 2019.

Ensemble of Supervised and Unsupervised Learning Models to Predict a Profitable Business Decision

1stMaryam Heidari
George Mason University
mheidari@gmu.edu

2nd Samira Zad
Florida International university
Szad001@fiu.edu

3nd Setareh Rafatirad
George Mason University
Srafatir@gmu.edu

***Abstract*—Real-Estate rent prediction in housing market analysis plays a key role in calculating the Rate of Return - a salient index used to evaluate real-estate investment options. Accurate rent prediction in real estate investment can help in generating capital gains and guaranty a financial success. In this paper, we carry out a comprehensive analysis and study of seven machine learning algorithms for rent prediction, including Linear Regression, Multilayer Perceptron, Random Forest, KNN, Locally Weighted Learning, SMO, and KStar algorithms. We train new model for the US territory, including three house types of single-family, townhouse, and condo. Each data instance in the dataset has 21 internal attributes (e.g., area space, price, number of bed/bathroom, rent, school rating, so forth). A subset of the collected features selected by filter methods for the prediction models. We also employ a hierarchical clustering approach to cluster the data based on two factors of house type, and average rent estimate of zip codes. The empirical results suggest that the rent prediction models based on lazy learning algorithms lead to higher accuracy and lower prediction error compared to eager learning methods.**

***Index Terms*—housing analytics, applied machine learning, rent prediction, data mining**

I. Introduction

Real estate rent prediction has a key role in calculating the Rate of Return- a measure used to evaluate the performance of an investment in the housing market. Rate-of-Return can measure the quality of a real estate investment over a time-period and has two important factors: *Net Present Value and Future Value*.Proper rental property investments can lead to a successful and profitable Rate of Return over time. However, such ventures can be very risky due to miscalculation or inaccuracy of algorithms used in rent prediction. Applying machine learning algorithms to perform house rent prediction is not a novel trend. Lambert and Greenland [1] investigate eager learning methods like Multi- Layer perceptron, and bagging REP trees to estimate the rental rate for both the land-owners and students interested in renting a place close to a university campus. The training set contains two property types: *i)* apartment and *ii)* condo. The coverage area of the training set is limited to three distant zip codes surrounding a university campus. The input features entertained in this work include proximity to university campus, apartment appliances (like Cable TV) and dimensions, the length of the apartment contract, and the date of the residence's constructions. The study reports bagging REP trees as the best rent prediction algorithm. However, the proposed global learning-based solution can generate a biased model due to the skewed data set, all located surrounding a university campus.Machine learning models can be used in social sciences [2]–[4] [5] [6] and provides more view about peoples opinion about housing market. Transfer models can provide more powerful insight about house price prediction [7]–[9]

Machine learning models have advanced applications in health [10]–[12], cyber security, computer hardware [13], [14] and computer science [15]–[17] [18], [19]and business [20], [21]. In the previous studies, the prediction models for real estate rent/price prediction are very generic and they don't differentiate according to the house type and/or zip code [22]–[24]. For instance, a generalized prediction model is proposed by [25] for city-wise scope of data, to predict rent and house prices. However, this can lead to inaccurate predictions. Rent behavior is different, even for the real estate properties which are in the same state/city or a close geo-spatial proximity from each other. For instance, the zip codes 22066 and 20190 are neighbors but they show a very different behavior in terms of the average rent price. In addition, statistical data shows that house type parameter affects the rent price due to internal factors like area space, number of bed/bathrooms, HOA fee, community factors, so forth.

The average rent price for a zip code depends on internal factors (like house type, number of bedroom/bathrooms, house price, area space, other) and external factors. In fact, external factors like crime rate and school ratings corresponding to a zip code impact the price of rent and are deal-breakers for many real estate investors [26]. In this paper, internal house properties as well as external attributes like walk score, transit score, crime rate, and school rating are entertained. Walk score indicates the errands that can be accomplished on foot or those that require a car to nearby amenities. Transit score indicates the connectivity (i.e., proximity to metro), access to jobs, and frequency of service [27]. Crime score indicates the rate of violent and non-violent incidents related to a zip code [28].

In the past studies, the impact of eager learning methods for real estate rent/price prediction has been investigated [1]. In general, unlike eager learning methods, lazy learning (or instance learning) techniques aim at finding the local optimal solutions for each test instance [29], [30]. Friedman, Kohavi, and Yun in [29], and Homayouni, Hashemi, and Hamzeh in [30] store the training instances and delay the generalization until a new instance arrives. Another work

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

carried out by Galv´an et al. [31], compares memory-based (or locally) learning vs neural network methods, and report the superiority of memory-based (i.e., lazy) learning method over Neural Network approach using multiple data sets of UCI machine learning data set repository, including Iris, Diabetes, Sonar, Vehicle, car, and balance. In this study, the two lazy learning algorithms, namely Integer Part and Atomic Radios outperformed eager learning algorithms, namely SMO and Naive Bayes algorithm. Our work is inspired by these research explorations.

We collected a Zillow data set of 600K real estate properties in Virginia State. In addition, transit score, walk score, and crime rate are collected from information sources like alltransit.cnt.org, walkscore.com, and crimereports.com respectively. Transit parameters entertained in the data collected from AllTransit data source clearly indicates the proximity to metro as a significant parameter in determining the transit score of a location.

The data set consists of three house types: *town-house*, *single family*, and *condo*. This study is motivated by the need to build models with respect to house type and zip code. To deal with the sparse (i.e., thin) data in every zip code, we divided the data set according to house type, and then applied K-means clustering to generate subsets of instances within zip codes with similar average rent prices. The clustering method uses the similarity measure of *average-rent* to compute the distance between the data points. The data samples in each cluster is later used to train a rent prediction model.

In this work, we study the impact of several machine learning methods on this data set by performing a comparative analysis of lazy vs. eager learning methods. We report the empirical results to show the superiority of lazy learning algorithms over eager learning methods in real-estate rent prediction for each house type and a subset of zip codes with similar average rent prices. We examine the performance of Linear Regression (LR), SMO, Multilayer Perceptron (MLP), and Random Forest (FR) algorithms (eager/globally-based learning) against KNN, locally weighted learning (LWL) and K-star (K*) algorithms (lazy/memory-based learning), using three performance evaluation metrics: i) R-squared , ii) Mean Absolute Error (MAE), and iii) Area Under ROC Curve (AU-ROC). The target variable is the rent price and the evaluation metrics show the variance between the predicted target variable and the actual rent price.

Our rent prediction algorithm uses a salient subset of data set attributes, which is determined during feature selection using Principal Component Analysis technique (PCA). PCA technique filters out unwanted features based on each house type (i.e., single family, town house, and condo), and independent of the learning algorithm applied on the data.

For imputation, we removed the observations with many missing attributes as the proportion of these instances to the entire data set was less than 3%.

The remainder of this paper is structured as follows: In section II related work is discussed. We describe the data set used in this paper for analysis in section III. Section IV describes our methodological framework including data preprocessing, data exploration and feature selection, building prediction models, and model evaluation. In section V, experiments and results are discussed. Finally, section VI gives the conclusion.

II. Related Work

Real estate rent/price prediction using machine learning techniques has already been studied in several works [22]–[25]. In [32], PSO-SVM algorithm is used for real- estate price prediction. In [33], [34], spatiotemporal dependencies between housing transactions is used to predict future house prices. However, this approach is limited by spatial autocorrelation, since the degree of similarity between observations is not solely based on the distance separating them.

Some of the previous work focus on hedonic price models as a method of estimating the demand and value in the housing market and determination of house prices [35], [36]. In these studies, rather than internal and external house features, economic submarkets are used in the prediction model which are defined in terms of the characteristics of neighborhoods or census units. In [22] a sample size of 200 houses of all house types in New Zealand were used in a hedonic price model and an artificial neural network (ANN) model, and shows that the eager method ANN outperforms the hedonic model. The problem with the hedonic approach is disregarding the differences between the properties in the same geographical area.

Park and Bae in [37] determine the house sales trends of the US housing market subject to the Standard & Poor's Case-Shiller home price indices and OFHEO which is the housing price index of the Office of Federal Housing Enterprise Oversight. The authors used and compared the classification accuracy of four methods C4.5, RIPPER, Naïve Bayesian, and AdaBoost, where RIPPER algorithm outperformed others.

We are now at the age where people in different fields are hacking their way into machine learning. Machine learning techniques have become available as commodities which can be used to perform prediction and classification tasks in various domains like real-estate rent/price prediction. Bin Khamis and Kamarudin in [20] compared the efficacy of the eager learning method Neural Network (NN) against the hedonic model Multiple-Linear Regression (MLR), and showed that NN outperforms MLR. However, Galv´an et al. in [5], reports the superiority of lazy learning methods over NN.

According to [38], eager learning methods can sometimes lead to suboptimal predictions because of deriving a single model that seeks to minimize the average error over the entire data set, whereas lazy learning can help improve prediction accuracy. While our study is inspired by [31], [38], we take our analysis to the next level, by comparing the impact of eager and lazy learning algorithms in the prediction accuracy of the generated models with respect to each house type and a subset of zip codes with similar average rent prices. We use a two- layer clustering technique, and a subset of internal and external real-estate property factors.

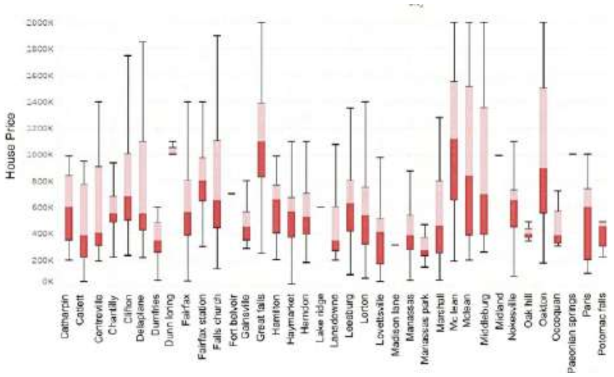


Fig. 1. House Price Distribution

III. Data set description

There are variety of housing data sources in the real estate market. Zillow API [39] delivers home details including historical data on sales prices, year of sale, tax information, number of bed/baths, so forth, for the US territory. In this paper, we used the Zillow API to collect a data set of residential housing data for the state of Virginia. The size of this data set contains about 4000 housing property records (including townhouse, single-family, and condo) with 21 attributes. In addition, external attributes, namely walk score, transit score, and crime rate are collected.

We collected transit score data from AllTransit data source [27]: their dataset is collected from 824 agencies, and it includes 662K stop locations and 13K routs. Transit and walk scores are collected per household, while crime rate is obtained for each zip code , normalized by the number of people living in that area [40] using Selenium tool with Python. Crime score data was normalized using Dickson method [40] indicated by equation (1):

$$Incident_{norm} = \frac{CrimeIncidents}{population} * 100,000 \tag{1}$$

We obtained zip code-wise population by collecting data from *www.moving.com*.

IV. Methodology

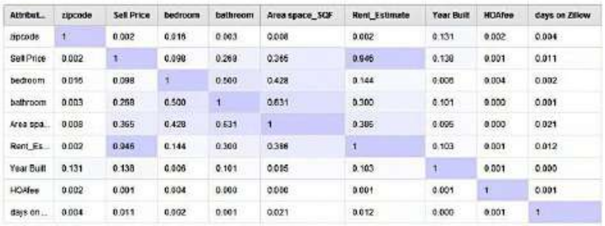
A. Data Preprocessing

One of the rudimentary principles in calibration of machine learning models when dealing with a biased data is to re-sample the data to balance them [41]. Some of the areas have much higher densities compared to other areas. To normalize the data, we re-sampled the data in zip codes with higher house prices due to their crowded density relative to the zip codes with lower house prices.

For imputing the missing values of external attributes, we used K-means clustering and KNN. The distances between data points is calculated with respect to each cluster centroid.

To reduce the dimensionality of the data set and enhance the generalization of the model, we perform feature selection by applying PCA (principle component analysis) to all 21 attributes of the data set. However, before applying PCA, attributes are normalized based on Min-Max Normalization, using equation (2):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$



Attribut...	zipcode	Sell Price	bedroom	bathroom	Area space_SQF	Rent_Estimate	Year Built	HOAfee	days on Zillow
zipcode	1	0.002	0.016	0.003	0.008	0.002	0.131	0.002	0.004
Sell Price	0.002	1	0.098	0.268	0.365	0.946	0.138	0.001	0.011
bedroom	0.016	0.098	1	0.500	0.428	0.144	0.006	0.004	0.002
bathroom	0.003	0.268	0.500	1	0.631	0.300	0.101	0.000	0.001
Area spa...	0.008	0.365	0.428	0.631	1	0.386	0.095	0.000	0.021
Rent_Es...	0.002	0.946	0.144	0.300	0.386	1	0.103	0.001	0.012
Year Built	0.131	0.138	0.006	0.101	0.095	0.103	1	0.001	0.000
HOAfee	0.002	0.001	0.004	0.000	0.000	0.001	0.001	1	0.001
days on ...	0.004	0.011	0.002	0.001	0.021	0.012	0.000	0.001	1

Fig. 2. Correlation matrix based on house features

B. Data Exploration

Figure 1 shows house Price Distribution for the state of Virginia We analyzed the correlations between various variables of the data set to identify the co-linearity between the variables. Discovering co-linearity between the data set variables and the target variable yields valuable insights about the dependent variables that affect the rent price. Figure 2 illustrates the co- linearity between the data set attributes in the VA data set. Since zip code is a nominal attribute, there is no collinearity between zip code and rent price. Also, there is a strong dependency between rent and internal attributes like number of bed/bathrooms, area space, year, and sale price. In addition, there is a correlation between average rent price and urban planning parameters (external attributes) of walk/transit-score and crime-rate. The general trend indicates a positive correlation between average rent and walk/transit score, and a negative correlation between the average rent and crime rate across multiple zip codes.

We hypothesize that there is a rent prediction model for every house type within a zip code/similar zip codes. To test this hypothesis, we carry out the following analysis:

Prior to attribute selection, to obtain a suitable representation of the data set, we apply PCA (principle component analysis) to the 21 data set attributes. The attributes consist of ZipID (a unique id for each house in the Zillow API), Number of bed/baths, floor size (the area of the house based on SQF), latitude and longitude (geographical location of each house), year built (the year of house construction), status (house type), zip code, house features (facilities in a house described by owner), estimated rent (basic amount of rent price for each house used as a class label in the prediction task), so forth.

Since zip code and house type are nominal attributes, there is no collinearity between the rent price and these two attributes.

C. Feature Selection

To identify important attributes, we apply PCA (principle component analysis) - which is a well-known and studied method- on three subsets of data samples, each subset covering a different house type across all zip codes in the state of Virginia. This is illustrated in Figure 3.

unlike town house and condo, features like HOA fee, walk score, and transit score show a very low variance for single family instances. The higher variance of transit score, especially for condos, explains the outpacing of median appreciation rates of condos compared to single family detached-houses in large metropolitan areas [42]. Number of bedrooms is found to be an important feature only when house type is single-family or town-house. Next, unlike town-house instances, average school rating is discovered to be an important feature for both single-family and condo instances. This can be explained due to sparsity of school rating for town-house instances in our data set. In our future work, we will employ data mining techniques to obtain this information for town-house instances.

We validated the above mentioned strategy by training our models based on a data set including all house types, and then based on each house type. We discovered that the latter approach leads to relatively higher accuracy and lower prediction error.

D. Data Clustering

First, we cluster the housing data set based on house type and zip code attributes, to eventually learn a model for each cluster. However, we observed that the some of the clusters are very sparse with the number of instances below 100, which could immensely affect the ability to train the prediction models [43], [44]. To increase the density of the training samples and facilitate the accuracy of prediction models at the same time, we carried out a different strategy, first, we divided the data set into three groups based on the house type attribute. We refer to these groups as status-clusters. Next, we calculated the average rent for every zip code in each status-cluster. Furthermore, we applied K-means clustering to cluster the content of each status-cluster based on the average-rent. Using this clustering technique, we increased the density of the training samples.

1) Problem Formulation: Given a $status - cluster_j$,$i = (sf, th, co)$ where sf=single-family, th= town-house, and co=condo with a set of observations $(o_1, o_2, \cdots, o_m)$ where each observation is a d_i dimensional vector, K-means clustering partitions the observations into $n(n = k, h, gs.t., n \ll m)$ sets S.

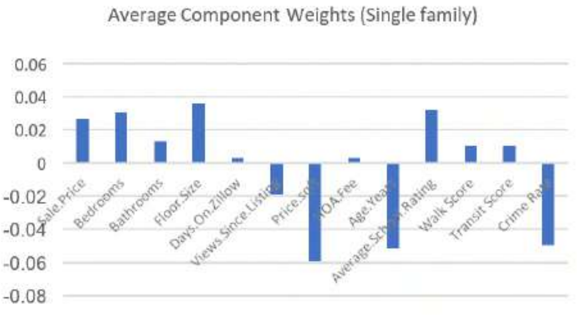


Fig. 3. The PCA plot for single family

Clustering the data points according average-rent implies organizing instances of similar zip codes inside the same cluster, e.g., cluster s_j. The next task is to train a model for each cluster s_j: $s_j \rightarrow Learner \rightarrow f_j$, such that f_j is a rent prediction model. In this study, we build rent prediction models with respect to each house type and a subset of zip codes with similar average rent prices, using two eager and two lazy learning algorithms.

2) Data Partitioning: In our experiments, we use 10-Fold Cross-Validation to partition the training data set into 10 equal parts. During each round of 10 iterations, we repeat the prediction by using one of the 10 parts as test data and the other 9 parts as training data to create a prediction model. Next, we select the model with the best accuracy. We then evaluate the rent prediction model on the test data set that covers 30% of the entire Virginia housing data set collected from Zillow website.

The important attributes determined during feature selection are used as input of the above-mentioned models to predict the target variable *rent price*.

E. Model Evaluation

The key comparison measure used for regression analysis and model evaluation described in this section uses three different measures: i) Mean Absolute Error (MAE), ii) R-squared, and iii) Area Under the Receiver Operating Characteristics curve (AUROC or AUC). MAE measures the accuracy of the prediction models over the test data set. R- squared (or the coefficient of determination) is a quadratic statistical scoring rule which shows how close the actual target data are to the fitted regression line. R-squared is used in the paper to show the variance between the predicted target variable and the actual rent price. Hence, the lower MAE and the higher R-squared, the better our model fits the data. The ROC curve [45] is a graphical technique to visualize the performance of prediction models and selecting the best model based on their performance. AUROC measures the accuracy of a prediction model. Based on these evaluation metrics, we calculate and compare the efficacy of the produced rent prediction models for seven machines learning algorithms MLP, RF, LR, SMO, LWL, KStar, and KNN based on the hierarchical clustering . For KNN's combination function, we used simple unweighted voting for K=3, based on Euclidean distance. The comparison of MAE and R-squared is illustrated in figure 4 and figure 5.

V. Experiments and Results

This section discusses the result of our experiments carried out to evaluate and compare the performance of MLP, RF, LR, SMO, LWL, KStar, and KNN algorithms. According to table I, KStar algorithm outperforms the other algorithms. KStar shows the lowest variations(R-squared) and highest accuracy (MAE) compared to other algorithms tested in this work. Based on the overall measure of the fit of the model, we compare the best of eager methods with that of lazy methods. Among the eager methods tested, LR and RF show the highest accuracy, and LR shows the lowest variance (R-squared),

TABLE I
MACHINE LEARNING MODELS PERFORMANCE

Algorithm	**Single-Family**			**Town-House**			**Condo**		
	R-squared	***MAE***	***AUROC***	***R-squared***	***MAE***	***AUROC***	***R-squared***	***MAE***	***AUROC***
MLP(eager)	1373.4	410	0.865	150.4	105.4	0.913	218.8	152.4	0.896
RF (eager)	543.2	322.7	0.924	250.48	109.7	0.937	525.1	109.3	0.914
LR (eager)	497.004	294.1	0.853	119.230	83.37	0.941	156.9	112.7	0.891
SMO (eager)	1114.76	342.02	0.881	234.06	98.02	0.891	178.2	254.1	0.746
LWL (lazy)	489.6	299.2	0.951	146.96	98.1	0.951	260.3	121.4	0.905
KStar(lazy)	150.0	91.7	0.971	86.4	49.3	0.981	194.3	103.2	0.965
KNN (lazy)	167.0	321.065	0.89	140.00	97.15	0.912	101.1	110.78	0.907

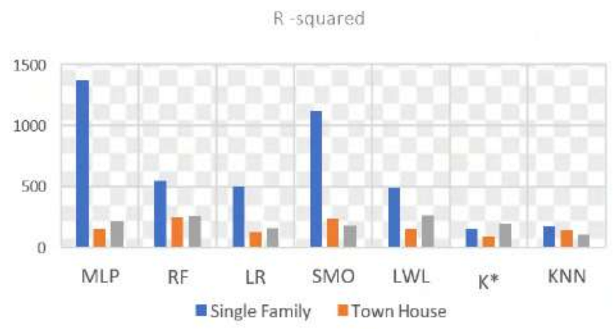


Fig. 4. Model performance for single family and townhouse

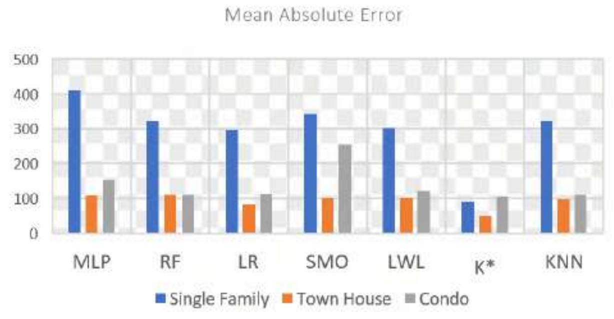


Fig. 5. Comparison of Model performance

while KNN and KStar show the minimum variance among the lazy methods. In addition, KStar shows the highest accuracy compared to other ML methods tested in this work. In fact, KStar algorithm decreased the prediction error by 69%, 41%, and 8% for single-family, town-house, and condo respectively, compared to LR method. Also, KStar decreased the prediction error by 71%, 55%, and 5% for single-family, town-house, and condo respectively, compared to RF method.

Finally, KStar decreased the prediction error by 71%, 49%, and 6.8% for single-family, town-house, and condo respectively, compared to KNN method. Based on figure 5 the low variations of MAE measure across different algorithms for the town-house records compared to single-family and condo is happening due to skewness of the data set: the number of single-family properties, and then condo, dominate the data set compared to town-house records, which indicates the skewness of the data set. In addition, according to Table I, AUROC for KNN and LWL are very close, while eager version of SMO shows the lowest AUROC among other examined algorithms. The results show that KStar regression model provides the best fit, and lazy learning methods overall, outperform the eager methods. All algorithms perform relatively well when it comes to the town-house data.

This result can be explained due to insufficient data in terms of internal/external features, and the number of observations. For instance, the single-family houses with the rent price above $9K in cities like Fairfax and Gainsville are very sparse and listed as "home-office by the owner to rent to doctors; these houses are rented as "home-office" with the medical equipment inside the rental property. We observed that for town-house records, the performance of the learning method LR is very close to KNN lazy learning method.

VI. CONCLUSION

Eager learning methods build a prediction model using the entire training set which is later used on the test instances to evaluate the performance of the model. While eager learning methods tend to extract the general properties of data and minimize the global error, lazy learning methods select the most appropriate samples in the learning process, and minimize the local error.

Based on our experimental study in the domain of real-estate, well-known Lazy learning methods can surpass eager learning algorithms in rent prediction problem for each house type (single family, town-house, and condo) in our housing data set. We examined KStar, KNN, and locally weighted learning models as the well-known representatives of lazy learning methods, and compared them with MLP, LR, SMO, and Random Forest eager learning algorithms. The results indicate that the behavior of eager learning methods can negatively affect the generalization capability of these models.

Given that deep learning methods work well on high dimensional data, we plan to employ natural language processing tools to extract metadata from property owner comments in Zillow website, and increase the number of attributes. We also plan to investigate the prediction accuracy of deep learning models like Neural Network.

REFERENCES

[1] J. Lambert and J. Greenland, "Is the price right?," *Prediction of Monthly Rental Prices in Provo, Utah*, 2015.

[2] M. U. Nisar, S. Voghoei, and L. Ramaswamy, "Caching for pattern matching queries in time evolving graphs: Challenges and approaches," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2352–2357, 2017.

[3] S. Voghoei, N. Hashemi Tonekaboni, J. G. Wallace, and H. R. Arabnia, "Deep learning at the edge," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 895–901, 2018.

[4] S. Voghoei, N. Hashemi Tonekaboni, D. Yazdansepas, and H. R. Arabnia, "University online courses: Correlation between students' participation rate and academic performance," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 772–777, 2019.

[5] M. Heidari, J. H. J. Jones, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[6] S. Zad and M. Finlayson, "Systematic evaluation of a framework for unsupervised emotion recognition for narrative text," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pp. 26–37, 2020.

[7] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[8] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network to recommend airline tickets by using online reviews," in *IEEE 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020*, 2020.

[9] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[10] M. Saadati, J. Nelson, and H. Ayaz, "Mental workload classification from spatial representation of fnirs recordings using convolutional neural networks," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2019.

[11] D. G. A. A. M. M. S. J. M. S. J. K. J. B. M.-C. H. Golnoush Asaeikheybari, Cory Hughart, "Precision hiv health app, positive peers, powered by data harnessing, ai, and learning," in *IEEE 2020 Second International Conference on Transdisciplinary AI (TransAI), TransAI 2020*, 2020.

[12] G. Asaeikheybari, J. Green, X. Qian, H. Jiang, and M.-C. Huang, "Medical image learning from a few/few training samples: Melanoma segmentation study," *Smart Health*, vol. 14, p. 100088, 2019.

[13] M. H. Gavgani and S. Eftekharnejad, "Critical component identification under load uncertainty for cascading failure analysis," in *2020 IEEE Texas Power and Energy Conference (TPEC)*, pp. 1–6, 2020.

[14] M. H. Gavgani and S. Eftekharnejad, "A graph model for enhancing situational awareness in power systems," in *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, pp. 1–6, 2017.

[15] N. H. Tonekaboni, S. Kulkarni, and L. Ramaswamy, "Edge-based anomalous sensor placement detection for participatory sensing of urban heat islands," in *2018 IEEE International Smart Cities Conference (ISC2)*, pp. 1–8, IEEE, 2018.

[16] N. H. Tonekaboni, L. Ramaswamy, and S. Sachdev, "A mobile and web-based approach for targeted and proactive participatory sensing," in *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 215–230, Springer, 2019.

[17] N. H. Tonekaboni, L. Ramaswamy, D. Mishra, O. Setayeshfar, and S. Omidvar, "Spatio-temporal coverage enhancement in drive-by sensing through utility-aware mobile agent selection," in *International Conference on Internet of Things*, pp. 108–124, Springer, 2020.

[18] N. Etemadyrad and J. K. Nelson, "A sequential detection approach to indoor positioning using rss-based fingerprinting," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1127–1131, IEEE, 2016.

[19] J. K. Nelson and N. Etemadyrad, "Prioritizing goals in cognitive sonar: Tracking multiple targets," in *2018 21st International Conference on Information Fusion (FUSION)*, pp. 1–6, IEEE, 2018.

[20] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a specific business domain," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[21] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[22] V. Limsombunchai and V. Limsombunchai, "House price prediction: Hedonic price model vs. artificial neural network," 2004.

[23] A. Khamis and N. K. K. Kamarudin, "Comparative study on estimate house price using statistical and neural network model," *International Journal of Scientific & Technology Research*, vol. 3, pp. 126–131, 2014.

[24] S. Basu and T. G. Thibodeau, "Analysis of spatial autocorrelation in house prices," *The Journal of Real Estate Finance and Economics*, vol. 17, no. 1, pp. 61–85, 1998.

[25] H. Yu and J. Wu, "Real estate price prediction with regression and classification cs 229 autumn 2016 project final report," 2016.

[26] RealEstate-US News Homepage. https://realestate.usnews.com/real-estate/articles/how-homicide-affects-home-values.

[27] "Alltransit homepage." https://alltransit.cnt.org/.

[28] "Crime in the united states 2011 : Violent crime." https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/violent-crime/violent-crime.

[29] J. H. Friedman, R. Kohavi, and Y. Yun, "Lazy decision trees," in *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96, Portland, Oregon, USA, August 4-8, 1996, Volume 1* (W. J. Clancey and D. S. Weld, eds.), pp. 717–724, AAAI Press / The MIT Press, 1996.

[30] H. Homayouni, S. Hashemi, and A. Hamzeh, "A lazy ensemble learning method to classification," *IJCSI International Journal of Computer Science Issues*, vol. 7, no. 5, pp. 344–349, 2010.

[31] I. M. Galván, J. M. Valls, M. García, and P. Isasi, "A lazy learning approach for building classification models," *International Journal of Intelligent Systems*, vol. 26, pp. 773–786, May 2011.

[32] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO," *Optik*, vol. 125, pp. 1439–1443, Feb. 2014.

[33] X. Liu, "Spatial and temporal dependence in house price prediction," *The Journal of Real Estate Finance and Economics*, vol. 47, pp. 341–369, Jan. 2012.

[34] M. Kuntz and M. Helbich, "Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging," *International Journal of Geographical Information Science*, vol. 28, pp. 1904–1921, May 2014.

[35] Z. Önder, V. Dökmeci, and B. Keskin, "The impact of public perception of earthquake risk on istanbul's housing market," *Journal of Real Estate Literature*, vol. 12, no. 2, pp. 181–194, 2004.

[36] E. Ozus, V. Dokmeci, G. Kiroglu, and G. Egdemir, "Spatial analysis of residential prices in istanbul," *European Planning Studies*, vol. 15, pp. 707–721, Apr. 2007.

[37] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," *Expert Systems with Applications*, vol. 42, pp. 2928–2934, Apr. 2015.

[38] G. I. Webb, *Lazy Learning*, pp. 571–572. Boston, MA: Springer US, 2010.

[39] "Zillow api homepage." https://www.zillow.com/howto/api/APIOverview.htm.

[40] D. S, "Crime in the united states 2011 : Violent crime." http://crimeanalystsblog.blogspot.com/2014/02/how-to-calculate-crime-rate.html, 3 February 2014.

[41] A. B. Sanjuán, *Model Integration in Data Mining: From Local to Global Decisions*. PhD thesis, 2012.

[42] K. R. Harney, "Condos may be appreciating faster than single-family houses." http://wapo.st/2OZc33S, 18 April 2017.

[43] G. Adomavicius and J. Zhang, "Stability of recommendation algorithms," *ACM Transactions on Information Systems*, vol. 30, pp. 1–31, Nov. 2012.

[44] X. Li, C. X. Ling, and H. Wang, "The convergence behavior of naive bayes on large sparse datasets," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 1, pp. 10:1–10:24, 2016.

[45] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," pp. 1–38, 03 2004.

RegPattern2Vec: Link Prediction in Knowledge Graphs

Abbas Keshavarzi
Department of Computer Science
University of Georgia
Athens, USA
abbas@uga.edu

Natarajan Kannan
Institute of Bioinformatics
University of Georgia
Athens, USA
nkannan@uga.edu

Krys Kochut
Department of Computer Science
University of Georgia
Athens, USA
kkochut@uga.edu

***Abstract*—Link prediction is an important task in many domains, including health sciences, biology, recommender systems, social networks, and many more. It is one of the problems residing within the intersection of knowledge graphs and machine learning. Link prediction aims to discover unknown links between entities in a graph using various techniques. However, due to the size of knowledge graphs today and their complexity, it is a challenging and time-consuming task. In this work, we present RegPattern2Vec, a method to effectively sample a large knowledge graph to learn node embeddings, while capturing the semantic relationships between graph nodes with minimum prior knowledge and human involvement. Our results show that the link prediction using RegPattern2Vec outperforms related graph embedding approaches on large-scale and complex knowledge graphs.**

***Keywords*—knowledge graph, link prediction, machine learning**

I. Introduction

Today, knowledge graphs (KG) are gaining popularity in many domains. KGs can be stored and represented using standardized vocabularies, including Resource Description Framework (RDF), Web Ontology Language (OWL), and various graph databases that are rapidly gaining popularity, nowadays. Today, KGs often are very large and represent vast amounts of actionable data and researchers and computing practitioners have turned to graph data mining to leverage the KG data even further.

Recently, machine learning (ML) has been gaining popularity due to numerous successful applications in many different domains, including graph mining [1], image and video processing [2], [3], text mining [4], reinforcement learning [5], and many others. It is also applicable to data mining of KGs. Software systems can automatically discover interesting patterns in KGs, while not being explicitly programmed to achieve that task. There are various types of learning algorithms, such as supervised and unsupervised learning, reinforcement learning, and many more. Classification and clustering are some of the most popular algorithms for machine learning on KGs. Experts take advantage of different algorithms to create recommendation systems for social media networks, entertainment libraries, find similarities in bibliographic networks, and many more.

KGs in nature are similar to Heterogeneous Information Networks (HIN) [6], where a variety of node and/or relation types between are used to represent data. This diversity of types provides benefits for learning, as compared to Homogenous Network, where types of nodes and relations are uniform. Finding appropriate methods to capture this extra information and use it in ML algorithms is a significant challenge.

Another major area, in the intersection of machine learning and KGs, is KG completion and, more specifically, link prediction. This problem has two aspects: predicting the missing links. KGs are usually populated automatically using variety of internal and external resources and due to the incompleteness of those resources, there might be some known links that are missing in populated KG. Here, one task of machine learning methods is to find those missing links and suggest that there should be connections between them.

As the graph analytics techniques are computationally expensive, especially on large graphs, researchers often aim to reduce the dimensionality of a graph into a low dimensional space. Graph embedding aims to preserve the structure of the graph while representing it as low dimensional vectors [7].

Based on [7], there are six different categories to generate vectors from a graph. These include matrix factorization, deep learning, edge reconstruction-based optimization, graph kernels, generative models, and hybrid models. In this work, we use a deep learning approach, in which random walks are used to sample the graph. This approach is based on a family of models from Natural Language Processing (NLP) called word2vec [8]; we specifically use the modified version of skip-gram model [9] in producing the vector embeddings. Skip-gram attempts to find the semantic similarity between words in a context by learning a meaningful representation for each word used in sentences in a corpus or documents. The main intuition is that we can discover the meaning of a word by understanding other words appearing close in a sentence. In the basic word2vec approach, the algorithm accepts a sentence and considers a window, usually of size 5 to 10, around the word of interest (center word) and generates training examples for a simple Neural Network (NN) with one hidden layer. The training examples are pairs of the center word and each of the words within the window size (context words). Then it trains the Neural Network to maximize the probability of a context word, given a center word. Then, the

Funding from NIH/NCI (U01CA239106) for this work is acknowledged.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

weights in the trained network are used as embeddings for each word in the corpus dictionary.

In this paper, we adopted this NLP method to our novel graph flattening approach using regular expressions to produce vector representation for the nodes in the graph. We formulate the link prediction as classification problem, using a model trained on the vector embeddings of the pairs of nodes connected by the link of interest. Our method, which we call RegPattern2Vec, shows high accuracy and discovers interesting possible links between unlinked nodes in the graph.

II. RELTED WORK

Past research includes several approaches to capturing the semantic relationship between graph nodes. Matrix factorization-based methods generate embeddings by factoring the matrix that represents the relations between nodes [10]. The matrix can be an adjacency or a Laplacian matrix, among other methods. Another technique that is used to generate vector embeddings, is Graph Kernels. Graph kernels are a measure of similarity of pairs of graphs. For example, [11] uses graph kernels for subtrees and similarity of instances in the original graph by counting common structures. Intuitively, vector embeddings of nodes with similar structure in a subtree are closer to each other.

Generative models are also popular as a graph representation learning method. The generative and discriminative models play the minmax game, where the generator approximates the connectivity of a graph and a discriminator calculates the probability of edge's existence. They are used to perform link prediction and node classification [12]. Finally, we discuss two approaches that can be classified as Deep Learning (DL) methods. DLs using random walks, such as metapath2vec [9], and DLs not relying on random walks [13] and [14], utilizing other techniques for computing vector embeddings. Whether employing minimization of Margin-Based Ranking Loss for entities or constructing a multilayer graph with structural similarity of all nodes in level of hierarchy, their goal is to translate the graph to a low dimensional space, where it can be used for applications such as link prediction or node classification. It is worth mentioning that there are other techniques such as using Convolutional Networks [15] and Autoencoders [16], which we do not discuss in this section because they fall into an entirely different type of methods. Generally, most GCN approaches suffer from scalability problems when the graph is large and dense due to the number of parameters and so are impractical to use. Therefore, different type of approach with a smaller number of parameters and hyperparameters will be required to large KGs such as approaches to sample the graph and learn representation in more efficient way, which will be discussed next.

Since in our method we rely on deep learning using random walks, we will focus on similar approaches in greater detail. As the computation of all possible walks on a graph is computationally expensive, researchers tend to choose random walks on the graph using some probability distribution. This would be sufficient for walks on homogenous networks. However, in heterogeneous networks with multiple types of nodes and edges, we need to differentiate types of nodes/edges when selecting the next node. Metapath2vec++ [9] is an approach that considers a fixed path of node types, which is called a meta-path. For instance, on a DBLP computer science bibliographic dataset [17], the meta-paths APA, APVPA, and OAPVPAO were chosen, where A represents the author, P paper, O the organization, and V the venue. These meta-paths are used to bias the random selection of the next node with the appropriate type in a random walk. Although some results of automatically discovering meta-paths have been published [18], usually domain experts are needed to choose the meta-paths of their interest for random walks. A domain expert should fully understand the KG organization. Although some tools for a KG schema discovery exist, such as KGdiff [19], due to the complexity of KGs and the hierarchy of concepts in them, it becomes difficult and time consuming to create appropriate meta-paths. Their selection should consider several aspects, e.g., the problem we want to solve (node classification, clustering, etc.), either the selected subgraph or the whole KG, and the meta-path coverage within it (number of nodes that can be reached using meta-paths/meta-graphs). Although the meta-path approaches on HINs were often used for various tasks, they are not useful for capturing more complex relations among entities. Each type of node must be explicitly defined or the meta-path does not capture the variation of attributes linked to the nodes. [20] proposed meta-graphs, which in a nutshell are meaningful combinations of meta-paths. For instance, if there are two meta-paths as APA, AVA, a possible meta-graph would be A-[P/V]-A. The use of meta-graphs as constrained to random walks was tested in [21], but the choice of meta-graphs where they can improve the overall model poses another challenge. To overcome these weaknesses, we introduce RegPattern2Vec, where a regular expression guides the random walks to sample sequences of nodes in a more efficient way, especially for large KGs. where other methods fail due to the lack of scalability. The embeddings produced by our representation learning captures all of the necessary characteristics of each node to be used for high accuracy link prediction.

III. PRELIMINARIES AND PROBLEM DEFINITION

In this section, we first introduce some preliminary concepts and then define the problem of Link prediction on KGs using Random Walks constrained by Regular Expressions. As of this writing, a single, commonly accepted definition of a knowledge graph does not exist, yet, and many researchers provide their own definitions. A good analysis is of KG definitions has been presented in [22], [23]. Here, we will use graph-based definition.

Knowledge graphs. A knowledge graph (KG) is a directed graph $\mathcal{G}: (\mathcal{V}, \mathcal{E})$ whose nodes $v_i \in \mathcal{V}$ are entities and edges $e_i \in \mathcal{E}$ are relations connecting the entities. Edges, usually referred to as triples of the form (v_i, e, v_j), represent some type of semantic dependence between the connected entities. Nodes have an associated type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{T}$, where $\mathcal{T}$ denotes a node type, while edges have an associated type mapping function $\varphi : \mathcal{E} \rightarrow \mathcal{R}$ where $\mathcal{R}$ denotes a relation type set.

Given a knowledge graph G, an edge with a relation type R connects source nodes of type S and target nodes of type T defines a meta edge $S \xrightarrow{R} T$. A set of all such meta edges for G is called a *schema graph* (sometime referred to as *meta-template*). In fact, schema graph is a directed graph defined over node types $\mathcal{T}$, with edges from $\mathcal{R}$, denoted as $G_S = (\mathcal{T}, \mathcal{R})$ [24].

Knowledge graphs are often represented as RDF [25] datasets, where nodes (entities) and relationships are represented using Uniform Resource Identifiers (URI). Nodes and relationships have assigned types, given as URIs, as well. Furthermore, these types may form type hierarchies. RDFS [26] is often used to define a schema for an RDF knowledge graph. Knowledge graphs are closely related to Heterogeneous Information Networks (HIN). In HINs, object (node) and relationship *types* both contain more than one element, that is, there are multiple labels for graph nodes and multiple labels for edges. In case that type sets are singletons, the Information Network is called a Homogeneous Information Network (all nodes in the network are of the same type and all edges are of the same type).

Despite the obvious similarity of KGs and HINs, there exist important differences between them. An important distinction is that in HINs, a relation of type $R \in \mathcal{R}$ uniquely determines the types of source and target nodes that can be connected by the relation R. In knowledge graphs, however, a relation of type R may connect nodes of many different source types and target types. Many other differences KGs and HINs exist, but they are not important for the research presented in this paper.

It has been shown that the results of various graph-embedding tasks are sensitive to the selection of a specific meta-paths [27]. In this paper, we propose a method of using regular expressions as a specification of a wide range of semantic relationships to be incorporated in random walks.

Regular Patterns on KGs. Let G be a knowledge graph, $G = (\mathcal{V}, \mathcal{E})$, with a node type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{T}$ and an edge type mapping function $\varphi : \mathcal{E} \rightarrow \mathcal{R}$. A Regular Pattern on G is a regular expression (pattern) [24] *r* formed over either set $\mathcal{T}$ or $\mathcal{R}$ as the alphabet.

We will not formally define regular expressions, since they are commonly used in computing, today. Briefly, a regular expression defines a set of strings (sequences, or words) over an alphabet; it defines a regular set [24]. We assume a standard format of regular expressions used in many programming languages today, for example in Python [28]. Here, we will only use a subset of possible regular expression constructs, including the concatenation, the alternative (**|**), repetitions of zero or more times (the Kleene star *****), one or more times (+), specified number of times ({n, m} n through m, and {n,} at least n), and the complement matching [^xy] (any symbol other than x or y). Note that a meta-path, as defined in the metapath2vec algorithm [9], can be regarded as a regular expression over $\mathcal{T}$ (or $\mathcal{R}$) since it can be regarded as a concatenation of node (or relation) types placed on the meta-path (again, a relation in HINs uniquely determines the source and target node types and vice-versa).

As an example, given node types $T_i \in \mathcal{T}$ in a KG, we could formulate a variety regular patterns over node types, for example, $T_1\ T_2\ T_3$, $T_1\ (T_2 \mid T_3)\ T_4$, $(T_1 \mid T_2)\ T_3$+ T_4, any many others. Similarly, given edge types $E_i \in \mathcal{R}$ in a KG, we could create regular patterns over edge types, such as $[\text{^}R_1]\ R_2\ R_3\text{*}\ R_4$ or $R_1\ (R_2 \mid R_3)\text{*}\ R_4$. Intuitively, a regular pattern defines a set of node (or edge) type sequences (a regular set), which we use to bias random walks on a KG to follow semantically relevant data.

Many knowledge graphs utilize complex hierarchies of node (entity) types such as Yago[29], DBpedia[30], and NELL[31]. Consequently, defining regular patterns based on node types is impractical, as they would require costly type inference (node types in actual sampled walks could be subtypes of those included in the defined regular pattern). Hence, using regular patterns on edge types may be a better choice.

In this paper, we focus on link prediction and so we rely on a specific general format of regular patterns for biasing the random walks. Assume that given a KG, we need to predict an edge (*h, r, t*), where h, t $\in \mathcal{V}$, r $\in \mathcal{E}$, $\phi(h) = H$, $\phi(\text{t}) = T$, and $\varphi(\text{r}) = R$. If the KG has a simple (non-hierarchical) structure of node types, our expression pattern can be based on node types and have a general format $\boldsymbol{H[\text{^}T]\text{+}\ H\ T}$. However, in a KG with a large node type hierarchy, we use edge-based patterns with the general format of $\boldsymbol{[\text{^}R]\{2,\}\ R}$. That is, at least 2 edges with relation types different than the one to be predicted followed by the edge relation type to be predicted.

The intuition behind the above regular (expression) pattern formats primarily comes from the observations of meta-paths and meta-graphs, where the similarity of two nodes is calculated based on the number of paths between them that follow a specific meta-path [32]. While some works [9], [21], [33] use symmetric meta-paths to calculate similarity between nodes of the same type, others [34] use more complex meta-paths for a different types of nodes. RegPattern2Vec follows the latter idea of finding the similarity of nodes with different types, but meta-paths and meta-graphs must be explicitly designed by domain experts and each such meta-path needs to be used in a separate experiment. In general, individual meta-paths cannot capture all possible semantically relevant connections between the nodes of interest. RegPattern2Vec, due to its use of regular patterns cover a large set of meta-path-like connections and takes advantage of a multitude of such semantic connections in one experiment.

We can explain the RegPattern2Vec using a simple example shown in Fig. 1. The graph contains 4 different relations: R1, R2 R3, and R4 (red, green, blue and yellow, respectively). The link that we want to predict is R2. Following the regular pattern, [^R2]{2,}R2, we have R1, R3, R4 ∈ **[^*R2*]**. An example walk following the pattern is $a1 \xrightarrow{R1} c1 \xrightarrow{R1} a2 \xrightarrow{R2} d1$. The intuition is that if we found two nodes (a1 and a2) where they link to another common node (c1), they are semantically related and node a1 might have the relationship R2 to d1, as well, which is useful for the link prediction task.

In the example nodes a1 and a2 have two common nodes b1 and c1 and plus the above path, $a1 \xrightarrow{R4} b1 \xrightarrow{R4} a2 \xrightarrow{R2} d1$ is also allowed based on the regular pattern. It is possible that the intermediate nodes have a relationship with other nodes, such as nodes c1, c2, and c3. These links might be considered as loops within the same type of nodes. If the nodes are the same or different type, a regular pattern can find such path and there is hyperparameter to control number of such possible loop. If this parameter is 1 the path $a1 \overset{R1}{\rightarrow} c1 \overset{R3}{\rightarrow} c2 \overset{R1}{\rightarrow} a3 \overset{R2}{\rightarrow} d2$ is allowed and if it was 2, a walk from node c1 can reach node c3 and the path $a1 \overset{R1}{\rightarrow} c1 \overset{R3}{\rightarrow} c2 \overset{R3}{\rightarrow} c3 \overset{R1}{\rightarrow} a4 \overset{R2}{\rightarrow} d3$ is also permitted. Here, a1 and a2 are more similar than a1 and a4 but it can be beneficial to capture those paths, as well. By this logic, a random walk constrained by a regular pattern can reach different paths to capture more links within the graph, if necessary.

IV. RegPattern2Vec

RegPattern2Vec relies on random walks to produce graph embeddings. Random walks on knowledge graphs are constrained to those matching a defined regular pattern.

A. Random Walks

Random walks in RegPattern2Vec are designed to sample an arbitrary number of walks. Their number can be controlled by parameters, such as "walk length" and "number of walks" (per starting node). Even though the knowledge graphs we use are defined as directed graphs, here, we treat them as undirected. We do this to be able to sample paths from all possible paths according to a defined distribution. Having an undirected knowledge graph and a regular expression pattern, a random walk can be started from any instance of the starting edge (or node) type in the pattern.

A regular pattern is converted to an equivalent Deterministic Finite Automaton (DFA) *M* [24] with the same input alphabet as the one used in the regular expression pattern definition. We will not introduce a formal definition of a DFA here but simply state that a DFA has a finite set of states, an input alphabet, a transition function δ, a starting state, and a set of final states. In our case, the state transition function is defined as δ: S×$\mathcal{R} \rightarrow S$ (or δ: S×$\mathcal{T} \rightarrow S$) and specifies state transitions based on edge (or node) types, depending on the regular pattern expression. We assume that δ is a partial function and for some states, transitions on some relation (or node) types may be undefined. We use the transition function δ of *M* to define the probabilities of node selections in our random walks.

It is obvious that if we repeat the walk from each node, we will discover more paths as the node might link to multiple nodes, which are allowed based on the regular pattern. We will call this parameter "number of walks". We will discuss how to choose the parameter and analysis of their impacts in the next section. As in each step there might be multiple choices, randomization will help the random walk to select a next node in each step. In the scale-free networks, where the degree distribution follows the power law, some nodes, often referred to as hubs, have high number of incoming/outgoing edges. It seems that such nodes would dominate the random walks as they have a higher probability of being reached when the next node selection follows the normal distribution. As the frequency of nodes appeared in the walk is the key point of representation learning, we bias the walks while using a regular expression pattern and its equivalent DFA *M*, using the formula below:

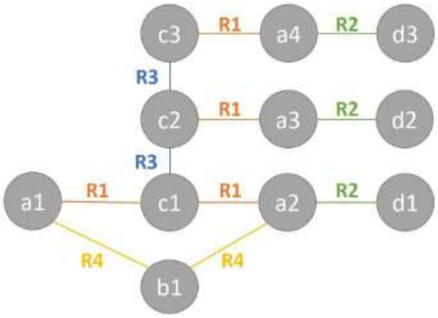


Fig 1. Illustration of random walks using regular expression.

$$P(v^{i+1} \mid v^i, M) = \begin{cases} \dfrac{\frac{1}{|N_{v^{i+1}}|}}{\sum_{v \in N_{v^i}} \frac{1}{|N_v|}} & (v^i, r^i, v^{i+1}) \in \mathcal{E}, M \text{ is in state } s_i \text{ and} \\ & \quad \delta(s_i, \varphi(r^i)) = s_{i+1} \\ 0 & (v^i, r^i, v^{i+1}) \in \mathcal{E}, \text{but } \delta(s_i, \varphi(r^i)) \text{ is undefined} \\ 0 & (v^i, r^i, v^{i+1}) \notin \mathcal{E} \end{cases} \quad (1)$$

(1)

Here, $|N_v|$ is the of degree of node v, v^i indicates the current node and $v^{i+1} \in N_{v^i}$ is the next candidate node, where N_{v^i} is the entire neighborhood of node v^i. Furthermore, s_i is the current state of *M*, that is, after processing the sequence of edge types $\varphi(r^1)\ldots\varphi(r^i)$, and *M*'s transition function from state s_i on edge type $\varphi(r^i)$ is defined and leads to state s_{i+1}. We can easily create a similar formula using a DFA *M* with node types, instead of relation types, as shown in (1). By fine tuning the previously mentioned parameters, this probability distribution will be sufficient to reach as many node as possible (of reachable nodes) to be included in walks, which results in more accurate vector embeddings.

B. Representation Learning

RegPattern2Vec converts the graph into the sequences of nodes and, from this point, we treat the nodes as words in sentences, as produced by random walks. These sentences are used as input to a model, similar to the one used in metapth2vec++ [9], for generating node embeddings. This model is an improved version of the original skip-gram model, as it takes into consideration types of edges (nodes). This allows the embeddings to capture the similarity of edges (nodes) based on their types (often considered as classes) along with their appearance of closely connected nodes, as required by the pattern.

C. Link Prediction

RegPattern2Vec formulates Link Prediction in KGs as a classification problem. Each existing link (or edge) of interest is represented as a vector of real numbers and is treated as a positive example for training the model. We can combine two vectors using Hadamard product and used the resulted vector as features for machine learning algorithm with label as positive. As the negative examples are typically not included in knowledge graphs, we create combinations of pairs of nodes that are not connected by edges in the graph and use them as negative examples, which as a common approach in the published research in this area. RegPattern2Vec uses an element-wise multiplication of vectors as the combination operation, which transforms the pairs of nodes to another space. These examples are used to train a classification model, such as Logistic Regression, which can be used for link prediction.

V. EXPERIMENTS

A. Datasets

In our experiments, we used two popular datasets, YAGO39K [35] and NELL [5]. YAGO39K contains a subset of the YAGO knowledge base [29], which includes data extracted from Wikipedia, WordNet and GeoNames. This subset contains 123,182 unique entities (nodes) and 1,084,040 edges, using 37 different relation types. A histogram of relation type distribution in the YAGO39K dataset is shown in Fig. 2. NELL is a knowledge graph mined from millions of Web documents and contains 49,869 unique nodes, 296,013 edges, using 827 relation types. In contrast to the Heterogeneous Information Networks, both datasets include many edges with the same relation type connecting nodes with many source types and/or many target node types.

B. Link Prediction Experiments

Following the work on link prediction on the YAGO dataset [35], we chose three different relation types namely *isLocatedIn*, *isCitizenOf*, and *isLeaderOf*. Based on the relation to be predicted, we split the KG data, as reported in Table 1. We need to extract some number of edges from each of the three types into three different test sets for three different tasks. To do so, we utilized minimum spanning tree to capture the minimum number of nodes that can be added to the test set while having the nodes in the training set. It is necessary, because our method requires that the node exist in the training data, although the node does not necessarily need to have the edge of interest in the training set. However, it can have other relations with other types of nodes. So, for each task, we extract the maximum number of edges of interest from the graph as the test set, while the remaining edges of interest and instances of other relations are combined to form the training data. We have already discussed how to use the edge of interest in creating the training data with positive and negative examples for a binary classification model. We can follow the same process to make examples for testing in order to evaluate the performance of our method.Following the work described in [5], we chose two relations *CompetesWith* and *playsAgainst* for our link prediction experiments with the NELL dataset. The cited work reported the best metapaths used to predict these relations and used these metapaths for comparison. We performed a similar process (as described above) in order to split the data into a train and test sets.

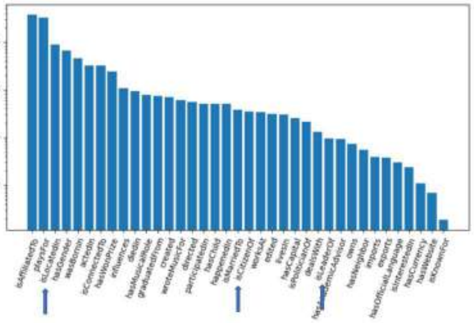


Fig 2. Distribution of relation types in the YAGO39K dataset.

TABLE 1. STATISTICS OF SPLIT OF DATA FOR DIFFERENT EXPERIMENTS

	isLocatedIn	isCitizenOf	IsLeaderOf
Train Set Graph	1,039,499	1,080,570	1,083,079
Train set edges	44,542	3,128	855
Test set edges	44,541	342	106

C. Performance

Having approximately a balance training set we train a logistic regression for binary classification. We evaluate our model using 10-fold cross-validation and test it using unseen set that we extracted from the graph.

To evaluate the performance of RegPattern2Vec on YAGO39, we chose metapath2vec as our baseline, shown in Fig 3, to demonstrate that RegPattern2Vec can cover more meta-paths without explicitly defining them and perform better, too. To get the best meta-paths for *isLocatedIn* and *isCitizenOf* relations, we chose the ones that achieved the best scores reported in the literature. However, were not able to find the best metapaths for *isLeaderOf*, and we designed them ourselves. After our experiments, Person $\xrightarrow{\text{isLeaderOf}}$ city $\xrightarrow{\text{isLocatedIn}}$ country $\xleftarrow{\text{isLocated}}$ city was the best meta-path. For example, a leader of state, is specified as leader of cities with the state. That information suggesting the earlier meta-paths to perform better than any other meta-paths. The regular pattern for three aforementioned links were defined as follow:

$$[\hat{}isLeaderOf\{2,\}\,isLeaderOf]$$

$$[\hat{}isLocatedIn]\{2,\}\,isLocatedIn]$$

$$[\hat{}isCitizenOf]\{2,\}\,isCitizenOf]$$

To run the experiments, we kept the same parameter settings for each of the method, when working on a specific relation prediction. The settings included the number of walks from each node, the maximum walk lengths, the Logistic Regression parameters, and the metrics to evaluate their performance.
To select the best algorithm for binary classification we examine two famous and popular algorithms, Logistic Regression and Random Forest, and we tested several experiments with both, and it seems that logistic Regression in our case the best performing method.

Therefore, all the experiments are performed with logistic Regression for evaluation purposes. RegPattern2Vec shown superior performance over metapath2vec method with best meta-paths possible. It is expected that RegPattern2Vec would perform better in link prediction tasks where it can obtain more semantics by exploring different path within the knowledge graph. In the case of the *isLeaderOf* relation, as the data does not contain information useful for predicting this relation, the performance is lower than for other relations.

For the evaluation of RegPattern2Vec on NELL datasets, we chose metapath2vec and used the best meta-paths reported in [5]

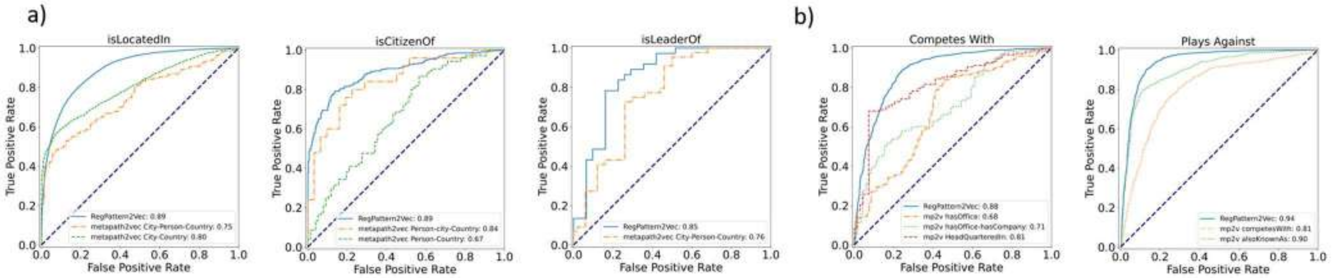


Fig 3. Comparing ROC of RegPattern2Vec with baseline on two datasets. a) Three relations on YAGO39K b) Two relations on NELL

for each of the relations. For *CompetesWith*, there were top five meta-paths. Our experiment showed that *HeadQuarteredIn* performs the best although it was ranked as the third in top 5. And for the *PlaysAgainst* relation, we selected two best meta-paths for evaluation purposes. We used the following regular expression patterns:

$$[\hat{}CompetesWith]\{2,\}\ CompetesWith$$

$$[\hat{}PlaysAgainst]\{2,\}\ PlaysAgainst$$

As shown in Fig. 3, RegPattern2Vec outperforms metapath2vec with different meta-paths for both relations in correctly predicting the unseen links between different nodes. The ROC shows that across most of the threshold the performance of RegPattern2Vec is significantly higher than metapath2vec with different metapaths.

D. Patterns Discovered of Random walk guided by Regular Pattern

Based on the dataset and underlying schema, RegPattern2Vec discovers different patterns in the data and uses them to accurately predict possible links in the KG. Fig. 4 shows the top 30 frequent meta-paths capture with RegPattern2Vec without explicitly specifying them. Although the number of possible relation sequences is very high, especially allowing for repetitions, some of the sequences can be seen frequently, based on the graph and they might significantly influence the vector embeddings. So, it is important to have a way to capture most of the patterns in the graph and consequently all path instances to allow the representation learning model to produce more accurate embeddings. This cannot be achieved by meta-paths as prior knowledge is needed to design each meta-path and they can be easily missed, especially when the schema of KG is unambiguous or not well-defined. Although just the frequency of each pattern in the data does not provide a good measure of how the pattern is important or useful for the link prediction problem, it can provide information about the graph itself and what are the frequent patterns in the graph. Then, the representation learning model can decide how frequent two specific nodes appear in the same context and provide a closer embedding for each of them, based on their local structures and neighbors.

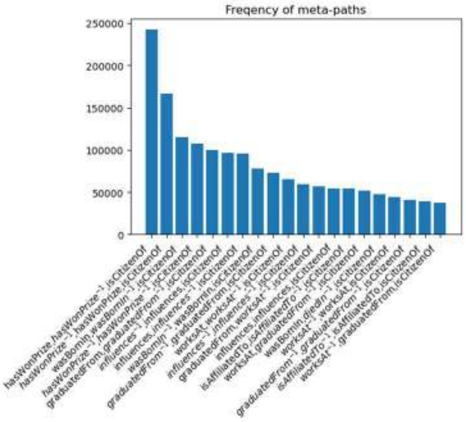


Fig 4. The top 30 most frequent relationship patterns discovered by RegPattern2Vec for *isCitizenOf* relation.

E. Effect of Hyper-parameters

In this section we studied the effect of two important hyper-parameters in the random walks on performance and elapse time namely the number of walk and length of walks. The effect of different choices of these two on the AUC ROC is demonstrated in Fig. 5a for *playsAgainst* relation on NELL dataset. As more nodes are connected to each other, we need to sample more paths by increasing the number of walks or walk length. On the other side, due to the large size of KGs, one of the challenges of learning the embeddings is scalability and efficiency. we showed that increasing the number of walks can improve the performance, as the random walks are able to trace more paths in the data. Fig 5b shows how the increase of this parameter affects the elapsed time of the random walk, in this case when experimenting with the *competesWith* relation on the NELL dataset in two cases of walk length 10, and 100. As demonstrated the elapsed time is linearly related to the hyper-parameters.

VI. Conclusion and Future work

In this work, we presented RegPattern2Vec, where a regular pattern guides the random walks in a knowledge graph to efficiently sample sequences of nodes to learn high quality

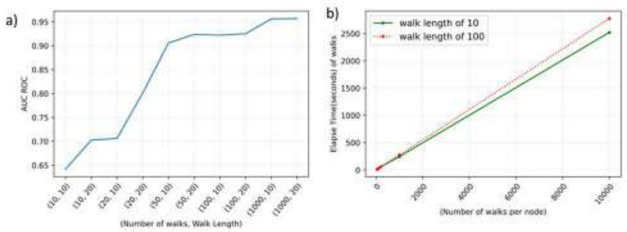


Fig 5. a) Effect of number of walks and walk length on AUC ROC b) Effect of number of walk per node on elapsed time of random walks with two different walk lengths 10 and 100.

embeddings for link prediction. We demonstrated link prediction using relation types, where the schema of knowledge graph is unknown, or node type hierarchy is complex. As a future direction for our work, we want to explore how to bias the random walks to favor the nodes or relations that might contribute more accurate link prediction results. Also, if the most frequent relations found in the patterns of walks would improve the results. We plan to achieve this by automatically tuning the bias score for each type of node (or relation), while training on the graph and checking the link prediction results to adjust the scores. The number of parameters to tune is related to the number of types or relations that we have in KG, which seems to be practical to a model to work with.

References

[1] I. Atastina, B. Sitohang, G. A. P. Saptawati, and V. S. Moertini, “A Review of Big Graph Mining Research,” 2018, Accessed: Feb. 20, 2018. [Online]. Available: http://iopscience.iop.org/article/10.1088/1757-899X/180/1/012065/pdf.

[2] A. R. Elias, N. Golubovic, C. Krintz, and R. Wolski, “Where’s The Bear?,” 2017 IEEE/ACM Second Int. Conf. Internet-of-Things Des. Implement., pp. 247–258, 2017, doi: 10.1145/3054977.3054986.

[3] M. Toutiaee, A. Keshavarzi, A. Farahani, and J. A. Miller, “Video Contents Understanding using Deep Neural Networks.”

[4] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, “Text mining and ontologies in biomedicine: Making sense of raw text,” Brief. Bioinform., vol. 6, no. 3, pp. 239–251, 2005, doi: 10.1093/bib/6.3.239.

[5] G. Wan, B. Du, S. Pan, and G. Haffari, “Reinforcement Learning based Meta-path Discovery in Large-scale Heterogeneous Information Networks,” Aaai, 2020, Accessed: Jun. 12, 2020. [Online]. Available: https://github.com/mxz12119/MPDRL.

[6] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, “A Survey of Heterogeneous Information Network Analysis,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 1, pp. 17–37, 2017, doi: 10.1109/TKDE.2016.2598561.

[7] H. Cai, V. W. Zheng, and K. C. C. Chang, “A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 9, pp. 1616–1637, 2018, doi: 10.1109/TKDE.2018.2807452.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality.” Accessed: Oct. 29, 2018. [Online]. Available: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[9] Y. Dong, N. V Chawla, and A. Swami, “metapath2vec: Scalable Representation Learning for Heterogeneous Networks,” 2017, doi: 10.1145/3097983.3098036.

[10] M. Belkin and P. Niyogi, “Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering,” 2002.

[11] P. Ristoski and H. Paulheim, “RDF2Vec: RDF graph embeddings for data mining,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9981 LNCS, pp. 498–514, doi: 10.1007/978-3-319-46523-4_30.

[12] H. Wang et al., “GraphGAN: Graph Representation Learning with Generative Adversarial Nets,” pp. 2508–2515, 2017, [Online]. Available: http://arxiv.org/abs/1711.08267.

[13] L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo, “struc2vec: Learning Node Representations from Structural Identity,” 2017, doi: 10.1145/3097983.3098061.

[14] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, “Learning Entity and Relation Embeddings for Knowledge Graph Completion.” Accessed: Oct. 02, 2019. [Online]. Available: www.aaai.org.

[15] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, “Modeling Relational Data with Graph Convolutional Networks,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 10843 LNCS, pp. 593–607, doi: 10.1007/978-3-319-93417-4_38.

[16] D. Wang, P. Cui, and W. Zhu, “Structural deep network embedding,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 1225–1234, 2016, doi: 10.1145/2939672.2939753.

[17] M. Ley, “Dblp computer science bibliography,” 2005. .

[18] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, “Discovering Meta-Paths in Large Heterogeneous Information Networks,” doi: 10.1145/2736277.2741123.

[19] A. Keshavarzi and K. J. Kochut, “KGdiff: Tracking the Evolution of Knowledge Graphs,” Proc. - 2020 IEEE 21st Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2020, pp. 279–286, 2020, doi: 10.1109/IRI49571.2020.00047.

[20] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li, “Meta structure: Computing relevance in large heterogeneous information networks,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, vol. 13-17-Augu, pp. 1595–1604, doi: 10.1145/2939672.2939815.

[21] D. Zhang, J. Yin, X. Zhu, and C. Zhang, “MetaGraph2Vec: Complex Semantic Path Augmented Heterogeneous Network Embedding,” 2018.

[22] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, “Convolutional 2D Knowledge Graph Embeddings.” Accessed: Dec. 23, 2019. [Online]. Available: www.aaai.org.

[23] W. L. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, and J. Leskovec, “Embedding Logical Queries on Knowledge Graphs.” Accessed: Aug. 20, 2019. [Online]. Available: https://papers.nips.cc/paper/7473-embedding-logical-queries-on-knowledge-graphs.pdf.

[24] M. Spiser, Introduction to the Theory of Computation. Cengage learning., 2012.

[25] R. Cyganiak, D. Wood, and M. Lanthaler, “RDF 1.1 Concepts and Abstract Syntax,” W3C Recommendation, 2014. https://www.w3.org/TR/rdf11-concepts/ (accessed Dec. 31, 2020).

[26] W3C, “RDF Schema 1.1,” 2014. Accessed: Dec. 31, 2020. [Online]. Available: https://www.w3.org/TR/rdf-schema/.

[27] X. Cao, Y. Zheng, C. Shi, J. Li, and · Bin Wu, “Meta-path-based link prediction in schema-rich heterogeneous information network,” Int. J. Data Sci. Anal., vol. 3, pp. 285–296, 2017, doi: 10.1007/s41060-017-0046-1.

[28] J. E. Friedl, Mastering Regular Expressions. O’Reilly Media, Inc., 2006.

[29] F. M. Suchanek, G. Kasneci, and G. Weikum, “Yago,” in Proceedings of the 16th international conference on World Wide Web - WWW ’07, 2007, p. 697, doi: 10.1145/1242572.1242667.

[30] J. Lehmann et al., “DBpedia-A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia LinkingLOD: interlinking knowledge bases View project DL-Learner View project DBpedia-A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia,” Semant. Web, vol. 1, pp. 1–5, 2012, doi: 10.3233/SW-140134.

[31] T. Mitchell et al., “Never-ending learning,” Commun. ACM, vol. 61, no. 5, pp. 103–115, 2018, doi: 10.1145/3191513.

[32] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, “Pathsim: Meta path-based top-k similarity search in heterogeneous information networks,” Proc. VLDB Endow., vol. 4, no. 11, pp. 992–1003, 2011, doi: 10.14778/3402707.3402736.

[33] C. Yang, M. Liu, F. He, X. Zhang, J. Peng, and J. Han, “Similarity modeling on heterogeneous networks via auto-matic path discovery,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11052 LNAI, pp. 37–54, doi: 10.1007/978-3-030-10928-8_3.

[34] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, “Meta-graph based recommendation fusion over heterogeneous information networks,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, vol. Part F1296, pp. 635–644, doi: 10.1145/3097983.3098063.

[35] X. Lv, L. Hou, J. Li, and Z. Liu, “Differentiating Concepts and Instances for Knowledge Graph Embedding.”

Design and Implementation of an Efficient Elliptic Curve Digital Signature Algorithm (ECDSA)

Yasin Genç
Dept. of Electrical-Electronics Eng.
Gazi University
Ankara, Turkey
yasin.genc@gazi.edu.tr

Erkan Afacan
Dept. of Electrical-Electronics Eng.
Gazi University
Ankara, Turkey
e.afacan@gazi.edu.tr

***Abstract*— Digital signatures are increasingly used today. It replaces wet signature with the development of technology. Elliptic curve digital signature algorithm (ECDSA) is used in many applications thanks to its security and efficiency. However, some mathematical operations such as inversion operation in modulation slow down the speed of this algorithm. In this study, we propose a more efficient and secure ECDSA. In the proposed method, the inversion operation in modulation of signature generation and signature verification phases is removed. Thus, the efficiency and speed of the ECDSA have been increased without reducing its security. The proposed method is implemented in Python programming language using P-512 elliptic curve and SHA-512 algorithm.**

Keywords—Elliptic Curve Digital Signature Algorithm (ECDSA), Elliptic Curve Cryptography, NIST P-521 Curve, Hash Function, Finite Fields, Inversion in modulation, Python.

I. Introduction

Information security is one of the issues that gain importance with developing technologies. Digital signature is a concept that has a key role in ensuring information security today. It is used in many areas such as health, banking, commerce, internet of things, electronic voting. Digital signatures are used to provide integrity, authentication, and non-denial [1].

ECDSA based on elliptic curve cryptography (ECC) is efficient compared to other signature algorithms such as Rivest Shamir Adleman (RSA) and digital signature algorithm (DSA) [2]. It was first accepted as the ISO standard (International Standards Organization) in 1998 and later the ANSI standard (American National Standards Institute) in 1999 and in 2000 as NIST (National Institute of Standards and Technology) and IEEE standards. It is accepted in international standards makes that is used widely.

ECDSA provides the same security level with smaller key size compared to other signature algorithms [3]. The security provided by RSA and DSA with 3072 bit key size is equal to the security provided by ECDSA with only 256 bit key size. Provides an advantage when resources such as processing power, storage space, bandwidth, and power consumption are limited [2]. Due to this feature, it is suitable for use in many applications such as Internet of Things, sensors, RFID, smartphones [4], smart cards, wireless devices [2].

There are many ECDSA variants in the literature and different algorithms are proposed. Some of the proposed methods want to increase speed, while others want to increase security [5-10]. In the proposed algorithm given in the study, the inversion operation in ECDSA is removed from the signature and signature verification phases without security level reduced. In addition, the multiplication operation is reduced in these phases**.** The inversion operation in modulation is added to the key pair generation phase. Since this phase is in the setup, it does not affect the signature and signature verification phases. Thus, a more efficient and security variant of ECDSA is designed. Designed algorithm as a variant of ECDSA is implemented in Python programming language using the P-521 elliptic curve recommended by NIST.

The rest of this paper is organized as follows. In the second section, hash functions are explained and SHA algorithms are compared. In the third section, ECC and its mathematical operations over finite fields are explained and also group order and elliptic curve discrete logarithm problem (ECDLP) are explained. In the fourth section, the ECDSA is introduced and key pair generation, signature generation and signature verification phases are explained. In the fifth section, the proposed algorithm are designed which is implemented in Python programming language. Finally, in the sixth section conclusions are given.

II. Hash Function

A one-way hash function $H(M)$, takes a variable length message, and produces a fixed length hash value h. Regardless of the length of the message M, even terabyte in size, the output value h is always fixed length. The message M cannot be found from the hash value h, denoted as one-way property. In this way, reverse engineering is not possible. This process is defined as follows [11].

$$h = H(M) \qquad (1)$$

The hash function can be used to ensure that the content of message M has not changed. If there is any change in the content, output value h will change. For this reason, it is used to ensure information security in many fields such as digital signature and digital forensic. The hash algorithm must provide to use safely one-way and collision resistant properties. Collision property is that the same output hash value h is produced for different input message values [12]. The collision case is defined as follows.

$$M_1 \neq M_2,\ \ H(M_1) = H(M_2) \qquad (2)$$

There are many different hash functions such as MD (Message Digests) family, SHA (Secure Hash Algorithm) family, RIPEMD (RIPE Message Digest) family, Whirlpool

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

family [17]. The most widely used today is the SHA (Secure hash algorithm) family. The properties of the different SHA algorithms are shown in TABLE I [12].

TABLE I. COMPARISON OF DIFFERENT SHA ALGORITHMS

Algorithm	Message Length (bits) (Maximum)	Block Size (bits)	Word Size (bits)	Hash (bits)
SHA 1	2^{64}-1	512	32	160
SHA 256	2^{64}-1	512	32	256
SHA 384	2^{128}-1	1024	64	384
SHA 512	2^{128}-1	1024	64	512

In this study, the SHA-512 algorithm is used. It is an improved version of the other SHA family [17].

III. Elliptic Curve Cryptography (ECC)

Elliptic curve cryptography (ECC) is a public key cryptography that was discovered independently of each other in 1985 by Neil Koblitz [13] and Victor Miller [14]. For ease of calculation, operations are performed on finite fields. The finite fields used in software and hardware applications may differ. Prime field (F_p) for software applications and binary field (F_{2^m}) for hardware applications are widely used [15]. The equation,

$$y^2 = x^3 + ax + b \ \ (mod\ p) \tag{3}$$

which is known as Weierstrass equation, where a and b are constant integers less than p that is a prime number, and satisfys the following condition.

$$4a^3 + 27b^2 \neq 0 \ \ (mod\ p) \tag{4}$$

Let define an elliptic curve $E(a,b)$ over prime field F_p. The values of a and b in equation (3) are selected as -1 and 0, respectively. $y^2 = x^3 - x$ is shown in Fig. 1.

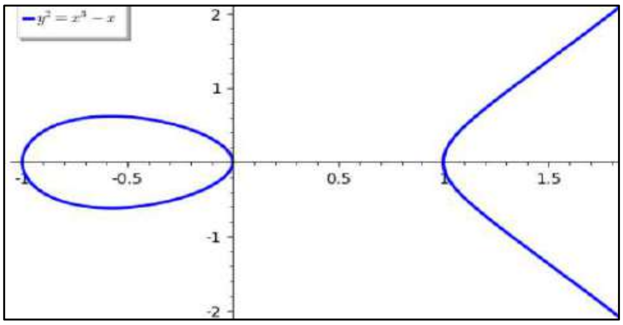


Fig. 1. $y^2 = x^3 - x$

A. Mathematical Operations of ECC

Operations on a finite field are based on the affine coordinate system which each point in the coordinate system is represented by the vector (*x*, *y*). Mathematical operations performed over finite fields are defined below [15].

1) Point addition

Let define two distinct points $P(x_1, y_1)$ and $Q(x_2, y_2)$ on the elliptic curve $E(a,b)$. The point addition calculation is given as follows.

$$R(x_3, y_3) = P(x_1, y_1) + Q(x_2, y_2) \tag{5}$$

- if $x_1 \neq x_2$, then the point addition is defined by

$$x_3 = \lambda^2 - x_1 - x_2 \tag{6}$$

$$y_3 = \lambda(x_1 - x_3) - y_1 \tag{7}$$

where

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1} \tag{8}$$

- if $x_1 = x_2$, the infinity point (O) is obtained.

2) Point Doubling

Let define two points $P(x_1, y_1)$ and $Q(x_2, y_2)$ on the elliptic curve $E(a,b)$, which are equal to each other i.e. $P = Q$. Point doubling can be defined as follows.

$$R(x_3, y_3) = P(x_1, y_1) + P(x_1, y_1) = 2P \tag{9}$$

where

$$P(x_1, y_1) = Q(x_2, y_2). \tag{10}$$

The calculation of the point doubling is given as follows.

$$x_3 = \lambda^2 - 2x_1 \tag{11}$$

$$y_3 = \lambda(x_1 - x_3) - y_1 \tag{12}$$

where

$$\lambda = \frac{3x_1^2 + a}{2y_1} \tag{13}$$

3) Point Subtraction

Let define two distinct points $P(x_1, y_1)$ and $Q(x_2, y_2)$ on the elliptic curve $E(a,b)$. The negative of the point Q can be defined as follows.

$$-Q(x_2, y_2) = Q(x_2, -y_2) \tag{14}$$

The subtraction of two points is given as follows.

$$P(x_1, y_1) - Q(x_2, y_2) = P(x_1, y_1) + Q(x_2, -y_2) \tag{15}$$

4) Scalar Multiplication

Let define any point $P(x_1, y_1)$ on the elliptic curve $E(a,b)$. The scalar multiplication $9P$ can be calculated as follows.

$$9P = 2(2(P+P)) + P \tag{16}$$

5) Inversion

Let *x* be a non-zero element and *y* be the inverse of *x* in modulo *p* (*x*, *y* ∈ F_p). The relation between *x* and *y* is given below.

$$xy = 1 \ (\mathrm{mod}\ p) \tag{17}$$

B. Group Order

Let define an elliptic curve $E(a,b)$ over finite field F_p. The number of points in $E(F_p)$ is called the order of $E(a,b)$ over prime field F_p which is denoted as $\#E(F_q)$ and is equal to n. It must be prime number and satisfy the following equation.

$$nG = O \text{ (infinity point)} \tag{18}$$

C. Elliptic Curve Discrete Logarithm Problem

The security of ECC is based on the difficulty of the elliptic curve discrete logarithm problem (ECDLP). Suppose we know the values of P and Q in Equation (19). Finding the value of k is quite difficult.

$$Q = kP \tag{19}$$

D. Recommended Elliptic Curves

There are elliptic curves recommended by NIST for U.S. Federal Government. Special curves are defined as over prime field F_p and over binary field F_{2^m}. Elliptic curves over F_p are called Edwards curves and Montgomery curves. The prime fields: P-192, P-224, P-256, P-384, P -521, W-25519, W-448, Curve25519, Curve448, E448, Edwards25519, Edwards448. Elliptic curves over binary field F_{2^m} are called Koblitz curves and pseudorandom curves. The binary fields: Curve K-163, Curve K-233, Curve K-283, Curve K-409, Curve K-571, Curve B-163, Curve B-233, Curve B-283, Curve B-409, Curve B-571 [16].

IV. Elliptic Curve Digital Signature Algorithm (ECDSA)

ECDSA was proposed in 1992 by Scott Vanstone which is based on the implementation of the digital signature algorithm (DSA) on the elliptic curves [3]. Its security is based on the elliptic curve discrete logarithm problem (ECDLP). ECDSA consists of three phases: key pair generation, signature generation and signature verification. Let examine these three phases below [3].

A. Key Pair Generation

1. Signer selects a random integer $d \in [1, n-1]$.
2. Compute $Q = dG$ hence $Q = (x_Q, y_Q)$.
3. Signer's public key is Q, and private key is *d.*

B. Signature Generation

1. Signer selects a random integer $k \in [1, n-1]$.
2. Compute $kG = (x_1, y_1)$.
3. Compute $r = x_1 \pmod{n}$. If r =0 then go to step 1.
4. Compute $k^{-1} \pmod{n}$.
5. Compute SHA-1(M) and convert this bit string to an integer h (*Hash*(M)=h).
6. Compute $s = k^{-1}(h + dr) \pmod{n}$. If s =0 then go to step 1.
7. The signature for the message M is (r, s).

C. Signature Verification

1. Verify r and s are integers and r, s ∈ $[1, n-1]$.
2. Compute $w = s^{-1} \pmod{n}$.
3. Compute SHA-1(M) and convert this bit string to an integer h (*Hash*(M)=h).
4. Compute $u_1 = hw \pmod{n}$ and $u_2 = rw \pmod{n}$.
5. Compute $(x_1, y_1) = u_1 G + u_2 Q$, $v = x_1 \pmod{n}$.
6. If $v = r$, the signature is valid, otherwise the signature is invalid.

D. Proof of ECDSA Signature Verification

Let's perform the steps below in order to proof the signature verification.

$$(x_1, y_1) = u_1 G + u_2 Q \tag{20}$$

If kG, hw, rw, w are put in equation (20).

1. $kG = u_1 G + u_2 Q$.
2. $kG = hwG + rwQ$.
3. $kG = hs^{-1}G + rs^{-1}Q$.
4. $kG = hs^{-1}G + rs^{-1}dG$.
5. $k = s^{-1}(h + dr)$.

Thus, signature verification has been proven.

E. Security of ECDSA

The security of ECDSA is based on the difficulty of solving the elliptic curve discrete logarithm problem (ECDLP). Suppose Eva is a black-hat hacker who want to forge our signature. She has domain parameters and our signature (*r, s*). If she gets hold of our private keys *d* and *k*, she can use our signature anywhere. Therefore, we must choose the elliptic curve parameters very carefully in accordance with the international standards.

If Eva starts from the signature (*r, s*) known to everyone, finds the equality between *k* and *d* according to the following equation (21). *h* is calculated by *h*=hash(*M*) and *r, s* are already known.

$$s = k^{-1}(h + dr) \tag{21}$$

$$k = s^{-1}(h + dr) \tag{22}$$

$$kG = (x_1, y_1) \tag{23}$$

$$Q = dG = (x_Q, y_Q) \tag{24}$$

Finding *k* and *d* are very diffucult due to ECDLP. The *k* value must be selected independently for each message when signing multiple messages. If the value of *k* is not selected differently, the value of *d* can be recovered. Suppose we assume the same value of *k* for two different messages. Since the same *k* value is used, *r* value will be the same. The obtained signature will be (r,s_1) and (r, s_2). The hash values h_1 and h_2 obtained for two different messages are different.

$$s_1 = k^{-1}(h_1 + dr) \tag{25}$$

$$s_2 = k^{-1}(h_2 + dr) \tag{26}$$

$$ks_1 - ks_2 = (h_1 + dr) - (h_2 + dr) \tag{27}$$

where

$$ks_1 - ks_2 = (h_1 - h_2) \tag{28}$$

$$k = (h_1 - h_2) / (s_1 - s_2) \tag{29}$$

$$d = (ks_1 - h_1) / r \tag{30}$$

First *k* is calculated form equation (29), then *d* from equation (30). However, this probability is ignored by choosing a random number.

V. Proposed Algorithm and Implementation

In the ECDSA, inversion operation in modulation is performed in both signature generation and signature verification. Inversion operation in modulation is a mathematically difficult operation. For this reason, signature generation and signature verification times are getting longer. The use of ECDSA in many areas depends on the speed of these processes. Security can be increased in ECDSA, but it is difficult to make the algorithm more efficient at the same security level. In the proposed method, the inversion in signature generation and signature verification phases is eliminated with the change in key pair generation algorithm. In the key pair generation algorithm, public key Q is calculated by taking the inversion of randomly selected integer *d*. The proposed method is given below. The computer used has an Intel (R) Core (TM) i7-7500U CPU processor and 8GB of RAM. Algorithms have been run 7 times with the "timeit" command and averaged.

A. Proposed Key Pair Generation Algorithm

1. Signer selects a random integer $d \in [1, n-1]$.
2. Compute $Q = (d^{-1}\, mod\ n)G$.
3. Signer's public key is Q, and private key is *d*.

B. Proposed Signature Generation Algorithm

1. Signer selects a random integer $k \in [1, n-1]$.
2. Compute $kG = (x_1, y_1)$.
3. Compute $r = x_1$ (mod n). If r =0 then go to step 1.
4. Compute SHA-512 *(M)* and convert this bit string to an integer *h* (*Hash(M)=h*).
5. Compute $s = d(k-h)$ (mod n). If s =0 then go to step 1.
6. The signature for the message *M* is (r, s).

C. Proposed Signature Verification Algorithm

1. Verify r and s are integers and *r, s* $\in [1, n-1]$.
2. Compute SHA-512 (*M*) and convert this bit string to an integer *h* (*Hash*(*M*)=*h*).
3. $u_1 = s$ (mod n) and $u_2 = h$ (mod n).
4. Compute (x_1, y_1)=$u_1Q + u_2G$, $v = x_1$ (mod n).
5. If $v = r$, the signature is valid, otherwise the signature is invalid.

D. Proof of Signature Verification

Let perform the steps below in order to proof the signature verification.

1. $s = d(k-h)$
2. $sd^{-1} = (k-h)$
3. $G(sd^{-1}) = G(k-h)$
4. $Gsd^{-1} = kG - hG$
5. $kG = hG + Gsd^{-1}$
6. $kG = hG + sQ$
7. $kG = u_2G + u_1Q$

Thus, signature verification has been proven.

E. Implementation

Firstly, let choose the NIST P-521 (p=$2^{521}-1$) curve over the prime field for implementation. The NIST P-521 domain parameters are shown in TABLE II. Then, let choose a message *M* that is "This is my signature".

TABLE II. NIST P-521 DOMAIN PARAMETERS

p	0x 1FF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF
a	-3
b	0X 51 953EB961 8E1C9A1F 929A21A0 B68540EE A2DA725B 99B315F3 B8B48991 8EF109E1 56193951 EC7E937B 1652C0BD 3BB1BF07 3573DF88 3D2C34F1 EF451FD4 6B503F00
n	0X 1FF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFA 51868783 BF2F966B 7FCC0148 F709A5D0 3BB5C9B8 899C47AE BB6FB71E 91386409
G_x	0X C6 858E06B7 0404E9CD 9E3ECB66 2395B442 9C648139 053FB521 F828AF60 6B4D3DBA A14B5E77 EFE75928 FE1DC127 A2FFA8DE 3348B3C1 856A429B F97E7E31 C2E5BD66
G_y	0X 118 39296A78 9A3BC004 5C8A5FB4 2C7D1BD9 98F54449 579B4468 17AFBD17 273E662C 97EE7299 5EF42640 C550B901 3FAD0761 353C7086 A272C240 88BE9476 9FD16650

Later, let's implement the key pair generation algorithm.

- Signer selects a random integer $d \in [1, n-1]$.

d=4764934111023554209164756839776691760950357761406156323126415632553003634680741850157564093178936351777068447570149848499099892358593730272618669107276817092.

- Compute $Q = (d^{-1}\, mod\ n)$*(G_x, G_y)*.

$\boldsymbol{d^{-1}(mod\ n)}$ = 2550551988987719524518416525813727106082231234352413662480938064692684876589607626489624103582323548791518046688576531225954939454358235923884330108721986044.

Q=(697589017079691619811651998078043754419665610919989743845304218683989330510289996737581789119023134658438591681299224078232547270027753749678932542964311 53 , 374013274480870730435871093074676389930487437769659703207735684537166084818688647166319034070356308955652745845814220304369974048179430507227238476796810886 1).

Later, let's implement the signature generation algorithm.

- Signer selects a random integer $k \in [1, n-1]$.

k=5427075795105267971775828753324518913843418964409127171774736678067416337955213290722244443903505208675745658209223805258285430440610821156685094504004397628.

- Compute $k(G_x, G_y) = (x_1, y_1)$.

$(\boldsymbol{x_1}, \boldsymbol{y_1})$=(2812016407114405561511487321626266331752141256561994586621921523149233387593949797763305037322611761926014035715659087053633252036152283695578449212012199543 , 1155945241278564416620287705448325378731925202754723632892005123564204682810543342701655677861367791756416406485264907787642904522849642300801043628621282942).

- Compute SHA-512 (This is my signature*)=M.*

ffb0b8b412770e7fbf5ab45b856c370cbee55dfa342ce9bb4823d948695b55ebe1a46d033ad079cc66f5a98f428f2a2844f3077b730f16372a65f182dd4de8aa.

- Convert this bit string to an integer *h (Hash(M)=h).*

h=13391588585886738632600139556386100265971188810783796321769392220422220347727593927717784174305529098666089325052246885816774225364695776567825092379011242.

- Compute $s = d(k-h)$ (mod n).

s=26075593207638998033271362760615191934391781587755738878810533023313352679412653560498372269275807196507024803695897036778704994547146254422036295390994498 73.

r=$\boldsymbol{x_1}$=28120164071144055615114873216262663317521412565619945866219215231492333875939497977633050373226117619260140357156590870536332520361522836955784492120121995 43.

- Message M is signed and the signature is (r, s).

Finally, let's implement the signature verification algorithm.

- $u_1 = s$ (mod n) and $u_2 = h$ (mod n).

$\boldsymbol{u_1}$=26075593207638998033271362760615191934391781587755738878810533023313352679412653560498372269275807196507024803695897036778704994547146254422036295390994498 73.

$\boldsymbol{u_2}$=13391588585886738632600139556386100265971188810783796321769392220422220347727593927717784174305529098666089325052246885816774225364695776567825092379011242.

- Compute $(x_1, y_1) = u_1Q + u_2G$ (mod n).

($\boldsymbol{x_1}$)=28120164071144055615114873216262663317521412565619945866219215231492333875939497977633050373226117619260140357156590870536332520361522836955784492120121995 43.

- $r = x_1$ (mod n). The signature has been verified.

VI. CONCLUSION

In this study, a more efficient and secure version of ECDSA is designed and implemented. The inversion operation that is included in both signature and verification in ECDSA is an expensive and time consuming operation in modular arithmetic. Execution times (µs) of mathematical operations in different elliptic curves defined on prime fields recommended by NIST are given in TABLE III. Note that inversion operation requires much more execution time than other operations. Please note that the duration of the multiplication is much higher compared to addition and subtraction. Hardware-based applications require more time to perform inversion operation in modular arithmetic than software-based applications.

Based on this challenge, the proposed method only includes inversion operation in the key pair generation algorithm. Since this algorithm is in the set-up phase of the parameters, it will not affect the signature and signature verification time.

TABLE III. EXECUTION TIME (µs) OF MATHEMATICAL OPERATIONS

Operation\Fields	$F_{P_{192}}$	$F_{P_{224}}$	$F_{P_{256}}$	$F_{P_{384}}$	$F_{P_{521}}$
Addition in Modulation	0.221	0.264	0.241	0.244	0.237
Subtraction in Modulation	0.217	0.228	0.223	0.228	0.225
Multiplication in Modulation	0.338	0.408	0.320	0.823	1.02
Inversion in Modulation	3.79	4.4	4.02	5.62	7.34
Point Addition	50.4	52.9	51	54.4	57.9
Point Doubling	64.1	65.8	65.6	69	74.8

Compared to other variants in the literature and ECDSA, the proposed method gives better results due to the shorter times and less mathematical operations. The comparison of the proposed method, ECDSA and other variants by number of operations is shown in TABLE IV. (A.M. : Addition in Modulation, Su.M : Subtraction in Modulation, M.M. : Multiplication in Modulation, I.M. : Inversion in Modulation, S.M. : Scalar Multiplication, P.A. : Point Addition, S.G. : Signature Generation, S.V. : Signature Verification, ⊕ : XOR)

TABLE IV. COMPARISON OF ECDSA AND VARIANTS

Algorithm	Phases	A. M.	Su.M	M. M	I. M.	S. M.	P.A
ECDSA	S.G.	1	-	2	1	1	-
	S.V.	-	-	2	1	2	1
Ref. [5]	S.G.	-	-	4	2	1	-
	S.V.	-	-	1	1	1	-
Ref. [6]	S.G.	-	1	2	1	-	-
	S.V.	-	-	2	1	2	1
Ref.[7]	S.G.	2+1(⊕)	-	3	-	2	-
	S.V.	1(⊕)	-	-	1	2	1
Ref. [8]	S.G.	2	-	3	1	2	-
	S.V.	1	-	1	1	2	1
Ref. [9]	S.G.	2	-	3	1	2	-
	S.V.	-	-	4	1	2	1
The Proposed Method	S.G.	-	1	1	-	1	-
	S.V.	-	-	-	-	2	1

Execution times (milliseconds: ms) of the ECDSA and the proposed method during the signature generation and verification phases are shown TABLE V. The proposed method provides 25% and 18% time advantage in the signature generation and verification phases, respectively.

TABLE V. COMPARISON OF EXECUTION TIMES

Algorithm	Signature Generation	Verification
ECDSA	**48**	**83**
The Proposed Method	**36**	**68**

The proposed method can be used in many software and hardware applications that require speed and security thanks to the features it provides. It is particularly suitable for constraints such as low power, limited memory and computing capacities that exist in areas such as the Internet of Things, which are increasingly becoming important today.

ACKNOWLEDGMENT

This work was supported by the TUBITAK 2211-E domestic doctoral scholarship program.

REFERENCES

[1] M. Al-Zubaidie, Z. Zhongwei, and J. Zhang, “Efficient and secure ECDSA algorithm and its applications: a survey,” arXiv preprint arXiv:1902.10313, 2019.

[2] S.A. Vanstone, “Next generation security for wireless: elliptic curve cryptography,” Computers & Security, pp.412-415, 2003.

[3] D. Johnson, A. Menezes, and S. Vanstone, “The elliptic curve digital signature algorithm (ECDSA),” International journal of information security, pp.36-63,2001.

[4] A.A. Imem, “Comparison and evaluation of digital signature schemes employed in NDN network,” arXiv preprint arXiv:1508.00184 , 2015.

[5] S. Lamba, M. Sharma, “An efficient elliptic curve digital signature algorithm (ecdsa),” International Conference on Machine Intelligence and Research Advancement. IEEE,. p. 179-183, 2013.

[6] H. Junru, “The improved elliptic curve digital signature algorithm,” Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, IEEE, Vol. 1, pp.257-259, 2011.

[7] Q.Zhang, Z. Li, & C. Song, “The Improvement of digital signature algorithm based on elliptic curve cryptography,” 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC) , IEEE, pp. 1689-1691, August, 2011.

[8] N. Mehibel, and M. Hamadouche, “A new enhancement of elliptic curve digital signature algorithm,” Journal of Discrete Mathematical Sciences and Cryptography, pp.743-757, 2020.

[9] M. Prabu, and R. Shanmugalakshmi, “A comparative analysis of signature schemes in a new approach of variant on ECDSA,” 2009 International Conference on Information and Multimedia Technology. IEEE, pp. 491-494, 2009.

[10] G. Sarath, D. C. Jinwala, and S. Patel, “A survey on elliptic curve digital signature algorithm and its variants,” *Computer Science & Information Technology (CS & IT)–CSCP* , pp. 121-136, 2014.

[11] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, “ Handbook of applied cryptography,” CRC press, 2018.

[12] A. Abidi, B. Bouallegue, and F. Kahri, “Implementation of elliptic curve digital signature algorithm (ECDSA),” *2014 Global Summit on Computer & Information Technology (GSCIT)*, Sousse, Tunisia, pp. 1-6, 2014.

[13] N. Koblitz, “Elliptic curve cryptosystem,” Mathematics of Computation, Vol. 48, No.177, pp.203-209, Jan 1987.

[14] V. S. Miller, “Uses of elliptic curves in cryptography,” Advances in Cryptography, Crypto85, ser. Lecture Notes in Computer Science, vol. 218, Springer-Verlag, pp.417-426, 1986.

[15] D. Hankerson, A. Menezes, and S. Vanstone, “Guide to Elliptic Curve Cryptography,” Springer-Verlag, 2004.

[16] L. Chen, D. Moody, A. Regenscheid, and K. Randall, “Recommendations for discrete logarithm-based cryptography: Elliptic curve domain parameters. No. NIST Special Publication (SP) 800-186 (Draft). National Institute of Standards and Technology,” 2019.

[17] J. Nakajima, and M. Matsui, “Performance analysis and parallel implementation of dedicated hash functions,” Advances in Cryptology-EUROCRYPT,vol.2332,Springer, pp.165-180, Berlin, Heidelberg, 2002.

Design of a Half-Wave Dipole Antenna for Wi-Fi & WLAN System using ISM Band

Rashedul Islam
Department of Electrical and Electronics Engineering (EEE)
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
saydulbabar147570@gmail.com

Fardeen Mahbub
Department of Electrical and Electronics Engineering (EEE)
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
mahbubfardeen1998@gmail.com

Shouherdho Banerjee Akash
Department of Electrical and Electronics Engineering (EEE)
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
akashbanerjee906@gmail.com

Sayed Abdul Kadir Al-Nahiun
Department of Electrical and Electronics Engineering (EEE)
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
nahiunkf42@gmail.com

***Abstract*—Due to the low cost and high data transmission speed, the demand for wireless communication systems (WLAN, Wi-Fi) is increasing day by day. Nowadays, high-speed data transmission can be provided to a whole area or a building using WLAN and Wi-Fi systems. Since better performances and efficiency are expected for WLAN and Wi-Fi systems in the future, considering these issues, a Half-wave Dipole Antenna has been designed with an operating frequency of 2.45 GHz (ISM-Band). The antenna has been developed using the Copper (annealed) and the Perfect Electrical Conductor (PEC) material in the CST Studio Suite 2019 Software. After completing the simulation of the antenna, a Return Loss of -31.019 dB has been successfully achieved. A Gain of 2.026 dBi and a Directivity of 2.032 dBi was also achieved, respectively. The antenna's efficiency was determined as 99.70%, which was calculated by using the value of Gain and Directivity. The other obtained performance parameters of the antenna are Surface Current, Bandwidth, etc. Considering all of the above performance parameters, the simulated Half-wave Dipole Antenna will be an ideal option for the WLAN and Wi-Fi systems.**

Keywords*— *Dipole Antenna, Wi-Fi, WLAN, Gain, VSWR, Surface Current, Return Loss, CST.

I. Introduction

In recent times, the improvement of technology has rapidly increased, and the use of the internet has also increased to a significant level in both developed and underdeveloped countries. For the requirement of high-speed data transfer, a variety of wireless networks, such as WiMAX, WLAN, Wi-Fi, etc., plays an important role [1]. WLAN (Wireless Local Area Network) refers to a form of computer network service that uses wireless networking to link multiple devices in a small area to form a Local Area Network (LAN) [2]. Wi-Fi refers to a wireless networking system that connects users to the internet, including computers, handheld devices (smartphones and wearables), computers (laptops and desktops), and other appliances [3]. Wireless communication networks have grown and continue to expand based on many technologies, including 2G/3G. Different types of frequency bands can be used for wireless communication such as Several frequency bands can be used to design the antenna, such as ISM-band (433.05-434.79 MHz, 866-868 MHz, 902-928 MHz, 2.4-2.4835 GHz, and 5.725-5.875 GHz), L-Band (1-2GHz), S-Band (2-4GHz), C-Band (4-8GHz) [4]. The most common frequency band is ISM-Band (2.4-2.5 GHz). Among them, ISM-Band is most commonly used. To have sufficient coverage for potential operating frequencies, modern antennas must comply with the ISM band or wideband requirements [5].

There are different types of antennas such as Monopole, Dipole, Microstrip Patch Antennas, etc., for wireless communication. Performances of these antennas are various according to the dimensions of the antenna and their substrate material. The antenna's size must be small enough to fit inside a wireless networking system with limited space. When it comes to commercial antennas for wireless networking, the most popular antenna is the dipole antenna [6]. One of the benefits of dipole antennas is that they can receive balanced signals essential for wireless communication. The antenna's two-pole nature allows a device to transmit signals from a wide range of frequencies. It also enables the system to resolve signal conflicts without sacrificing transmission efficiency [7]. Low-frequency bands are more preferable because achieving better performances with high frequencies is a big challenge. The drawback of using a high-frequency band is that the path distortion or free space loss is high in this type of band, causing the signal's degradation to interference plus noise ratio (SINR) [8]. This paper proposes to design a half-wave dipole antenna for Wi-Fi and WLAN using the ISM band to achieve a higher Bandwidth and a Return Loss ($S_{1,1}$) to communicate faster.

II. Design and Methodology

A half-wave dipole antenna was designed to operate at 2.45 GHz for improved performance in the Wi-Fi and WLAN Communications Systems genre. The proposed model's geometric orientation in free space was depicted in Figure-1. From the Geometry, it can be seen that the intent of feeding is a distance between the two arms of the antenna, known as the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

feeding gap denoted by g. L indicates the antenna's total length; antenna arm thickness is designated by D. The dipole's radiation resistance is 73 Ohm, which corresponds to the line's impedance [9].

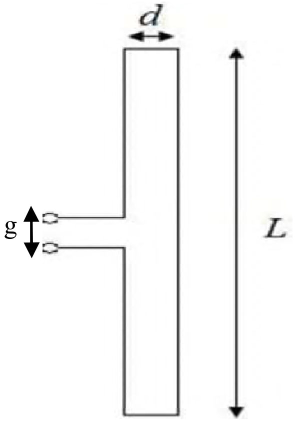


Fig.1. Geometry of the proposed Half-Wave Dipole Antenna

The frequency spectrum is first defined by the maximum and minimum magnitudes of the frequency, which are respectively 1GHz and 10GHz. Then the resonant frequency, *f* = 2.45GHz, was selected. Then the wavelength (λ) was calculated by using the given equation (1) [9].

$$\boldsymbol{\lambda = \frac{c}{f}} \qquad \textbf{(1)}$$

The total length of the antenna was calculated by using the resonant frequency, *f* using the following equation (2) [9].

$$\boldsymbol{L = \frac{143}{f}} \qquad \textbf{(2)}$$

By using the value of the total length, L, the antenna's feeding gap, g was calculated by using the following equation (3) [9].

$$\boldsymbol{g = \frac{L}{200}} \qquad \textbf{(3)}$$

The dipole radius was calculated by using the following equation (4) [9].

$$\boldsymbol{R = \frac{\lambda}{1000}} \qquad \textbf{(4)}$$

Table-I represents the input parameter values of the designed Half-Wave Dipole Antenna.

TABLE I. INPUT PARAMETERS OF THE DESIGNED HALF-WAVE DIPOLE ANTENNA

Input Parameters with their symbols	**Dimensions**	**Unit**
Radius of Dipole (R)	53.74	mm
Length of Feed (F)	10	mm
Length of Dipole (L)	54.64	mm
Input Impedance (Z)	73	Ω
Wavelength (λ)	109.28	mm

Figure-2 and Figure-3 depict the side view and the perspective view of the half-wave dipole antenna.

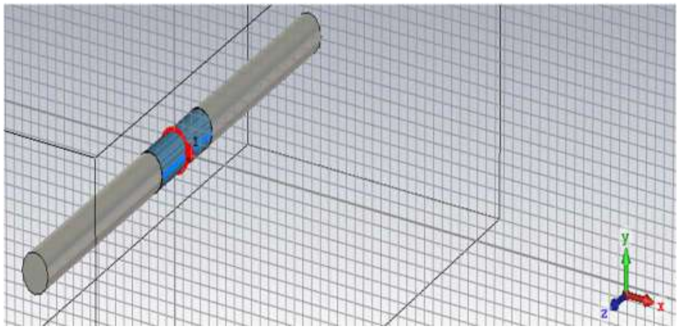


Fig.2. Side view of the Half-Wave Dipole Antenna

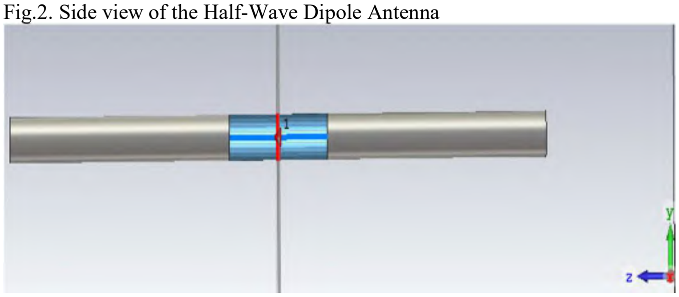


Fig.3. Perspective view of the Half-Wave Dipole Antenna

III. RESULT ANALYSIS

A. Return Loss

Return Loss($S_{1,1}$) in telecommunication networks refers to the magnitude of the signal that returns or bounces from the antenna due to a discontinuity of the optical fiber or transmission line grid. It is expressed in decibels (dB) [10]. It is better to have a larger Return Loss ($S_{1,1}$) because the antenna absorbs less energy while the Return Loss($S_{1,1}$) is minimal. Figure-4 depicts the proposed antenna's Return Loss ($S_{1,1}$) where the obtained Return Loss($S_{1,1}$) has a value of -31.019dB and Bandwidth has a value of 0.5384GHz at the operating frequency of 2.4345GHz. At the operating frequency of 7.5176 GHz Return Loss ($S_{1,1}$) of -12.506dB and a Bandwidth of 0.1171GHz were also achieved.

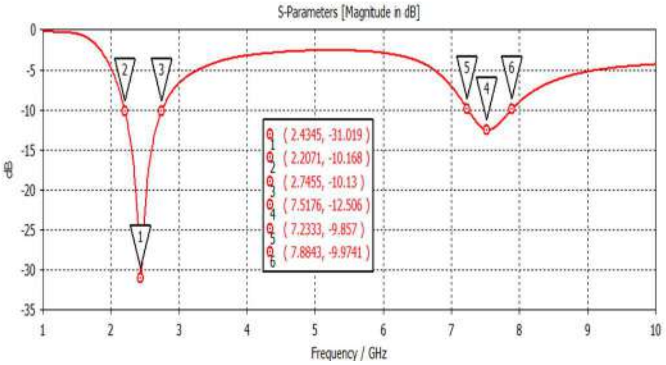


Fig.4. Return Loss ($S_{1,1}$) in Free Space

B. VSWR

The transmission line's impedance and the antenna's impedance must align to a considerable degree to transmit the necessary power to the desired antenna [11]. VSWR is crucial in this regard. The Voltage Standing Wave Ratio (VSWR), also known as SWR, is a measurement of how quickly radiofrequency power is transferred from a given power source to a load via a transmission cable. Figure 5 depicts the proposed antenna model's obtained VSWR, which is 1.0585 at the operating frequency of 2.4495GHz.

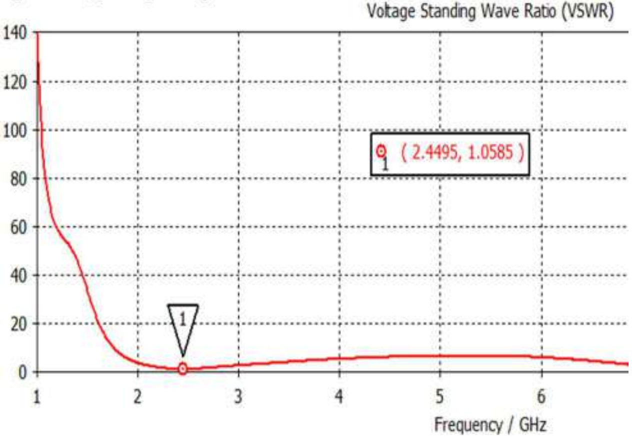


Fig. 5. VSWR in Free Space

C. Gain & Directivity

For evaluating the efficiency of an antenna, Gain and Directivity are two essential parameters. The Gain here denotes the amount of power transferred to the main beam, while the Directivity denotes the calculation of concentration radiation in a specific direction. The Gain denotes the power transferred to the main beam, and the Directivity is the calculation of the concentration radiation in a given direction [12]. Figure 6 and Figure 7 depict the 3D Gain and Directivity, with an obtained Gain of 2.026dBi and a Directivity of 2.031dBi at the operating frequency of 2.45GHz. The antenna's efficiency was measured to be 99.70% by using the formula

Efficiency = [Gain/Directivity] * 100%

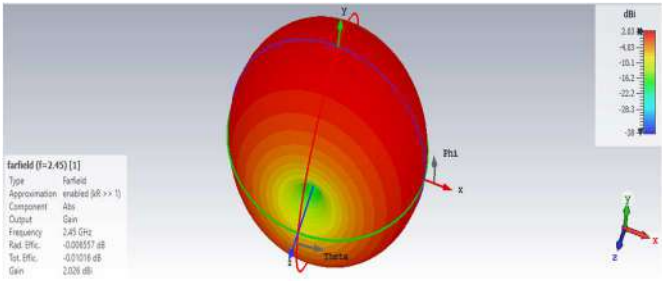


Fig.6. Gain of the Antenna in Free Space

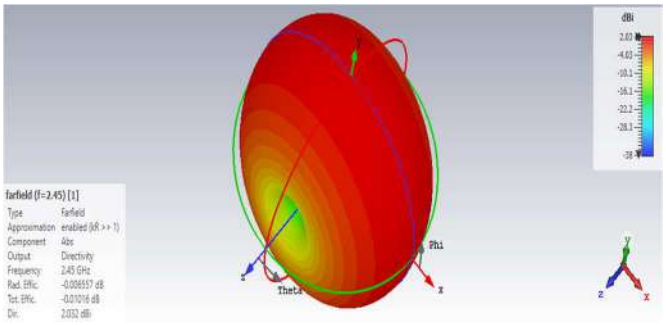


Fig.7. 3D Directivity in Free Space

D. Surface Current

Figure 8 depicts the proposed antenna model's surface current with an obtained value of 12.7A/m at the operating frequency of 2.45 GHz.

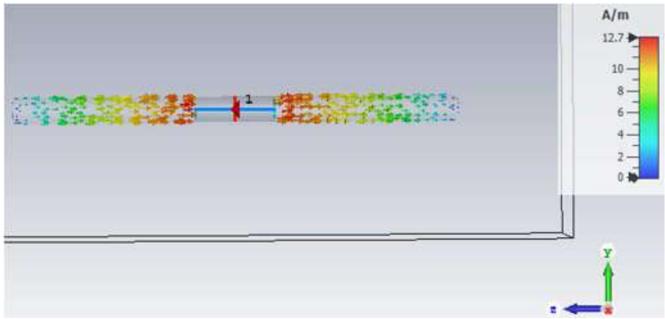


Fig.8. Surface Current in Free Space

E. Farfield Radiation (Polar Form)

Figure-9, Figure-10, and Figure-11 depict the Farfield Radiation (Polar Form for Phi 0^0, 90^0, 180^0) at the operating frequency of 2.45 GHz. In both of the cases, the Main lobe magnitude is 2.03 dBi.

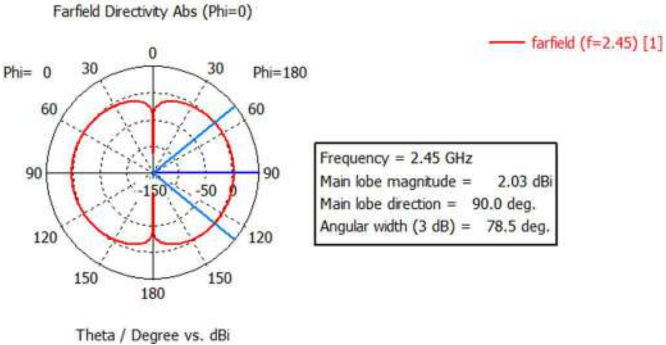


Fig.9. Farfield Radiation (Polar Form, at Phi=0^0)

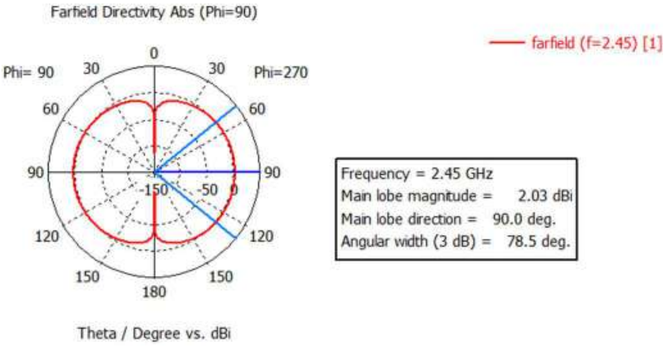


Fig.10. Farfield Radiation (Polar Form, at Phi=90^0)

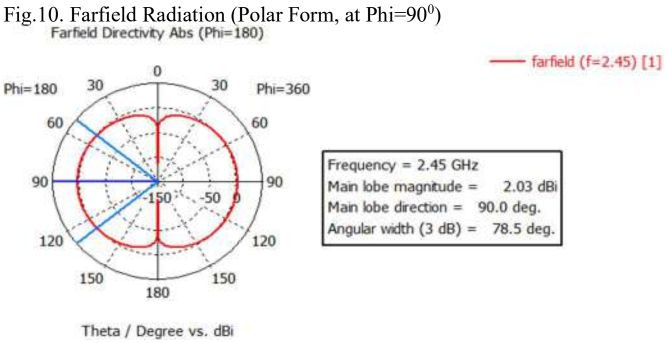


Fig.11. Farfield Radiation (Polar Form, at Phi=180^0)

Table II represents the summary of the obtained output parameters.

TABLE II. SUMMARY OF THE OBTAINED PARAMETERS

Antenna Output Parameter	*Obtained Values*
$S_{1,1}$	-31.019dB
Bandwidth	0.5384GHz
Gain	2.026
VSWR	1.0585
Efficiency	99.70%
Directivity	2.032 dBi
Surface Current	12.7 A/m

IV. COMPARATIVE STUDY

Various antenna parameters like Gain, Bandwidth, Return Loss [$S_{1,1}$], from many previous research works have been analyzed. Compared to the other reference paper mentioned in Table III, a better ($S_{1,1}$) parameter after conducting the simulation on CST at 2.45 GHz has been obtained. Since the value of the Return Loss is higher, the proposed model will accept more RF (Radiofrequency) energy compared to the other model.

TABLE III. COMPARISON OF THE ANTENNA MODEL WITH PREVIOUS RESEARCH PAPER

Return Loss ($S_{1,1}$) (dB)	*Bandwidth (Max) (GHz)*	*Gain (dBi)*	*Comment*	*Ref.*
-23	1.79	3.1	Low Gain, High Bandwidth	[13]
-19.89	0.2	2.07	Low Gain, Low Bandwidth	[14]
-	2.5	3.7	Low Gain, Low Bandwidth	[15]
-18.86	0.23	4.01	Low Gain, Low Bandwidth	[16]
-27	0.4	4.59	Low Bandwidth, Low Gain	[17]
-31.02	0.54	2.03	High Efficiency, Low Gain, Low Bandwidth	Proposed Antenna

V. CONCLUSION

For the sake of high-speed data transmission and for using the internet more conveniently, a Dipole Antenna has been designed for WLAN and Wi-Fi systems with an operating frequency of 2.45 GHz (ISM-band) ranging from 1 to 10 GHz in CST Studio Suite Software. Furthermore, the proposed model's different parameters are depicted in this article. Besides, the designed antenna model has an ($S_{1,1}$) value of -31.019 dB with a bandwidth of 0.5384 GHz. The antenna model's size has been kept compact to conserve space. The efficiency of the antenna is 99.70% which shows its effectiveness. To overcome Wi-Fi and WLAN systems' future challenges, The Return Loss ($S_{1,1}$), Gain, Bandwidth, and other obtained performance parameters of the antenna can be easily upgraded. A summary of all the performance Parameters was given in Table-ii, and a comparison of these parameters with some previous researches was mentioned in Table-iii. Taking both of these factors into consideration, it can be said that the proposed model may be an outstanding solution for Wi-Fi and WLAN communication systems, with its promising future.

REFERENCES

[1] Feng, B., Lai, J., Zeng, Q. and Chung, K.L., 2018. A dual-wideband and high gain magneto-electric dipole antenna and its 3D MIMO system with metasurface for 5G/WiMAX/WLAN/X-band applications. *IEEE Access*, *6*, pp.33387-33398.

[2] Mahbub, F., Akash, S.B., Al-Nahiun, S.A.K., Islam, R., Hasan, R.R. and Rahman, M.A., 2021, January. Microstrip Patch Antenna for the Applications of WLAN Systems using S-Band. In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1185-1189). IEEE.

[3] Nie, N.S., Yang, X.S., Chen, Z.N. and Wang, B.Z., 2019. A low-profile wideband hybrid metasurface antenna array for 5G and Wi-Fi systems. *IEEE Transactions on Antennas and Propagation*, *68*(2), pp.665-671.

[4] Karthick, M., 2015, February. Design of 2.4 GHz patch antennae for WLAN applications. In 2015 IEEE Seventh National Conference on Computing, Communication and Information Systems (NCCCIS) (pp. 1-4). IEEE.

[5] Aravindraj, E. and Ayyappan, K., 2017, January. Design of slotted H-shaped patch antenna for 2.4 GHz WLAN applications. In 2017 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.

[6] S. Genovesi, A. Monorchio and S. Saponara, "Compact triplefrequency antenna for sub-GHz wireless communications," IEEE Antennas and Wireless Propagation Letters, Vol. 11, pp. 14-17, 2012.

[7] Ding, H., Li, G. and Xu, M., 2019, October. Overview of Research on Broadband of Dipole Antenna. In 2019 IEEE 19th International Conference on Communication Technology (ICCT) (pp. 835-838). IEEE.

[8] Petrov, V., Komarov, M., Moltchanov, D., Jornet, J.M. and Koucheryavy, Y., 2017. Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas. IEEE Transactions on Wireless Communications, 16(3), pp.1791-1808.

[9] Balanis, C.A., 2016. Antenna theory: analysis and design. John wiley & sons.

[10] Kang, J.S., Kim, J.H., Kang, N.W. and Kim, D.C., 2012, July. Antenna measurement using S-parameters. In 2012 Conference on Precision electromagnetic Measurements (pp. 658-659). IEEE.

[11] Alzein, H., Milbrandt, J., Kaddour, A.S., Menudier, C., Thevenot, M. and Monediere, T., 2019, July. Study of Active VSWR in a Reduced BFN Antenna Array. In 2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (pp. 1145-1146). IEEE.

[12] Mahbub, F., Islam, R., Al-Nahiun, S.A.K., Akash, S.B., Hasan, R.R. and Rahman, M.A., 2021, January. A Single-Band 28.5 GHz Rectangular Microstrip Patch Antenna for 5G Communications Technology. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 1151-1156). IEEE.

[13] Remsha, M., Manoj, M., Shameena, V.A. and Mohanan, P., 2019, March. SIR based Broadband Dipole Antenna for LTE/WiMAX/WLAN Applications. In 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC) (pp. 1-3). IEEE.

[14] Malek, N.A., Sabri, N.A.C., Islam, M.R., Mohamad, S.Y. and Isa, F.N.M., 2019, November. Design of Hybrid Koch-Minkowski Fractal Dipole Antenna for Dual Band Wireless Applications. In 2019 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE) (pp. 1-5). IEEE.

[15] B. Feng, Q. Zeng, and M. Wang, "Substrate integrated magneto-electric dipole antenna for wlan and wimax applications," in 2016 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), July 2016, pp. 1–2.

[16] Patel, H. and Upadhyaya, T., 2021. Surface Mountable Compact Printed Dipole Antenna for GPS/WiMAX Applications. Progress In Electromagnetics *Research Letters*, *96*, pp.7-15.

[17] Dakhli, S., Smari, M., Floc'h, J.M. and Choubani, F., 2020, June. Design of Frequency-Reconfigurable Triband Dipole Antenna Using Capacitive Loading. In 2020 International Wireless Communications and Mobile Computing (IWCMC) (pp. 1342-1346). IEEE.

Magnet Integrated Shirt for Upper Body Posture Detection Using Wearable Magnetic Sensors

Mary Farnan
Department of Electrical
and Computer Engineering
University of St. Thomas
St. Paul, MN, USA
mary.farnan@stthomas.edu

Emily Dolezalek
Department of Electrical
and Computer Engineering
University of St. Thomas
St. Paul, MN, USA
emily.dolezalek@stthomas.edu

Cheol-Hong Min
Department of Electrical
and Computer Engineering
University of St. Thomas
St. Paul, MN, USA
cmin@stthomas.edu

***Abstract*— COVID-19 is a global pandemic that has caused an increase in remote work. Sitting in various positions at home without the proper back support is undesirable and can cause chronic back pain and other undesired side effects. Therefore, a new non-intrusive method to continuously monitor back postures in homes is proposed. A shirt is designed with integrated magnets. A magnetic sensor would be placed above the body's sternum, and magnets will be implemented on a shirt. The sensor will help ascertain the back posture position (straight or curved) and help provide feedback to mend the posture if deformed. In this paper, the initial results using the proposed system are presented using a wearable sensor system with a magnet integrated garment that can continuously monitor the varying sitting postures throughout daily lives. In addition, we discuss how the lower body posture affects the magnetic recording otherwise not detectable using the accelerometer-based systems currently available on the market.**

Keywords—activity detection, posture detection, magnetic sensor

I. Introduction

Wearable sensing through clothing minimizes the need for bulky devices, enabling a pragmatic approach to long-term monitoring of the body. Minimizing the wearable system allows for body movements to be captured easier without affecting everyday activities. In order to maximize the effectiveness, a loose shirt is implemented as the proposed platform in the wearable solution.

With the rise of remote work due to the COVID-19 crisis, the proposed system aims to detect the spinal postures of people working at home. Make-shift home offices often lack ergonomic furniture that impedes healthy posture [1]. Without proper desks or chairs, many remote workers find themselves on couches or kitchen tables hunched over their laptops, causing an increase in neck and lower back pain [2]. Additionally, the use of non-ergonomic equipment may increase Musculoskeletal (MSK) disorders [2]. MSK disorders are defined by pain and limitations in mobility, dexterity, and overall body functionality [3]. The number of people with MSK disorders and lower back pain will increase in the near future due to the COVID-19 crisis.

Many studies suggest an ideal home workstation setup where ergonomic principles are unattainable, creating a footrest when feet do not touch the floor when sitting [4]. Another study investigated adjustable desks and their relationship to healthy postures [4]. Other studies have investigated different sensing techniques to monitor back posture [5, 6, 7]. A research group investigated the use of conductive silicon to movement [8, 9]. Current advertised products to detect slouching only focus on forwarding or backward spinal postures. The proposed wearable system can detect the left and right and the forward and backward spinal postures.

Presented are the results of using a magnetic sensor to detect upper body posture while remotely working from home. First, the design of the garment-integrated system is introduced. Then, results from the posture monitoring sensor system are analyzed and presented. Finally, current progress and the next steps of the research are declared.

II. Sensors and System Overview

The objective is to detect the relative position of small magnets on a shirt using a magnetic sensor. The position will be used to determine upper body postures. Therefore, the proposed system will have a grid of magnets integrated onto a shirt and record the sensor's output [10, 11].

The magnetic sensor chosen for the proposed wearable system is Mbientlab's MetaMotionR (MMR). The wearable sensor provides real-time and constant monitoring of environmental data. The spinal position is extracted using x, y, and z-axis data collected. The x-axis data corresponds to left and right movement, and z-axis data corresponds to the front and backward movement, while the y-axis data corresponds to up and down movement of the upper body.

A grid of nickel-plated neodymium magnets (D42-N42) is placed in a 3x3 grid on a T-shirt. The magnets are placed on the shirt's front for the proposed system, as shown in Figure 1. Nine pockets placed 2 inches apart horizontally and 4 inches apart vertically are sewed onto the shirt for the magnets to be placed. Consistent magnetic polarity is ensured by fixing the magnets to cuttings of cardstock before placed in the pockets. The MMR sensor is placed on the body beneath the shirt and at 2 inches below the top row's middle magnet.

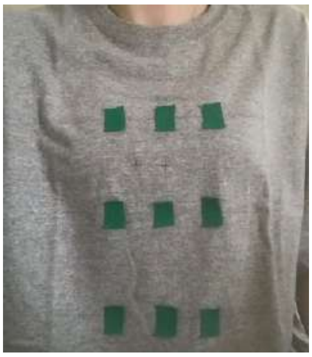


Figure 1. Subject wearing the magnet integrated shirt.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

III. Data Collection and System Testing

The MMR sensor can configure sampling frequencies for all data types using the mobile phone application. For all experiments, the sensor recorded the magnetometer data at a frequency of 25 Hz. The data recorded was placed through a lowpass Butterworth filter to eliminate unwanted noise. All data captured from the MMR was processed in MATLAB.

A. Initial Test Procedure and Results

Two protocols were designed and conducted to monitor upper body postures. For both protocols, the participants began and ended with a straight posture. Furthermore, the participants held each position for 15 seconds intervals. A straight posture is established with the head and upper body aligned with the spine. In the first protocol, the participants leaned left and right at 20 degrees. In all trials, leaning left increases the magnetic flux and leaning right decreases magnetic flux in the x-direction. It is determined that the x and y fields are beneficial when analyzing changes in side-to-side postures.

In the second protocol, participants pose their own version of a slouching posture by leaning forward of approximately 20 degrees or more. In all trials, the magnetic flux increases in the x-direction when the participant leaned forward. Changes in the y- and z-directions represent the shirt hanging away from the body, which creates a distance between the MMR sensor and the magnetic grid.

B. Further Testing Procedures and Results

In the initial testing, the magnetic flux does provide a consistent value to indicate straight posture. But, as it can be seen from figures 1 through 3, there exists a slight amount of noise due to the movement of the person. The Z-field of the sensor data, which detects the magnetic field variation with respect to the distance from the chest to the shirt, is bit noisier. This is due to the folding and creasing of the shirt that occurs after the movement. The magnetic flux shows the leaning direction of the participant. Therefore, the magnetic flux is beneficial when analyzing changes in side-to-side postural behavior. To determine when a participant is in a neutral posture, accelerometer and gyroscope measurements are added to the wearable system. The MMR sensor also provides accelerometer and gyroscope measurements. The DC offset with respect to the gravity present in the accelerometer measures the acceleration forces on the body to determine the position and monitor the movement of the body while the gyroscope measures the change in rotation angle of the body. In all trials, the accelerometer, magnetometer, and gyroscope are set to a sampling frequency of 25 Hz. The same two protocols as the initial testing were performed with the participant sitting and standing. Furthermore, each posture position is held for a 15 second interval. Although the accelerometer data is less noisy on the static posture data, its limitations are that it cannot detect the postural changes of the lower body while the magnetic sensor can, as it detects the changes due to a coupling of the lower body movement with the shirt movement, which in turn changes the magnetic fields detected by the magnetometer.

In both protocols, the participants began and ended with a straight posture. In the first protocol, the participant leaned left and right at 20 degrees. Figures 2, 3, and 4 show the magnetic flux, acceleration, and angular velocity strengths detected when the participant was sitting. The change in magnetic flux shows the side-to-side postural changes. Leaning left increases the magnetic flux and leaning right decreases the magnetic flux in the x-direction. The fluctuations in the y- and x-direction of the accelerometer are used to analyze when the participant has a straight, neutral posture. Leaning left decreases the change in acceleration and leaning right increases the change in acceleration in the x- and y-directions.

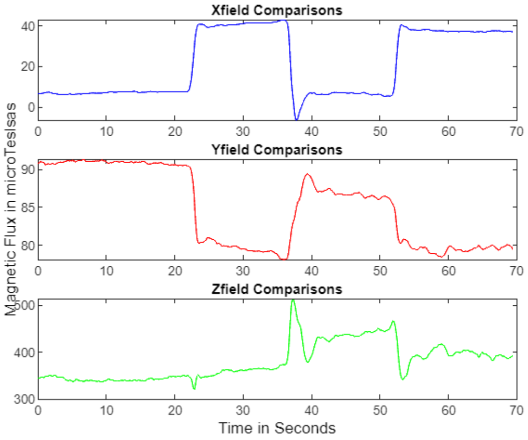


Figure 2. Magnetometer readings of the participant leaning straight, forward, and straight while sitting.

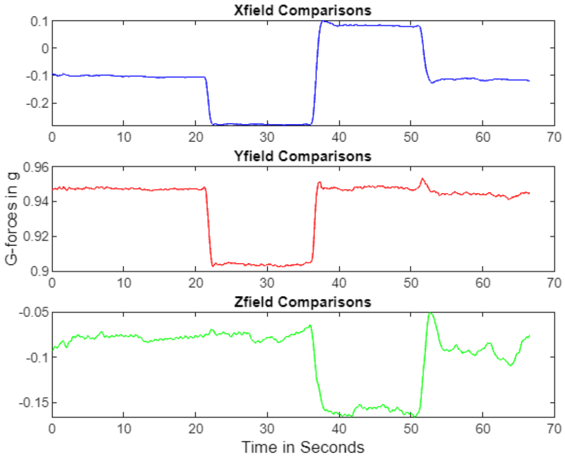


Figure 3. Accelerometer readings of the participant leaning straight, forward, and straight while sitting.

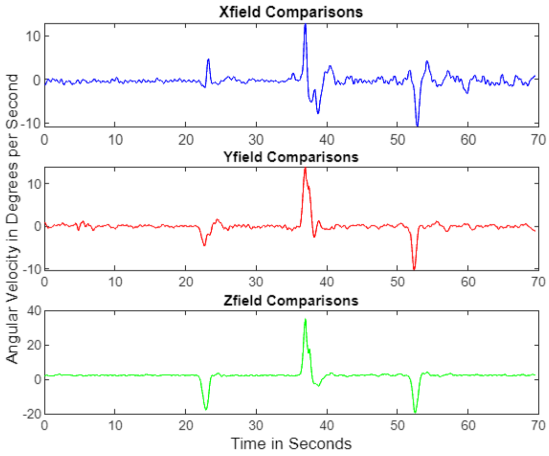


Figure 4. Gyroscope readings of the participant leaning straight, forward, and straight while sitting.

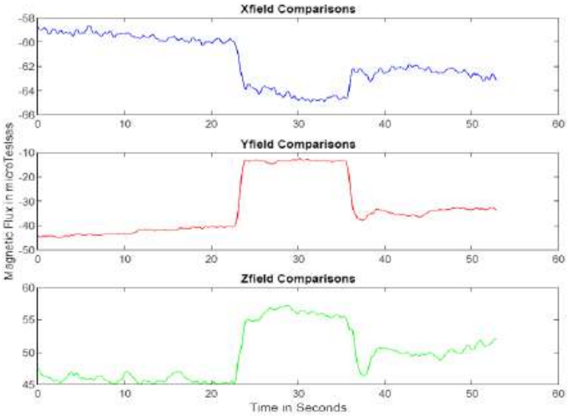


Figure 5. Magnetometer readings of the participant leaning straight, forward, and straight while standing.

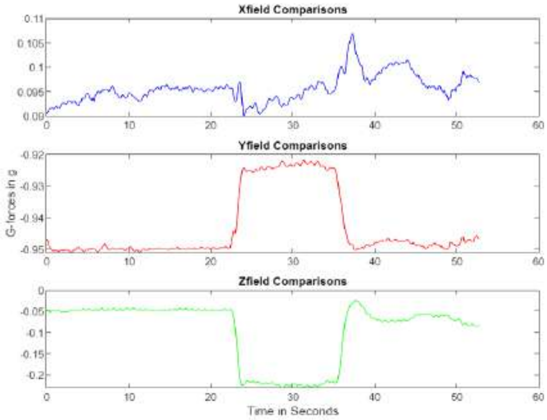


Figure 6. Accelerometer readings of the participant leaning straight, forward, and straight while standing.

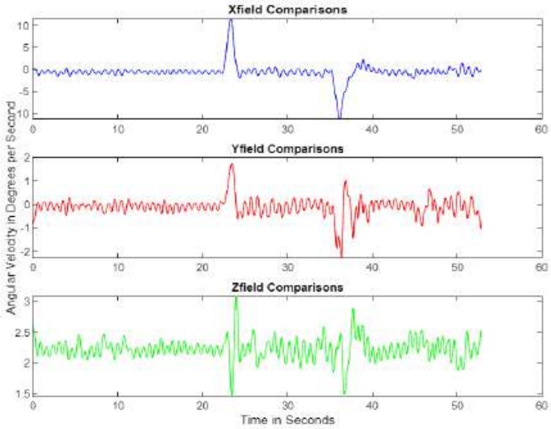


Figure 7. Gyroscope readings of the participant leaning straight, forward, and straight while standing.

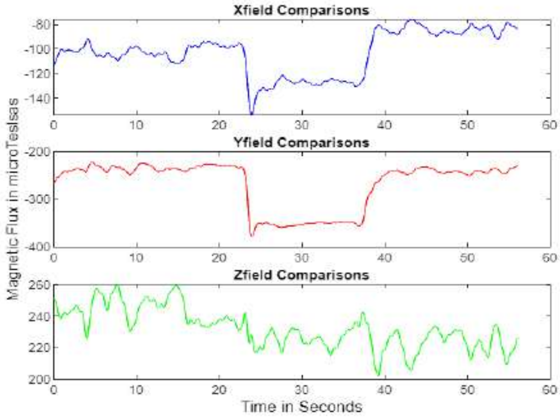


Figure 8. Magnetometer readings of the participant sitting straight, crossing right leg over left leg, and sitting straight.

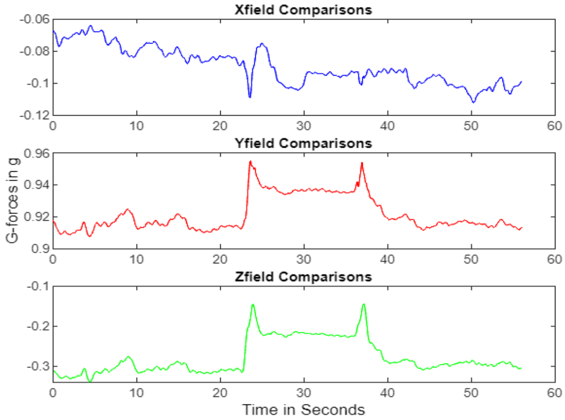


Figure 9. Accelerometer readings of the participant sitting straight, crossing right leg over left leg, and sitting straight.

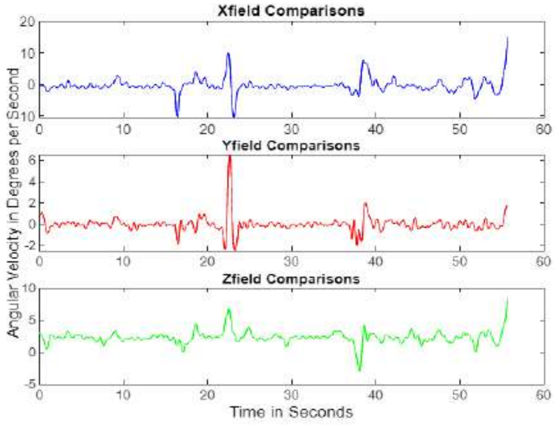


Figure 10. Gyroscope readings of the participant sitting straight, crossing right leg over left leg, and sitting straight.

In all trials, the accelerometer has an initial value approximately zero-G for x and z axis while the y-axis show near one-G for sitting straight up postures. When the participants return to the straight posture after altering postures, the values of the beginning and ending straight posture simulations align. Therefore, accelerometer shows when a participant has a straight, neutral posture. The exact value of the straight posture will change depending on the participant and where the shirt falls on the body. The fluctuations in the y- and z- directions of the gyroscope are useful when determining a change in posture position occurs. Leaning left produces a short decrease in the angular velocity and leaning right produces a sharp increase in the y- and z-directions. Changes in the angular velocity is indicative of the participants holding a constant posture.

The second protocol to simulate a slouching posture was then performed. In this protocol, the participants leaned forward at 20 degrees. Figures 5, 6 and 7 display the changes in magnetic flux, accelerometer/position, and angular velocity of when the participant was standing. When the participant is leaning forward, the magnetic flux is increased in the y- and z-directions. The accelerometer data shows an increase in the y-direction when the participants leaned forward. The x and z directions decrease in the accelerometer data when leaning forward. The accelerometer shows participants holding a straight posture due to its initial and ending value being the same. The initial value will slightly change depending on the participants; however, it is observed that the value is approximately zero. The gyroscope data has a sharp increase in the angular velocity in x- and y-directions when the participant leans forward. This is indicative of the sensor becoming closer to the ground as the participants lean forward.

Throughout all trials, changes in magnetic flux corresponds to the direction the participant has leaned. The addition of the accelerometer indicates when a participant has a straight posture. The gyroscope indicates when and how quickly a change in posture occurs. All measurements indicate how long the participant is holding the posture; however, the gyroscope provides the most promising results in identifying the instance in which a posture is held, although this information can also be extracted from the magnetometer.

C. Common Posture Testing and Results

The initial two protocols gave baseline results of straight posture compared to a slouching left, right, or forward posture. The results were then used to observe the most common types of posture while sitting at a desk. In all trails, the participants held each position for a 15 second interval. The sampling frequencies of the magnetometer, gyroscope, and accelerometer was set to 25 Hz. For all trials, the participant began and ended with a straight posture.

For the first simulation, the participants crossed the right leg over left leg. Figures 8, 9, and 10 show the change in magnetic flux, acceleration, and angular velocity when the participant crosses the right leg. It is observed that the magnetic flux decreased in the x- and y-directions. The z-direction of the magnetometer is noisy due to the movement of the shirt when adjusting position. The accelerometer data shows an increase in the y- and z-directions when crossing the right leg. The accelerometer shows the participant holding a non-straight posture due to the constant increase in the y- and z-directions. The gyroscope data has sharp increases in the angular velocity in the x-, y- and z-directions. This is indicative of the quick change in posture position, which can be tracked via sensor data analysis.

For the second simulation, the participants crossed their left leg over the right leg. It was observed that behaviors of the accelerometer, and gyroscope data values were the same as the observed data patterns of crossing the right leg over the left leg. For all trails regarding the participants crossing the right or left leg over the opposite leg, it was observed that the magnetometer can detect the change in the crossed legs. Depending on where the participants cross their leg, near the knee or further up the opposite leg, the values of the magnetometer data will change; however, the data patterns of the values remain the same. The folding or creasing of the shirt will also change depending on the position of the crossed leg. This is indicative of how noisy the data becomes in the z-direction. This noise due to the creasing and folding of the shirt allows the system to distinguish the lower body postures. For accelerometers, when the lower body position change occurs, there are instantaneous changes but the change in value is not maintained throughout the duration of the posture because the accelerometer responds to the acceleration and when the sensor does not move, it responds to gravity only. The gyroscope data is again an indicative of the instance when the adjustment in position occurs.

Other experiments to simulate common postures at a desk involve various ways of crossing one or both legs. Trials simulating the tucking of a leg under the bottom of the participants as well as placing the bottom of one foot on the chair and leaning on the same leg were conducted. Furthermore, methods involving the movement of an arm in relation to the upper torso, such as the participants leaning the head on the hand with the elbow on the desk, were examined. In all simulations, the magnetometer, accelerometer, and gyroscope data values were observed and analyzed. Again, the changes in the magnetic flux observed are indicative of the position the body. In summary, the magnetometer data corresponds to how far the participants have leaned as well as the sensor placement in relation to the grid of magnets on the shirt. Furthermore, the magnetometer is affected by the folding of the shirt. It is observed that the folding or creasing on the shirt will provide some noisy data when altering positions. The accelerometer helps determines when the participants are holding a neutral posture and when a change of posture occurs. The gyroscope is indicative of the time it takes to change posture position and when the participant holds the posture steady.

D. Current Research Progress

Currently experiments are being completed to allow the system to automatically detect the posture position and provide feedback to the participant. The experiments previously discussed were manually identified and labeled by the research team. An algorithm to determine how far the participants are leaning is being created: it will correlate the angles of leaning to specific postures. This will allow the algorithm to self-identify posture positions and provide feedback to the participants. The feedback will suggest ways in which the participants can change their posture positions. If the participants are leaning in a specific direction for a given amount of time, the system will identify the poor posture behavior, alter the participants of the behavior, and give directions to the participants to correct their posture. The system will recommend the participants to be in specific positions for a given amount of time to elevate the stress on their backs from the poor posture behaviors. The algorithm will provide the feedback recommendations from the data collected in the previous experiments. It will learn and adjust to the different postural positions and provide the most optimal feedback to the participants.

Allowing the system to have a feedback recommendation allows for participants to monitor their own postural behaviors when working from home. This will help elevate the stress on backs when sitting for long periods of time without proper posture due to lack of ergonomical furniture. For the purpose of this document, however, only the data collected and analyzed which proved constant throughout all trials from the magnetometer, accelerometer, and gyroscope changes were discussed.

IV. Conclusion

Through the initial experiments, an unique inexpensive wearable system using magnets, magnetometers, accelerometers, and gyroscope measures to monitor poor postural behaviors has been developed. The system provides a continuous monitoring of the spine in side-to-side and forward leaning positions. The wearable system is washable and can be used multiple times, allowing for an easy access to work at home. With the increase in remote work due to the COVID-19 crisis, the wearable system is an in-home health monitoring tool that can be used on participants in all ages. Furthermore, with more research and development, the wearable system may be used for long-term clinical monitoring on Spine Curvature Disorder (SCD), a medical condition that affects the shape of the spine.

REFERENCES

[1] A. Moretti, F. Menna, M. Aulicino, M. Paoletta, S. Liguori, and G. Iolascon, "Characterization of home working population during covid-19 emergency: A cross-sectional analysis," August 2020.

[2] K. Katella, "How to Set Up Your 'Pandemic' Home Office the Right Way," *Yale Medicine*, July 2020.

[3] "Musculoskeletal conditions," *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions.

[4] "Ergonomics 101: Working from Home During Coronavirus," *University of Nevada, Las Vegas*.

[5] S. Saito, M. Miyao, T. Kondo, H. Sakakibara, and H. Toyoshima, "Ergonomic Evaluation of Working Postures of VDT Operation Using Personal Computer with Flat Panel Display," in *Industrial Health, 1997. Apr. 35(2),* pp. 264-270.

[6] E. W. Bakker, A. P. Verhagen, C. Lucas, H. J. Koning, B. W. Koes, "Spinal Mechanical Load: A Predictor of Persistent Low Back Pain? A Prospective Cohort Study," in *European Spine Journal, v. 16(7), July 2007*, pp. 933-941.

[7] A. El-Metwally, J.J. Salminen, A. Auvinen, G. Macfarlane, M. Mikklesson, "Risk Factors For Development of Non-Specific Musculoskeletal Pain In Preteens and Early Adolescents: A Prospective 1-Year Follow-Up Study," in *MBC Musculoskeletal Disorders*, 2007, 8:46, pp. 3990–3993.:46

[8] A. Tognetti, F. Lorussi, R. Bartalesi, S. Quaglini, M. Tesconi, G. Zupone, and D. De Rossi, "Wearable Kinesthetic System For Capturing and Classifying Upper Limb Gesture in Post-Stroke Rehabilitation", *Journal of Neuroengineering Rehabilitation*, vol. 2:8, 2005.

[9] D. De Rossi, F. Carpi, F. Lorussi, A. Mazzoldi, R. Paradiso, E. P. Scilingo, and A. Tognetti, "Electroactive fabrics and wearable biomonitoring devices," *AUTEXRes.J.*vol.3, no.4, pp.180–185, 2003.

[10] S. Wielgos, E. Dolezalek, and C. Min, "Garment Integrated Spinal Posture Detection Using Wearable Magnetic Sensors", *Proceedings of the IEEE 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4030-4033,* July 2020.

[11] E. Dolezalek, M. Farnan, and C. Min, "Magnet Integrated Shirt for Spinal Posture Detection," *Proc, of 2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS),* August 2020.

Error Performance Index Based PID Tuning Methods for Temperature Control of Heat Exchanger System

V. Bharath Kumar, Dasa Sampath, V. N. Siva Praneeth, Y. V. Pavan Kumar
School of Electronics Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA
bharath.18bec7093@vitap.ac.in, sampath.18bec7101@vitap.ac.in, praneeth.18bev7022@vitap.ac.in, pavankumar.yv@vitap.ac.in

***Abstract* - Heat exchangers are commonly used in industries for the transfer of heat from one fluid to another fluid or from one form to another form. In this process of heat transfer, there is a need for continuous monitoring and control of the system to achieve desired heat levels. Normally, a robust PID controller is used for this purpose. To tune these PID controllers, many conventional methods are available. Of these methods, the OLTR (Open Loop Transient Response) methods are frequently used to obtain the gain parameters of the controllers. However, these methods are limited with respect to adjustment of steady-state error. As the heat exchanger systems are so sensitive to transients, any minor error can deviate the system behavior. In such systems, EPI (Error performance Index) methods provide a better system response by effectively nullifying the steady-state errors. Hence, this paper presents the design of PID controller gain parameters using various EPI methods. MATLAB/Simulink software is used to model the entire heat exchanger system and various control methods. The results justify that a better response is achieved with the use of ITAE based EPI method.**

Keywords - Heat exchanger system, manual setting of gains, PID controller, EPI Tuning methods, Performance parameters, MATLAB/Simulink

I. Introduction

The heat exchanger is generally used in industries for the transfer of heat energy from one liquid to another liquid as shown in Fig. 1. It is equally used for the cooling process also as mentioned in [1]. Some of the applications of heat exchanger systems are extensively refrigerating systems, air conditioning systems (AC's), sewage management systems, petroleum extracting, internal combustion engines (ICE) and refineries. In heat sinks, the heat generated from the electronic components and mechanical components is absorbed and transferred to the coolant fluid with the help of a heat exchanger [2]. In all these applications, there is a need for continuous supervision and control over the system. The common way to achieve this is by using a PID controller [3]. Obtaining the gain parameters for the PID controller that suits the system is a difficult task and there are many tuning methods available to compute gain values. Of all these available methods, the OLTR method is the oldest method and most commonly used [4]. These methods use the open-loop response of the system to obtain gain parameters [5]. The gain values that are obtained using these methods may not be effective for the closed-loop system and many methods in this group give only proportional gain and integral gain values to control the system [6], [7]. The differential gain parameter which plays an important role in controlling the system is ignored by the methods of this group.

In EPI methods, there are procedures to compute all the three gain parameters [8], [9]. These methods use the error between the reference (set value) and the present value of the output to compute gain values. This method is well suited to the considered heat exchanger system as the difference in temperature is a crucial factor to generate the output. The transfer function of a heat transfer system is given by (1).

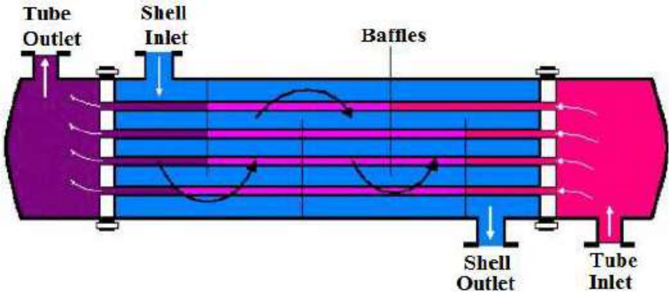


Fig. 1. Simplified representation of heat exchanger module.

$$G(s)=\frac{1}{38s+1}e^{-15s} \tag{1}$$

II. EPI Methods Based PID Tuning for the Control of Heat Exchanger System

A. Design of PID controller by manual setting up of gains:

The design of the PID controller includes proper tuning of gain parameters. These gain parameters are designed by changing the parameters manually according to system requirement or can be computed by some predefined tuning methods [10]- [12]. Many tuning methods are available to set PID gains and out of them, EPI methods are mainly chosen to control many system parameters. The PID controller block consists of proportionality gain, integral gain constant as well as derivative gain constant. All resultants of these three constants are summed up by using an adder block. The output of the PID controller will be a control signal which is a sum of all three gain constants as shown in Fig. 2. The mathematical notation of the controller is shown in (2). The values of K_P, K_I and K_D determine the dynamic performance of the system.

$$U(t)=K_p e_p(t)+K_I\int_0^t e_p(t)dt+K_D\frac{de(t)}{dt} \tag{2}$$

Where,

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

$U(t)$ = Output of PID controller (control signal)

K_p = Constant of Proportionality

$K_I = \frac{K_P}{T_I}$ = Constant of Integration

$K_D = K_P \times T_D$ = Constant of Differentiation

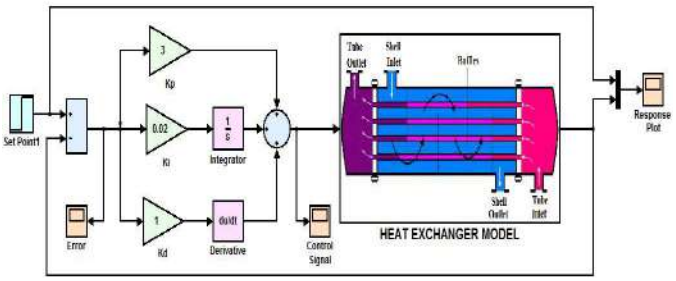


Fig. 2. MATLAB/Simulink model to represent the manual setting of PID gain.

The advantage of increasing the value of K_P is the decrease of error in steady-state and reduced rise time. It has a disadvantage of increased signal peak overshoot and change in settling time of the signal. Increasing the value of K_I creates an advantage of the decrease in rise time and error, but it affects the peak overshoot and settling time. Increasing the value of K_D reduces the peak overshoot of the signal and settling time, but results in a minor decrement in rise time and does not affect the error. Hence, in general for any systems which require a quick response, a PD controller is highly preferred, and for the systems which are more sensitive to steady-state error, a PI controller is preferred. At present, in many parts of industries, the PID controller is preferred because of having flexible control among all performance parameters. So, for a system to be stable and to acquire the desired response, an appropriate controller need to be designed [13].

B. Controller design with EPI tuning methods:

Error performance indices methods help to attain an improved performance and stability for the heat exchanger module. The steps which are to be followed to calculate the gain parameters are described as follows [14].

- **Step-1:** This step includes removal of negative feedback loop and controller of the system so that no control action is provided to the system. The feedback loop is connected between output of the system and the input of the controller.
- **Step-2:** Here, the step response of the system is probed and the response obtained is termed as process reaction curve.
- **Step-3:** In this step, a tangent is drawn to the open-loop reaction curve at the contact point of inflection. Inflection point is termed as the point at which the tangent and reaction curve have maximum contact area. At this point, the slope of the reaction curve starts decreasing. From the marked points, lag time (L_t), process reaction time (T_p) and stationary gain (K) are computed.
- **Step-4:** This step includes the calculation of PID gain parameters (K_P, K_I, K_D) by using (3)-(5). The constant values x_1, x_2, x_3, y_1, y_2, y_3 are given in Table. I for the calculation of K_p, T_I, T_D.

Where 'L_t' is the time interval between origin and the point at which the tangent line crosses the x-axis (time axis). Also, 'T_p' is defined as the time interval between point of intersection of tangent with time axis and the point on the time axis at which the tangent crosses the desired input as shown in Fig. 3. 'K' is stationary gain. Fig. 4 shows the simple flowchart representation of the process used to find gain parameters using EPI tuning.

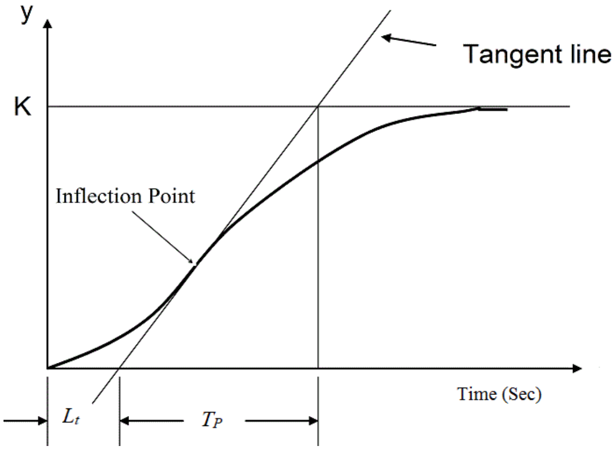
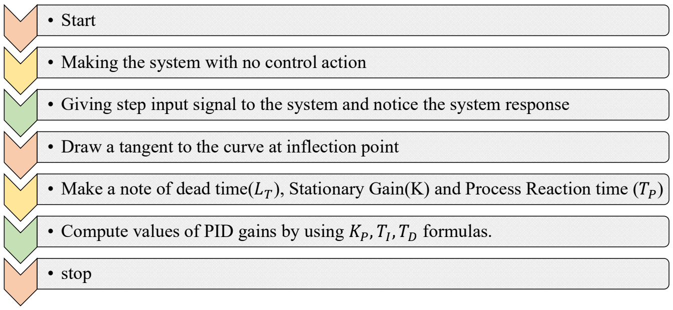


Fig. 3. Process reaction curve.

$$K_P = \frac{x_1}{K}\left(\frac{L_t}{T_p}\right)^{y1} \tag{3}$$

$$T_I = \frac{T}{x_2 + y_2\left(\frac{L_t}{T_p}\right)} \tag{4}$$

$$T_D = x_3 T\left(\frac{L_t}{T_p}\right)^{y3} \tag{5}$$



- Start
- Making the system with no control action
- Giving step input signal to the system and notice the system response
- Draw a tangent to the curve at inflection point
- Make a note of dead time(L_T), Stationary Gain(K) and Process Reaction time (T_P)
- Compute values of PID gains by using K_P, T_I, T_D formulas.
- stop

Fig. 4. Steps for computing tuning parameters using EPI methods.

TABLE I. PARAMETERS FOR CALCULATING PID GAINS USING EPI METHODS

Constants	ISTSE	ITAE	ISTE	ISE
x_1	1.042	0.965	0.968	1.048
y_1	-0.897	-0.85	-0.904	-0.897
x_2	0.987	0.796	0.977	1.195
y_2	-0.238	-0.1465	-0.253	-0.368
x_3	0.385	0.308	0.316	0.489
y_3	0.906	0.929	0.892	0.888

III. Time-Domain Performance Indices

When a closed-loop system response changes with respect to time, it is termed as the dynamic performance of the system. There are certain performance parameters as mentioned in [15], which are helpful to validate the system response. They are described as follows. Fig. 5 represents all the performance indices, which are helpful to analyze the stability of a system.

- **Delay-Time(T_d):** In transient state, the time required for the system response to outstretch from 0 – 50% of the desired value for the initial period of time is defined as Delay Time.
- **Rise-Time(T_r):** In transient state, the time required by system response to outstretch from 10 – 90% of the desired output for the initial period of time is said to be Rise-Time.
- **Settling-Time(T_s):** In transient state, the time required by the system to fix in a band of (2 – 5%) tolerance is termed as settling time.
- **Peak-Overshoot(M_p):** The deviation of the output response well above the desired output value (setpoint value) is considered as the Peak Overshoot. It is calculated by (6).

$$M_P(\%) = \frac{\textit{Peak value - Desired value}}{\textit{Desired value}} \times 100 \qquad (6)$$

- **Steady State Error(SSE):** The deviation of the system response from the desired value in steady-state is considered as steady-state error.

The simulink model for the heat exchanger system considered for the analysis is shown in Fig. 6. It contains step input signal block which is also termed as a desired response, summation block which calculates the error, PID block which is used to control the system and scope which is used to visualize the output response of the system.

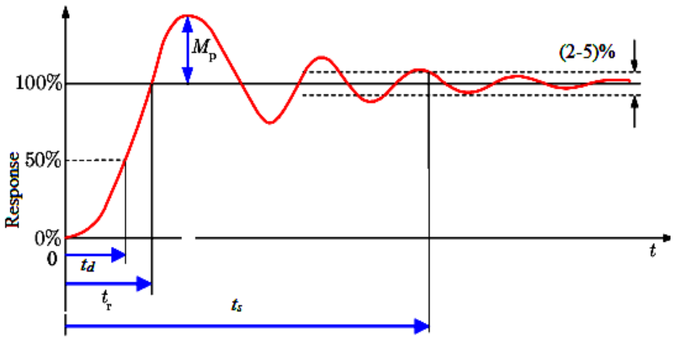


Fig. 5. Performance parameters for a first-order system response.

IV. Simulation Results

The following simulations represent the output responses of the system achieved by applying PID gains to the controller. controller gains for various tuning methods are represented in Table. II. Performance indices are used to justify the system stability in the time domain. For a system that has less performance indices values are said to be a stable system. ISE and ISTE methods give more oscillations in their transient time as shown in Fig. 7 and Fig. 8. For ISTSE and ITAE methods, the responses recorded less oscillations in transient time as shown in Fig. 9 and Fig. 10. Comparative visualization of all the system responses is shown in Fig. 11, where it is easy to compare responses in terms of performance indices.

TABLE II. Controller Gains For Various EPI Methods

S. No	Tuning Methods	K_p	K_i	K_d	T_i	T_d
1	ISE	3.1713	0.06636	26.607	47.788	8.391
2	ISTSE	2.954	0.05133	16.016	57.543	5.422
3	ISTE	3.153	0.05565	20.576	56.654	6.596
4	ITAE	2.755	0.070.	13.987	39.176	5.077

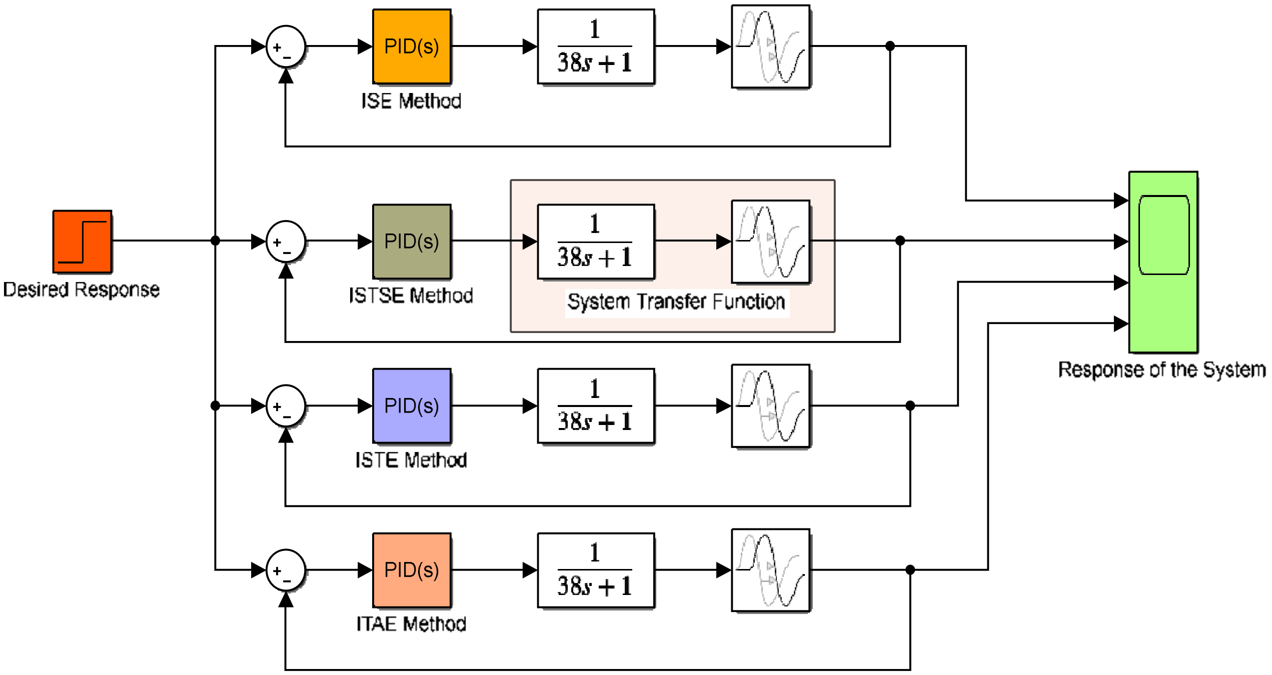


Fig. 6. MATLAB/Simulink model for heat exchanger system with PID controller designed with various EPI methods.

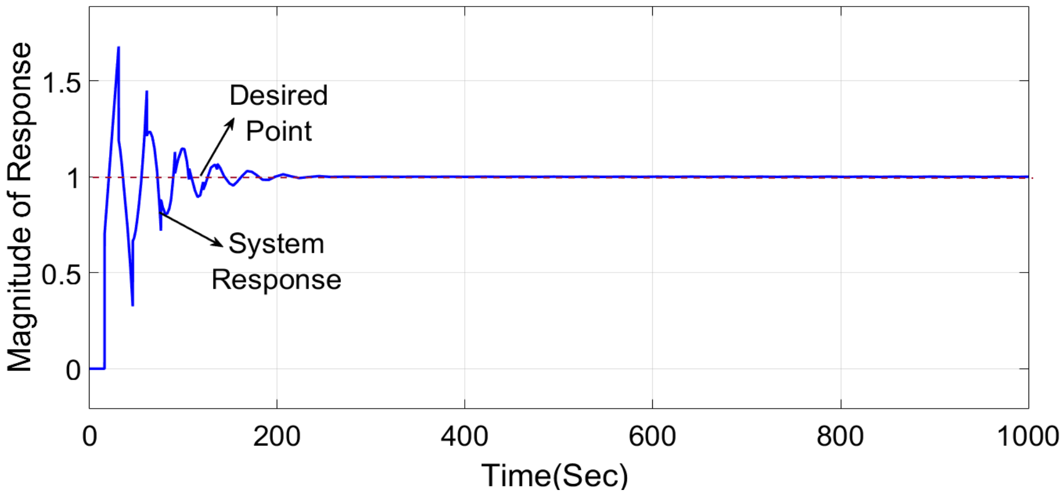


Fig. 7. Response obtained by using ISE tuning method

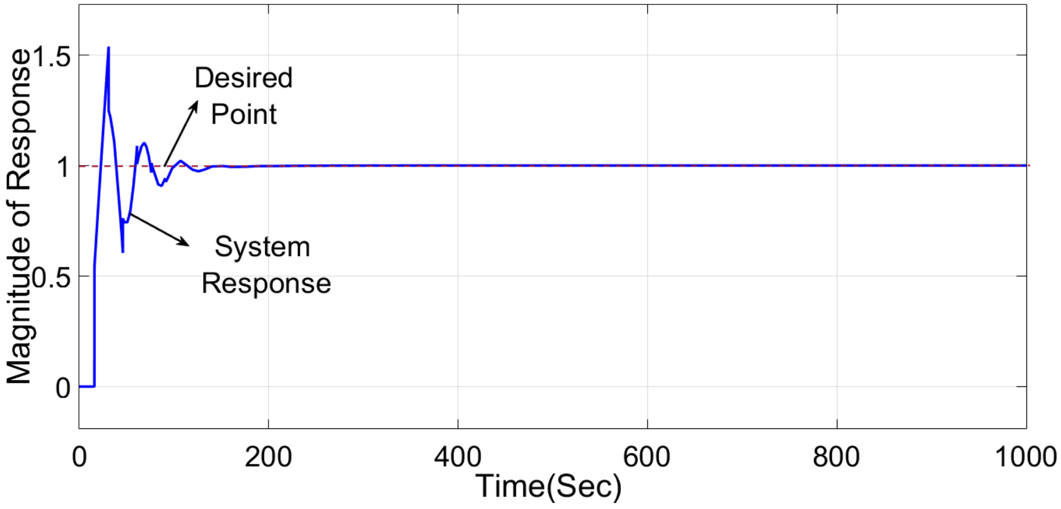


Fig. 8. Response obtained by using ISTE tuning method

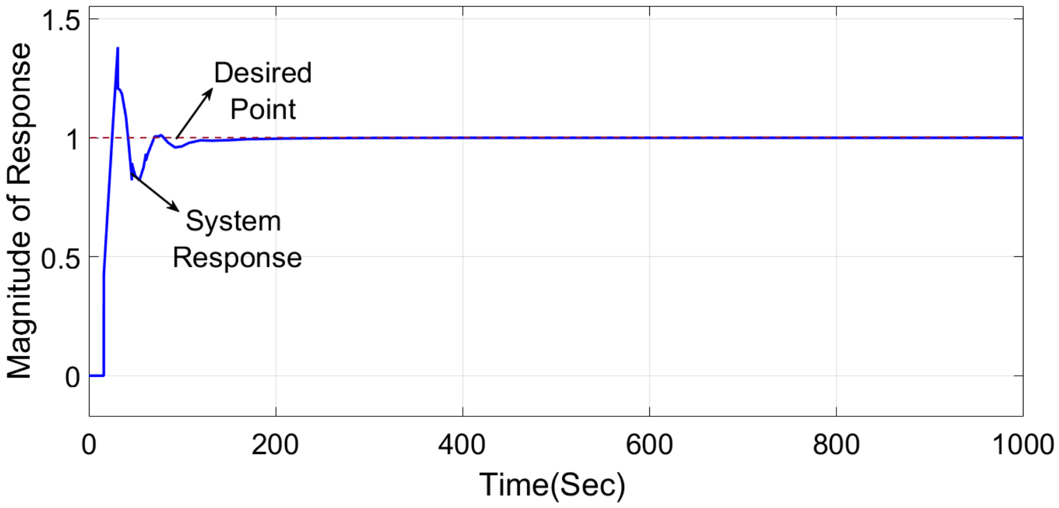


Fig. 9. Response obtained by using ISTSE tuning method

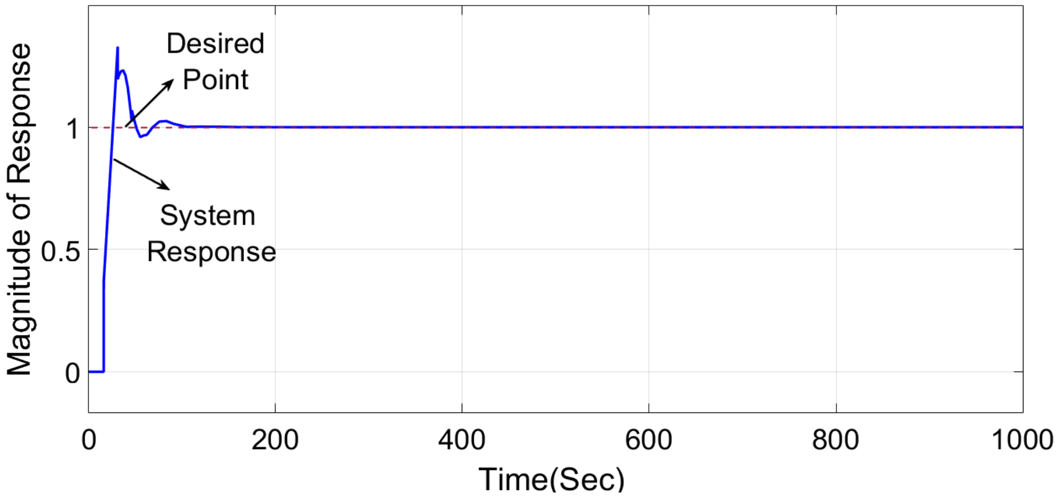


Fig. 10. Response obtained by using ITAE tuning method

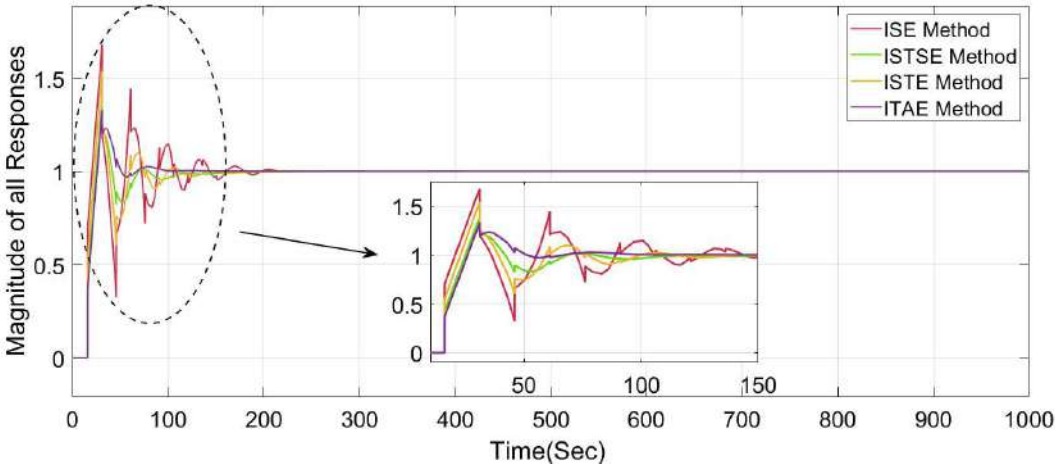


Fig. 11. Comparative system responses obtained by all EPI tuning methods.

TABLE III. Time Domain Performance Indices of the System Simulated with Various EPI Methods

S. No.	**EPI Tuning Method**	**Time Domain Performance Indices**				
		Delay-Time/ Lag Time (T_d)	***Rise-Time(T_r)***	***Settling-Time(T_s)***	***Peak-Overshoot(M_p)***	***Transient Nature of the System***
1.	ITAE Method	18.1	25.6	88	24.81	Oscillatory
2.	ISE Method	16.01	19.05	200	40.11	Oscillatory
3.	ISTE Method	16.02	21.23	132	34.64	Oscillatory
4.	ISTSE Method	17.2	24.95	109	27.53	Oscillatory

V. Conclusion

This paper implements and evaluates all the Error performance index based PID tuning methods for heat exchanger control system. Various time-domain performance results are obtained for the comparative analysis. These are computed for each EPI method and tabulated as shown in Table. III. By comparing all these obtained values, the following conclusions are made.

- ITAE method takes less time to settle, less percentage of peak overshoot and no steady-state error when compared with other EPI methods.
- ISE method has shown degraded performance for this system as it has a high settling time and peak overshoot. Further, it is having high oscillatory behavior in the transient state.

Thus, among all the EPI methods, ITAE is recommended as the superior method for the design of the PID controller for the heat exchanger system.

REFERENCES

[1] M. Arie, D. Hymas, F. Singer, A. Shooshtari and M. Ohadi, “Performance Characterization of a Novel Cross-Media Composite Heat Exchanger for Air-to-Liquid Applications,” 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), Las Vegas, NV, USA, 2019.

[2] S. Gupta, R. Gupta and S. Padhee, "Parametric system identification and robust controller design for liquid–liquid heat exchanger system," in IET Control Theory & Applications, Vol. 12, pp. 1474-1482, 2018.

[3] Murthy B. V., Pavan Kumar Y. V., Kumari U. V. R., “Application of neural networks in process control: automatic/online of pid controller gains for ±10% disturbance rejection,” IEEE International Conference on Advanced Communication Control and Computing Technologies, Ramanathapuram, India, 2012.

[4] V. Bharath Kumar, Godavarthi Charan, and Y. V. Pavan kumar, “Design of Robust PID Controller for Improving Voltage Response of a Cuk Converter”, Innovation in Electrical and Electronic Engineering, Delhi, India, Vol.661, pp. 301-318, 2020.

[5] R. Karthik, A. S. Hari, Y. V. Pavan Kumar and D. J. Pradeep, "Modelling and Control Design for Variable Speed Wind Turbine Energy System," International Conference on Artificial Intelligence and Signal Processing (AISP), Amaravati, India, pp. 1-6, 2020.

[6] Busra Ozgenc, M. Sinasi Ayas, and Ismail H. Altas, “A Hybrid optimization approach to design optimally tuned PID controller for an AVR system,” International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2020.

[7] Anto E. K., Asumadu J. A., Okyere P.Y., “PID control for improving P&O-MPPT performance of a grid-connected solar PV system with ziegler-nichols tuning method,” IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, China, pp. 1847-1852, 2016.

[8] Wen Tan, Jinzhen Liu, Ton Gwen Chen, and Horacio J. Marquez, “Comparison of some well-known PID tuning formulas, Computers & Chemical Engineering, Vol.30, pp.1416-1423, 2006.

[9] Vinayambika S. Bhat, Akshitha G. Shettigar, Nikhitha, Nidhi Dayanand, and K. P. Vishal Kumar, “Analysis of PID Control Algorithms for Transfer Function Model of Electric Vehicle”, International Journal of Recent Technology and Engineering (IJRTE), vol.8, pp.1022-1026, 2019.

[10] A. H. Hazza, M. Y. Mashor, and M. C. Mahdi, "Performance of Manual and Auto-Tuning PID Controller for Unstable Plant - Nano Satellite Attitude Control System," 6th International Conference on Cyber and IT Service Management (CITSM), Parapat, Indonesia, pp. 1-5, 2018.

[11] A. Kumar and S. Pan, “A PID controller design method using stability margin with transient improvement criteria,” 4th International Conference on Electrical Energy Systems (ICEES), Chennai, India, pp. 506-510, 2018.

[12] Md Nishat Anwar, “Design of PID Controller for High-order Process via IMC Scheme in Frequency Domain”, International Conference on Computational and Characterization Techniques in Engineering & Sciences (CCTES), Lucknow, India, pp. 54-58, 2018.

[13] Sumukh Surya and D. B. Singh, “Comparative study of P, PI, PD and PID controllers for operation of a pressure regulating valve in a blow-down wind tunnel”, IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Manipal, India, pp. 1-3, 2019.

[14] K. Sandeep Rao, V. N. Siva Praneeth, and Y. V. Pavan Kumar, “Fuzzy Logic-Based Intelligent PID Controller for Speed Control of Linear Internal Combustion Engine,” Innovation in Electrical and Electronic Engineering, Delhi, India, Vol.661, pp. 505-521, 2020.

[15] K. Sandeep Rao, V. Bharath Kumar, V. N. Siva Praneeth, and Y. V. Pavan Kumar, “Fuzzy Logic Theory-Based PI Controller Tuning for Improved Control of Liquid Level System”, Intelligent Algorithms for Analysis and Control of Dynamical Systems, Jaipur, India, pp. 1-12, 2020.

Industrial Heating Furnace Temperature Control System Design Through Fuzzy-PID Controller

V. Bharath Kumar, K. Sandeep Rao, Godavarthi Charan, Y. V. Pavan Kumar
School of Electronics Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA
bharath.18bec7093@vitap.ac.in, sandeep.18bec7094@vitap.ac.in, charan.18bec7075@vitap.ac.in, pavankumar.yv@vitap.ac.in

***Abstract* - The industrial heating furnace (IHF) is a system that requires continuous monitoring and control over the temperature. A small deviation in the temperature may create a huge impact on the system. So, there is a need for proper control for IHF. The conventional way is using PID (Proportion + Integral + Derivative) controller as the process controller in industries. Many standard tuning algorithms are present to compute gain values for the PID controller. The drawback of these conventional controllers is poor disturbance rejection as these are tuned offline to the system so that they cannot address the disturbances that occur while the system is working. So, there is a need for an artificial intelligent controller that can rectify the disturbance that occurs while the system is running. This can be achieved by using Fuzzy logic based controller that feeds the gain values externally to the PID controller according to the disturbance. The entire system design is simulated with the help of MATLAB/Simulink software. The outcome depicts that the proposed fuzzy-based controller has the best dynamic performance, rapidity, and good robustness.**

Keywords - Industrial Heat Furnace, Tuning Methods, Proportional-Integral-Derivative, Fuzzy Logic, MATLAB/Simulink.

I. INTRODUCTION

The device that is used for heating the materials in industries is defined as a furnace. Heat will be generated in the furnace by mixing up fuels with air. Industries use these types of furnace equipment for melting up of metals, for combining two metals, forging (shaping) of metals, galvanizing, enamelling (coating of glass on the metals) etc. Some of the furnaces types are household furnaces, metallurgical furnaces, and industrial furnaces.

The use of Household furnaces is to generate heat inside the house. Generally, a household furnace is installed by providing the movement of fluid, which may be through hot water, steam or air. Metallurgical furnaces are used in factories for casting metal. In these, furnaces are used for heating and reheating of the metals which are used in: Slitting mills, including tinplate works and Rolling mills and industrial furnaces or direct fired heaters, are used for generation of heat in the industries. These furnaces are also used to serve the reactor which provides heat for the reaction. These types of furnaces have many designs. The designs vary because of fuel used and purpose of heating and the method of initiating combustion air. A simplified representation of an industrial process furnace is as shown in Fig.1.

In IHF the control aspect of temperature is the key part of the industrial process. So this control of the temperature can be achieved using a conventional PID controller [1]. For obtaining the gain values for this PID controller many traditional tuning methods are available [2]. Of these available tuning methods, some of the methods use only the open-loop response of the system which may not be appropriate when these values are used for closed-loop response of the system and another group of methods requires the sustained oscillations for calculating the gain values as it is not possible for all kinds of systems [3]-[5].

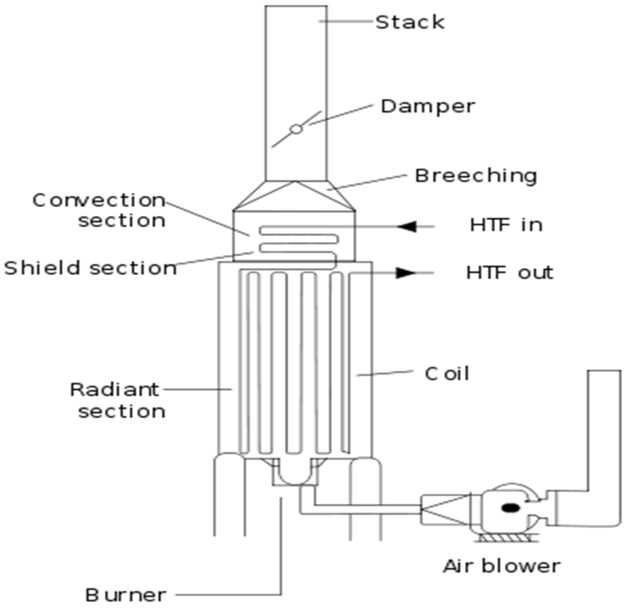


Fig. 1. Simplified representation of an industrial process furnace

In the same way, certain limitations are associated with each of the traditional tuning methods. So, picking up the correct tuning method for the targeted system becomes a difficult task and also the main imperfection of these traditional methods is that the rejection of disturbance is very poor and also these cannot address the online disturbance i.e. disturbances that occur while the system is running [6]-[8]. To overcome all the above-discussed problems there is a necessity of going to a fuzzy logic-based PID (FLP) controller which uses the fuzzy theory for obtaining the gain values according to the disturbance and feed them to the conventional PID externally [9], [10]. So, in this paper FLP controller is used to control IHF. The IHF can be qualitatively modelled using equations (1) - (2). The transfer function for this system can be obtained from [11].

$$H(s) = \frac{K}{TS+1} e^{-\tau s} \qquad (1)$$

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

$$H(S) = \frac{11}{50\,S + 1} e^{-8S} \qquad (2)$$

Where, K = 11, which represents the static gain; T = 50 sec, which represents the time constant; τ = 8 sec, which represents the lag delay time.

II. Implementation of Conventional PID Controller

Many industrial applications use PID controller for controlling different processes like temperature flow, pressure flow, liquid level etc. PID controller sets the system response to our desired response by choosing appropriate K_p, K_i, K_d gains values. Many conventional tuning rules are present to calculate the gain values but selecting a particular tuning rule for a PID controller is a tough task [5]. Moreover, all these methods are offline and need to set up manually to the controller. Proportional, Integral and Derivative gain are added up to obtain controller output as shown in (3).

$$U(t) = K_p e(t) + \frac{K_p}{T_i}\int_0^t e(t)dt + K_p T_d \frac{de(t)}{dt} \qquad (3)$$

Where,

U (t) = controller's output signal given to system
K_P = Proportional-Gain
$K_I = K_P / T_i$ = Integral-Gain
$K_D = K_P \times T_D$ = Derivative-Gain
E = Error = Desired value (DV) – Obtained Value (OV)
T = Time or instantaneous time

The following are some of the PID tuning rules for designing gain values of the PID controller.

Ziegler and Nichols introduced two methods namely Ziegler-Nichols (ZN) closed-loop method and Modified Ziegler-Nichols (MZN) method in the year 1942 [3]. While designing their controller tuning algorithms they used a phenomenon called allowable stability which mentions that the ratio of two succeeding peaks of the system response is close equals to 1/4 as shown in Fig.2. Modified Ziegler-Nichols (MZN) is another tuning method for designing gain values of closed-loop systems which are introduced by Ziegler and Nichols. The main focus is to initiate this method to provide good stability and better time-domain specifications when compared with the ordinary ZN tuning method.

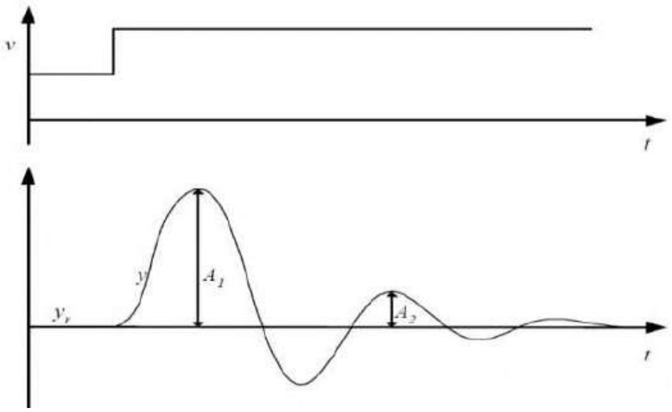


Fig. 2. Allowable stability condition for ZN method

Tyreus-Luyben's method was originated in the year 1992 which depends on oscillations in Ziegler-Nichols Method but with some minute changes in formulas for calculating PID gains. This method gives better stability in contrast to the Ziegler Nichols method [7]. The output response of the system which uses this method will have less oscillatory and less subtle uncertainty. Shinskey PID Controller is another tuning method for a closed loop control system. The main scope of this method is to yield good stability and better time-domain specifications and Zhaung-Atherton PID Controller is another tuning method for the closed loop control system. The main objective of this method is to provide good stability and better time domain specifications. So the formulas for designing the controller gains for various methods are given in Table. I.

The entire system design with the conventional controller is shown in Fig.3 where the desired value is considered as a step signal and controller gains are taken as shown in Table. II.

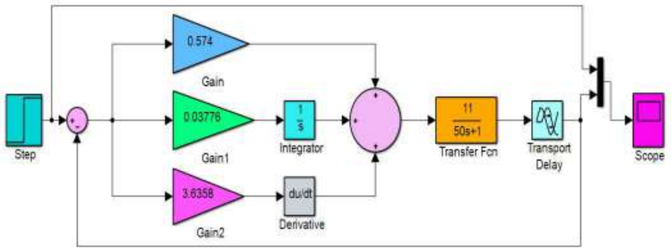


Fig. 3. System design with conventional PID controller

TABLE I. Design Formulas For Various PID Tuning Methods

Parameters	Zeigler-Nichols	Modified-ZN	Tyreus-Luyben	Shinskey Method	Zhaung & Atherton
T_i	$\frac{T_c}{2}$	T_c	$2.2 \times T_c$	$1.43 \times T_c$	$\frac{0.05 \times T_c}{36.3 \times K_c + 1}$
T_d	$\frac{T_c}{8}$	$\frac{T_c}{3}$	$\frac{T_c}{6.3}$	$0.52 \times T_c$	$0.13 \times T_c$
K_p	$0.6 \times K_c$	$0.2 \times K_c$	$0.45 \times K_c$	$\frac{0.95 \times T_c}{11 \times t}$	$0.51 \times K_c$
K_i	$\frac{K_p}{T_i}$	$\frac{K_p}{T_i}$	$\frac{K_p}{T_i}$	$\frac{K_p}{T_i}$	$\frac{K_p}{T_i}$
K_d	$K_p \times T_d$	$K_p \times T_d$	$K_p \times T_d$	$K_p \times T_d$	$K_p \times T_d$

TABLE II. PID GAINS WITH DIFFERENT TRADITIONAL METHODS

S.No	Tuning methods	K_P	T_I	K_I	T_D	K_D
1.	Ziegler-Nichols PID Controller	0.574	15.2	0.03776	3.8	3.6358
2.	Modified Ziegler-Nichols PID Controller	0.1913	30.4	0.00629	10.13	1.9885
3.	Tyreus-Luyben PID Controller	0.43	66.88	0.00643	4.825	2.0747
4.	Shinskey PID Controller	0.3281	11.44	0.0286	4.16	1.3648
5.	Zhuang-Atherton PID Controller	0.4879	54.312	0.00898	3.952	1.9281

III. IMPLEMENTATION OF PROPOSED FUZZY LOGIC BASED PID CONTROLLER

Fuzzy logic is the most effective controller for nonlinear, time-variant systems and also it doesn't require the transfer function of the system for obtaining gain values [12]. The IHF considered in this paper is a nonlinear system as the temperature of the furnace should be regulated constantly even in the case of disturbances that occur externally. So, to achieve this, the FLP controller is used for controlling the IHF in this paper. The fuzzy logic controller converts all the input and output values given to it in the range of 0 to 1. Thus, converted inputs and outputs are mapped using the fuzzy inference structure (FIS) as shown in Fig.4. In fuzzy logic control (FLC) all the crisp values are converted into the fuzzified values (between 0 to 1) and this process is called fuzzification. In this fuzzification process, many membership functions such as trapezoidal, triangular, signum etc., membership functions (MFCN) are available. These MFCN's are to be taken according to the type of the system. Mamdani and Sugeno are the two conventional fuzzification processes that are available. The defuzzification process starts after the processing of input data is done. The defuzzification process is quite opposite of the fuzzification process. In this, the fuzzified values are converted back into the crisp values. In the case of the defuzzification process, there are many methods such as the centre of sums, first of maxima method etc. are available. The relation between the input and output MFCN is done by creating the set of rules. The efficacy and the intelligence of the FLC depend on the rules that are fed to it. By using the FLP, it takes the supremacy of both the conventional PID controller and the intelligent FLC in controlling the system [13]. The gain values according to the disturbance are fed to the conventional PID controller so that it can address the disturbances efficiently [14] - [16]. So, this paper proposes the FLP controller developed using the gaussian MFCN and Mamdani method. MATLAB/Simulink model of the system with Fuzzy PID controller is shown in Fig.6 and the ranges of each MFCN is described in Table. IV. The Table. III and Fig. 5 shows the rules for mapping input and output where P1>P2>P3.

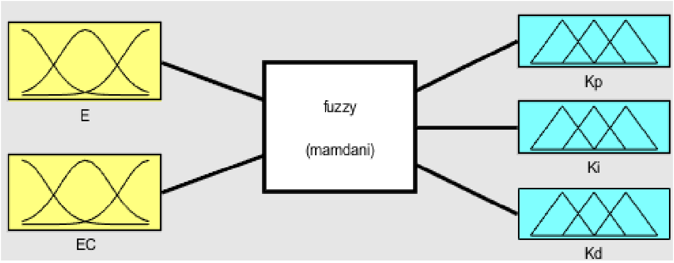
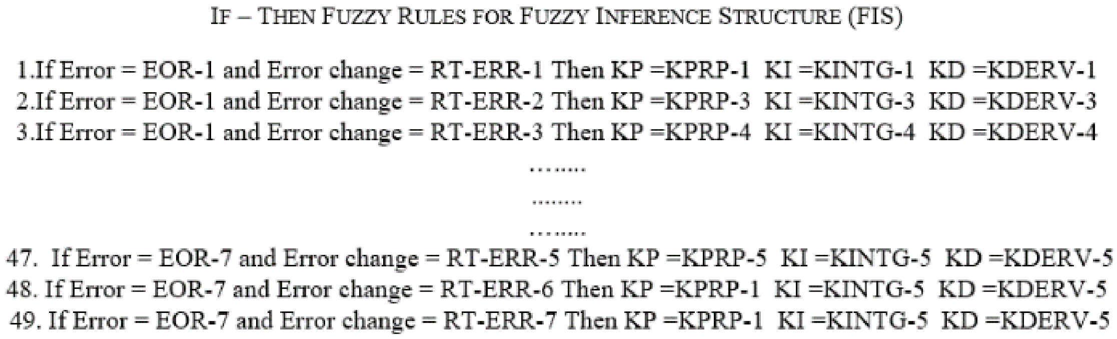


Fig. 4. FIS structure of FLC


IF – THEN FUZZY RULES FOR FUZZY INFERENCE STRUCTURE (FIS)

1.If Error = EOR-1 and Error change = RT-ERR-1 Then KP =KPRP-1 KI =KINTG-1 KD =KDERV-1
2.If Error = EOR-1 and Error change = RT-ERR-2 Then KP =KPRP-3 KI =KINTG-3 KD =KDERV-3
3.If Error = EOR-1 and Error change = RT-ERR-3 Then KP =KPRP-4 KI =KINTG-4 KD =KDERV-4

……..
……..
……..

47. If Error = EOR-7 and Error change = RT-ERR-5 Then KP =KPRP-5 KI =KINTG-5 KD =KDERV-5
48. If Error = EOR-7 and Error change = RT-ERR-6 Then KP =KPRP-1 KI =KINTG-5 KD =KDERV-5
49. If Error = EOR-7 and Error change = RT-ERR-7 Then KP =KPRP-1 KI =KINTG-5 KD =KDERV-5

Fig. 5. Statements used to develop the FIS

TABLE III. RULES USED FOR IMPLEMENTING THE PROPOSED FUZZY LOGIC CONTROLLER

INPUT-1 INPUT-2	EOR-1	EOR-2	EOR-3	EOR-4	EOR-5	EOR-6	EOR-7
RT-ERR-1	KINTG- 3	KINTG -3	KINTG- 3	KINTG -3	KINTG -2	KINTG- 1	KINTG-4
RT-ERR-2	KINTG 3	KINTG 3	KINTG 2	KINTG 2	KINTG 1	KINTG-4	KINTG-7
RT-ERR-3	KINTG 3	KINTG- 3	KINTG- 2	KINTG- 1	KINTG-4	KINTG-7	KINTG-6
RT-ERR-4	KINTG- 3	KINTG- 2	KINTG- 1	KINTG-4	KINTG-7	KINTG-6	KINTG-5
RT-ERR-5	KINTG- 2	KINTG- 2	KINTG-4	KINTG-7	KINTG-6	KINTG-5	KINTG-5
RT-ERR-6	KINTG- 1	KINTG-4	KINTG-7	KINTG-6	KINTG-6	KINTG-5	KINTG-5
RT-ERR-7	KINTG-4	KINTG-7	KINTG-6	KINTG-5	KINTG-5	KINTG-5	KINTG-5

TABLE IV. Ranges Used For Implementing Fuzzy Logic Controller

MFCN		Ranges	
Input Variable	Error	EOR-1	[0.8496-5.937]
		EOR-2	[0.849 -4]
		EOR-3	[0.85 -2.03]
		EOR-4	[0.8495 0]
		EOR-5	[0.849 2]
		EOR-6	[0.8495 4]
		EOR-7	[0.8509 6]
	Rate of error	RT-ERR-1	[0.8493 -6]
		RT-ERR-2	[0.8493 -4]
		RT-ERR-3	[0.8493 -2]
		RT-ERR-4	[0.8493 0]
		RT-ERR-5	[0.8493 2]
		RT-ERR-6	[0.8493 4]
		RT-ERR-7	[0.8509 6]
Output Variable	K_p	KPRP-1	[0.8493 -6]
		KPRP-2	[0.8493 -4]
		KPRP-3	[0.8493 -2]
		KPRP-4	[0.8493 0]
		KPRP-5	[0.8493 2]
		KPRP-6	[0.8493 4]
		KPRP-7	[0.8509 6]
	K_i	KINTG-1	[0.8493 -6]
		KINTG-2	[0.8493 -4]
		KINTG-3	[0.8493 -2]
		KINTG-4	[0.8493 0]
		KINTG-5	[0.849 2]
		KINTG-6	[0.8493 4]
		KINTG-7	[0.8509 6]
	K_d	KDERV-1	[0.8493 -5.968]
		KDERV-2	[0.8493 -4]
		KDERV-3	[0.8493 -1.937]
		KDERV-4	[0.8493 0]
		KDERV-5	[0.8493 2]
		KDERV-6	[0.8493 4]
		KDERV-7	[0.851 6.03]

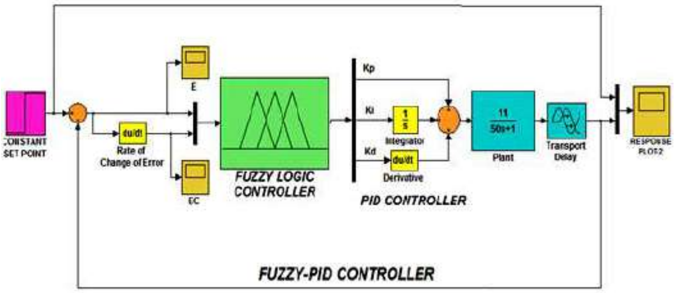


Fig. 6. System design with FLP Controller

IV. Results and Analysis

The results are segregated into time-domain response analysis and frequency-domain response analysis. These results discuss the system behaviour under the transient state and the steady state as described follows.

A. Time-Domain Response Analysis

The following results show the time domain system response with various conventional PID controller and FLP controller. The time-domain specification will help to judge whether the system response is better or not. Various time-domain performance indices like Rise time (RT), Delay Time (DT), Settling Time (ST), peak overshoot (PO) and transient behaviour for all the system responses are listed in Table.5. The responses which are having less settling time with no peak overshoot and smooth transient behaviour will be considered as a good response. The output response of the system with the ZN method, MZN method and Shinskey method has a high Peak overshoot. In transient state, all the system responses with conventional tuning methods show oscillations except Fuzzy PID and Tyreus-Luyben as shown in Fig.7 and Fig.8.

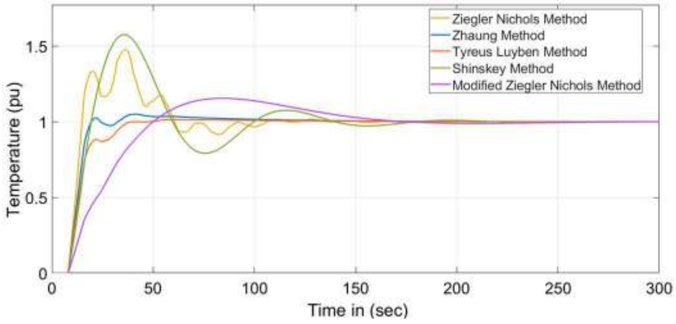


Fig. 7. Comparison of various traditional PID methods for the system

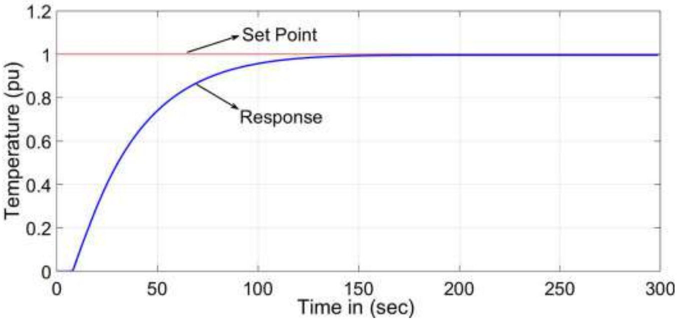


Fig. 8. System output with FLP controller.

TABLE V. Time Domain Perfoemance Indices of Conventional vs Proposed Fuzzy PID controller

S. No	Tuning methods	Delay-Time (D_T) in sec	Rise-Time (R_T) in sec	Settling-Time (S_T) in sec	Peak-Overshoot (P_O) in %	Transient Behavior
1	Ziegler-Nichols PID Controller	3.59	5.544	109.14	47.25	Oscillatory
2	Modified Ziegler-Nichols PID Controller	15.87	33.155	170.194	15.81	Oscillatory
3	Tyreus-Luyben PID Controller	5.363	20.698	39.665	1.5	No Oscillations
4	Shinskey PID Controller	11.48	12.455	175.3	57	Oscillatory
5	Zhuang-Atherton PID Controller	9.6	8.86	106.3	7	Oscillatory
6	Fuzzy-PID Controller	24.4	65.754	129.2	No Overshoot	Smooth

B. Frequency-Domain Response Analysis

The most commonly used methods for analysing the stability of the system in the frequency domain is the bode plot. Bode plot analysis is a process of analysing the transient-state and the steady-state response of the system with the help of frequency and phase. The stability of the system is given can be decided with the help of gain margin (GM) and phase margin (PM). So, to analyze these parameters, the bode plots are drawn as given through Fig. 9 to Fig. 14. The observations from these plots are given as follows.

- Fig 9 shows the bode plot for the ZN-PID method and from this, it can be observed the PM value are 138, -180 and the GM value is 0.
- Fig 10 shows the bode plot for the MZN PID method and from this, it can be observed the PM value are 152, -180 and the GM value is 0.
- Fig 11 shows the bode plot for the Zhaung-Atherton PID method and from this, it can be observed the PM value is -180 and the GM value is 0.
- Fig 12 shows the bode plot for the Shinskey-PID method and from this, it can be observed the PM value are 119, -180 and the GM value is 0.
- Fig 13 shows the bode plot for the Tyreus-Luyben PID method and from this, it can be observed the PM value are 152, -180 and the GM value is 0.
- Fig 14 shows the bode plot of the FLP method and from this, it can be observed the PM value is -180 and the GM value is 28.2.

Normally, the response with high PM and GM are considered as the more stable system. Hence, it can be concluded that the proposed FLP method is superior in terms of GM.

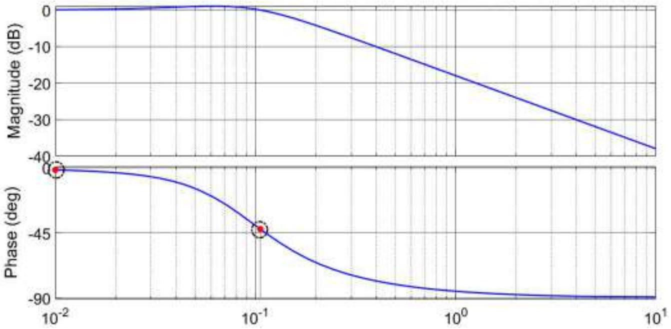


Fig. 9. Bode plot for ZN-PID controller

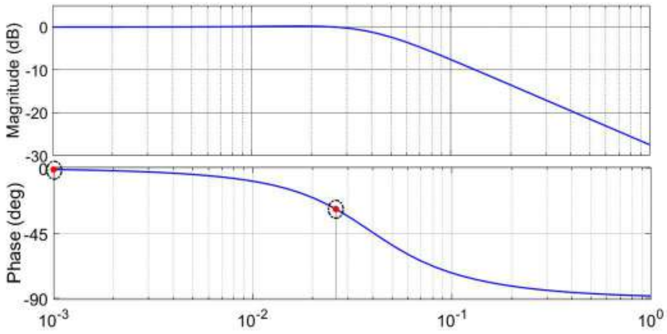


Fig. 10. Bode plot for MZN PID controller

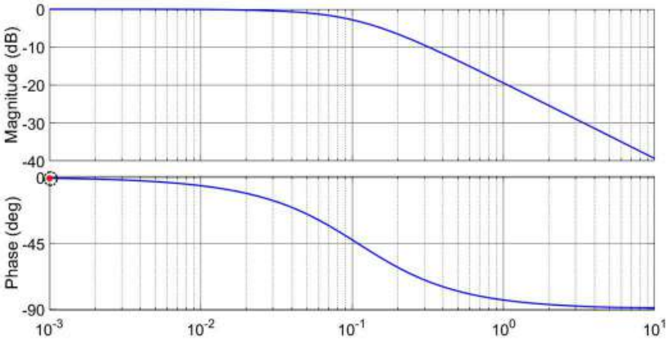


Fig. 11. Bode plot for Zhaung-Atherton PID controller

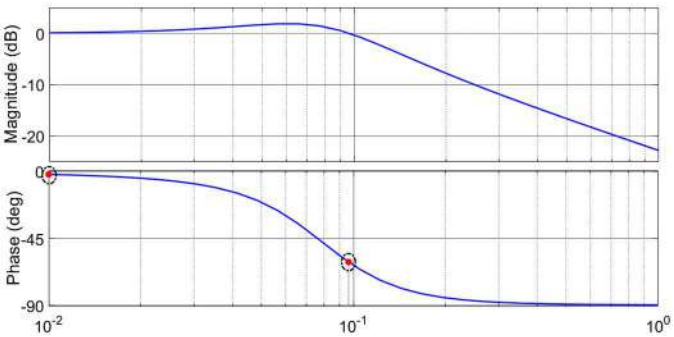


Fig. 12. Bode plot for Shinskey-PID controller

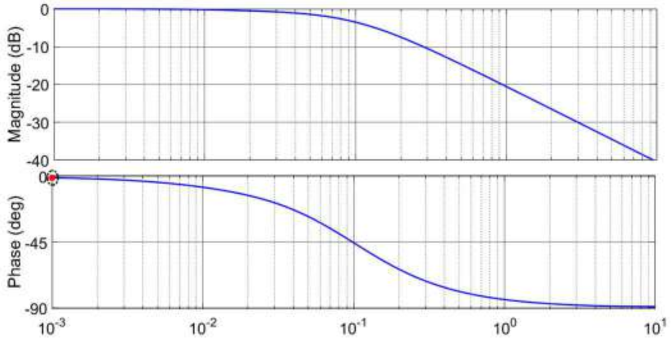


Fig. 13. Bode plot for Tyreus-Luyben PID controller

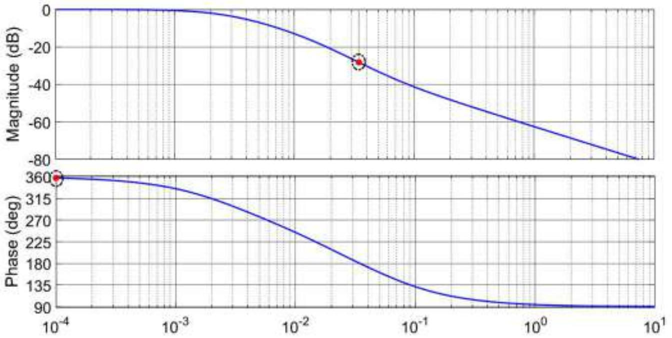


Fig. 14. Bode plot for FLP controller

V. Conclusion

With the use of the traditional PID controller for controlling the IHF system, there are some drawbacks such as poor rejection of disturbance, picking of the appropriate method according to the system, etc. These are addressed by using the proposed FLP controller. With the use of the proposed FLP for the IHF system, the following achievements are obtained.

- As the IHF system has non-linear characteristics, the controller design using traditional methods is not effective. Whereas, the use of the proposed FLP controller has produced better output.
- The proposed FLP controller produces a smooth response of the system with no peak overshoot wherein case of using the traditional controller, the system produces an oscillatory response with peak overshoot.
- The proposed FLC has a higher capability to reject the disturbance.

References

[1] Prusty S. B., Padhee S., Pati U. C., Kamala K. M., "Comparative performance analysis of various tuning methods in the design of pid controller," Michael faraday IET International Summit: MFIIS-2015, Kolkata, India, pp.43-48, 2015.

[2] R. Karthik, A. Sri Hari, Y. V. Pavan Kumar, D. John Pradeep, "Modelling and Control Design for Variable Speed Wind Turbine Energy System", 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), Amaravati, India, pp.1-6, 2020.

[3] V. Bharath Kumar, Godavarthi Charan, Y. V. Pavan kumar, "Design of Robust PID Controller for Improving Voltage Response of a Cuk Converter", Innovation in Electrical and Electronic Engineering, Delhi, India, Vol. 661, pp. 3011-318, 2020.

[4] Anto E. K., Asumadu J. A., Okyere P.Y., "PID control for improving P&O-MPPT performance of a grid-connected solar PV system with ziegler-nichols tuning method," IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, China, pp. 1847-1852, 2016.

[5] Ajila T Yimchunger, Debasis Acharya, Dushmanta Kumar Das, "Particle Swarm Optimization based PID-Controller Design for Volume Control of Artificial Ventilation System", 2020 IEEE Calcutta Conference (CALCON), Calcutta, India, pp. 278-282, 2020.

[6] Lin J. M., Kao P. F., Cho K. T., "Ziegler-Nichols based intelligent controller design of a SPM system," International Conference on Automatic Control and Artificial Intelligence (ACAI), Xiamen, pp. 2272-2275, 2012.

[7] Tasoren A. E., Orenbas H., Sahin S., "Analyze and comparison of different PID tuning methods on a brushless DC motor using atmega328 based microcontroller unit," International Conference on Control Engineering & Information Technology (CEIT), Istanbul, Turkey, pp.1-4, 2018.

[8] Srinivas P., Lakshmi K. V., Kumar V. N., "A comparison of PID controller tuning methods for three tank level process," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, No. 1, pp. 6810-6820, 2014.

[9] T. Anitha, G. Gopu, M. Nagarajapandian, P. Arun Mozhi Devan, "Hybrid Fuzzy PID Controller for Pressure Process Control Application", 2019 IEEE Student Conference on Research and Development SCOReD, Malaysia, pp. 129-133, 2019.

[10] Rohit Bhimte, Kalyani Bhole, Pritesh Shah, "Fractional Order Fuzzy PID Controller for a Rotary Servo System", 2nd International Conference on Trends in Electronics and Informatics, Tirunelveli, India, pp.538-542, 2018.

[11] Hongfu Zhou, "Simulation on Temperature Fuzzy Control in Injection Mould Machine by Simulink," Guangdong Provincial Science and Technology Founding of China, pp.123-128, 2007.

[12] Kartik Sharma, Dheeraj Kumar Palwalia, "A modified PID control with adaptive fuzzy controller applied to DC motor", International Conference on Information, Communication, Instrumentation and Control, Indore, India, pp.1-6, 2017.

[13] T. Anitha, G. Gopu, M. Nagarajapandian, P. Arun Mozhi Devan, "Hybrid Fuzzy PID Controller for Pressure Process Control Application", IEEE Student Conference on Research and Development, Bandar Seri Iskandar, Malaysia, pp.129-133, 2019.

[14] Roshan Bharti, Rishika Trivedi, Prabin k. Padhy, "Design of Optimized PID Type Fuzzy Logic Controller for Higher Order System", 5th International Conference on Signal Processing and Integrated Networks, Noida, India, pp.760-764, 2018.

[15] K. Sandeep Rao, V. N. Siva Praneeth, Y. V. Pavan Kumar, "Fuzzy Logic-Based Intelligent PID Controller for Speed Control of Linear Internal Combustion Engine", Innovation in Electrical and Electronic Engineering, Delhi, India, Vol. 661, pp. 505-521, 2020.

[16] Samarth Singh, R. Mitra, "Comparative analysis of robustness of optimally offline tuned PID controller and Fuzzy supervised PID controller", 2014 Recent Advances in Engineering and Computational Sciences, Chandigarh, India, pp.1-6, 2014.

Context-aware Authorization Model for Smartphones

Moeiz Miraoui

High Institute of Scientific and
Technological Studies

University of Gafsa ,Tunisia

College of computer at Al-Lith
Umm Al-Qura University
Makkah, KSA
mfmiraoui@uqu.edu.sa

***Abstract*—Smartphones are considered as the most pervasive electronic devices all over the world. These hand-held devices are carried with people wherever they go. Thanks to the availability of communication infrastructures anywhere, people depend on these devices to accomplish some tasks related to their work (resp. study), indispensable social media tool and a storage tool for a lot of user's private content such as photos, videos, messages, emails, etc. For that, smartphones should be protected against unauthorized access. Current access control models to services and applications of the smartphones are static (fixed) even though the user context is dynamic, and some access control model could not be appropriate in some contexts for the security and safety of both user's privacy and smartphone. Moreover, existing access control models are not dedicated to personal devices like smartphones but oriented to multi-users and network devices. In this paper I propose a context-aware access control model for smartphones which adapts dynamically to the user's context and provide appropriate access control mechanism to improve the security and safety of both user's privacy and smart phones.**

***Keywords*—smartphones, access control, context, authorization, dynamic, permission.**

I. Introduction

People nowadays rely on their smartphones to accomplish a lot of their everyday tasks, social communication and storing private contents. Smartphones are becoming a precious device for them which should be protected against unauthorized access and should be kept secure all the time. Current access control mechanisms to services and applications of the smartphone are mostly static (set once by the user) even though the user context is dynamic and access control model could not be appropriate in some contexts for the security and safety of both user's privacy and smartphone. Some safe and trusted contexts such as sitting alone at home do not require a powerful mechanism for access control to smartphone and its applications because the security risks (e.g., shoulder surfing or the use by non-trusted person) are minim. On the other hand, if the user is in a public place such as a cafe, the smart phone access system should consider the risk of using the smartphone in such a place like shoulder surfing when user access some services to keep his/her privacy away from the eyes of nosy people. Also reinforce the security of the smartphone when there is a possibility that the user's smartphone is used by a stranger which could be happen frequently when the user forgets to take his smartphone with him/her and leaves it with companions for whatever reason. In addition, mostly access control mechanisms are not oriented to personal devices like smartphones but to multi-users devices and networked equipment. In this paper I propose an approach to adapt the access control system to the changing context of the user to provide appropriate and secure access to smartphone and its applications. The context-aware access control adaptation is done using contextual information especially location and surrounding people.

The rest of this paper is organized as follows: Section 2 presents previous work on access control models. Section 3 discusses the risks of using a smartphone related to location and nearby people. Section 4 describes the proposed approach for context-aware access control and Section 5 concludes the paper.

II. Previous work

In the field of information security, Access control is a fundamental security mechanism to regulate who can access an asset and what can be done with it. Access control is a selective restriction of access to computer devices, networks, applications and data. Its main goal is to minimize the security risks of an authorized access by ensuring that users are who they claim they are and have been granted the appropriate permissions. Generally, the access control system process consist of three basic steps:1) identification is the act of uniquely identify a person (resp. a subject) which performed generally by providing a user name, smart card or anything else that uniquely identify the user. 2) authentication which consists of proving a claimed identity by providing the appropriate credentials such as password, a personal identification numbers (PINs), a security token, biometric scans or other authentication factors. 3) authorization that specifies access right/privileges to resources and what can be done according to permission granted.

Among the first models of access control is the Discretionary Access Control (DAC) [1] which is based on an access control list that describes what operations can be done on a given object by each subject. Another model is the

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

mandatory access control in which a central authority regulates access rights to resources according to a multiple level of security. The role-based access control model was introduced by [2] to enforce the previous access control ones where access to resources is granted according to the job function of a user (resp. group of users) inside an organization instead of identities of individual users. Wang et al.[3] proposed the Attribute-based access control (ABAC) model in which access to resources was granted according to the attributes of users, systems and environmental conditions. Das et al [4] developed the MITHRIL framework [5] which included an access control component called the MithrilAC that used the Attribute-Based Access Control standard. The aim of the MithrilAC component was to capture the behavior of an application in a particular context and user's access control needs in order to refine the access control policy. Arfaoui et al. [6, 7] proposed a new context-aware attribute-based access control scheme which took into account the dynamic change of context. the proposed approach was oriented to IoT devices which have limited processing capabilities. They combined the user's attributes with contextual information to check if the user satisfies or not the access policy. They used the CP-ABE (ciphertext-policy attribute-based encryption) scheme and embed the contextual information into access structures using contextual tokens. Kayes et al. [8] Presented a survey of security, privacy and access control research which used contextual information and oriented to cloud and fog networks. They proposed a Fog-Based Context-Aware Access Control (FB-CAAC) framework for accessing data from distributed cloud data centres along with IoT devices and fog computational nodes. Kayes et al. [9] proposed a Context-Aware Access Control (CAAC) Policy Framework based on the role-based access control approach which used a formal policy model to specify user-role and role- permissions assignment based on relevant contextual information. The policy used a context specification language and ontology for modeling context-aware access control (CAAC) policies. Ternka et al. [10] made an interesting survey on authentication and authorization methods for internet of things in which they showed how context-awareness extends security. Some other work like [11, 12, 13] were oriented to context-aware authentication systems rather than to the whole access control process.

III. RISKS RELATED TO LOCATION AND NEARBY PEOPLE

Nowadays, the best known and used technologies for wireless communication are Bluetooth and Wi-Fi and all smartphones are equipped with both wireless technologies. Several risks related to our pervasive use of smartphones need to be addressed and wireless security is just one often-overlooked piece of the mobile security. In a public place or in a crowded untrusted place, hackers can potentially gain access to a smartphone through either Bluetooth or Wi-Fi protocols. They can eavesdrop on our conversations and even steal our personal and business' data stored on smartphones. Some well-known attacks related to Bluetooth protocol are: 1) Bluebugging or Bluejacking attacks which are somehow old and recently are easily prevented. 2) Bluesnarfing attack where it is possible to access personal content and even copy the content of the smartphone. Alike, common Wi-Fi attacks include: 1) session Hijacking where hackers use sniffer tools to look through transmitted packets and find the information needed, generally occur when using an open Wi-Fi network. 2) Evil Twin attack where a fake access point with the same name of the one used by the user is setup by the attacker to capture all the data traffic made by the user's smartphone. 3) Man in the middle attack where the hacker listen and modify transmitted data between the user's smartphone and other parties. This attack can be initiated using DHCP spoofing among others methods.

To avoid the aforementioned attacks on wireless protocol, user's should avoid using public Wi-Fi networks or use a connection encrypted by a VPN (Virtual Private Network). Keep Bluetooth settings to non-discoverable or invisible in public places and switch it off when it is not in use. Also require user approval for devices pairing with the smartphone.

Due to the user's mobility, most smartphones services and applications can be accessed even in public places resulting in a high probability of a social engineering attack called shoulder surfing which consists of observing other people's information without their consent over their shoulders. This attack is generally used to obtain information such as passwords and other confidential data. To protect against such attack, the smartphone should demand a two-factors authentication (preferably one of them biometric) to access sensitive applications in public places. It is also possible that the smartphone notify the user to be with his/her back to the wall or to angle the smartphone screen so that other people cannot see what he/she is typing or watching. These are the two common ways to protect from prying eyes.

IV. CONTEXT-AWARE ACCESS CONTROL MODEL

A smartphone provides a set of services to the user and allow him/her to access and use several applications. In order to access the smartphone services and applications, user should be granted authorization to use the smartphone which generally locked for security purpose an require an authentication method which mostly a password, a PIN (personal identification number), a pattern drawing, fingerprint etc. But some users allow a direct access to their phone which is not advised at all in some contexts. Once user has access to smartphone, he/she could access most applications which do not require new authentications since user was already authenticated previously except some applications which demand a new authentication at each access. This situation exposes the smartphone to many risks related to access to personal data (resp. credentials) and the possibility of it being stolen if another person accesses the user's smartphone or sees the content of the smartphone by another person during normal use by the owner of the smartphone.

A. Applications classification

In order to provide a personal authorization system, smartphone user should provide his/her preferences in term of which permission should be granted to which applications. For this reason, applications (including access to smartphone) are classified in two categories: sensitive and normal.

It is up to the user to classify each application to either sensitive or normal application. The classification is done according to whether the application contains private data or user's credentials. For example, the clock application could be classified as a normal one because it does not contain any user's private data or credentials and access such application does not present any risk. However, a banking application should be classified as a sensitive one because it contains both private data and user's credentials. One example of possible applications classification is given by Table 1.

TABLE I. APPLICATIONS CLASSIFICATION

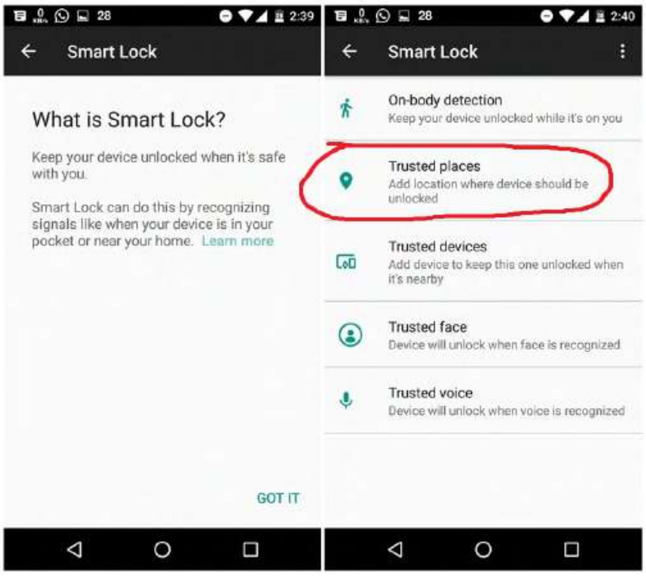
Application	**Type**	
	Sensitive	***Normal***
Access to smartphone (unclock)	■	
Clock		■
Social media (facebook, what's up, …)	■	
Calender		■
Banking	■	
Photo/video galery	■	
Phone book	■	
Games		■
Health	■	
e-mail	■	
Weather		■
SMS	■	
…	…	…

B. *Locations classification*

User and his/her smartphone can be in multiple locations which can be classified mainly in two types of location:

- Private location: set of places where risks of accessing smartphone and its applications are minim. It is somehow a trusted location. The best example of such location is home where user is not afraid that his/her smartphone can be used by a stranger and steal or view his/her personal data (resp. credentials). Such locations do not require high level permissions.
- Public location: the set of places where there are a lot of strange and non-trusted people and present a high security risk for smartphone accessing. These types of locations generally refer to public places and can be subdivided into two types:
    - ➢ Trusted public place: security risks on smartphone are not raised in such place. it is a trusted place. The ideal example of a trusted public place is the user's office at work.
    - ➢ Non-trusted public place: user does not has a private corner in such places and surrounded with a lot of strange people. These places present a high risks on smartphones.

A somewhat similar classification that allows the user to choose his trusted places to automatically unlock his/her smartphone without going through the authentication step is provided by the Google Smart Lock for android devices [14] (see Figure 1). However such application does classification into either trusted places or non-trusted places. In addition, this application is specific for only automatic unlock of smartphones and not for unlocking or accessing other applications.



What is Smart Lock?

Keep your device unlocked when it's safe with you.

Smart Lock can do this by recognizing signals like when your device is in your pocket or near your home.

On-body detection
Keep your device unlocked while it's on you

Trusted places
Add location where device should be unlocked

Trusted devices
Add device to keep this one unlocked when it's nearby

Trusted face
Device will unlock when face is recognized

Trusted voice
Device will unlock when voice is recognized

Fig. 1. Google Smart Lock [13]

C. *User's situations classification*

Due to the permanent user mobility, context of use of smartphone is ever changing. One important information which affect the context of use of smartphone is the user's nearby people. The main risk generated by this context element is the shoulder surfing attack especially when the user is surrounded by some people and trying to access sensitive (resp. private) data through his/her smartphone. This context element is not independent but influenced by the user's location. Even though in a same situation where the user is alone, it differs either his alone in a private place or alone in a public place. So granting permission to user should consider both his/her situation and location. There are two main situations:

- Alone: In this situation, the risk of shoulder surfing attack is practically absent.
- Not alone (surrounded): in this situation the occurrence of a shoulder surfing attack is strongly possible.

D. *Smartphone's use context*

In all, there are six smartphone's use contexts related to locations and user's nearby people which are the result of the cartesian product of the set of possible user's locations and user's situations (Fig.2 shows all possible smartphone’s use context).:

- **C1**: The user is alone in a private place.
- **C2**: The user is not alone in a private place.
- **C3**: The user is a lone in a trusted public place.
- **C4**: the user is not alone in a trusted public place.
- **C5**: The user is alone in a non-trusted public place.
- **C6**: The user is not alone in a non-trusted public place.

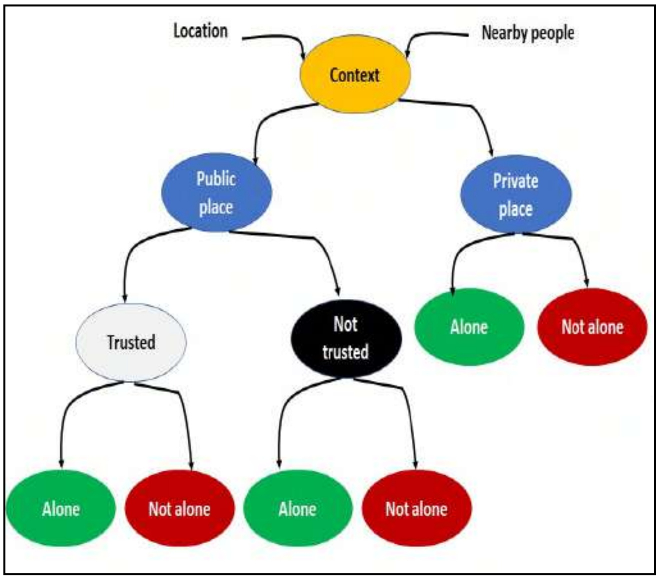


Fig. 2. Set of smartphone use context

E. Context-aware authorization model

In each of the aforementioned contexts, the user is vulnerable to some threats. Generally we can classify these threats into four classes (T1, T2, T3, T4) as follows:

- Use of public Wi-Fi network **(T1)**: the use of a public WiFi network can make the smartphone vulnerable to some attecks as mentioned in section III.
- Open Bluetooth **(T2)**: if the user's smartphone is open it could be vulnerable to some kind of attacks related to Bluetooth
- Visible content **(T3)**: when the smartphone content is visible, shoulder surfing attack may occur by nosy people and prying eyes
- Available smartphone **(T4)**: if user forget to carry his smartphone with him where ever he goes, it could be then accessed by someone else.

To every context of use is assigned a risk degree from the following list: {Zero, Low, Average, High, Very high} depending on how many type of the above risks present in that context.

In context C1, when the user is alone in a private place and using his private Wi-Fi network with trusted people, normally there is no threats and the risk degree will be zero in such context. Context C2 is alike to context C1 except the user is not alone, so the threat of accessing his/her smartphone by someone else or watching his/her personal content over his/her shoulder is present. So the risk degree will be Average. When the user in alone in a trusted public place, the smartphone will be connected to a trusted Wi-Fi network and controlled environment (case of office). The smartphone is vulnerable to one possible threat related to open Bluetooth connection where people generally omit to close their Bluetooth in a trusted place. In this case the risk degree is low because the possibility of making Bluetooth attacks in a trusted place is minim. In context C4, when the user is not alone in a trusted public place, the smartphone can be vulnerable to both shoulder surfing attack and access by other people in addition to possible Bluetooth related attack. In this case the risk degree is high. When the user is alone in a non-trusted public place (context C5), the smartphone is vulnerable to both Wi-Fi related attacks and Bluetooth related attacks. The risk degree will be then average. Finally, in context C6 when the user is not alone in a non-trusted public place, the smartphone is vulnerable to the four possible threats. and the risk degree of such context is very high. Table II shows the risk degree of each possible context of use.

TABLE II. RISK DEGREE OF EACH CONTEXT

	T1	**T2**	**T3**	**T4**	**Risk degree**
C1					Zero
C2			■	■	Average
C3		■			Low
C4		■	■	■	High
C5	■	■			Average
C6	■	■	■	■	Very high

Authorization to access smartphone or its applications will basically depend on two factors: context of use and sensitivity of the applications. The authorization system will consist of a simple if-then statements where the antecedent (the if-clause) is composed of the context of use and (logic and) the type of demanded application whereas the consequent (the then-clause) is the appropriate authorization. This rule-based authorization adaptation system will provide an acceptable decision speed which is required for the usability of the smartphone and it does not consume much resources of the smartphone which are limited in most cases.

Normal applications of the smartphone are granted a direct full access without any authentication after an authenticated access to the smartphone. For sensitive applications, access will be granted according to the current context of use and its degree of risk. The latter will determine the access method using a one-factor, two-factor or three-factor authentication to enforce the security of the smartphone.

In case of the usage of sensitive applications within the six aforementioned contexts, the access method will be chosen according to the risk degree of each context of use which composed of five degrees namely: zero, low, average, high and very high. The authorization method will adapted to current context of use as follows:

- In context C1with zero risk degree, the access to the smartphone and its application is direct with full permissions.
- In contexts C2 and C5 with an average risk degree, the access to the smartphone will be granted by unlocking the smartphone using any method (passcode, PIN, graphic pattern, fingerprint, ...) and sensitive applications will be accessed using a one factor authentication method. Normal applications will be accessed directly without authentication.
- In context C3 with a low access degree, the access to the smartphone is granted after unlocking the smartphone using any method and access to all applications is direct without additional authentication.
- In context C4 with a high risk degree, Access to smartphone is granted after unlocking it and access to sensitive applications will be granted after a two-factor authentication (e.g. password + fingerprint).
- In context C6 with a very high risk degree, access to smartphone is granted after unlocking it and access to sensitive applications will be granted after a three-factor authentication.

The set of rules that allow to use the appropriate access method is given by TABLE III.

TABLE III. ACCESS METHODS RULES OF ADAPTATION

Context	*Unlock*	*Sensitive App.*	*Normal App.*
C1	Automatic	Direct access	Direct Access
C2 & C5	Explicit unlock	One-factor authentication	Direct Access
C3	Explicit unlock	Direct Access	Direct Access
C4	Explicit unlock	Two-factor authentication	Direct Access
C6	Explicit unlock	Three-factor authentication	Direct Access

The explicit unlock means that the user should authenticate before using the smartphone. Two-factors and three-factor authentication are obviously considered heavy solution for the smartphone usability, but security has a price and users should invest in their privacy and security.

V. CONCLUSION AND FUTURE WORKK

Smartphones are carried with users where ever their go and we rely on them to perform several tasks in our everyday lives. Moreover we store in them a variety of private contents. The security of smartphones is becoming a must and a permanent concern for users. smartphone should be protected against unauthorized access. Unfortunately existing methods of access control do not consider the ever changing context of user and do not adapt dynamically to this context in order to provide an appropriate access method. In this proposed I proposed a context-aware access control method based on two main elements of context namely location and nearby people. The proposed approach change the access method according to the current context and the type of application to be accessed. The future work will consist on detailing sensitive applications and provide a more appropriate access method.

REFERENCES

[1] L. Qiu, Y. Zhang, F. Wang, M. Kyung, H.R. Mahajan, National Computer Security Center. Citeseer; Philadelphia, PA, USA: 1985. Trusted computer system evaluation criteria.

[2] R.S. Sandhu, E.J. Coyne, H.L. Feinstein, C.E. Youman, Role-Based Access Control Models. IEEE Comput. 1996;29:38–47. doi: 10.1109/2.485845

[3] L. Wang, D. Wijesekera, S. Jajodia, A logic-based framework for attribute based access control; Proceedings of the 2004 ACM Workshop on Formal Methods in Security Engineering; Washington, DC, USA. 25–29 October 2004

[4] K.D. Prajit, J. Anupam, F. Tim, Personalizing Context-Aware Access Control on Mobile Platforms, IEEE 3rd International Conference on Collaboration and Internet Computing (CIC), IEEE, 2017

[5] P. K. Das, "Context-dependent privacy and security management on mobile devices," 2017.

[6] A. Arfaoui, C. Soumaya, K. Ali, M. S. M. Senouci, Context-Aware Adaptive Remote Access for IoT Applications IEEE Internet of Things Journal, Volume: 7, Issue: 1, IEEE 2020

[7] A. Arfaoui, C. Soumaya, K. Ali, M. S. M. Senouci, H. Mohamed, Context-Aware Adaptive Authentication And Authorization in Internet of Things, ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China

[8] A. S. M. Kayes, K. Rudri, H. Iqbal, Sarker, Md. Saiful Islam, P. A. Watters, A. Ng, M. Hammoudeh, S. Badsha, I. Kumara, A Survey of Context-Aware Access Control Mechanisms for Cloud and Fog Networks: Taxonomy and Open Research Issues, Sensors (Basel) 2020 May; 20(9): 2464.

[9] A. S. M. Kayes, J. Han, W. Rahayu, T. Dillon, M. Saiful Islam, A. Colman, A Policy Model and Framework for Context-Aware Access Control to Information Resources, The Computer Journal, Volume 62, Issue 5, May 2019, Pages 670–705

[10] M. Trnka , T. Cerny , and N. Stickney, Survey of Authentication and Authorization forthe Internet of Things, Hindawi Security and Communication Networks, Volume 2018, pp. 1-17, 2018

[11] Z. Liu, R. Bonazzi, privacy-based adaptive context-aware authentication system for personal mobile devicess, Journal of mobile multimedia, Vol. 12, No. 1&2, pp. 159-180, 2016

[12] M Miraoui, S El-etriby, A Context-aware Authentication Approach for Smartphones, 2019 International Conference on Computer and Information Sciences (ICCIS), Al-Jouf, KSA

[13] E. Hayashi, S. Das, S. A. Amini, J. I. Hong, I. Oakley, CASA: Context-Aware Scalable Authentication, SOUPS '13: Proceedings of the Ninth Symposium on Usable Privacy and Security, pp.1-10, July 2013

[14] https://nerdschalk.com/how-to-automatically-unlock-your-android-phone-using-smart-lock-functions/

Analysis of the Impact of Traffic Density on the Compromised CAV Rate : a Multi-Agent Modeling Approach

Manal El Mouhib
Faculté des Sciences
Université Mohammed Premier
Oujda, Morocco
m.elmouhib@ump.ac.ma

Kamal Azghiou
École Nationale des Sciences Appliquées
Université Mohammed Premier
Oujda, Morocco
k.azghiou@ump.ac.ma

Abdelwahed Tahani
Faculté des Sciences
Université Mohammed Premier
Oujda, Morocco
wtahani@yahoo.fr

***Abstract*—Connected and Autonomous Vehicles (CAVs) combine both the characteristics present in autonomous vehicles and those already implemented in large parts in connected vehicles. Therefore, CAVs expose a larger attack surface. In this work, we propose a threat model concerning the attacks that can take place in the community of CAVs by exploiting the Vehicle to Vehicle (V2V) communication channels. Then, we simulate the proposed model using NetLogo (1000 simulation for each scenario). The adoption of multi-agent-based simulation allows us to abstract the nature of the attack, which leaves us to focus on the macroscopic effect of this on different traffic densities. The got results show that the more a CAV is engaged in dense road traffic, the more it is likely to be compromised. Finally, in order to reduce the effect of this category of cyberattacks, we are introducing countermeasures to be adopted in the development of security policy within the context of Intelligent Transportation System (ITS).**

***Index Terms*—connected and autonomous aehicles (CAVs), cyberattack, cybersecurity, intelligent transportation system (ITS), multi-agent-based simulation (MABS), Threat Model, V2X, V2V.**

I. Introduction

The capacity of a vehicle to exchange information with its environment according to the V2X paradigm is at the heart of the development of intelligent transport systems. This being effective thanks to the sensors available to the CAV allowing it to extract useful information concerning its ambient environment. Such information can be road signs, start lines or obstacles to avoid. A CAV can use the cellular or ad-hoc networks at its disposal to share relevant information with other components of the intelligent transport system.

802.11p gave emergence to two more or less similar protocols: DSRC adopted by North America and C-ITS used in Europe. Adoption of one of these standards allows moving vehicles to establish communications with other vehicles nearby and with road infrastructure. In this approach, a CAV can transmit messages step by step without having to resort to the cellular network. Such interactions are characterized by very low latencies and ranges of the order of a kilometer. The main applications using this type of communication are those related to CAV safety, such as collision avoidance and overtaking assistance systems.

The decentralized nature of V2V communications results in a significant attack surface [1], [2]. Sybil is an example of an attack that the intelligent transport system may be the target [3]. This attack consists of forging false identities of CAVs in order to make the rest of the system believe the presence, in a zone, of a certain number of these. The aim is to manipulate the behavior of the intelligent transport system according to a logic previously set by the attacker. Forced routing of packets to a destination other than that intended by the source is another type of attack that CAV systems can suffer from. We know this type of attack as black holes.

The safety of the CAV relies on the reliability and security of communications carried out according to the V2V paradigm [4], [5]. This requires a minimum of reciprocal trust between the various stakeholders. The authors of [6], [7] have already pointed out the threat of computer attacks targeting CAVs because of their interaction with their environment. For example, an attacker can block a CAV just by sending it malicious packets [6]. Many works have noted that the systems embedded in CAVs are vulnerable to remote attacks or those targeting its sensors [8], [9]. Therefore, it is crucial to consider these attacks and the spread of their underlying malwares as an emerging phenomenon in the CAVs community when they will hold a considerable part of road traffic.

Malware spreading in V2V ad-hoc networks is not the same as the one in the computer networks. Thus, applying the classical approaches as they exist to protect a CAV from other CAVs, may not be a good countermeasure. We can summarize the major differences in: (i) time and traffic conditions involve a dynamic topology of the network; (ii) Driving way or mobility is another factor that can affect the V2V links configuration which can bring more or less cybersecurity risks to the targeted vehicle; (iii) Target pointing is another reason that can distinguish CAVs networks from computers ones.

Two models can help to characterize a V2V network

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

topology: the CAV mobility and the V2V communication models. The CAV mobility model deals with the movement of CAVs and defines the neighborhood of each CAV. This model depends on traffic density and road topology. The V2V communication model deals with data transmission and its underlying constraints such as interferance and transmission power. To study Malware propagation behavior in a such Networks, one needs to know all their topological details.

In what follows, our contribution is threefold: (i) we set up a threat model for traffic of CAVs, in a four-lane highway, in which one of the CAVs is malicious. (ii) We simulate our model by adopting the multi-agent-based approach to consider emerging phenomena and this using the NetLogo software. After carrying out 1000 simulation runs for each scenario, we visualize the results as tables and plots using Matlab. (iii) Finally, we introduce draft solutions aimed at clarifying the logic of the distribution of roles on the different stakeholders in order to reduce the intensity of cyber attacks having traffic density as a catalyst.

We can summarize the structure of this paper, as follows: We find just after the introductory section (sec. I), A section dedicated to the review of the literature related to the subject, here discussed (sec. II). At, the end of this section we give the research gap treated. The next section (sec. III) presents the adopted threat model. It determines the threat agent, the attack surface and the attack methods. Section IV concerns the simulation scenarios adopted and the different parameters considered. Quite after, section V discusses the main results. The last section (sec. VI) is dedicated to the conclusion of this work.

II. Literature review

A. Automotive cyber-security incidents

According to [10], the profitability of cybercrime, including the (ITS) sector, exceeds that of global illicit drug trafficking by 200 billion dollars (600 billion dollars against 400 billion dollars).

Thieves in India could steal 41,118 vehicles in 2019, according to a statement from Gurugram police. This has been possible using only inexpensive electronic devices [11]. Once the vehicle is stolen, its GPS service is deactivated, its license plates changed and a new registration certificate is issued.

Researchers stopped a Tesla Model X by attacking its Advanced Driver Assistance System (ADAS) by inserting an image of the stop sign in a roadside advertisement. The autopilot saw the image as a real sign. Likewise for Mobileye 630 Pro. Also, this type of attack can be applied remotely using a portable projector thus triggering a braking or a deviation, of the vehicle, towards oncoming traffic which can be fatal [12].

Security researchers from the Sky-Go team could open the doors and remotely start the engine of a Mercedes-Benz E-Class by exploiting 19 vulnerabilities [13].

B. Related works

We can find in the research literature various works that model the dynamics of malware spreading in computer networks. The most significant ones are based on [14], which addresses the epidemic spreading in a homogeneous population of individuals applying mathematical theories. The authors in [15] explain that medium access control mechanism reduces the virus spreading in WSN and they demonstrate that the higher is the node density or the node communication radius the higher is the number of infected nodes. Authors in [16] provide a controlled epidemic model with alert and develop an optimal control problem based on the elaborated model. That is a solution for dynamic containment of malware, at a lower cost. Khayam and Radha in [17] establish a model describing dynamics of worms spreading in WSN that has two categories of nodes: Suscptible or Infected. They defined Suscptible and infected as to be, respectively, a node that can be infected when coming in contact with infectious node and a node which has received a worm payload.

In [18], the authors provide a two-layer model to describe network to network propagation of malware. They establish that the exponential distribution is the most likely distribution in the early stages while power law with a short exponential tail and the power law are, respectively, the more adequate distributions for the late and final stages. They have performed real-world experiments on malware data sets to validate their theoretical findings.

The authors in [19] have studied worms spreading over VANETs. They have considered low-density traffic and the congested situations to build a model based on stochastic infectious disease modeling. The established model allows to investigate worms spreading parameters in the considered context. The same work analyzes scenarios for preemptive and interactive patching.

The work [20] provides an alternative way to deal with contagions in vehicular environments by considering epidemic's blocking as a separate process from the curing one. The adopted approach consists on using distributed infrastructure rather than the fixed one to block the spreading of malicious software. To validate their work, the authors setup a simulation to emphasis the performance of the followed method.

After modeling the mobility, the communication channel, the MAC and the worm propagation, the authors in [21] analyzed the spreading of worms in the case of urban traffic under some assumptions. They utilized Monte Carlo simulation to reveal the impact of transmission range, velocity, vehicles density and MAC on worm spreading in VANETs. The authors, finally, determine the relation of the infection rate with network parameters.

Most of the works mentioned above focus on the microscopic aspects of the security of CAVs or their underlying ad hoc networks facilitating cooperative driving. However, when one considers several CAVs new case studies emerge thus revealing new challenges manifesting themselves just at the macroscopic scale. The impact of road traffic density on the number of CAVs compromised by a malicious CAV forming part of the same traffic is an example which, to the state of our knowledge, has not yet been dealt with in the research literature.

III. Threat Model

A. Threat Agent

The first component to consider in our threat model is the threat agent. In this work, a threat agent is any CAV whose controller has sufficient motivation to attack any other CAVs by exploiting the attack surface offered to him. These motivations can be financial, such as organized crime, industrial competition, or the theft of resources, or ideological, as with terrorism and hacktivism. Also, a threat agent must have all the knowledge and resources necessary to complete successfully his attack. In the simulation section (sec. IV), a black hat CAV (black CAV) represents a threat agent.

B. Attack Surface

Because of its aspects of autonomy and cooperation, a CAV exhibits a fairly large attack surface. We can distinguish the classes of vulnerabilities that can be exploited locally from those that can be exploited remotely. In the latter's context, we can cite, among others, the updating of maps, the communication of the CAV with the road infrastructure (V2I) and the communications taking place between CAVs according to the V2V paradigm. It is the latter that we will consider further. We will be interested in attacks originating from communications according to the V2V paradigm.

C. Attack methods and Attack Potential

Based on the STRIDE framework [22], the attack methods that can be implemented within the framework of the considered attack surface fall under the following four categories: (i) Spoofing, (ii) Tampering, (iii) Denial of service, and (iv) Information disclosure.

Regardless of the attack as specified by STRIDE, an attack always needs a window of opportunity i.e. the time needed to be completed, the financial means, the knowledge regarding the target CAV, the required expertise and the means in terms of equipment. We propose to capture most of these factors by the simulation parameters proposed in TABLE I.

IV. Agent based simulation of V2V penetration rate

The interest of multi-agent modeling lies in its ability to provide models considering different levels of abstraction. The agents taken apart are of a lower level of abstraction compared to that of a system formed by these agents. It describes the causes and rules of behavior at a local level. So instead of describing the dynamics of a system by mathematical equations, the multi-agent paradigm dissects the global system into several agents and makes it possible to deduce the dynamics of the system from the interactions taking place between the different agents and with their environment. We recognize at this level the notion of emergence [23].

In this part of our work, we use NetLogo as a simulator which is a well-known, in the scientific researchers community, as a cross-platform software that allows modeling conforming to the multi-agent paradigm [24]. Turtles are the agents in NetLogo. A script written in a language specific to NetLogo describes the dynamics of the system and the behavior of each turtle. A user can interact with NetLogo using a graphical user interface to configure the simulation parameters of a system with buttons linked to the script and to visualize its evolution using outputs gadgets such as monitors. Two procedures that a NetLogo-based simulation cannot get rid of: the setup and the go procedures. The first procedure concerns the initial parameters of the simulation and the second the evolution of the system by one step.

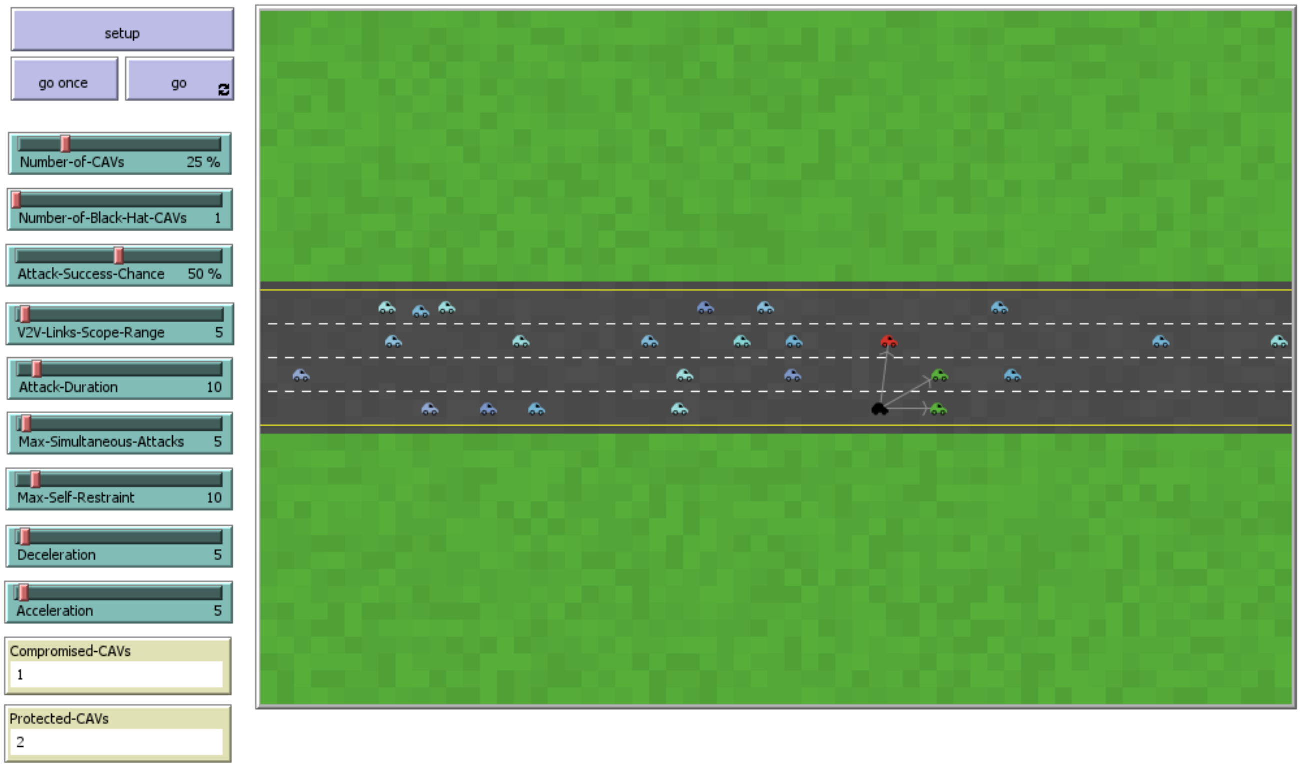
A. Simulation Parameters

In this section, we propose to simulate the impact that a black hat CAV would have on CAV traffic as a function of its density. For this, we consider the following simulation parameters: (i) The duration of the attack which presents the time slots during which the black hat CAV tries to compromise one or more target CAVs. (ii) The maximum number of attacks that a CAV can attempt to carry out simultaneously. This parameter is variable during a simulation run and varies between one attack and the maximum number of this one fixed by the user. (iii) The chance of an attack to succeed. (iv) The maximum self-restraint which reflects the tendency of a CAV to change the lane, this parameter is deduced from the number of brakes that a CAV has performed. (v) The deceleration takes place if one CAV is hampered by another one. (vi) acceleration is a parameter which reflects the increase in speed when the context permits. (vii) The number of CAVs is the number of this one without counting the CAV of the attacker. We assume for this parameter that the section of highway considered is saturated if one hundred CAVs occupies it. (viii) The purpose of the V2V link scope range parameter is to consider the physical limits of the communication channels between two CAVs and the number of these that a CAV wishes to put in place for cooperative driving. (ix) The last parameter specifies the number of CAVs used to perform the attack. We fixed it at one throughout the simulation.

TABLE I
Simulation parameters values

Simulation Parameter	Value
Attack-Duration	10
Max-Simultaneous-Attacks	$up - to$ 5
Attack-Success-Chance	50%
Max-Self-Restraint	50
Deceleration	5
Acceleration	5
Number-of-CAVs	25 $(resp.$ $50; 100)$
V2V-Links-Scope-Range	5
Number-of-Black-Hat-CAVs	1

TABLE I summarizes the values adopted for each parameter during the simulation. Note that the Max-Simultaneous-



Number-of-CAVs 25 %
Number-of-Black-Hat-CAVs 1
Attack-Success-Chance 50 %
V2V-Links-Scope-Range 5
Attack-Duration 10
Max-Simultaneous-Attacks 5
Max-Self-Restraint 10
Deceleration 5
Acceleration 5
Compromised-CAVs 1
Protected-CAVs 2

Fig. 1. The Graphical Interface screenshot of the performed Simulation

Attacks parameter is free to take any value between one and five. In addition, the Number-of-CAVs parameter takes the values 25%, 50% and 100% respectively, which corresponds to the simulation scenarios carried out.

Fig. 1 is a screenshot of the graphical interface provided by NetLogo to allow the user to control the simulation parameters. In the world part of this interface, there is a highway occupied by twenty-five CAVs. CAVs whose colors tend towards blue are CAVs susceptible to be attacked by black CAV. While the red and green CAVs are CAVs already attacked by the latter and which were, respectively, compromised and resistant.

B. Simulation Logic

When the simulation is launched, the black hat CAV tries to attack other CAVs and this according to a logic dictated by the values assigned to the different parameters. Apart from the number of CAVs we do varied here. All other parameters are fixed. For example, the CAV cannot exceed a well-defined range through the parameter *V2V-Links-Scope-Range*. Also, the number of attacks in parallel cannot be higher than that fixed by the parameter *Max-Simultaneous-Attacks*. After interaction with the black CAV, a target CAV is compromised (red CAVs) or protected (Green CAVs). The choice between these two states (red or green) is made randomly with the contribution of a weighting by the value of the *Attack-Success-Chance* parameter.

One CAV is black and the other CAVs are blue is the first step in every simulation run. When we execute the *go* procedure the black CAV tries to compromise the blue CAVs according to the logic mentioned in the previous paragraph. The consumption of a *tick* characterizes each step. In our case an attack attempt can last up to ten ticks after which a target CAV is either compromised or protected. However, it should be noted that once a CAV is subjected to an attack, it is not in the next *ticks* of a simulation run. Finally, a simulation run ends if all the CAVs are visited by the black one or if a certain maximum number of *ticks* is set by the user.

V. Discussion

A. Simulation Results

TABLE II
SIMULATION RESULTS SUMMARY

T. Den. (%)	Rate Mean	Std. Dev.	95% C. I.
25	0.3002	0.1068	[0.2935, 0.3068]
50	0.3516	0.0802	[0.3467, 0.3566]
100	0.3907	0.0498	[0.3876, 0.3938]

To get the results in TABLE II, we ran the simulation a thousand times (1000 runs) for each traffic density. At the end of each simulation run, we measure the rate of CAVs having been compromised. As shown by the figures in TABLE II, the more the traffic density increases, the more

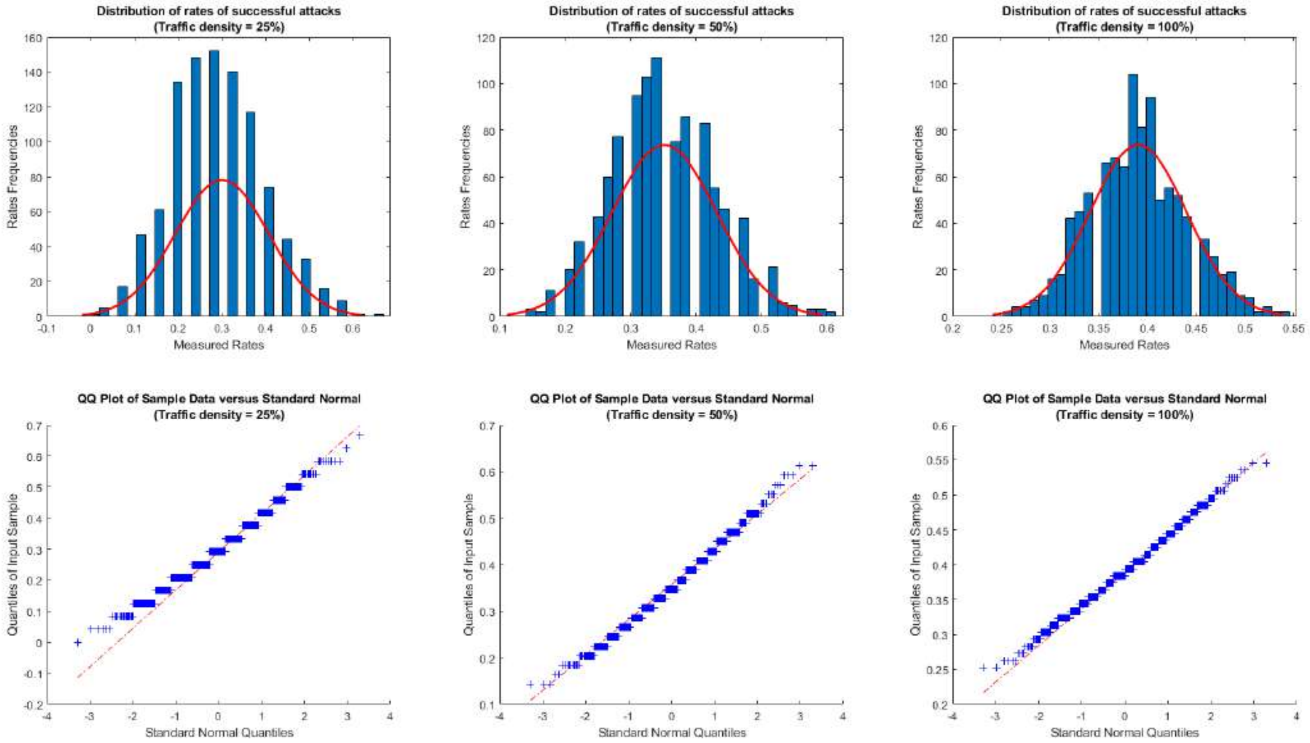


Fig. 2. The graphics showing the obtained results for the three densities

the rate of compromised CAVs increases. Also, we notice that the difference between the different mean of the rate of compromised CAVs tends to shrink despite the increase in traffic density. For example, the deviation of 25% in the range of traffic density less than 50% induces a deviation of the order of 0.05 regarding the sample mean rate, while a deviation of 50% in that greater than 50% only introduces a difference of 0.04 in terms of the sample mean rate. For each calculated sample mean, using the Matlab software, we associate it with a 95% confidence interval specifying the possible values that the true mean can take.

In Fig. 2, we represent the histograms associated with each simulation scenario and their respective Quantile-Quantile plots. We note, from the histograms that the distributions of the rates corresponding to the different scenarios take the form of a Normal centralized on the means specified in TABLE II and having a width dictated by the standard deviations mentioned in the same table (TABLE II).

QQ plots inform us of the degree to which the quantiles of the sample data fit well with that of a Standard Normal. We can see that for a traffic density equal to or greater than 50%, the fitting of the quantiles resulting from the data generated by the simulation with those of a Standard Normal, is more consequent than that for low densities.

If we inspect the Q-Q plots, we can see that for the traffic density of 25%, the small values of the sample data deviate a lot from those corresponding to the theoretical distribution. We can attribute this to the fact that during a simulation, the black hat CAV can always find several CAVs to compromise, greater than the number theoretically predicted by the normal distribution. This is mainly because of the non-uniform nature of the movements of the CAVs. While, sample data values greater than 0.6 are lower than those that enter the normal distribution because of their scarcity in the low-density case.

For the Q-Q plots corresponding to the traffic densities of 50% and 100%, we notice the reduction in the deviations of the sample data from the normal. In fact, more values of the compromised CAV rate are occupied as the traffic density becomes greater and the simulation time has virtually no impact on these values.

B. Lessons to learn

The results got during the simulation highlight the close relationship that exists between the density of CAVs in a road and the rate of successful attacks considering a single attacking CAV. To develop countermeasures, this may require the intervention of several stakeholders who share responsibilities according to the following logic: (i) Road traffic management can play a major role in reducing attacks between CAVs, and therefore limit the spread of malware within the CAV community. In fact, the more the road traffic management system keeps low and regular traffic densities, the lower the rate of attacks or malware propagation. (ii) The incentive to promote public transport and transport sharing are other solutions allowing the reduction of the attack rate because they have a direct impact on the reduction of the traffic density.

(iii) Designers and developers of CAV control software and protocols should know the more communication links a CAV establishes with its counterparts, the greater the risk of the latter being compromised. So establishing the bare minimum of V2V links would compensate for the effect of traffic density.
(iv) The responsibility of CAV manufacturers to make their product difficult to compromise will have a direct impact on the reduction of attacking CAVs and make the target CAVs more resistant.

VI. Conclusion and Further Works

During this work we produced a threat model to specify the study considered here. Then we used NetLogo which is a simulator dedicated to performing agent-based simulations. We have fixed all the simulation parameters except for the density of CAVs in order to study the impact of this on the evolution of the rate of compromised CAVs when the latter are targets of cyberattack originated from other CAV.

The results show that the more dense the traffic in which a CAV is engaged, the more it is susceptible to being attacked by another CAV. However, the results show that the evolution of the rate in question is not necessarily monotonous. In fact, the velocity of change in the rate of successful attacks decreases while approaching 100% density.

In order to reduce the impact of traffic density on the security of target CAVs, we have proposed draft solutions to which several stakeholders such as CAV manufacturers, Governments, standardization Bodies, and researchers can contribute, each according to their perspective. Basically, these draft solutions proposed revolve around regulating the density of traffic and compensating for the effect that the density of CAVs has on the rate of successful attacks by optimizing the number of V2V links required that a CAV must establish with its counterparts.

As future work, we intend to examine other situations by varying other parameters such as those related to the way of driving, with the structure of roads, and those of intelligent traffic management.

References

[1] M. Raya and J.-P. Hubaux, “Securing vehicular ad hoc networks,” *Journal of computer security*, vol. 15, no. 1, pp. 39–68, 2007.

[2] H. Hasrouny, A. E. Samhat, C. Bassil, and A. Laouiti, “Vanet security challenges and solutions: A survey,” *Vehicular Communications*, vol. 7, pp. 7–20, 2017.

[3] J. R. Douceur, “The sybil attack,” in *International workshop on peer-to-peer systems*. Springer, 2002, pp. 251–260.

[4] K. AZGHIOU, M. El MOUHIB, and A. BENALI, “Perspective on the reliability behavior of intelligent transport systems during the transition phase from legacy vehicles to autonomous and connected ones: four-road intersections as a case study,” in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2020, pp. 1–6.

[5] K. Azghiou, M. El Mouhib, M.-A. Koulali, and A. Benali, “An end-to-end reliability framework of the internet of things,” *Sensors*, vol. 20, no. 9, p. 2439, 2020.

[6] A. Lang, J. Dittmann, S. Kiltz, and T. Hoppe, “Future perspectives: The car and its ip-address–a potential safety and security risk assessment,” in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2007, pp. 40–53.

[7] P. Papadimitratos, “?on the road?-reflections on the security of vehicular communication systems,” in *2008 IEEE International Conference on Vehicular Electronics and Safety*. IEEE, 2008, pp. 359–363.

[8] J. Petit and S. E. Shladover, “Potential cyberattacks on automated vehicles,” *IEEE Transactions on Intelligent transportation systems*, vol. 16, no. 2, pp. 546–556, 2014.

[9] J. Cui, L. S. Liew, G. Sabaliauskaite, and F. Zhou, “A review on safety failures, security attacks, and available countermeasures for autonomous vehicles,” *Ad Hoc Networks*, vol. 90, p. 101823, 2019.

[10] S. Tengler. The top twenty unspoken automotive cybersecurity questions and their risks. [Online]. Available: https://www.forbes.com/sites/stevetengler/2020/09/01/the-top-twenty-unspoken-automotive-cybersecurity-questions-and-their-risks/?sh=5c78816c457d

[11] Upstream. Car theft using chinese gadgets in gurugram. [Online]. Available: https://upstream.auto/research/automotive-cybersecurity/?id=4720

[12] ——. Hacking driver assistance systems using depthless objects. [Online]. Available: https://upstream.auto/research/automotive-cybersecurity/?id=4870

[13] S. TEAM, “Security research report on mercedes benz cars,” 360 Sky-Go Security Team, Tech. Rep., 2020.

[14] W. O. Kermack and A. G. McKendrick, “A contribution to the mathematical theory of epidemics,” *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[15] W. Ya-Qi and Y. Xiao-Yuan, “Virus spreading in wireless sensor networks with a medium access control mechanism,” *Chinese Physics B*, vol. 22, no. 4, p. 040206, 2013.

[16] T. Zhang, L.-X. Yang, X. Yang, Y. Wu, and Y. Y. Tang, “Dynamic malware containment under an epidemic model with alert,” *Physica A: Statistical Mechanics and its Applications*, vol. 470, pp. 249–260, 2017.

[17] S. A. Khayam and H. Radha, “A topologically-aware worm propagation model for wireless sensor networks,” in *25th IEEE international conference on distributed computing systems workshops*. IEEE, 2005, pp. 210–216.

[18] S. Yu, G. Gu, A. Barnawi, S. Guo, and I. Stojmenovic, “Malware propagation in large-scale networks,” *IEEE Transactions on Knowledge and data engineering*, vol. 27, no. 1, pp. 170–179, 2014.

[19] S. A. Khayam and H. Radha, “Analyzing the spread of active worms over vanet,” in *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*, 2004, pp. 86–87.

[20] P. Basaras, I. Belikaidis, L. Maglaras, and D. Katsaros, “Blocking epidemic propagation in vehicular networks,” in *2016 12th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*. IEEE, 2016, pp. 1–8.

[21] J. Wang, Y. Liu, and K. Deng, “Modelling and simulating worm propagation in static and dynamic traffic,” *IET Intelligent Transport Systems*, vol. 8, no. 2, pp. 155–163, 2014.

[22] Microsoft. The stride threat model. [Online]. Available: https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)?redirectedfrom=MSDN

[23] J. Fromm, *The emergence of complexity*. Kassel university press Kassel, 2004.

[24] U. Wilensky, “Netlogo (and netlogo user manual),” *Center for connected learning and computer-based modeling, Northwestern University. http://ccl. northwestern. edu/netlogo*, 1999.

Efficient CPW Fed UWB Antenna with Triple Notch Band Characteristics

Srijita Chakraborty[a], Dr. N.N.Pathak[b] ,Dr.Mrinmoy Chakraborty[b]

[a]*Institute of Engineering & Management,Kolkata, W.B., India,* [b]*Dr. B.C. Roy Engineering College, Durgapur*

Abstract: A novel CPW fed UWB antenna with 4GHz/5.48GHz/7.5Ghz triple notch band features has been proposed. The antenna which has a dimension of 35.4mm×28.8mm, is composed of pentagonal radiating stub with CPW framework. The ultra wide band from 3.03GHz to 11.1GHz frequency range has been obtained and by inserting three simple L shaped regular slots at the ground plane and the stub and optimized in order to achieve desired stop band characteristics . The triple frequency band notched characteristics has been achieved at IMT (International Mobile Telecommunications) advanced system 4th generation mobile communication system (3.4-4.2GHz), WLAN (5.15-5.3GHz) and X Band satellite communication(downlink) (7.25-7.75GHz). The antenna shows broad bandwidth, good impedance match and omnidirectional radiation patterns in the entire frequency range.

Keywords: UWB antenna, Notch characteristics, L slot, CPW Fed

I. INTRODUCTION

Ultra-wideband (UWB) technology has received much attention due to some important features such as low cost, low complexity, low spectral power density, high precision ranging and have become the most potential candidate for short-range high speed wireless communication systems. As a key component of the UWB systems, the antennas with ultra-wide bandwidth have been widely investigated by both industry and academia since the Federal Communications Commission (FCC) released 3.1 to 10.6 GHz unlicensed band for radio communication. To meet UWB requirements several coplanar waveguide-fed and microstrip-fed planar antennas have been proposed till date.

Using frequency band rejected characteristics structure, several UWB antennas have been attempted to solve interference problems. Many techniques, such as an isolated slit inside a patch, two open-end slits at the top edge of a T-stub, two parasitic strips, embedded semi-circular annular parasitic strip; and a pair of T-shaped stubs inside an elliptical slot, inclusion of an additional small radiating patch, L-shaped quarter wavelength resonators coupled to the patch have been used to design band notched antennas [1,2,3].

It is difficult to monitor the bandwidth of the notch-bands in a limited space when designing a dual or triple notch band antenna. The key challenges in achieving an effective triple band-notch UWB antenna with balanced VSWR characteristics is strong coupling between the band-notch characteristic designs for adjacent frequencies. [4,5,6,7].

To realize ultra wide band, complementary antenna design technique can be implemented where the patch and ground are complementary to one another with respect to their positions.. It can be observed that ground plane effect is reduced in case of complementary structures and thus ultra wide band is realizable.

In this paper, a novel pentagonal CPW fed antenna has been proposed which consist of a pentagonal patch with microstrip feed and rectangular ground plane. After realization of the ultra wide band , three simple L shaped slots are inserted at the ground plane and patch to notch the application frequency bands viz. IMT (International Mobile Telecommunications) advanced system 4th generation mobile communication system (3.4-4.2GHz), WLAN (5.15-5.3GHz) and X Band satellite communication(downlink) (7.25-7.75GHz).

II.ANTENNA DESIGN AND ANALYSIS

The final optimized design of the proposed antenna was simulated using the electromagnetics simulator tool as shown in Fig. 1. The simulated antenna was printed on the FR4 substrate with thickness of 1.6 mm, relative dielectric constant of 4.4, and loss tangent of 0.002. The radiating patch is fed by microstrip fed line. The simulated antenna is found to exhibit omnidirectional radiation pattern, stable gain and good impedance match throughout the entire bandwidth. Detailed dimensions of the CPW fed UWB antenna are shown in Fig.1.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

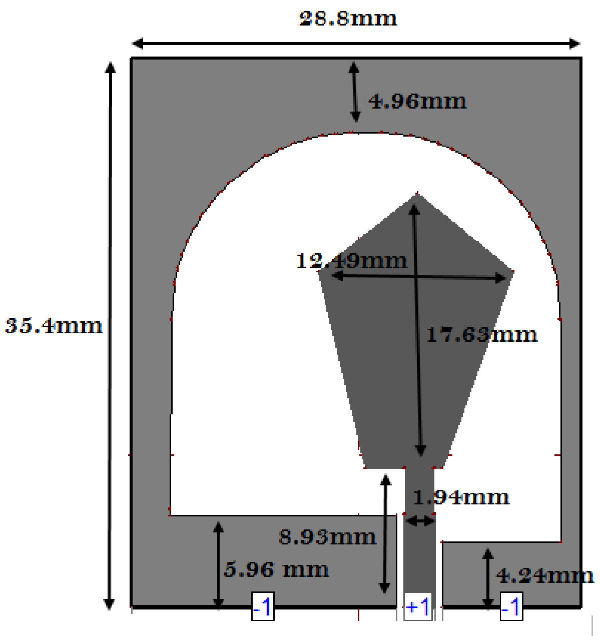


Fig 1. Pentagonal UWB antenna

The antenna is found to give perfect UWB characteristics throughout the entire frequency range.

Capacitive loaded transmission line resonator can be realized using microwave loop resonators. Microwave loop resonators can essentially behave as slow wave structures confining the current density within. Thus slow wave loop resonators can be effectively implemented to notch any desired application frequency bands. It is a general observation that with the increase of the resonator width the frequency notch bandwidth increases and vice versa. The notch characteristics can be implemented by shifting the resonator to the optimum position and modifying the structures.

In this proposed paper we have introduced L shaped slot resonators to notch the frequency bands viz. IMT (International Mobile Telecommunications) advanced system 4th generation mobile communication system (3.4-4.2GHz), WLAN (5.15-5.3GHz) and X Band satellite communication(downlink) (7.25-7.75GHz). Here initially the length of the resonator has been varied to get the notched frequency band and by varying the resonator width the bandwidth is further adjusted. The width of all the slots are 0.2mm. Detailed dimensions of the slots are shown in Fig.2.One of the main problems that is generally encountered in UWB antenna with triple notch characteristics is mutual electromagnetic coupling between the resonators. Electromagnetic coupling is a phenomenon where an electromagnetic field in one resonator results in electrical charge in another. A transferring of electromagnetic properties from one location to another takes place without any physical contact which tend to modify the notch bands. To minimize coupling the resonators must be placed in an optimum position.

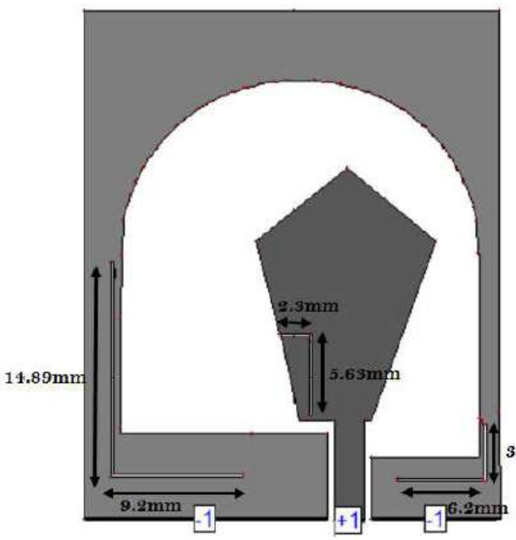


Fig 2: UWB antenna with L shaped slot resonators

III. RESULTS AND DISCUSSION

S11 Versus Frequency Graph

The S11 versus frequency graph of the UWB antenna is given in Fig. 3, where it can be seen that the antenna resonated from 3GHz to 11.2 GHz, covering the entire UWB frequency range. Fig. 4 demonstrates the impedance versus frequency variation of the CPW fed UWB antenna. It is observed that at resonance, the resistive part of impedance of the antenna varies between 40-60 ohms; while the reactive part varies from -10 to +10ohms for optimal performance.

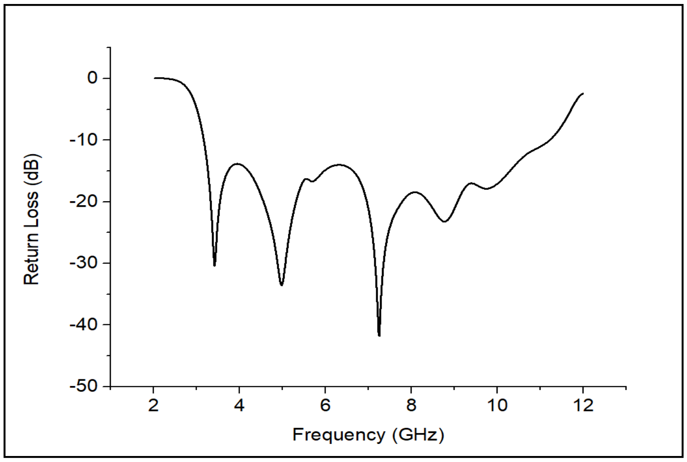


Fig 3: Return loss versus frequency of UWB antenna

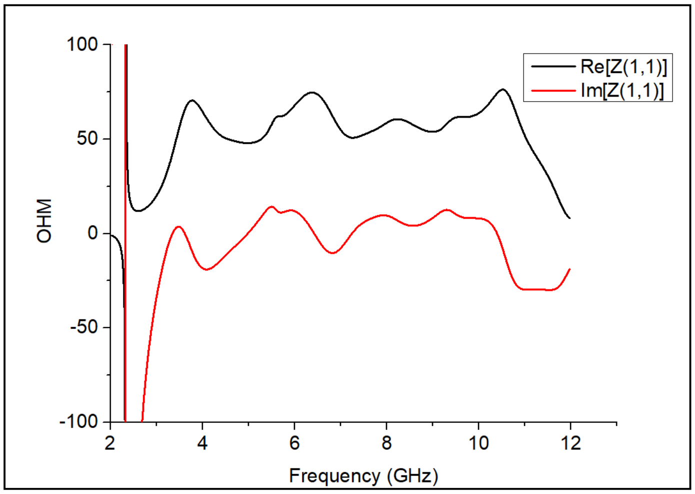


Fig 4: Impedance loss versus frequency of UWB antenna

Therefore at the resonating frequency the UWB microstrip antenna is perfectly matched for optimum performance. Fig. 5 displays the gain versus frequency variation of the CPW fed UWB antenna and it has been seen that throughout the UWB band, the overall gain is more than 3dBi.

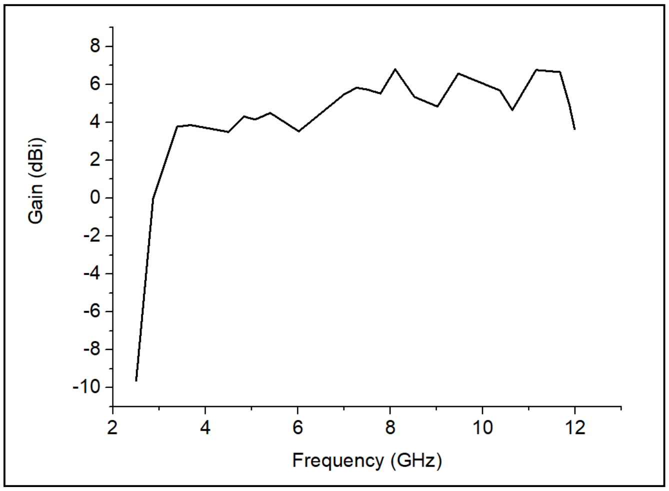


Fig. 5. Gain versus frequency variation of UWB antenna

In the proposed UWB antenna, the first slot is introduced in the ground plane to obtain the first notch (IMT frequency band) with notch bandwidth 3.4GHz to 4.2Ghz and VSWR 5.1 dB. Second slot is introduced at the patch to obtain the second notch (WLAN band) with notch bandwidth 5.15GHz to 5.3Ghz and VSWR 5dB. Third slot is introduced in the ground plane to obtain the third notch (X band communication downlink) with notch bandwidth 7.23GHz to 7.80Ghz and VSWR 5dB.

Extensive study has been conducted on individual L shaped slots by varying the resonator width. It has observed that as the resonator width is increased the notch bandwidth gradually increases confirming the theory of resonators. The variation of resonator width has been plotted in figure. Moreover it has been seen that the slots are placed at an optimum position so as to minimize the coupling effect. Figure 6 and figure 7 demonstrates the return loss and VSWR characteristics of the UWB antenna with triple notch band characteristics. Figure 8 demonstrates the impedance versus frequency curve for pentagonal UWB antenna with three notches.

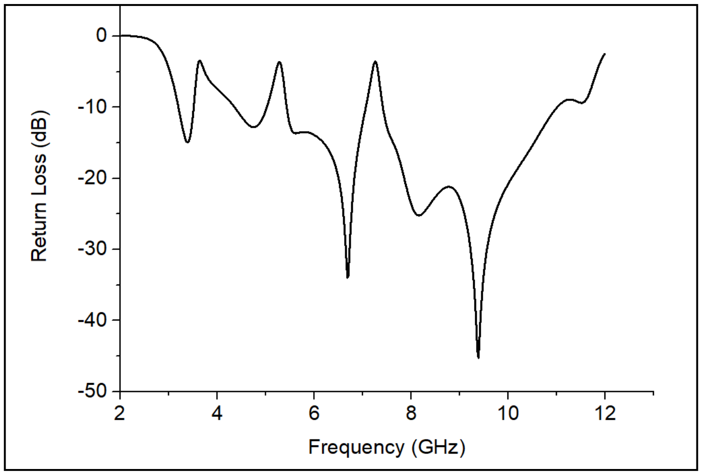


Fig : 6 Return loss versus frequency curve for pentagonal UWB antenna with three notches

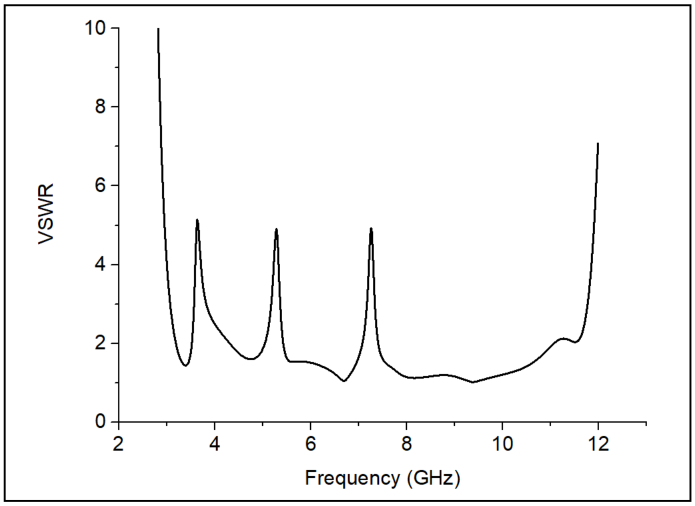


Fig : 7 VSWR versus frequency curve for pentagonal UWB antenna with three notches

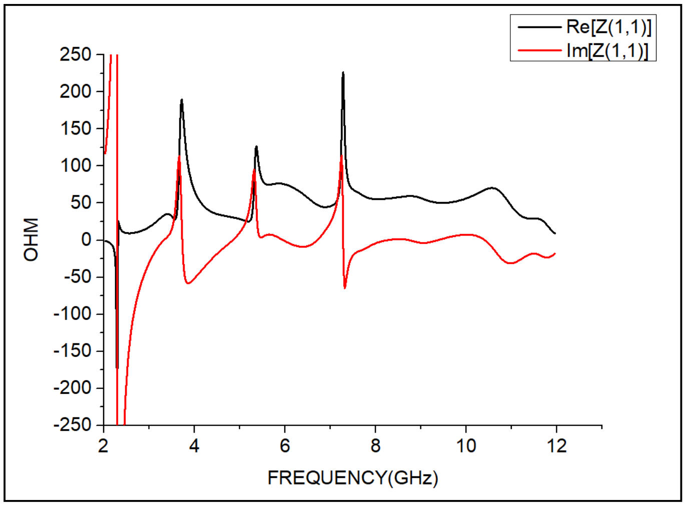


Fig : 8 Impedance versus frequency curve for pentagonal UWB antenna with three notches

IV. CONCLUSION

In this work, a CPW fed pentagonal planar ultra-wideband (UWB) antenna with 4GHz/5.2GHz/7.5GHz triple band notch characteristics has been proposed. Three L shaped slots are placed in to achieve triple notched bands of 3.4-4.2 GHz, 5.15-5.3 GHz, and 7.24-7.80 GHz. It has been seen that changing of parameters of the band notch structure influences only the notched bands, the rest of the UWB frequency band remains unaffected. Thus the UWB antenna with triple notch band characteristic can find immense application in practical UWB systems.

ACKNOWLEDGMENT

The proposed research work is done in Dr. B. C. Roy Engineering College, Durgapur, India and the authors would like to acknowledge the contribution of all concerned.

REFERENCES

[1] Mrinmoy Chakraborty, Biswarup Rana, P.P. Sarkar, Achintya Das,"Design and Analysis of a Compact circularMicrostrip Antenna with slots using Defective Ground Structure", Elsevier Science Direct Procedia Technology 4 (2012) 411 – 416

[2] Srijita Chakraborty, Mrinmoy Chakraborty, Srikanta Pal ,"Design and Analysis of a Compact circularMicrostrip Patch Antenna with Defected Ground Structure", Proceeding of 4th International Conference On Technical And Managerial Innovation in Computing And Communications in Industry And Academia, IEMCON2013,pp. 281-283

[3] S. Bhunia, D. Sarkar, S. Biswas,P. P. Sarkar, B.Gupta,and K. Yasumoto,"Reduced Size Small Dual and Multi-Frequency Microstrip Antennas",MICROWAVE AND OPTICAL TECHNOLOGY LETTERS / Vol. 50, No. 4, April 2008

[4] Srijita Chakraborty, Mrinmoy Chakraborty, Biswarup Rana, N.N.Pathak, Achintya Das "Design and Analysis of a Compact circular Microstrip Antenna with Slots and Defected Ground Structure", Proceedings of National Conference on Electronics and Communication Systems,2013,Section-3, Chapter 2, Page 150

[5] M Chakraborty, S Chakraborty, PS Reddy, S Samanta," High performance DGS integrated compact antenna for 2.4/5.2/5.8 GHz WLAN band," Radioengineering 26 (1), 71-77

[6] Mrinmoy Chakraborty, Srijita Chakraborty, Srikanta Pal," High Performance DGS Based Compact Microstrip Patch Antenna", Proceedings of 1st International Science & Technology Congress 2014

[7] Mrinmoy Chakraborty Srijita Chakraborty, Biswarup Rana, Rajiv Kumar, Monalisa Pal, N.N.Pathak," Design and Analysis of a Compact Circular Microstrip Antenna with Slots and Defected Ground Structure", Proceedings of National Conference on Electronics and Communication Systems, 2013, Section-3, Chapter 2, Page 150

Derivative Based Kalman Filter and its Implementation on Tuning PI Controller for the Van de Vusse Reactor

Atanu Panda
Department of Electronics and Communication Engineering
Institute of Engineering and Management
Kolkata, West Bengal, India
atanu.panda@iemcal.com

Parijat Bhowmick
Department of Electrical and Electronic Engineering
University of Manchester
Manchester, UK
parijat.bhowmick@manchester.ac.uk

Soham Kanti Bishnu
Department of Electronics and Communication Engineering
Institute of Engineering and Management
Kolkata, West Bengal, India
soham.kanti.bishnu@iemcal.com

Sanjay Bhadra
Department of Electrical Engineering
University of Engineering and Management, Kolkata
West Bengal, India
sanjay.bhadra@uem.edu.in

Arijit Ganguly
Department of Electrical Engineering
University of Engineering and Management, Kolkata
West Bengal, India
arijit.ganguly@uem.edu.in

Malay Gangopadhyaya
Department of Electronics and Communication Engineering
Institute of Engineering and Management
Kolkata, West Bengal, India
malay.ganguly@iemcal.com

***Abstract*—This work focusses on predictive PI (PPI) control law employing on a class of stable, nonlinear benchmark process. To facilitate controller parameter(s) updation, two different types of derivative based Kalman filter (KF) strategies like Extended Kalman filter (EKF) and Ensemble Kalman filter (EnKF) techniques were taken into consideration. The servo-regulatory performance of the PPI controller was found satisfactory even in presence of white Gaussian noise. From the extended simulation studies, it can be inferred that EnKF-PPI control logic implemented on the nonlinear dynamical systems are having slightly better performance over EKF-PPI control law. Demonstration and practical utility of the PPI control method in the presence of process uncertainty or process-model mismatch scenario have also been investigated in this work.**

***Keywords*— Van de Vusse Reactor, parameter estimation, PPI, EKF, EnKF.**

I. Introduction

Unitary controls schemes like *P/PI/PID* control laws have drawn significant attention over wide range of control theory and applications due to its simplicity and cost effective nature [1-3]. To facilitate better closed loop performances using *PID* control law, many researchers have put their efforts [9, 11]. For the time-varying processes, offline *PID* controller tuning rule is not sufficient enough to provide adequate performances. To enhance the transient performances online tuning methodologies would be more appropriate for the process whose parameters are uncertain or the case where plant dynamics exhibits under multi-operating regions [13]. Self-tuning control (*STC*), Gain schedule (*GS*) scheme, model reference adaptive controller (*MRAC*) are the commonly used adaptive control schemes used for a class time-varying stable/unstable systems [2, 16]. However, the tuning methodologies for the classical adaptive *PI* (*CA-PI*) control methods are broadly be classified in some major categories like *PID0*, *PID1*, *PID2* and *PID3* [8]. While synthesizing *PI* control rule, it comes under *PID1* or *PID2* category.

As a well-established and widely-used example of online tuning methodologies, predictive *PID* control theories took attention on several realistic applications over last two decades [5]. The *PPI* control strategies were studied in several engineering applications like predictive set point algorithm[4], nonlinear state-feedback based approach [5], robust type of *PPI* control scheme [10], state-space method using pole placement technique [11], event based *PPI* controller [12], *PPI* control rule without using any addition filter [13], fractional order *PPI* [14], using *EKF* [15, 18] and Unscented Kalman filter (*UKF*) [17] methods, Grey-Wolf optimization logic [19] and so on.

This work deals with extension of Pandey et al. (2021) with imposing state and input constraints in order to address desire trajectory tracking and disturbances attenuation problems in the presence of white Gaussian noise for a class of finite dimensional, nonlinear processes. Thus, the aim of this paper exploits designing of derivative based Kalman filter approaches, where the controller parameters were updated by the predictive mechanism. A Typical structural characteristic based benchmark industrial process were presented and analyzed in this theoretical foundation. Hence, the motivation of the developed work extends with employing an efficient, computationally less complex type of control methodologies, which preserves satisfactory set-point tracking performances and would be sufficient enough to attenuate noises and able to cope up with parametric uncertainties. Performance analysis of different *PI* control efforts using derivative based estimators were discussed well. Since, the salient features of the proposed *PPI* control logics are summarized below:

1) Able to preserve desired closed loop performances with complex nonlinear processes by adjusting controller parameters while process moves to the new operating regions.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

2) The methodologies can be implemented on a wide range of linear/nonlinear, stable/unstable processes.
3) Alike other types of *PI* control laws, *PPI* method practices explicit considerations of input/state constraints.
4) The developed techniques promises desired set-point tracking performance or regulatory compliance even in the presence of observational noise.
5) They also facilitate satisfactory robust performances even with model-plant mismatch scenario.

The organization of the research work is mentioned as: Section *I* emphasizes academic literatures and more significantly the motivation of the proposed scheme. Section *II* addresses derivative based Kalman filter strategies used to formulate control framework. Details of the case based example specifying with servo-regulatory compliance, elimination of observational noise, performance comparison and assessing robustness with different control strategies were discussed in section *III*. Section *IV* offers conclusion of this control method.

Basic Preliminaries and Assumptions: Here, $u \in \mathrm{U} \subseteq R^{m}$ represents the manipulated variable(s), $x \in \mathrm{X} \subseteq R^{n}$ denotes the state(s), $pv \in \mathrm{PV} \subseteq R^{r}$ signifies the process output(s), d is the added nonlinear component or external perturbation(s) $(d \in R^{n_d})$. It should be noted that u, x, pv and d are time-varying signals. However, under some necessary conditions and convexity assumptions, the proposed control schemes are established well.

Assumption1: Consider the vector $f : R^{n} \times R^{m} \times R^{n_d}$ is continuous in $u$and d and is locally Lipschitz in x and it follows the condition $f(0,0,0) = 0$.

Assumption2: The vector field $g : R^{r}$ is also continuous in$x$. Predetermined set point $(pv_{sp} \in R^{r})$ and the process output (pv) are assumed either as constant or slow time-varying, bounded and piecewise signal $(pv_{\min} \le pv_{sp}(t) \le pv_{\max}, pv_{\min} \le pv(t) \le pv_{\max})$ for all $t \ge 0$ with having finite number of discontinuities.

Assumption3: Both $u(t)$ and $d(t)$ are bounded and piecewise continuous signals for all $t \ge 0$ with having finite number of discontinuities.

II. Control Schemes

A. Classical adaptive PI control scheme:

The classical adaptive *PI* control rule is expressed by the common equation as mentioned below [3]:

$$u(t) = u(t-1) + (k_c * \text{ƛ}e(t)) + (T_s * (k_c / \tau_i) * e(t)) \quad (1)$$

Here, k_c and τ_i denote the proportional-gain and integral-time-constant of the *CA-PI* control rule respectively. T_s specifies sampling instance. e represents measurement error and can be expressed by:

$$e(t) = y_{sp}(t) - y(t) \quad (2)$$

$$\text{ƛ}e(t) = e(t) - e(t-1) \quad (3)$$

B. Proposed adaptive PI control law using EKF tuning logic:

Firstly, the *PI* controller tuning parameter(s) $(k_c ; \tau_i)$ were updated/estimated by *EKF* technique and we have named this scheme as *EKF-PI* shortly. The basic steps involve with updation/estimation of the controller parameters are given as follows [6,15]:

Initialization of the controller parameters:

$$\hat{\mathbf{\Theta}}_0^a = E[\mathbf{\Theta}_0^a] \quad (4)$$

$$\mathbf{P}_{\Theta_0} = E[(\mathbf{\Theta}_0^a - \hat{\mathbf{\Theta}}_0^a)(\mathbf{\Theta}_0^a - \hat{\mathbf{\Theta}}_0^a)^T] \quad (5)$$

Time update equations:

$$\hat{\mathbf{\Theta}}^f(t) = \hat{\mathbf{\Theta}}^a(t-1) \quad (6)$$

$$\mathbf{P}_{\Theta}^f(t) = \mathbf{P}_{\Theta}(t-1) + \mathbf{Q} \quad (7)$$

Computation of Kalman gain:

$$\mathbf{K}_{EKF,\Theta} = \mathbf{P}_{\Theta}^f(t)\mathbf{J}_{\Theta}^T(\mathbf{J}_{\Theta}\mathbf{P}_{\Theta}^f(t)\mathbf{J}_{\Theta}^T + \mathbf{R})^{-1} \quad (8)$$

$\mathbf{J}_{\Theta}$ denotes Jacobian of the measurement matrix and is described by:

$$\mathbf{J}_{\Theta} = \left.\frac{\partial h(\mathbf{x}(t-1), \mathbf{\Theta})^T}{\partial \Theta}\right|_{\left(\Theta = \hat{\Theta}^f(t)\right)} \quad (9)$$

Hence, the measurement update equations are as follows:

$$\hat{\mathbf{\Theta}}^a(t) = \hat{\mathbf{\Theta}}^f(t) + \mathbf{K}_{EKF,\Theta}(\mathbf{pv} - h(\mathbf{x}(t-1), \hat{\mathbf{\Theta}}^f(t))) \quad (10)$$

$$\mathbf{P}_{\Theta}^a(t) = (\mathbf{I} - \mathbf{K}_{EKF,\Theta}\mathbf{J}_{\Theta})\mathbf{P}_{\Theta}^f(t) \quad (11)$$

C. Proposed adaptive PI control law using EnKF tuning logic

In another approach, $(k_c ; \tau_i)$ were estimated/updated by *EnKF* technique (we have named this algorithm as *EnKF-PI* shortly). Let's assume, $\overline{\mathbf{\Theta}}^f$ and $\overline{\mathbf{\Theta}}^a$ represent forecast and posterior ensemble mean of the estimated parameter(s) $\hat{\mathbf{\Theta}}^f$ and $\hat{\mathbf{\Theta}}^a$ respectively whereas $\mathbf{P}^f$ and $\mathbf{P}^a$ corresponds to the covariance's of the forecast and analysis respectively. Skipping the filter initialization (mentioned in equation: 4-5), time update equations can be written as [21]:

$$\hat{\mathbf{\Theta}}^f(t) = f(\hat{\mathbf{\Theta}}^a(t-1)) + \mathrm{w} \quad (12)$$

$$\mathbf{P}^f H^T \equiv \frac{1}{N-1}\sum_{k=1}^{N}(\hat{\mathbf{\Theta}}^f - \overline{\mathbf{\Theta}^f})(H\hat{\mathbf{\Theta}}^f - \overline{H\mathbf{\Theta}^f})^T \quad (13)$$

$$H\mathbf{P}^f H^T \equiv \frac{1}{N-1}\sum_{t=1}^{N}(H\hat{\mathbf{\Theta}}^f - \overline{H\mathbf{\Theta}^f})(H\hat{\mathbf{\Theta}}^f - \overline{H\mathbf{\Theta}^f})^T \quad (14)$$

$$\mathbf{K}_{EnKF,\Theta} = \mathbf{P}^f H^T (H\mathbf{P}^f H^T + \mathbf{R})^{-1} \quad (15)$$

$$\mathbf{pv}_i(t) = \mathbf{pv}(t) + v \quad (16)$$

$$\hat{\mathbf{\Theta}}^a(t) = \hat{\mathbf{\Theta}}^f(t) + \mathbf{K}_{EnKF,\Theta}(\mathbf{pv}_i(t) - H\hat{\mathbf{\Theta}}^a(t)) \quad (17)$$

We have prepared mean square error (MSE) and control effort (CE) chart for identifying close loop performances of the said controllers [17, 20].

$$MSE = \frac{1}{t}\sum_{i=1}^{t}\left\| pv_{sp,i} - pv_i \right\|^2 ; \text{ for } i = 1 \ldots N \quad (18)$$

$$CE = \frac{1}{t}\sum_{i=1}^{t}\left\| u_i(t) - u_i(t-1) \right\| ; \text{ for } i = 1 \ldots N \quad (19)$$

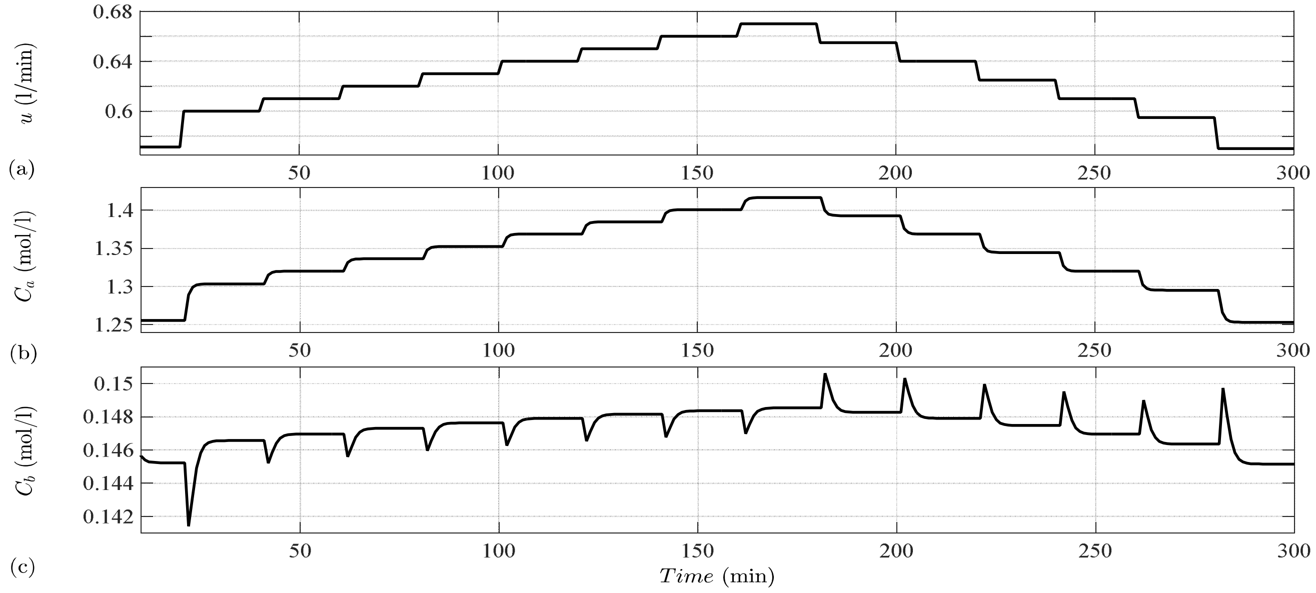


Fig. 1. Open loop response of the Van de Reactor; (a) variation in manipulated variable, *u*; (b) process output, C_b; (c) variation in C_b

III. Process Description and Simulation Studies

The processes taken for the simulation study is Van de Vasse Reactor.

A. Process description:

Let us consider, constituent ‘a’ decomposes into product ‘b’ in an isothermal *CSTR*. Product ‘b’ further converted into product ‘c’ and finally transformed into product ‘d’. The reaction chain can be expressed as [7, 18]:

$$a \longrightarrow b \longrightarrow c$$
$$2a \longrightarrow d$$

The material balance equation of the Van de Vusse Reactor mentioned in [7] has been considered in this work. Concentration of product ‘b’ is the process output. Disturbance was introduces through changes in the concentration of ‘a’ in the inlet flow (C_{af}). Table I mentions the system parameter values for the benchmark

Van de Vusse Reactor. Using 1st principle equation, the process is described by [7]:

$$\frac{dC_a}{dt} = \frac{q}{V}(C_{af} - C_a) - k_1 C_a - k_3 C_a^2 \tag{20}$$

$$\frac{dC_b}{dt} = k_1 C_a - k_2 C_b - \frac{q}{V} C_b \tag{21}$$

All the simulation executes by considering equations [20-21]. Sampling time was taken as 0.1 min. A constrained on the controller output $(10^{-h} < u < 400^{-h})$ were imposed. An equilibrium point of interest $(\bar{C}_a = 3, \bar{C}_b = 2.84, \bar{u} = 55.7)$ [7] was considered for the entire simulation studies.

TABLE I. Standard Values of the System Parameters: Van de Vusse Reactor

Process variable	Normal operating condition
Concentration of A (C_{af})	10 mol/l
Reactor volume (V)	1 l
Reaction rate constant (k_1)	100/h
Reaction rate constant (k_2)	50/h
Reaction rate constant (k_3)	10 l/mol/h
Inlet flow rate (F)	55.7 l/h

B. Open loop study:

In order to assess open loop study with Van de Vusse Reactor, a stepwise changes (combination of positive and negative steps) in the manipulated variable (u) have been introduced (see Fig. 1(a)). It should be noted that manipulated variable (u) can be obtained based on the ratio of inlet flow rate and volume of the Reactor. The evolution of process output (C_b) and another state variable (C_a) are shown in Fig. 1(b-c) respectively.

C. Servo-regulatory Response

In order to identify the tracking capability of all the above mentioned control schemes, below pattern of set-point (variation (as shown in Fig. 2(d)) has been introduced.

$$pv_{sp} = \begin{cases} 3 & \text{for } \ 0 \le t < 25 \\ 3.2 & \text{for } \ 0 \le t < 100 \\ 3.15 & \text{for } t \ge 100 \end{cases} ;\ C_{af} = \begin{cases} 10 & \text{for } \ 0 \le t < 65 \\ 9 & \text{for } \ 65 \le t < 160 \\ 9.8 & \text{for } \ t \ge 160 \end{cases} \tag{22}$$

To measure load rejection capability of the PPI control techniques, sequence of negative and positive stepwise changes in concentration of A with the above magnitude were introduced (see Fig. 2(a)). Table II represents the parameters associated with EKF/EnKF estimation strategies. Fig. 2(d) infers that the PPI controllers are sufficient enough to bring back the process variable at desired value. It was also concluded that the said control schemes performs satisfactory to eliminate load disturbances. The variation in the manipulated variables were portrayed in Fig. 2(e). The evolution of the $(k_c; \tau_i)$ were presented in Fig. 2(b-c).

D. Performance assessment with PPI control approaches

In order to analyse performances of PPI control methods for the servo-regulatory phase (both with and without presence of co-related noise), MSE (see Table III) and CE

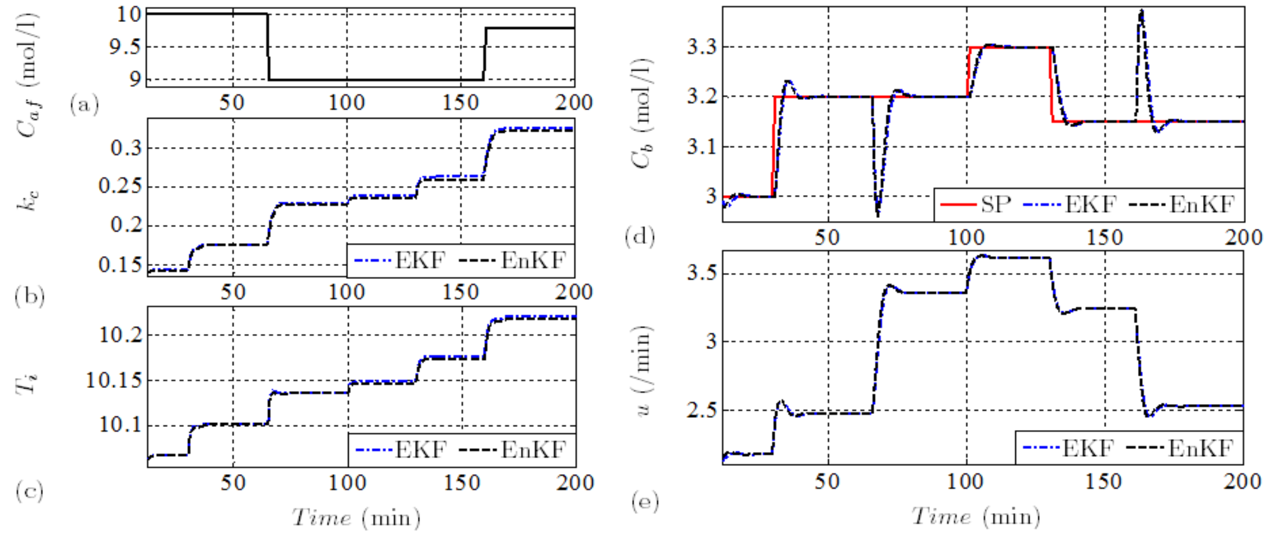


Fig. 2. Servo-regulatory response of Van de Vusse Reactor: (a) variation of C_{af} ; (b) evolution of k_c ;(c) evolution of τ_i (d) process output, C_b ; (e) variation of controller output, u

(see Table IV) chart has been prepared. It should be noted that, servo-regulatory performance chart for the Van de Vusse Reactor was prepared based on Fig. 2(d). However it was observed that EnKF-PI scheme is having slightly better performance (MSE, CE) over EKF-PI control method. It can also be inferred that, the closed loop performances deteriorates in the presence of Gaussian noise (noise to signal ratio (NSR) = 0.01). Toward the end, it was concluded that, both EKF-PI and EnKF-PI control strategies preserves satisfactory set-point tracking performances and able to attenuate load disturbances.

E. Impact of process-model mismatch

In order to identify goodness of PPI control strategies in the presence of observational noise (NSR=0.01), step like changes on different process parameters (C_{af} and k_1) of the following pattern have been considered (reported in Fig. 3(a), 3(c)). From Fig. 3(b), it can be inferred that

$$C_{af} = \begin{cases} 10 & \text{for } 0 \le t < 75 \\ 10.3 & \text{for } t \ge 75 \end{cases}; \; k_1 = \begin{cases} 0.8333 & \text{for } 0 \le t < 125 \\ 0.75 & \text{for } t \ge 125 \end{cases} \tag{23}$$

PPI controller brings the process variable back to the desired value even in case of parametric uncertainty. The evolution of the manipulated variable has been shown in Fig. 3(d). From the result and discussion, it was observed that the PPI control method provides acceptable performances even in the presence of model-process mismatch scenario.

TABLE II. PARAMETERS ASSOCIATED WITH EKF/ENKF ESTIMATION TECHNIQUE

Parameter	Value
Observational noise covariance (R)	$\begin{pmatrix} 0.0015 & 0 \\ 0 & 0.005 \end{pmatrix}$
System noise covariance (Q)	$\begin{pmatrix} (0.035)^2 & 0 \\ 0 & (0.05)^2 \end{pmatrix}$
Initial value of parameter vector $\hat{\boldsymbol{\Theta}}_0^{\mathrm{a}}(0\,\|\,0)$	$\hat{\Theta}_0^{\mathrm{a}}(0\,\|\,0) = [0.1 \quad 10]^{\mathrm{T}}$
Initial value of error covariance $\mathrm{P}(0\,\|\,0)$	$\begin{pmatrix} (0.005)^2 & 0 \\ 0 & (0.05)^2 \end{pmatrix}$

TABLE III. MSE COMPUTATION FOR SERVO-REGULATORY (S-R) COMPLIANCE WITH (NSR = 0.01) AND WITHOUT PRESENCE OF OBSERVATIONAL NOISE FOR PPI CONTROLLER

Control schemes	(S-R) level	(S-R) level (with measurement noise)
EKF-PI	6.2958*e-08	1.8364*e-06
EnKF-PI	6.0749*e-08	1.7942*e-06

TABLE IV. CE COMPUTATION FOR SERVO-REGULATORY (S-R) COMPLIANCE WITH (NSR = 0.01) AND WITHOUT PRESENCE OF OBSERVATIONAL NOISE FOR PPI CONTROLLER

Control schemes	(S-R) level	(S-R) level (with measurement noise)
EKF-PI	9.5107*e-02	7.3842*e-01
EnKF-PI	9.3924*e-02	7.3615*e-01

IV. CONCLUSION

In this work, a successful deployment of EKF and EnKF logic based time-varying PI control approaches were reported. From the realistic simulation, it was observed that servo-regulatory response (both with and without presence of measurement noise) of the PPI control laws with Van de Vusse Reactor promises satisfactory results. In a detail analysis, it was investigated that the PPI control strategy takes less time reach steady-state condition. From the settling time point of view, EnKF-PI shows slightly better over EKF-PI control approach. With introducing parametric uncertainty, it was investigated that the PPI controller facilitate desired closed loop performances even in presence of white Gaussian noise. From the performance comparative charts (MSE and CE), it was inferred that EnKF-PI controller results slightly better over EKF-PI scheme. From the overall analysis, it was concluded that the PI control scheme poised with derivative based predictor approaches preserve acceptable performances. In the future scope, the proposed EKF-PI and EnKF-PI schemes can be utilised to design a set-point tracking controller for nonlinear Negative Imaginary (NI) systems [22, 23, 24, 25, 26, 27 and 28] with actuator saturation. Moreover, this type of stochastic nonlinear model-based control scheme can be extended for Micro/Smart-grid systems [29].

IV. References

[1] D. Rivera, M. Morari, and S. Skogestad, “Internal model control: PID controller design,” Ind. Eng. Chem. Res., vol. 25, no. 1, pp. 252 - 265, 1986.

[2] B. W. Bequette, “Nonlinear control of chemical processes: a review,” Ind. Eng Chem Res., vol.30, no. 7, pp. 13911 - 413, 1991.

[3] Astrom K.J., T. Hagglund, Hung C.C. and Ho W.K., “Automatic tuning and adaptation in. PID controllers - A survey,” IFAC Proc. vol. 25, no. 14, pp. 371 - 376, 1992.

[4] B. Foss and S.O. Wasbo (1993), “Adaptive Predictive PI-Control of an Unknown Plant,” Autom., vol. 30, no. 4, pp. 593 - 598, 1994.

[5] D. Angeli, A. Casavola, and E. Mosca, “Predictive PI-control of linear plants under positional and incremental input saturations,” Autom., vol. 36, no. 10, pp. 1505 - 1516, 2000.

[6] S. Haykin, "Kalman Filtering and Neural Networks". John Wiley & Sons, 2004.

[7] H. Kashiwagi and L. Rong, “Identification of Volterra Kernels of Nonlinear Van de Vusse Reactor,” Trans. Cont. Autom. Syst. Engg. vol. 4, no. 2, pp. 109 - 113, 2002.

[8] R. C. Panda, C.-C. Yu, and H.-P. Huang, “PID tuning rules for SOPDT systems: Review and some new results,” ISA Trans., vol. 43, no. 2, pp. 283 - 295, 2004.

[9] M. W. Foley, N. R. Ramharack, and B. R. Copeland, “Comparison of PI Controller Tuning Methods,” Ind. Eng. Chem. Res., vol. 44, no. 17, pp. 6741 - 6750, 2005.

[10] M. Shamsuzzoha and S. Skogestad, “Robust Predictive PI Controller Tuning,” J. Process Cont., vol. 20, no. 10, pp. 1220 - 1234, 2010.

[11] S. Thomsen, N. Hoffmann and F.W. Fuchs, “PI Control, PI-Based State Space Control, and Model-Based Predictive Control for Drive Systems With Elastically Coupled Loads - A Comparative Study,” IEEE Trans. Indus. Elect., vol. 58, no. 8, pp. 3647 - 3657, 2011.

[12] P. Airikka, “Event-Based Predictive PI Control for Mobile Crushing Plants,” IFAC Proc., vol. 45, no. 28, pp. 66 - 71, 2012.

[13] P. Airikka, “Robust Predictive PI Controller Tuning,” IFAC Proc., vol. 47, no. 3, pp. 9301 - 9306, 2014.

[14] E. Edet and R. Katebi, “On Fractional Predictive PID Controller Design Method,” IFAC Proc. vol. 50, no. 1, pp. 8555 - 8560, 2017.

[15] O. Saleem and U. Omer, “EKF-based self-regulation of an adaptive nonlinear PI speed controller for a DC motor,” Turkish J. Electr. Eng. Comput. Sci., vol. 25, no. 5, pp. 4131 - 4141, 2017.

[16] S. Chandrasekharan, R.C. Panda, B.N. Swaminathan, A. Panda, “Operational control of an integrated drum boiler of a coal fired thermal power plant,” Energy, vol. 159, pp. 977 - 987, 2018.

[17] A. Panda and R. C. Panda, “Adaptive nonlinear model-based control scheme implemented on the nonlinear processes,” Nonlin. Dyn., vol. 91, no. 4, pp. 2735 - 2753, 2018.

[18] S. Bhadra, A. Panda, P. Bhowmick, S. Goswami and R.C. Panda, “Design and application of nonlinear model-based tracking control schemes employing DEKF estimation,” Opt. Cont. Appl. Meth. vol. 40, no. 5, pp. 938 - 960, 2019.

[19] N. Bounasla and S. Barkat, “Optimum Design of Fractional Order PI^{α} Speed Controller for Predictive Direct Torque Control of a Sensorless Five-Phase Permanent Magnet Synchronous Machine (PMSM),” J. Europ. Syst. Autom., vol. 53, no. 4, pp. 437 - 449, 2020.

[20] I. Pandey, A. Panda and P. Bhowmick (2021), “Kalman Filter and its Application on Tuning PI Controller Parameters”, in Proc. of IEEE 11th Annual Comp. Comm. Work. Conf., pp. 1551 - 1556, 2021, NV, USA.

[21] P. Wang, L. Yang, H. Wang, D.M. Tartakovsky and S. Onori, “Temperature estimation from current and voltage measurements in lithium-ion battery systems,” J. Ener. Stor., vol. 34, 102133, 2021.

[22] P. Bhowmick and S. Patra, “On input-output negative-imaginary systems and an output strict negative-imaginary lemma,” In Proceedings of 2nd IEEE Indian Control Conference, pp. 176 - 181, Hyderabad, India, January 2016.

[23] P. Bhowmick and S. Patra, “On LTI output strictly negative-imaginary systems,” Systems & Control Letters, vol. 100, pp. 32 - 42, February 2017.

[24] P. Bhowmick and S. Patra, “An observer-based control scheme using negative-imaginary theory,” Automatica, vol. 81, pp. 196 - 202, July 2017.

[25] P. Bhowmick and S. Patra, “On decentralized integral controllability of stable negative-imaginary systems and some related extensions,” Automatica, vol. 94, pp. 443 - 451, August 2018.

[26] P. Bhowmick and A. Lanzon, “Output strictly negative imaginary systems and its connections to dissipativity theory,” In *Proceedings of 59th IEEE Conference on Decision and Control*, pp. 6754 – 6759, Nice, France, December 2019.

[27] P. Bhowmick and S. Patra, “Solution to negative-imaginary control problem for uncertain LTI systems with multi-objective performance,” Automatica, vol. 112, pp. 108735 (1 - 9), February 2020.

[28] P. Bhowmick and A. Lanzon, “Applying negative imaginary systems theory to non-square systems with polytopic uncertainty,” Automatica, vol. 128, pp. 109570(1 - 18), June 2021.

[29] J. Hu and P. Bhowmick, “A consensus-based robust secondary voltage and frequency control scheme for islanded microgrids,” International Journal of Electrical Power & Energy Systems, vol. 116, pp. 105575(1–11), March 2020.

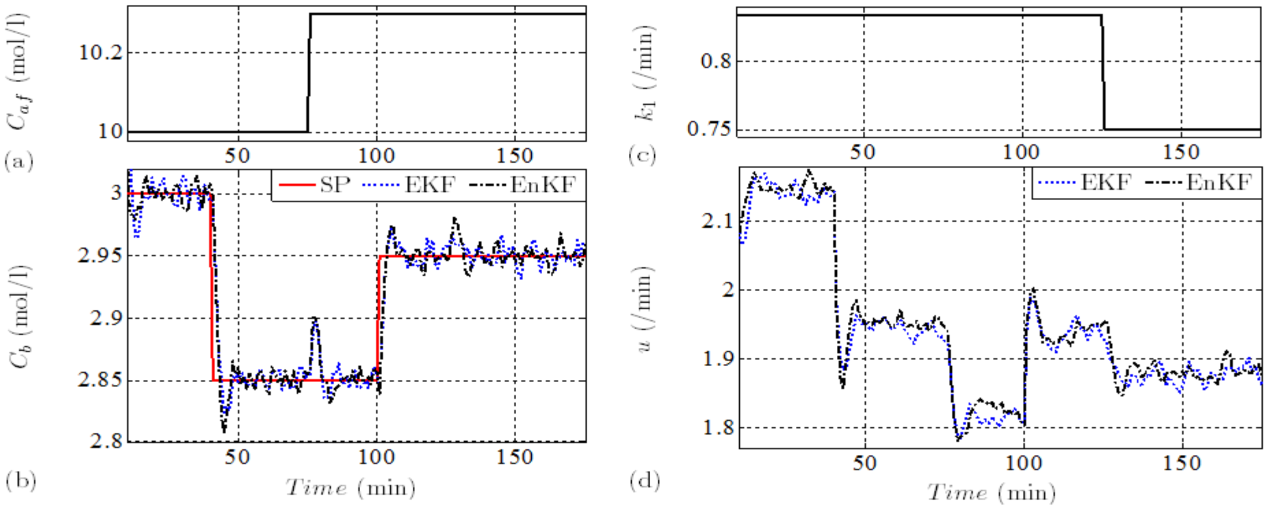


Fig. 3. Robustness assessment with different control scheme: (a) variation of C_{af}; (b) variation of k_1 ; (c) process output, C_b; (d) variation of control output, u

An optimal type-1 servo control mechanism for flight-path-rate-demand lateral missile autopilot

Parijat Bhowmick
Dept. of Electrical & Electronic Engg.
University of Manchester
Manchester, U.K.
parijat.bhowmick@manchester.ac.uk

Atanu Panda
Dept. of Electronics & Communication Engg.
IEM, Kolkata
Kolkata, India
atanu.panda@iemcal.com

Arijit Ganguly
Dept. of Electrical Engg.
UEM, Kolkata
Kolkata, India
arijit.ganguly@uem.edu.in

Sanjay Bhadra
Dept. of Electrical Engg.
UEM, Kolkata
Kolkata, India
sanjay.bhadra@uem.edu.in

Soham Kanti Bishnu
Dept. of Electronics & Communication Engg.
IEM, Kolkata
Kolkata, India
soham.kanti.bishnu@iemcal.com

Malay Ganguly
Dept. of Electronics & Communication Engg.
IEM, Kolkata
Kolkata, India
malay.ganguly@iemcal.com

***Abstract*—This brief proposes an optimal type-1 servo scheme for autopilot design for a class of rear-controlled, flight-path-rate-demand, guided, lateral missiles. The rear-controlled missiles pose significant challenges in the autopilot design due to their non-minimum phase (NMP) characteristic and open-loop instability. The control objective is to design a flight-path rate demand, two-loop, lateral missile autopilot minimising the initial undershoot in the time response (due to the presence of NMP zero). An existing linearised model of a rear-controlled missile has been considered in this paper. The proposed scheme employs a reduced-order observer for generating some of the states of a missile that are difficult to measure in practice. The LQR principle has been used to find the optimal feedback gains based on the given time-domain performance criteria. An exhaustive Matlab simulation study has been carried out to demonstrate the effectiveness of the proposed autopilot scheme in achieving the desired control objectives. The simulation results suggest that the optimal type-1 servo scheme performs better than the conventional frequency-domain design methodologies.**

***Index Terms*—Missile autopilot, non-minimum phase system, reduced-order observer, generalized matrix inverse, type-1 servo control, LQR.**

I. Introduction

Missile autopilot design has always been a critical task since most missiles exhibit non-minimum phase (NMP) behaviour, and they have open-loop unstable dynamics. Moreover, the missiles are highly nonlinear. For these reasons, autopilot design has drawn serious attention from the control and aerospace community. The effectiveness of an autopilot relies heavily upon the modelling of a missile configuration and the accuracy of the sensors. The control scheme for an autopilot needs not only to be effective and reliable but also fast enough. In that respect, a highly advanced and complex controller may offer outstanding performance but at the cost of increased processing time. Hence, a control engineer often seeks to choose a trade-off between the complexity of a control algorithm and its executing time. It has been investigated that for a class of rear-controlled, guided missiles (see Fig. 1[1]), linear gain-scheduled autopilots ensure a fair stability margin and good dynamic performance [1]–[3].

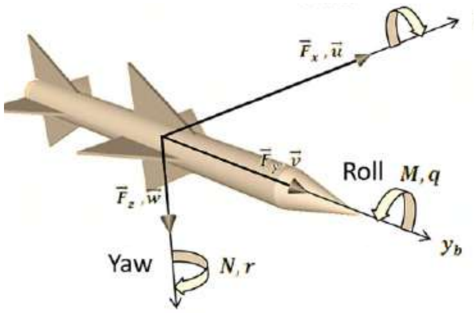

$\overline{F}_x, \vec{u}$
$\overline{F}_y, \vec{v}$
Roll
M, q
y_b
$\overline{F}_z, \vec{w}$
Yaw
N, r

Fig. 1. Coordinate system of a rear-controlled guided missile.

The articles [1]–[3] did pioneering research on developing frequency-domain autopilot design methodologies relying on the linearised missile dynamics. [1] first proposed a two-loop flight-path-rate-demand autopilot (see Fig. 2) for a class of rear-controlled, guided missile. The terminology ‘two-loop’ [1] was given after the fact that the proposed autopilot scheme was comprised of two loops – the outer-loop provides the flight-path-rate feedback and the inner-loop gives the body-rate (pitch rate in this paper) feedback. [1] and [2] adopted a linearised model of a rear-controlled, tactical, homing missile from [12] and developed three different autopilot tuning methodologies utilising the classical frequency-domain techniques. However, the authors later realised that a two-loop autopilot configuration had no direct control over the missile’s body-rate. In the cases of tactical missiles, the body-rate must be controlled in addition to the flight-path-rate to keep the

[1] The image has been taken from internet sources through Google.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

body-rate demand within the prescribed limit and reduce the body acceleration.

To overcome this limitation, the same group of authors introduced a ‘three-loop’ autopilot configuration in [3] by inserting an additional loop (called the middle-loop) that provides rate gyro feedback (refer to Fig. 3). This middle-loop integrates the body-rate error signal and thereby facilitates meeting the body-rate demanded by the outer-loop. It has been observed that the ‘three loop’ autopilot offers a larger DC-gain and a lower high-frequency gain in contrast to the two-loop configuration. These two factors significantly improve the steady-state performance and reduce the initial undershoot in the step response without sacrificing the speed of response [3]. In recent times, [5] has developed an improved version of the classical three-loop lateral missile autopilot scheme proposed in [3] using a frequency-domain approach. [5] has also done a valuable survey on a variety of three-loop autopilot configurations available in the existing literature. Subsequently, in [6], the authors have exploited a PI compensator to eliminate the small steady-state error in the flight-path-rate that appears in the classical two-loop and three-loop autopilots.

Drawn by the advancements and the limitations mentioned above, this paper aims to design a reduced-order observer-based, optimal, type-1 servo scheme for tuning a flight-path-rate-demand lateral autopilot for a class of rear-controlled, tactical missiles. The proposed scheme involves a forward-path PI controller in addition to the state feedback to enforce more substantial control over the flight-path-rate ($\dot{\gamma}$) and the body-rate (q). The LQR technique has been exploited to find the optimal feedback gains, including the integral gain. An iterative algorithm is also provided to choose the weightage matrices $Q > 0$ and $R > 0$ required to run the LQR command. Finally, an exhaustive Matlab simulation study has been done, which underpins the usefulness and effectiveness of the proposed type-1 servo scheme for tuning the flight-path-rate-demand autopilot over the classical two-loop and three-loop autopilots given in [1]–[3].

Notation and symbols

We will now introduce the mathematical symbols used in this paper to denote the missile parameters. $\dot{\gamma}$ and q denote respectively the flight-path rate and the pitch rate of the missile. ω_a is the natural frequency of the actuator and ζ_a denotes its damping ratio. ω_a indicates the weather cock frequency. τ_a is the incidence lag of the airframe. η represents the elevator deflection. σ is a quantity whose inverse determines the frequencies of the NMP zeros associated with the linearised model of the autopilot configuration considered in Fig. 2 and Fig. 3. k_p, k_q and k_b stand for the control gains used in the autopilot schemes shown in Fig. 2 and Fig. 3.

II. Technical Background and Problem Formation

This section provides the backbone for developing the main results of the paper. We will first present the two-loop and three-loop configurations of the classical flight-path-rate-demand autopilot for a class of rear-controlled, tactical, guided missile. After that, we will formulate the research problem addressed in this paper.

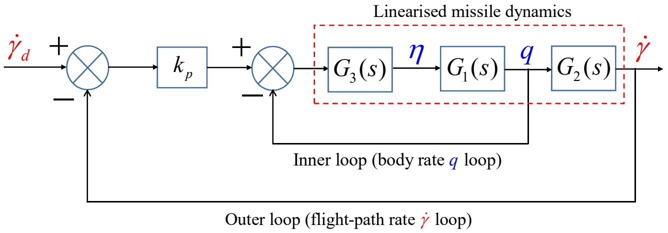

Linearised missile dynamics
$\dot{\gamma}_d$
+
−
k_p
+
−
$G_3(s)$
η
$G_1(s)$
q
$G_2(s)$
$\dot{\gamma}$
Inner loop (body rate q loop)
Outer loop (flight-path rate $\dot{\gamma}$ loop)

Fig. 2. The classical two-loop autopilot configuration [1], [2].

Fig. 2 shows the classical two-loop flight-path-rate-demand autopilot, as developed in [1] and [2]. The linearised missile dynamics consists of three transfer function blocks: $G_1(s) = \dfrac{k_b\omega_b^2(s\tau_a+1)}{s^2+\omega_b^2}$, $G_2(s) = \dfrac{1-\sigma^2 s^2}{\tau_a s+1}$ and $G_3(s) = \dfrac{k_q\omega_a^2}{s^2+2\zeta_a\omega_a s+\omega_a^2}$. The two-loop configuration was later upgraded to a three-loop configuration in [3], known as the three-loop flight-path-rate-demand autopilot, as depicted in Fig. 3. The three-loop scheme significantly strengthens the two-loop scheme since the former provides a direct control over the body-rate (q) via the middle loop.

We now derive a state-space description $\Sigma : \begin{cases} \dot{x} = Ax + Bu, & x_0 = x(0); \\ y = Cx; \end{cases}$ of the open-loop model of the linearised missile dynamics based on the component transfer functions $G_1(s)$, $G_2(s)$ and $G_3(s)$ included in Fig. 2. The A, B and C matrices are given below in (1).

State-space description of the linearised missile dynamics

$$A = \begin{bmatrix} -\dfrac{1}{\tau_a} & \dfrac{(1+\sigma^2\omega_b^2)}{\tau_a} & \dfrac{-k_b\sigma^2\omega_b^2}{\tau_a} & -k_b\sigma^2\omega_b^2 \\ \dfrac{-(1+\omega_b^2\tau_a^2)}{\tau_a(1+\sigma^2\omega_b^2)} & \dfrac{1}{\tau_a} & \dfrac{k_b\omega_b^2\tau_a(1-\frac{\sigma^2}{\tau_a^2})}{(1+\sigma^2\omega_b^2)} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\omega_a^2 & -2\zeta_a\omega_a \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ k_q\omega_a^2 \end{bmatrix}, \quad \text{and} \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{1}$$

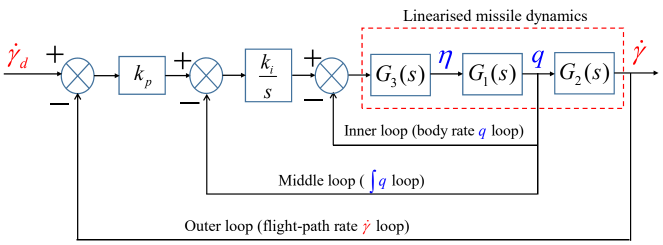

Linearised missile dynamics
$\dot{\gamma}_d$
+
−
k_p
+
−
$\frac{k_i}{s}$
+
−
$G_3(s)$
η
$G_1(s)$
q
$G_2(s)$
$\dot{\gamma}$
Inner loop (body rate q loop)
Middle loop ($\int q$ loop)
Outer loop (flight-path rate $\dot{\gamma}$ loop)

Fig. 3. The classical three-loop autopilot configuration [3].

The state vector $x \in \mathbb{R}^{4\times 1}$ is comprised of four states *viz.* $x_1 = \dot{\gamma}(t)$ denotes the flight-path-rate, $x_2 = q(t)$ represents the pitch-rate, $x_3 = \eta(t)$ indicates the angle of elevator deflection and $x_4 = \dot{\eta}(t)$ signifies the rate of the change of elevator deflection. For simplicity of the presentation, from now onwards, we will omit the explicit dependence of the time t. The states-space model Σ has two outputs $y = \begin{bmatrix} \dot{\gamma} \\ q \end{bmatrix}$ and one input u.

A. Problem statement

Given the linearised dynamics (1) of a rear-controlled, tactical missile, design a reduced-order observer-based, optimal, type-1 servo control scheme such that the autopilot configuration shown in Fig. 4 can i) perfectly track the commanded flight-path-rate $\dot{\gamma}_d$, ii) keep the pitch-rate (q) demand within the prescribed limit, iii) keep the peak undershoot below 5% and iv) reduce the peak overshoot and undershoot of the elevator deflection (η) and the rate of elevator deflection ($\dot{\eta}$).

III. Main Results

This section caters the main contribution of the paper. A reduced-order observer-based, optimal, type-1 servo control scheme (shown in Fig. 4) is proposed for tuning the flight-path-rate-demand lateral missile autopilot taking the inspiration from [8]. The terminology 'type-1' has been chosen to recall its connection with the classical servo control mechanism used for constant input reference tracking for type-0 plants.

The scheme shown in Fig. 4 utilises an output feedback on the basis of $\dot{\gamma}$ in addition to the state feedback and a PI controller in the forward path to enable perfect flight-path-rate tracking. We assume that the flight-path-rate ($\dot{\gamma}$) and the pitch-rate (q) are measures by onboard accelerometer and rate gyro sensors. A reduced-order observer is therefore employed to generate the remaining states $x_3 = \eta$ and $x_4 = \dot{\eta}$. Note that according to the practical guidelines on aerospace system design, we try to reduce the number of sensors as i) it is often difficult to find appropriate and accurate sensors, ii) sensor-measurement is the primary source of high-frequency noise, and iii) the onboard sensors and the associated signal-conditioning units require sufficient space and they increase the overall weight of the missile resulting in a reduced capacity of the payload.

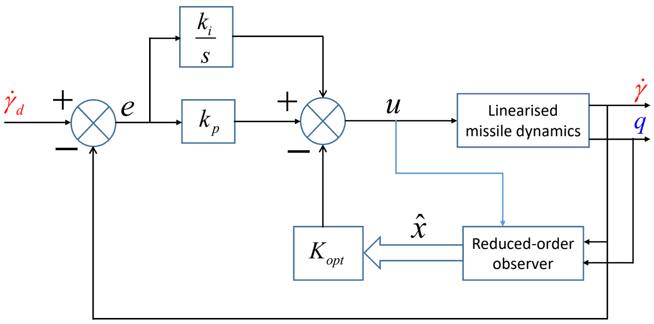


Fig. 4. A reduced-order observer-based, optimal, type-1 servo scheme for a class of flight-path-rate-demand lateral missile autopilot.

In this paper, we have used a Generalised Matrix Inverse (GMI)-based reduced-order observer introduced in [4] instead of the commonly used reduced-order Luenberger observer [10], [11]. Note that although this GMI-based reduced-order observer design methodology [4] is unconventional, it is equivalent to the Luenberger observer and in some cases, the former offers certain advantages over the latter. It was shown in [9] that Luenberger observer works only when the C matrix has the particular form $\begin{bmatrix} I & \vdots & 0 \end{bmatrix}$, otherwise, a coordinate transformation is required to transform the C matrix into the said form. However, the GMI-based reduced-order observer does not have this limitation.

The GMI-based reduced-order observer is governed by the following state-space equations:

$$\begin{cases} \dot{\hat{q}} = \hat{A}\hat{q} + \hat{J}y + \hat{B}u, & \hat{q}_0 = \hat{q}(0) \\ \hat{x} = \hat{C}\hat{q} + \hat{D}y \end{cases} \tag{2}$$

where $\left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right]$ is the given minimal state-space realisation of a plant with $\operatorname{rank}[C] = r \leq m$, $\hat{x} \in \mathbb{R}^{\hat{n}}$, $\hat{n} \leq n$, is the desired the observed state vector and

$$\begin{cases} \hat{A} = \left(L^g AL - MCAL\right), \\ \hat{J} = \left(L^g AC^g - MCAC^g\right) + \left(L^g AL - MCAL\right)M, \\ \hat{B} = L^g B - MCB, \\ \hat{C} = L, \\ \hat{D} = C^g + LM, \end{cases}$$

where the matrix $L \in \mathbb{R}^{n\times k}$ is chosen such that L has full column-rank and satisfies the relationship $LL^g = I_n - C^g C$ via [4] and $M \in \mathbb{R}^{(n-r)\times m}$ plays the role of the observer-gain matrix. The symbols L^g and C^g denote the generalised matrix inverse [13] of the matrices L and C respectively.

Let x_i denote the state of the integrator placed in the forward path of the proposed scheme in Fig. 4. We have from Fig. 4

$$\dot{x}_i = r - \dot{\gamma} = r - x_1 = r - C_1 x \tag{3}$$

where $r = \dot{\gamma}_d$ is the flight-path-rate command and $C_1 =$

$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$. The error signal $e(t) = \dot{x}_i$. Now, we obtain the combined state-space model of the augmented system (i.e. the combination of the reduced-order observer-based state feedback, output feedback and the PI controller) as

$$\begin{bmatrix} \dot{x} \\ \dot{x}_i \\ \dot{\hat{q}} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} A - Bk_pC_1 \\ -BK_{opt}\hat{D}C \end{pmatrix} & Bk_i & -BK_{opt}\hat{C} \\ -C_1 & 0_{1\times 1} & 0_{1\times 2} \\ \begin{pmatrix} \hat{J}C - \hat{B}K_{opt}\hat{D}C \\ -\hat{B}k_pC_1 \end{pmatrix} & \hat{B}k_i & (\hat{A} - \hat{B}K_{opt}\hat{C}) \end{bmatrix} \times \begin{bmatrix} x \\ x_i \\ \hat{q} \end{bmatrix} + \begin{bmatrix} Bk_p \\ 1 \\ \hat{B}k_p \end{bmatrix} r, \tag{4}$$

$$Y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ x_i \\ \hat{q} \end{bmatrix} \quad \text{and} \tag{5}$$

$$\hat{x} = \hat{C}\hat{q} + \hat{D}y = \hat{C}\hat{q} + \hat{D}Cx. \tag{6}$$

The designed control input is given by

$$u = k_i x_i + k_p e - K_{opt}\hat{x} \tag{7}$$

according to the proposed type-1 servo scheme, shown in Fig. 4, where $K_{opt} \in \mathbb{R}^{1\times 4}$ is the optimal feedback gain to be determined by the LQR algorithm. The type-1 servo scheme also facilitates to find k_i and K_{opt} together $[K_{opt} \ \vdots \ -k_i]$ by applying the LQR algorithm on the augmented system model $\left(\begin{bmatrix} A & 0_{4\times 1} \\ -C_1 & 0_{1\times 1} \end{bmatrix}, \begin{bmatrix} B \\ 0_{1\times 1} \end{bmatrix} \right)$ subject to a given choice of $Q = Q^\top > 0$, $Q \in \mathbb{R}^{5\times 5}$, and $R = R^\top > 0$, $R \in \mathbb{R}^{1\times 1}$ (refer to [14]). We have also given an iterative process (Algorithm 1) for finding the LQR weightage matrices Q and R based on the given time-domain performance specifications.

IV. Matlab simulation results: A comparative study among the classical two-loop, three-loop and the proposed type-1 servo scheme

This section presents a comparative study on the simulation responses achieved by the classical two-loop and three-loop autopilot and the proposed type-1 servo scheme. The missile parameter values (tabulated in Table I) have been taken from [3]. Based on this data set, we compute the state-space matrices (A, B, C) of the open-loop model of the linearised missile dynamics (1), as given below:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} -2.77 & 2.8894 & 1.1860 & 0.4269 \\ -50.6161 & 2.77 & -508.388 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -32400 & -216 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -3888 \end{bmatrix} u.$$

Algorithm 1 Procedure to design the control law for TGFT problem involving multiple leaders

```
1: Initialisation: Choose $Q = \text{diag}\{q_{11}, q_{22}, q_{33}, q_{44}, q_{55}\} = \text{diag}\{10, 10, 10, 10, 10\}$, $R = 1$ and $k_p = 2$;
2: Calculate $K_{opt}$ and $k_i$ using the LQR command in Matlab;
3: for $i = 1, 2, \cdots$ do
4:     Take $\dot{\gamma}_d = 1$ and obtain the step response of the scheme shown in Fig. 4 based on the augmented state-space model (4)–(6) taking the data from Table I;
5:     if the maximum undershoot of the $\dot{\gamma}$-response $> 3\%$ of $\dot{\gamma}_d$ or settling time of the $\dot{\gamma}$-response $> 0.25$ s or the percentage peak overshoot of the $q$-response $> 30\%$ then
6:         Choose a different $Q$ from the set of combinations $\big(\{1, 80\} \times \{1, 80\} \times \{1, 50\} \times \{1, 50\} \times \{1, 20\}\big)$ and a $k_p$ from the range $\{1, 10\}$;
7:         Calculate $K_{opt}$ and $k_i$ using the LQR command in Matlab based on the new choice of $Q$ and $k_p$;
8:         back to Step 3 and repeat Steps 4 & 5;
9:     else
10:        Construct the control law $u$ as per (7);
11:    end if
12: end for
```

TABLE I
Model parameters of the missile (taken from [3]).

Parameter	Value
τ_a	0.36 sec
σ^2	0.00029 sec^2
ω_b	11.77 rad/sec
ω_a	180 rad/sec
ζ_a	0.6
ν	470
k_p	5.51
k_b	-10.6272 /sec
k_q	-0.07
k_i	22.2

We then find the state-space matrices of the overall closed-loop scheme (Fig. 4) described via (4)–(6). As mentioned earlier, the proposed scheme has one control input u and two outputs $\dot{\gamma}$ and q. The augmented state vector $X = \begin{bmatrix} x^\top & x_i & \hat{q}^\top \end{bmatrix}^\top \in \mathbb{R}^{7\times 1}$.

Following Algorithm 1, we obtain the LQR weightage matrices Q and R. Based on these Q and R, we find the required set of optimal gain values K_{opt}, k_i and k_p. After feeding all the computed numeric values into the Matlab-Simulink model of the type-1 servo autopilot scheme, we perform a unit-step simulation by choosing the reference command $\dot{\gamma}_d = 1$. The same unit-step simulation is performed for the two-loop and three-loop configurations shown in Fig. 2 and Fig. 3, respectively. Finally, three sets of simulation responses are plotted on the same graph (Fig. 5–Fig. 8), corresponding to each of the four states $\dot{\gamma}(t)$, $q(t)$, $\eta(t)$ and $\dot{\eta}(t)$.

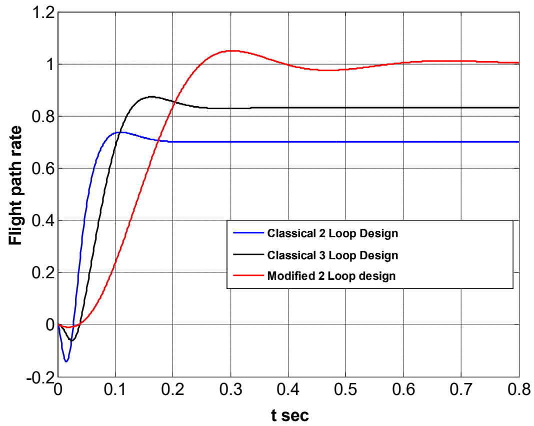


Fig. 5. Comparison among the flight-path-rate responses ($\dot{\gamma}$) achieved by the two-loop, three-loop and type-1 servo schemes.

Fig. 5 shows the comparative plot obtained by the two-loop, three-loop and type-1 servo autopilot schemes, subject to unit-step command (i.e. $\dot{\gamma}_d = 1$). Note that the Red-coloured plots in each of these four figures (designated by 'Modified 2 loop design') indicate the responses achieved by the type-1 servo autopilot. Fig. 5 reveals that the flight-path-rate $\dot{\gamma}$ has entered into the 2% tolerance band with respect to the commanded signal $\dot{\gamma}_d$ within only 0.22 sec and the percentage peak overshoot is less than 5%. Most importantly, the peak undershoot of the $\dot{\gamma}$-response (subject to $\dot{\gamma}_d = 1$) is now only -0.00984 in contrast to -0.1428 and -0.0544 in the cases of the classical two-loop and three-loop schemes, respectively. Although the settling time 0.22 sec is slightly more than 0.17 sec and 0.20 sec achieved by the two-loop and three-loop schemes, respectively, the reduction of the peak undershoot in the case of type-1 servo scheme is much stronger than the classical ones. A similar performance improvement is observed in the remaining three figures: Fig. 6, Fig. 7 and Fig. 8 show the comparative plots of the pitch-rate (q), angle of elevator deflection (η) and the rate of the change of elevator deflection ($\dot{\eta}$), respectively. The performance improvement in terms of the reduction in the peak overshoot and undershoot, increase in the speed of response (decrease in the rise time and peak time), reduction in the settling time, nullification of the steady-state error, etc., are now quantified based on the comparative plots given in Fig. 5, Fig. 6, Fig. 7 and Fig. 8 and tabulated in Table II. The comparative study in Table II recommends a remarkable improvement in both the steady-state and transient performance (in all four states, including the outputs $\dot{\gamma}$ and q) achieved by the proposed type-1 servo scheme over the classical two-loop and three-loop techniques.

V. Conclusions

This paper proposes a reduced-order observer-based, optimal, type-1 servo configuration for tuning a flight-path-rate-demand autopilot for a class of rear-controlled, tactical, guided missiles. The linearised missile configuration exhibits NMP behaviour and has open-loop unstable dynamics, which

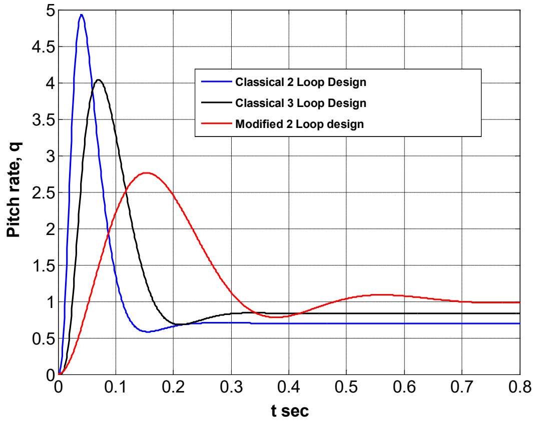


Fig. 6. Comparison among the pitch-rate responses (q) achieved by the two-loop, three-loop and type-1 servo schemes.

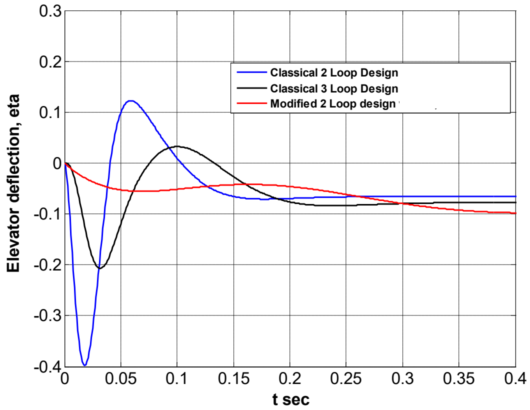


Fig. 7. Comparison among the elevator deflection responses (η) achieved by the two-loop, three-loop and type-1 servo schemes.

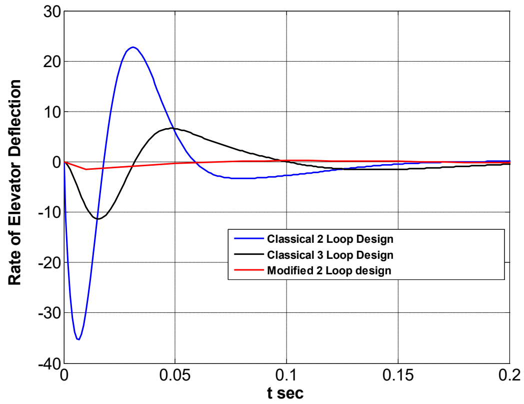


Fig. 8. Comparison among the rates of elevator deflection ($\dot{\eta}$) achieved by the two-loop, three-loop and type-1 servo schemes.

TABLE II
COMPARATIVE STUDY ON THE TIME-DOMAIN PERFORMANCES ACHIEVED BY THE CLASSICAL TWO-LOOP [1], [2], THREE-LOOP [3] AND THE PROPOSED REDUCED-ORDER OBSERVER-BASED, OPTIMAL, TYPE-1 SERVO AUTOPILOT SCHEMES

Serial No.	Parameters	Classical two-loop autopilot (**I**) [1], [2]	Classical three-loop autopilot (**II**) [3]	Type-1 servo scheme (**III**)	Percentage comparison $\left(\frac{\mathbf{I-III}}{\mathbf{I}}\right) \times 100$ %	Percentage comparison $\left(\frac{\mathbf{II-III}}{\mathbf{II}}\right) \times 100$ %
1	Peak undershoot of $\dot{\gamma}$	-0.1428	-0.0544	-0.00984	93.12% improved	82% improved
2	Steady-state value of $\dot{\gamma}$	0.7016	0.8319	1.00	42.53% improved	20.20% improved
3	Settling time of $\dot{\gamma}$	0.17 sec	0.2 sec	0.22 sec	29% increased	10% increased
4	Peak overshoot of q	4.928	4.04	2.761	44% reduced	31.65% reduced
5	Steady-state value of q	0.7015	0.8319	1.00	42.53% improved	20.20% improved
6	Peak undershoot of η	-0.40	-0.2056	-0.05571	86.0% reduced	72.9% reduced
7	Peak overshoot of η	0.1223	0.0322	Nil	significant improvement	significant improvement
8	Steady-state value of η	-0.06601	-0.07828	-0.09425	42% increased	20.40% increased
9	Peak undershoot of $\dot{\eta}$	-35.29	-10.03	-1.573	96% reduced	84.32% reduced
10	Peak overshoot of $\dot{\eta}$	22.74	6.58	0.2238	99% reduced	96.59% reduced

offer non-trivial challenges in the autopilot design. An in-depth Matlab simulation study has been performed, which shows significant improvement in the transient and steady-state performance of the proposed type-1 autopilot scheme over the classical two-loop and three-loop design methodologies [1]–[3]. In the future scope, Negative Imaginary (NI) theory can be applied to develop a robust autopilot design methodology following the ideas given in [15]–[20]. Moreover, cooperative control of multiple missiles can be developed using the scheme proposed in [21].

ACKNOWLEDGEMENT

The first and the fourth authors would like to express their sincere thanks and deepest gratitude to ex-Prof Gourhari Das, Dept. of Electrical Engg., Jadavpur University, Kolkata, India for his mentorship and continuous support.

REFERENCES

[1] G. Das, K. Dutta, T. K. Ghosal and S. K. Goswami, "Structured Design Methodology of Missile Autopilot," *Journal of the Institution of Engineers (India),* pp. 49–59, November 1996.

[2] G. Das, K. Dutta, T. K. Ghosal and S. K. Goswami, "Structured Design Methodology of Missile Autopilot – II," *Journal of the Institution of Engineers (India),* vol. 79 pp. 28–34, November 1998.

[3] G. Das, K. Dutta, T. K. Ghosal and S. K. Goswami, "Structured Linear Design Methodology for Three-Loop Lateral Missile Autopilot," *Journal of the Institution of Engineers (India),* vol. 85, pp. 231–238, February 2005.

[4] G. Das and T. K. Ghosal, "Reduced-order observer construction by generalized matrix inverse," *International Journal of Control,* vol. 33, no. 2, pp. 371–378, 1981.

[5] L. Defu, F. Junfang, Q. Zaikang and M. Yu, "Analysis and improvement of missile three-loop autopilots," *Journal of Systems Engineering and Electronics,* vol. 20, no. 4, pp. 844–851, 2009.

[6] W. Qiu-qiu, X. Qun-li and C. Chun-tao, "Study on autopilot PI compensator optimal parameter design with phase lag constrain," In *Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems,* pp. 865–869, Xiamen, China, October 2010.

[7] T. Çlimen, "A generic approach to missile autopilot design using state-dependent nonlinear control," *IFAC Proceedings Volumes,* vol. 44, no. 1, pp. 9587-9600, 2011.

[8] P. Bhowmick and G. Das, "Modification of classical two loop lateral missile autopilot design using reduced order observer (DGO) and LQR," In *Proceedings of the 3rd IEEE International Conference on Computing Communication and Networking Technologies,* pp. 1–6, Coimbatore, India, July 2012.

[9] P. Bhowmick, "Optimal Design of Inverted Pendulum and Missile Autopilot for Comparison between classical Luenberger Observer and Generalized-Inverse based reduced-order Observer," *Master's Thesis,* Jadavpur University, Kolkata, India, June 2012.

[10] D. G. Luenberger, "An Introduction to Observers," *IEEE Transactions on Automatic Control,* vol. AC-16, no. 6, pp. 596–602, December 1971.

[11] D. G. Luenberger, "Observing the State of a Linear System," *IEEE Transactions on Military Electronics,* vol. 8, no. 2, pp. 74-80, 1964.

[12] P. Garnell and D. J. East, *"Guided Weapon Control Systems"*, 1st edition, Pergamon press, 1977.

[13] F. A. Graybill, *"Introduction to Matrices with Applications in Statistics"*, 1st edition, Wadsworth Publishing Company, Inc., California, 1969.

[14] E. Hendrics, O. Jannerup and P. H. Sorensen, *"Linear Systems Control – Deterministic and stochastic Methods"*, 1st Edition, Springer-Verlag Berlin Heidelberg, 2008.

[15] P. Bhowmick and A. Lanzon, "Applying negative imaginary systems theory to non-square systems with polytopic uncertainty," *Automatica,* vol. 128, pp. 109570(1–18), June 2021.

[16] P. Bhowmick and S. Patra, "Solution to negative-imaginary control problem for uncertain LTI systems with multi-objective performance," *Automatica,* vol. 112, pp. 108735 (1–9), February 2020.

[17] P. Bhowmick and S. Patra, "On decentralized integral controllability of stable negative-imaginary systems and some related extensions," *Automatica,* vol. 94, pp. 443–451, August 2018.

[18] P. Bhowmick and S. Patra, "An observer-based control scheme using negative-imaginary theory," *Automatica,* vol. 81, pp. 196–202, July 2017.

[19] P. Bhowmick and S. Patra, "On LTI output strictly negative-imaginary systems," *Systems & Control Letters,* vol. 100, pp. 32–42, 2017.

[20] P. Bhowmick and S. Patra, "On input-output negative-imaginary systems and an output strict negative-imaginary lemma," In *Proceedings of 2nd IEEE Indian Control Conference,* pp. 176-181, Hyderabad, India, 2016.

[21] J. Hu and P. Bhowmick, "A consensus-based robust secondary voltage and frequency control scheme for islanded microgrids," *International Journal of Electrical Power & Energy Systems,* vol. 116, pp. 105575 (1–11), March 2020.

A Study and Optimization of Different Probe Positions for Different Feeding Techniques using Particle Swarm Optimization

Sutapa Ray, Soham Kanti Bishnu, Agniva Chatterjee, Malay Gangopadhyaya
Electronics and Communication Engineering Department,
Institute of Engineering & Management,
Kolkata,India
agnivachatterjee2014@gmail.com

Abstract— **Being a promising alternative to traditional antennas, liquid antennas have gained high attention in the recent years due to their reconfigurability. As the probe feed position alters the impedance matching and electrical length, the performance like S11, gain and BW will change. This paper tries to illustrates the outcome of different probe positions for different probe types used in the cylindrical liquid antenna which is further optimized using Particle Swarm Optimization. CST Studio suite have been used to simulate the results. 4GHz frequency was maintained throughout the experiment by changing the feed lines and keeping the inner radius of the cylinder, height of the liquid and dielectric constant to excite the different modes of the antenna for multiple applications.**

Keywords— *liquid antenna, probe, particle swarm optimization*

I. INTRODUCTION

This Modern era of wireless communication require an antenna to have multiple functions within a limited space. Multiple attempts have been tried to upgrade the reconfigurability of the metallic antennas in the areas of the radiation pattern, operating frequency and bandwidth .The performance of the low cost radio frequency switches was not good due to presence of metal strip biasing circuits which would be difficult to embed with antennas.[1,3] Again , the high performance RF switches are highly priced. Thus, dielectric resonators were used as radiating elements. The popularity of these antennas increased as they can switch between functions within a single structure without resorting multiple antennas.[2] Since, liquids have high conformability, reconfigurability, improved electromagnetic coupling, thus water having high permittivity is used in the antenna construction. Here, for different Probe types different Probe positions were taken then optimized using Particle Swarm Optimization available in CST.[4] We have stated the effect of changing the probe positions on the working of the dielectric liquid antenna.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

II. DIFFERENT PROBE POSITIONS

Here, in this paper we have used three different probe positions for three different feed types in the liquid antenna , namely Probe feed, Conical feed and Disc feed .

A. Probe Feed

The most common feeding method is the co- axial feeding method.[5] The inner cylindrical conductor of the co-axial connector is adhered to the cylindrical tube of the antenna and the outer part is connected to ground.[6] The radius of the dielectric fluid used here is 6.94084 mm.

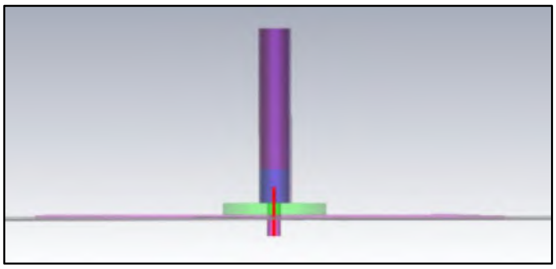


Figure 1: Probe feed

B. Conical Feed

The conical feed is mostly used in reconfigurable water based antenna[7]. The conical feed is projected from the liquid into to cylindrical tube containing the water.[8] The radius of the water used here is 8.45 mm which is the highest among the three[9]. This antenna has a wider aperture which provides greater directivity.

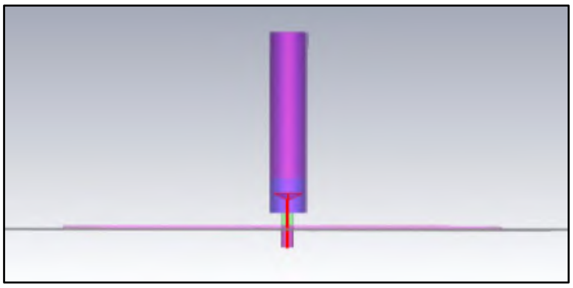


Figure 2: Conical feed

C. Disc Feed

The disc feed is attached to cylindrical tube that contains the water in it.[10] This type of antenna is widely used in satellite communications and also helps to improve the transmittance and signal reception on all the planes. The radius of the water used here is 6.78 mm The circular shape provides higher directivity and a wider aperture[11].

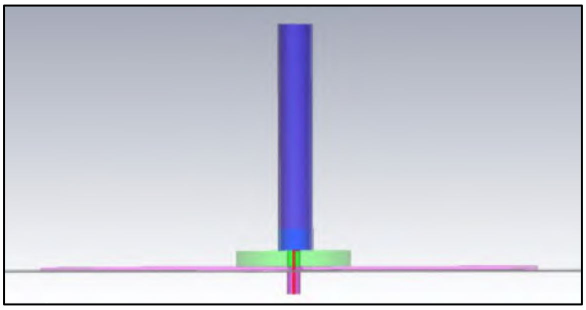


Figure 3: Disc feed

TABLE 1: DIFFERENT PROBE POSITIONS FOR DIFFERENT PROBE TYPES

Type of Feed	Radius of liquid(mm)	X-Position	Y-Position
Probe	6.94084	-0.35	-0.68
Disc	6.78	-0.44	-0.97
Conical	8.45	-0.32	-0.9

III. RESULTS

A. Probe Feed

A bandwidth of approximately 0.9967 GHz is achieved and the value of the S-parameter is approximately -24.04 dB at the operating frequency of 4 GHz. Here, the directivity which is the main lobe magnitude is 15.4 dB (V/m). In the far-field radiation pattern the angular width of 38.8 degrees is observed. The main lobe direction is directed towards 63.0 degrees and the sidelobe level is observed to be -1.1 dB. At 4 GHz the maximum gain over frequency is optimized which is 1.45 as shown in Fig 6.

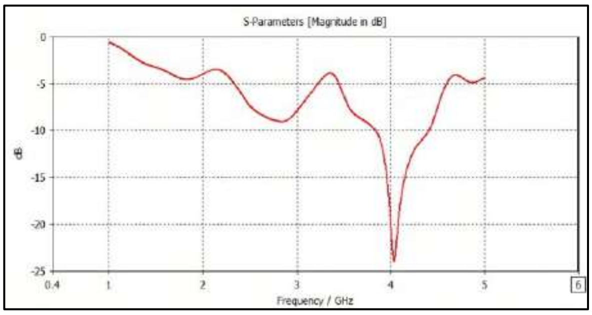


Figure 4: S- Parameter

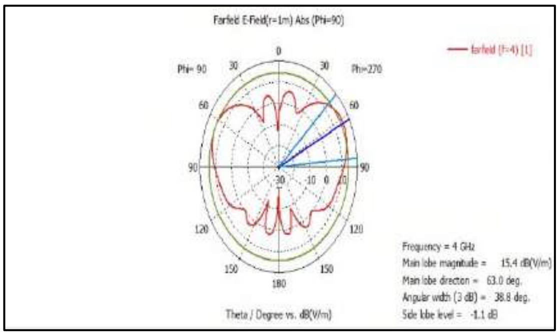


Figure 5: E- Field(4GHz)

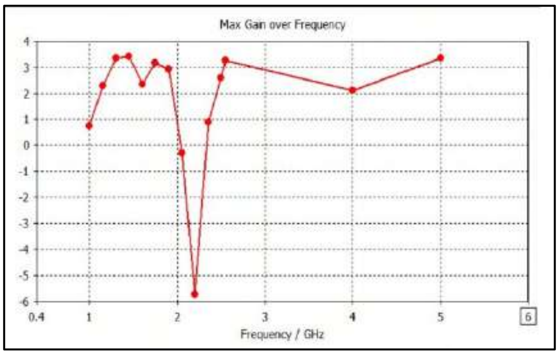


Figure 6: Maximum Gain Over Frequency

B. Conical Feed

A bandwidth of approximately 0.994 GHz, which is the lowest among all the feeds, is achieved and the value of the S-parameter is approximately -14.512 dB at the operating frequency of 4 GHz. Here, the directivity of the radiation pattern is 15.6 dB (V/m). In the far-field radiation pattern the angular width of 39.1 degrees is observed. The main lobe direction is directed towards 58.0 degrees and the sidelobe level is observed to be -6.5 dB. At 4 GHz the maximum gain over frequency is optimized as shown in Fig 9.

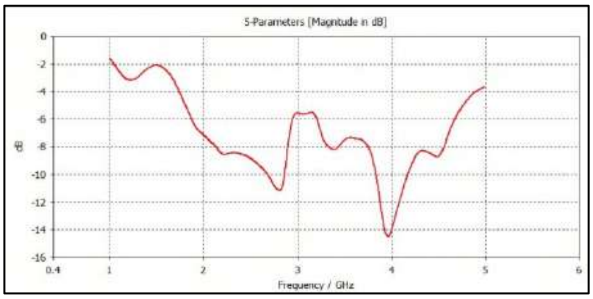


Figure 7: S- Parameter

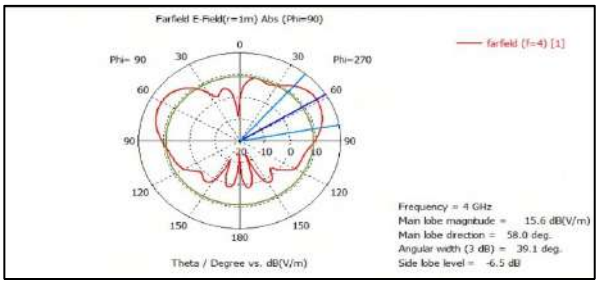


Figure 8: E- Field(4GHz)

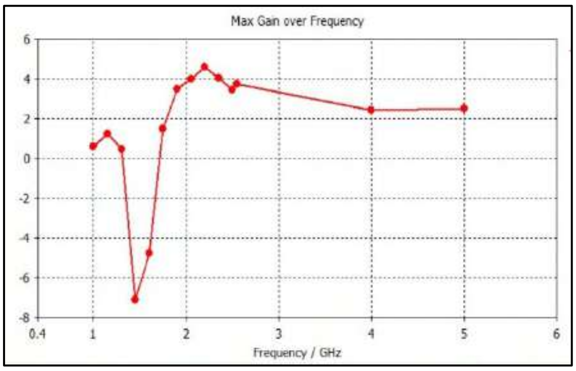


Figure 9: Maximum Gain Over Frequency

C. Disc Feed

A bandwidth of approximately 1.0049 GHz is achieved and the value of the S-parameter is approximately -33.26 dB at the operating frequency of 4 GHz. Here, the directivity which is the main lobe magnitude is 13.7 dB (V/m). In the far-field radiation pattern the angular width of 42.7 degrees is observed. The main lobe direction is directed towards 57.0 degrees and the sidelobe level is observed to be -1.0 dB. At 4 GHz the maximum gain over frequency is optimized as 2.03 shown in Fig 12.

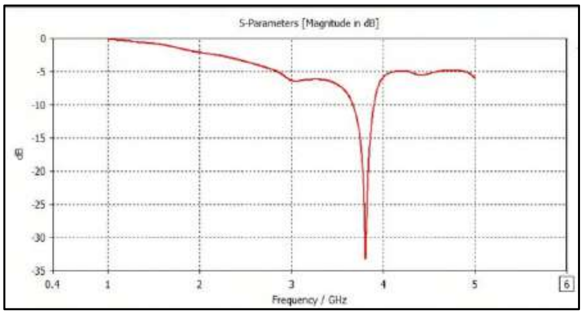


Figure 10: S- Parameter

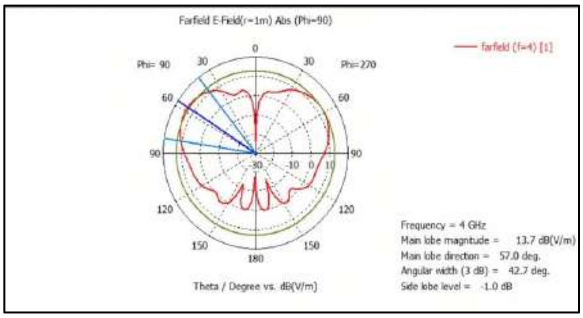


Figure 11: E- Field(4GHz)

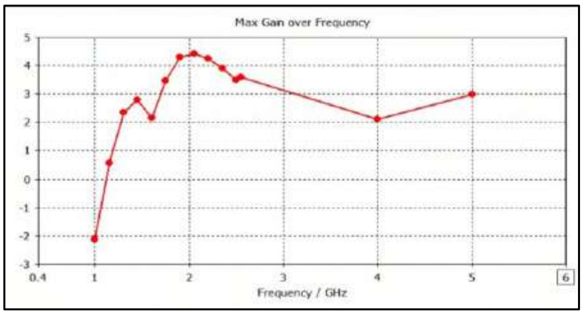


Figure 12: Maximum Gain Over Frequency

IV. COMPARISON OF RESULTS

The simulation results are assembled and compared in the table below: (all the results in comparison table are approximate value)

TABLE I COMPARISON TABLE

	Disc Feed	**Probe Feed**	**Conical Feed**
Bandwidth(GHz)	1.0049	0.9967	0.994
S-parameter(dB)	-33.26	-24.04	-14.512
Directivity(dBi)	13.7	15.4	15.6
Angular Width(deg.)	42.7	38.8	39.1
Side Lobe Level(dB)	-1.0	-1.1	-6.5
Main Lobe Direction(deg.)	57.0	63.0	58
Maximum Gain Over Frequency	2.03	1.45	2.19

V. CONCLUSION

In this paper we have reached the following conclusions from the three different probe positions for three different probe types and compared their bandwidth, S-parameter, directivity, angular width, maximum gain frequency, and sidelobe level. These are mentioned below:

1.The disk feed type has the highest angular width and the probe feed type has the least. Thus , this type of disc feed type antenna can be used for beam forming technique.

2. The conical feed has the highest directivity that is the magnitude of the main lobe is maximum while disc feed has the lowest directivity.

3. The conical feed has the highest gain over frequency and the probe feed has the least.

4. The disc feed has the highest bandwidth among the three. Thus, it can be used in the narrow band applications.

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to Prof. Bhaskar Gupta, Jadavpur University, Kolkata for his continuous support.

REFERNCES

[1] Ray, S., Bishnu, S. K., Chatterjee, A., & Gangopadhyay, M. (2020). Resonant Frequency Optimization of Cylindrical Liquid Antenna Using Particle Swarm Optimization

Algorithm. 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). doi:10.1109/iemtronics51293.2020.9216406

[2] Paracha, K. N., Butt, A. D., Alghamdi, A. S., Babale, S. A., & Soh, P. J. (2019). *Liquid Metal Antennas: Materials, Fabrication and Applications. Sensors, 20(1), 177.* doi:10.3390/s20010177

[3] Motovilova, E., & Huang, S. Y. (2020). A Review on Reconfigurable Liquid Dielectric Antennas. Materials, 13(8), 1863. doi:10.3390/ma13081863

[4] Chen, Z.; Wong, H. Liquid dielectric resonator antenna with circular polarization reconfigurability. IEEE Trans. Antennas Propag. 2017, 66, 444–449

[5] Mondal, K., Pandey, K., Singh, S. K., & Pal, T. (2015). Effect of feeding locations on bandwidth, gain and resonance frequency of the patch antenna. 2015 International Conference on Microwave and Photonics (ICMAP). doi:10.1109/icmap.2015.7408731

[6] Udit Raithatha, S. Sreenath Kashyap," Microstrip Patch Antenna Parameters, Feeding Techniques & Shapes of the Patch – A Survey", International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015,(ISSN 2229-5518)

[7] B.T.P.Madhav, 1J.Chandrasekhar Rao, 1K.Nalini, 2N.Durga Indira ," Analysis of Coaxial Feeding and Strip Line Feeding on the Performance of the Square Patch Antenna",(ISSN:2229-6093 B.T.P.Madhav et al, Int. J. Comp. Tech. Appl., Vol 2 (5), 1352-1356)

[8] Yujian Li and Kwai-Man Luk, ― A water dense dielectric patch antennas‖ IEEE access., Vol 3, pp. 274-280, March 2015.

[9] Borda-Fortuny, C.; Cai, L.; Tong, K.F.; Wong, K.K. Low-Cost 3D-Printed Coupling-Fed Frequency Agile Fluidic Monopole Antenna System. IEEE Access **2019**, 7, 95058–95064

[10] Li, G.; Huang, Y.; Gao, G.; Yang, C.; Lu, Z.; Liu, W. A broadband helical saline water liquid antenna for wearable systems. Int. J. Electron.

[11] Xing, L.; Zhu, J.; Xu, Q.; Yan, D.; Zhao, Y. A Circular Beam-Steering Antenna With Parasitic Water Reflectors. IEEE Antennas Wirel. Propag. Lett. 2019, 18, 2140–2144.

Diseased Surface Assessment of Maize *Cercospora* Leaf Spot Using Hybrid Gaussian Quantum-Behaved Particle Swarm and Recurrent Neural Network

Ronnie Concepcion II
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
ronnie_concepcionii@dlsu.edu.ph

Elmer Dadios
Manufacturing Engineering and Management Department
De La Salle University
Manila, Philippines
elmer.dadios@dlsu.edu.ph

Jonnel Alejandrino
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
jonnel_alejandrino@dlsu.edu.ph

Christan Hail Mendigoria
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
christan_mendigoria@dlsu.edu.ph

Heinrick Aquino
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
heinrick_aquino@dlsu.edu.ph

Oliver John Alajas
Electronics and Communications Engineering Department
De La Salle University
Manila, Philippines
oliver_alajas@dlsu.edu.ph

***Abstract*—*Cercospora zeae-maydis* (CERCZM) is a destructive fungus that is strengthened by hot tropical weather and high humidity such as in the Philippines, resulting to recurrent adverse impacts of having maize *Cercospora* leas spot disease, and quantification of leaf damage is essential for plant phenotyping in understanding pathogen interaction. Visual detection of this disease often results in subjective classification. To address this challenge, the integration of computer vision and computational intelligence is employed in detecting healthy and damaged corn leaves and predicting the surface damage percentage due to maize *Cercospora* leaf spot. Dataset with 583 images contains matured healthy and diseased corn leaves that were grown outdoor and individually captured by a digital camera. Graph-cut segmentation through lazysnapping segmented the vegetation pixels and CIELab thresholding segmented healthy and diseased regions. Spectro-textural-morphological leaf signatures were extracted and selected using combined neighborhood component analysis and ReliefF resulting in R, H, *a, Cb, Cr, entropy, and whole leaf area. MobileNetV2 exhibited the best performance in classifying maize leaf health status. Gaussian quantum-behaved particle swarm optimized recurrent neural network (GQPSO-RNN) bested other feature-based machine learning and deep transfer image networks in predicting maize *Cercospora* leaf spot surface damage percentage with R^2 of 0.949, RMSE of 6.290, and inference time of 3 seconds. This developed seamless MobileNetV2-GQPSO-RNN model provides reliable disease detection and quantitative assessment on the maize leaf surface in on-field phenotyping.**

Keywords*—*computational intelligence, computer vision, corn leaf disease, model optimization, particle swarm optimization

I. Introduction

Crops are susceptible to pathogens that may cause outbreaks if left unchecked. To ensure that plant life is preserved, inspecting common parts of plants, such as leaves, for the manifestation of diseases should be considered. Maize or corn, otherwise known as *Zea mays L.*, is one of the most utilized crops in the animal and food industries in the Philippines [1]. It has become an essential food in several countries, surpassing the demand for wheat and rice. Leaf ailments are a hindrance to corn production. Determining leaf ailments using human judgment is not an efficient method in monitoring crops [2-3].

Similarities such as manifestations of spots that mature into yellowish lesions and can spread through the leaves of corn crops may branch out to surrounding plants, increasing its infectivity. These symptoms are caused by the *Cercospora* leaf disease (CLD). There are different variants of CLD such as bacterial leaf growth, northern corn leaf blight (NCLB), southern corn leaf blight (SCLB), and gray leaf spot (GLS) [4]. In general, *Cercospora zeae-maydis* (CERCZM) is characterized by gray leaf spot, reduced fruit size, necrotic areas in leaves, and early senescence of the whole plant. All of these variants are commonly caused by *Cercospora beticola* and can reduce a plant's vigor, making it defoliated [5].

Convolutional neural network (CNN) is capable of diagnosing early symptoms of corn crop diseases in a real-time and non-destructive manner [6-8]. Employing genetic algorithm support vector machine (GA-SVM) distinguishes the difference of each disease from one another [2]. Multifractal detrended fluctuation analysis (MF-DFA) utilized the gray variance values of blemishes seen on crop leaves with CLD [9]. Stripping the layers of crop leaves [10] and applying trifloxystrobin and epoxiconazole mixtures reduced the damage done by CLD and provided high crop yield [11]. Crop rotation, tillage residue of CLD on equipment, stressful environments, and use of chemicals are factors that should be considered to manage CLD properly [12]. Mapping genetic traces of CLD can aid in identifying early signs of the disease, preventing an outbreak from occurring [13]. Screening of plant phenotype is important to avoid abnormalities in growing crops, and thus, increasing total harvest yield [14]. Quantitative trait locus (QTL) mapping can be done to identify genetic characteristics of *Cercospora* in a different population of corn crops [4]. Monitoring plant life using ultrasonic sensing sprayers provide the optimum height and placement of sprayers to protect corn crops and increase total harvest yield [15]. Complex leaf shapes and varying colors by using integrating local threshold and seeded region growing (LTSRG) [16]. Information about leaf color can be mapped by converting

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

RGB to Cb-Cr space [17]. Computer vision (CV) technology has been widely used in classifying and identifying crop diseases through histogram-oriented gradient (HOG), segmented fractal texture analysis (SFTA), and local ternary patterns (LTP) [18-19]. Discolorations and spots are then identified to determine the presence of *Cercospora* infection [20].

Despite the abovementioned studies with technological advancements, the severity of damage due to corn leaf disease has not been explored and modeled yet. Subjective assessment of damage percentage is also evidently using human bare eyes which is not ideal for scientific phenotyping. Likewise, most plant imaging devices are bulky and placed inside a laboratory which requires transporting of leaf samples from the field that adds up to the delay in real-time phenotyping.

In this study, corn leaf quality was classified into healthy and defective in terms of the presence of *Cercospora* leaf spots, and the defective leaves have undergone surface damage computational assessment using computational intelligence (CI) models by calculating the damaged percentage in terms of the leaf area. Quantum particle swarm optimization (GQPSO) was used to improve the model performance of the recurrent neural network (RNN). GQPSO was chosen and explored because it enables both linear and nonlinear and bounded and unbounded problems to generate the best solution possible. Likewise, artificial bee colony (ABC) and genetic algorithm (GA) were also explored but resulted into convergence issue. The developed model is a novel approach to assessing corn *Cercospora* leaf spot.

II. Materials and Methods

Corn *cercospora* leaf spot is quantitatively evaluated based on its damage on the visible surface of the leaf structure. The developmental architecture of detecting corn leaf quality and assessing the percentage of the diseased surface using deep transfer image network and feature-based machine learning models is shown in Fig. 1. The model employed in this study emphasizes the degree of impact of corn leaf disease by predicting the leaf spot surface defect percentage. Images that are classified with healthy maize leaf results in 0% diseased surface percentage. MATLAB R2020a is the only computational intelligence programming platform used in computer vision, image feature extraction, leaf signatures selection, model optimization, and development of prediction models.

A. Cultivar Data Description and Environmental Condition

Zea mays convar. Saccharate var. rugosa, also known as sweet corn, was cultivated in open field agriculture in Bukidnon, Philippines from September to November 2019 through a monoculture approach. Environment temperature during cultivation ranges from 23 – 33°C. Matured 90-day old health and diseased corn leaves that are infected by *Cercospora zeae-maydis* characterized by rough gray spots were captured using a consumer-grade digital camera with an aspect ratio of 1:1 having an image spatial resolution of 256 x 256 pixels. A total of 583 leaf images was analyzed coming from 283 diseased and 300 healthy leaves. The used dataset is based on [21].

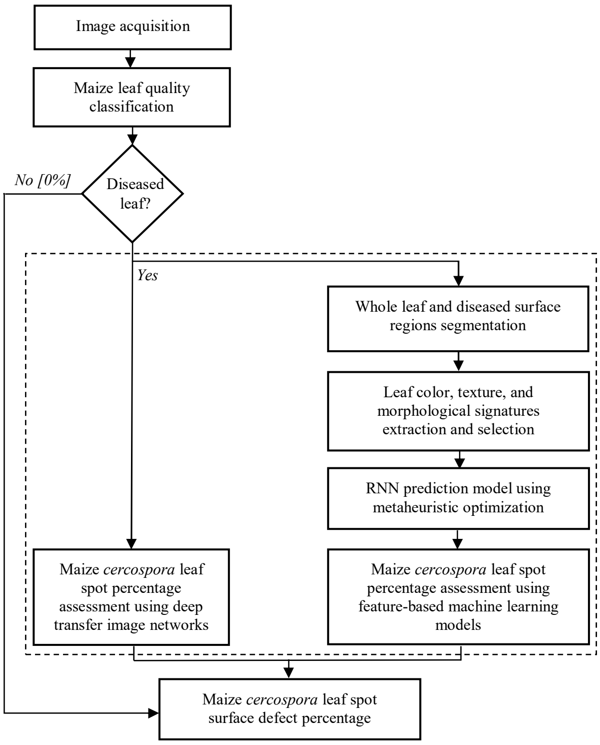


Image acquisition
Maize leaf quality classification
Diseased leaf?
No [0%]
Yes
Whole leaf and diseased surface regions segmentation
Leaf color, texture, and morphological signatures extraction and selection
RNN prediction model using metaheuristic optimization
Maize *cercospora* leaf spot percentage assessment using deep transfer image networks
Maize *cercospora* leaf spot percentage assessment using feature-based machine learning models
Maize *cercospora* leaf spot surface defect percentage

Fig. 1. The developmental architecture of detecting maize leaf quality and assessing the percentage of the diseased surface using deep transfer image network and feature-based machine learning models

B. Leaf Segmentation, Spectro-Textural-Morphological Feature Extraction and Selection

Classified diseased maize leaf images were furtherly assessed by undergoing graph-cut segmentation of the whole leaf and diseased region to compute the damaged percentage concerning the captured leaf area (Fig. 2). Graph-cut segmentation is based on lazysnapping algorithm (Fig. 2b) that transforms RGB image to CIELab color space where L ranges from 0 to 100, a* is from -86.1827 to 98.2343, and b* is from -107.8602 to 94.4780. It employs marking of foreground and background pixels which is, in this case, are vegetative and non-vegetative elements, respectively. It categorizes a pixel based on K-means clustering characterization on edge signatures of the neighboring pixels with a superpixels value of 421.

Both whole leaf and diseased region segmented images require feature extraction for the development of a computational assessment model. Morphological area is extracted from whole leaf image ($A_{wholeleaf}$) which is then used in (1) in yielding the diseased or damaged surface percentage (DSP). For the diseased regions, spectral (R, G, B, H, S, V, L, a*, b*, Y, Cb, Cr), textural (contrast, correlation, energy, entropy, homogeneity) [22], and morphological area parameters were extracted. The morphological area of the diseased region ($A_{diseasedregion}$) is then divided by $A_{wholeleaf}$ to get the ratio of diseased and whole leaf region (1). This 18-feature vector is purposively reduced to several highly significant feature vectors based on neighborhood component analysis (NCA) and ReliefF. NCA is a multidimensional reduction based on Mahalanobis distance while ReliefF ranks the 18 features using an intelligent Relief algorithm. Combined characterization of NCA and ReliefF resulted in a 7-feature vector, [R, H, a*, Cb, Cr, entropy, whole-leaf area], which results in significant relation in predicting damaged surface percentage (Fig. 3).

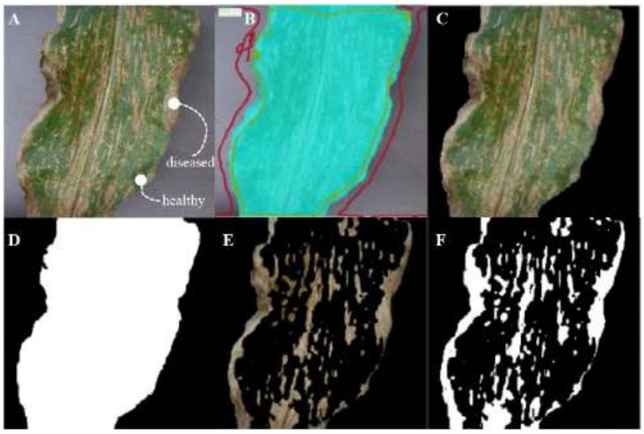


Fig. 2. Process of graph-cut segmentation of whole and diseased leaf regions: (a) raw leaf image indicatng healthy and diseased regions, (b) lazysnapping masking, (c) annotated whole leaf image, (d) whole leaf binary image, (e) diseased regions, and (f) binary diseased regions

$$DSP = (A_{diseasedregion} / A_{wholeleaf}) \text{ x } 100 \qquad (1)$$

C. Feature-Based Machine Learning Regression for Diseased Surface Percentage Prediction Modeling

Diseased leaf region percentage affected by *Cercospora zeae-maydis* is predicted using generalized processing regression (GPR), recurrent neural network (RNN), regression tree (RTree), and regression support vector machine (RSVM). GPR is optimized using constant basis function with a beta of 854.1724 and sigma value of 5.3574, squared exponential kernel function, exact predict method, exact fit method, and rando active set method. Elman RNN is optimized using QGPSO and is discussed in the next section. RTree is optimized using a minimum leaf size of 6. RSVM is optimized using box constraint of 0.21536, 0.035471 kernel scale, 0.31729 epsilon value, -3.0948x10^3 bias, and sequential minimal optimization (SMO) as solver. Hyperparameters of GPR, RTRee, and RSVM were enhanced using Bayesian optimization. Stratified sampling was employed in distributing images with 56%-24%-20% partitions for training, validation, and testing image set.

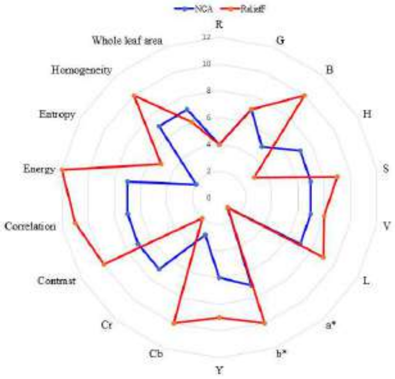


Fig. 3. Weight spectrum of spectro-textural-morphological leaf signatures as characterized by neighborhood component analysis and ReliefF

D. Parameter Optimization Using Gaussian Quantum-Behaved Particle Swarm

Gaussian quantum-behaved particle swarm optimization (GQPSO) is a quantum-based PSO with Gaussian probability distribution acting on its mutation process [23-25]. This Gaussian distribution improves the divergence of this bioinspired algorithm and is an advanced approach in mitigating local optima premature convergence. Instead of position and velocity in classical PSO, in the quantum model of GQPSO, each involved particle state is depicted by wave function *ψ(x, t)*, where *x* is the particle position at time *t*, of the standard Schrödinger equation [23]. In this study, GQPSO is performed using these chronological steps: initialization of swarm positions using Gaussian probability distribution, particle fitness evaluation, updating of the global best particle, comparing particle fitness to current best value, comparing current best particle to global best, updating particle positions using Gaussian distribution, and repeating the cyclic evolution. The algorithm stops only if the maximum number of generations is exhausted. Other optimization algorithms such as artificial bee colony (ABC) and genetic algorithm (GA) were also explored and resulted in divergence issue, making GQPSO the algorithm fit for this application.

The topology of the RNN fitness function based on the linear regression of its corresponding hidden artificial neurons and mean square error (MSE) network performance is shown in (2). The three hidden layers of RNN, N_1, N_2, and N_3, are the variables that will be generated using GQPSO. Regression coefficients α, β, γ, and δ have values of 1.432x10^{-3}, 1x10^{-7}, 1.3x$10^{-6,}$ and 1x10^{-6}, respectively.

$$MSE = f\,(N_1, N_2, N_3) = \alpha - \beta N_1 - \gamma N_2 + \delta N_3 \qquad (2)$$

In this study, the potential function configured to Delta Potential Well with 3-dimensional space, 250 swarm particles, 0 to 500 boundaries, and a non-negative constant of 0.9. GQPSO finds the global optimum of (2) at N_1, N_2, and N_3 values of 140, 100, and 30 with a smoothing fitness curve up to 100 iterations (Fig. 4).

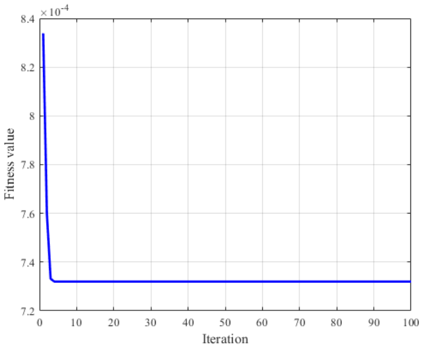


Fig. 4. Optimization curve using Gaussian quantum-behaved particle swarm optimization characterized by fitness value and iteration performance

E. Deep Transfer Image Regression for Diseased Surface Percentage Prediction Modeling

Deep transfer regression networks of MobileNetV2, ResNet101, and InceptionV3 were configured using SGDM with a minimum batch size of 20, maximum epochs of 7, 0.0001 initial learn rate, piecewise learn rate schedule, 0.1 learn rate drop factor, 20 learn rate drop period, 'every-epoch' shuffling method, and validation frequency of 3. Raw leaf images were utilized as no further image processing is required except for data augmentation to make sure that the network will not overshoot. The optimized feature-based machine learning models and image-based deep learning networks were evaluated based on root mean square (RMSE), R^2, mean absolute error (MAE), and inference time throughout training to testing stages.

III. Results and Discussions

A. *Leaf Phenotype Analysis*

Correlation analysis confirmed that the extracted a* and Cr of diseased maize leaf image has a strong positive influence on diseased surface percentage. R, S, V, and b* has very weak positive significance to DSP (Table 1). All numerical Haralick textural features have a very strong negative or positive correlation with DSP. Additionally, all spectro-textural-morphological features of healthy maize leaf images exhibit no significance with the diseased surface percentage at $\alpha \leq 0.05$.

TABLE I. Correlation Analysis of NCA-ReliefF Extracted from Pearson Correlation Analysis of Diseased and Healthy Maize Leaves

Leaf Health Status	R	H	a*	Cb	Cr	Entropy	Whole Leaf Area
Diseased	0.196	-0.655	0.784	-0.185	0.880	0.038	-0.086
Healthy	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*

[a] *ns* means not significant

B. *Maize Crop Leaf Quality Classification Using Deep Transfer Network*

Matured maize leaves were classified healthy and defective using deep transfer image classification networks of MobileNetV2, ResNet101, and InceptionV3 (Fig. 5). For 117 test samples, diseased maize leaf samples were arranged from 1 to 57 and the remaining samples are the healthy leaf images without a hint of *Cercospora zeae-maydis*. MobileNetV2 misclassified 1 healthy maize leaf as defective or diseased (Fig. 5c). On the other hand, both ResNet101 and InceptionV3 misclassified 1 diseased leaf as a healthy leaf as depicted by green and blue lines respectively (Fig. 5d). These very intelligent classification performances resulted in the same accuracy, fall-out, precision, specificity, recall, f1-score, MCC, and Hamming loss (Table 2). However, MobileNetV2 bested ResNet101 and InceptionV3 in discriminating maize leaf quality in terms of inference time with 76.25% faster than ResNet101 and 57.23% faster than InceptionV3.

C. *Feature-Based Maize Leaf Diseased Surface Percentage Prediction*

MobileNetV2-classified diseased matured maize leaves exhibiting the existence of *Cercospora zeae-maydis* were computationally assessed using GPR, RNN, RTree, and RSVM. A numerical subscript was suffixed to each machine learning model to denote the number of spectro-textural-morphological features used as descriptors such as $RSVM_{18}$ for RSVM using 18 predictors (Fig. 6a). 18-feature predictors are the original unreduced vector while the 7-feature vector is the utilized the NCA-ReliefF selected leaf elements which are R, H, *a, Cb Cr, entropy, and whole leaf area (Figures 3 and 6a). GQPSO-optimized RNN_7 performed the most accurate and sensitive predictive feature-based machine learning model used in this study in determining diseased surface percentage as reflected by lowest RMSE and MAE, highest R^2 value of 0.949, and shortest inference time of 3 seconds (Table 3). The accuracy of GQPSO-RNN_7 rose by 350.544% from its RNN_{18} performance. On the other hand, GPR_7 is also comparable with GQPSO-RNN_7 as it exhibited an R^2 of 0.945 and RMSE and MAE values close to GQPSO-RNN_7 in the testing phase. Despite the 115.545% improvement in R^2 factor of GPR_7 from GPR_{18}, its resulting value missed 0.004 to be coequal with GQPSO-RNN_7. Correspondingly, RTree variants performed the worst predictions, and RSVM exhibited the longest inference time that is 54 multiples that of the GQPSO-RNN_7. Evidently, the use of intelligently selected combinations of color, texture, and morphological features significantly resulted in improved prediction performance, thus, making MobileNetV2 the most reliable model in this application.

D. *Image-Based Maize Leaf Diseased Surface Percentage Prediction*

In contrast with GPR, RNN, RTree, and RSVM, MobileNetV2, ResNet101 and InceptionV3 utilized the raw maize leaf images as network input for prediction of diseased surface percentage (Fig. 6). ResNet101 bested out MobileNetV2 and InceptionV3 in predicting DSP in terms of lower RMSE and MAE and strong positive R^2 of 0.9333. However, the inference time of ResNet101 is more than twice of MobileNetV2 and 1.5x of the InceptionV3 model (Table 3). This makes ResNet101 the slowest deep transfer image network to predict DSP and MobilenetV2 is the fastest one. It should be taken into consideration that in this study, accuracy and sensitivity is what weighs more than that of inference time.

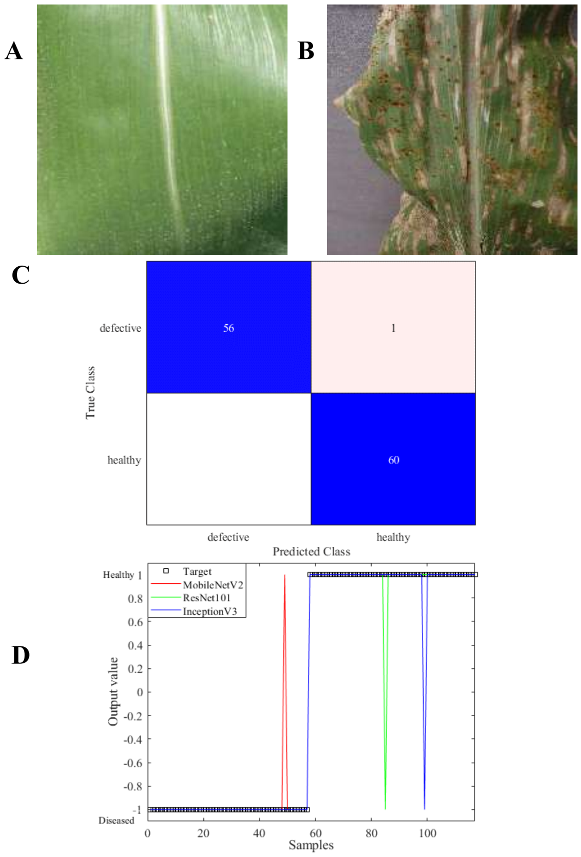


Fig. 5. Samples of (a) healthy and (b) diseased maize leaves, (c) confusion matrix, and (d) comparison of classification output of each deep transfer image network

After series of successful explorations in predicting the diseased surface percentage of maize *Cercospora* leaf spot using both feature-based and image-based computational intelligence models, it can be determined that GQPSO-RNN7 exhibited the most accurate, sensitive, and quickest model amongst GPR, RTree, and RSVM with and without using NCA-RelieF selected features, and MobileNetV2, ResNet101, and InceptionV3. In this study computer vision, through a low consumer-grade RGB camera, was employed to extract maize leaf signatures to mitigate subjective detection of the presence of fungi in crop leaf which is the challenge encountered by [2-3]. It is interesting that among 18 unreduced feature vector, S, Cb, Cr, correlation, entropy, and homogeneity of the healthy maize leaf corresponds to smaller numerical value compare to diseased leaves. With this, there is a common implication that 60% of the texture properties can independently distinguish healthy maize leaf. The a* value for a healthy leaf is 227.712% smaller than that of the disease, thus, making it profound that a healthy leaf exhibits greener vegetation pixels. The green color component from the RGB color space dictates that healthy maize leaves have an average of 131.167 G value and is 142.825% higher than the diseased leaf. It does make sense as maize *cercospora* leaf can be detected with gray, red, and darker brown spots that later on promotes necrosis.

The developed seamless MobileNetV2-GQPSO-RNN model exhibits a non-destructive approach in crop leaf phenotyping which only requires an RGB image as input for it evaluate the rendered maize leaf disease unlike [7], [9]. It is also comparable that other studies focused on using one leaf trait only as input to their phenotype model [16]-[17] but this developed framework also qualified to use combinations of leaf signatures from color, texture, and morphological components to increase the performance accuracy and sensitivity of the combined computer vision and computational intelligence.

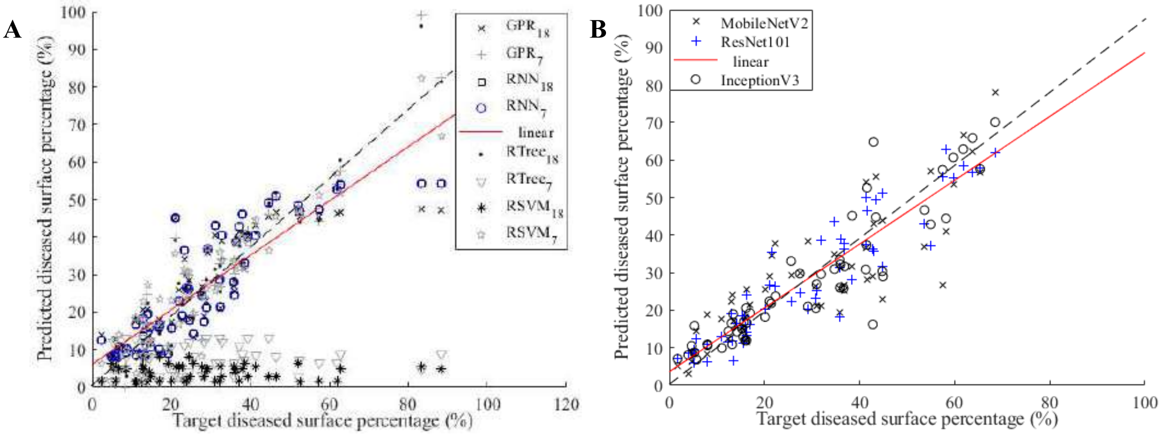


Fig. 6. Regression curves of (a) feature-based machine learning and (b) image-based deep transfer image regression networks in predicting diseased surface percentage of maize leaf

TABLE II. EVALUATION OF MAIZE CROP LEAF HEALTH CLASSIFICATION USING DEEP TRANSFER IMAGE NETWORKS

Model	Training	Validation	Accuracy								
	Accuracy	Accuracy	Accuracy	Fall-out	Precision	Specificity	Recall	F1-score	MCC	Hamming Loss	Inference Time (s)
MobileNetV2	95.730	99.150	99.145	0.008	0.992	0.992	0.991	0.991	0.983	0.009	1349
ResNet101	95.730	99.150	99.145	0.008	0.992	0.992	0.991	0.991	0.983	0.009	5679
InceptionV3	95.730	99.150	99.145	0.008	0.992	0.992	0.991	0.991	0.983	0.009	3154

TABLE III. EVALUATION OF MAIZE CROP LEAF DISEASED SURFACE PERCENTAGE PREDICTION USING COMPUTATIONAL INTELLIGENCE REGRESSION MODELS

odel	Training			Validation			Testing			
	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	Inference Time (s)
Using Image-Based Deep Transfer Networks										
MobileNetV2	9.7910	0.952	10.845	6.6883	0.898	7.225	19.8094	0.8555	7.3522	493
ResNet101	10.6110	0.958	7.020	5.6112	0.938	5.014	13.7413	0.9333	5.3461	1172
InceptionV3	7.5010	0.952	8.474	6.3201	0.922	6.009	15.8644	0.9115	5.4266	802
Using Feature-Based Machine Learning Models										
GPR_{18}	9.549	0.907	6.093	10.331	0.911	6.665	10.336	0.847	6.880	66.295
GPR_7	4.796	0.970	3.661	5.108	0.971	3.860	6.593	0.945	5.208	3.000
RNN_{18}	8.722	0.921	5.883	11.301	0.903	7.140	18.685	0.271	13.939	3.000
RNN_7	5.115	0.966	4.002	5.143	0.970	3.897	6.290	0.949	5.090	3.000
$RTree_{18}$	3.251	0.986	2.268	30.510	0.015	21.823	27.809	0.053	21.151	42.148
$RTree_7$	3.251	0.980	2.831	34.715	-0.089	25.841	30.298	0.000	23.864	34.580

IV. Conclusion

Maize leaf health status was classified into healthy or diseased using MobileNetV2. ResNet101 and InceptionV3 exhibited the same network performances with MobileNetV2 except that MobileNetV2 classifies leaf health status with the shortest time. The degree of evidence of *Cercospora zeae-maydis* in maize leaf was characterized by combined neighborhood component analysis and ReliefF, resulting in R, H, *a, Cb Cr, entropy, and whole leaf area as the most significant features in predicting diseased surface percentage. Lazysnapping provided exquisite segmentation of *Cercospora* leaf spot regions from other healthy vegetive and non-vegetative leaf regions. Gaussian quantum-behaved particle swam optimization significantly increased the sensitivity and accuracy of RNN by determining the optimal number of hidden artificial neurons. GQPSO-RNN with reduced predictors bested out other feature-based and deep transfer machine learning networks in predicting DSP. This low-cost plant imaging approach can be employed by low-income agriculturists who want to assess their crops effectively and quantitatively by just acquiring inexpensive camera and following the computational intelligence model development stated above. The developed seamless model of MobileNetV2-GQPSO-RNN is highly essential for computational phenotyping of maize leaf *Cercospora* leaf spot and other leaf diseases. For future studies, it is recommended to cultivate different species of corn suitable in Philippine environmental conditions and employ the developed computer vision approach for on-site leaf *Cercospora* disease phenotyping.

Acknowledgment

The authors would like to acknowledge the support of the Engineering Research and Development for Technology (ERDT) of the Department of Science and Technology (DOST) of the Philippines and the Intelligent Systems Laboratory of the De La Salle University, Philippines.

References

[1] A. C. de Lira, G. M. Mascarin, and Í. Delalibera Júnior, “Microsclerotia production of Metarhizium spp. for dual role as plant biostimulant and control of Spodoptera frugiperda through corn seed coating,” Fungal Biol., vol. 124, no. 8, pp. 689–699, 2020.

[2] A. Waheed, M. Goyal, D. Gupta, A. Khanna, A. E. Hassanien, and H. M. Pandey, “An optimized dense convolutional neural network model for disease recognition and classification in corn leaf,” Comput. Electron. Agric., vol. 175, no. January, p. 105456, 2020.

[3] R.S. Concepcion, S.C. Lauguico, R.R. Tobias, E.P. Dadios, A.A. Bandala, and E. Sybingco. “Estimation of Photosynthetic Growth Signature at the Canopy Scale Using New Genetic Algorithm-Modified Visible Band Triangular Greenness Index,” International Conference on Advanced Robotics and Intelligent Systems (ARIS), 2020.

[4] Y. Qiu, J. Cooper, C. Kaiser, R. Wisser, S. X. Mideros, and T. M. Jamann, “Identification of loci that confer resistance to bacterial and fungal diseases of maize,” G3 Genes, Genomes, Genet., vol. 10, no. 8, pp. 2819–2828, 2020.

[5] K. Braga, L. H. Fantin, J. M. T. Roy, M. G. Canteri, and A. A. de Paiva Custódio, “Development and validation of a diagrammatic scale for the assessment of the severity of bacterial leaf streak of corn,” Eur. J. Plant Pathol., vol. 157, no. 2, pp. 367–375, 2020.

[6] T. Wiesner-Hanks et al., “Image set for deep learning: Field images of maize annotated with disease symptoms,” BMC Res. Notes, vol. 11, no. 1, pp. 10–12, 2018.

[7] M. Agarwal, S. K. Gupta, and K. K. Biswas, “Development of Efficient CNN model for Tomato crop disease identification,” Sustain. Comput. Informatics Syst., vol. 28, p. 100407, 2020.

[8] S. Mishra, R. Sachan, and D. Rajpal, “Deep Convolutional Neural Network based Detection System for Real-time Corn Plant Disease Recognition,” Procedia Comput. Sci., vol. 167, pp. 2003–2010, 2020.

[9] F. Wang, J. W. Li, W. Shi, and G. P. Liao, “Leaf image segmentation method based on multifractal detrended fluctuation analysis,” J. Appl. Phys., vol. 114, no. 21, 2013.

[10] H. Kaur et al., “Leaf stripping: an alternative strategy to manage banded leaf and sheath blight of maize,” Indian Phytopathol., vol. 73, no. 2, pp. 203–211, 2020.

[11] P. D. Carpane, A. M. Peper, and F. Kohn, “Management of Northern Corn Leaf Blight using Nativo (Trifloxistrobin + Tebuconazole) Fungicide Applications,” Crop Prot., vol. 127, p. 104982, 2020.

[12] D.L. Elliot, and R.E. Valentine, “Recurrent Neural Networks for Moisture Content Prediction in Seed Corn Dryer Buildings,” 23rd International Conference on Tools with Artificial Intelligence, 2011.

[13] H. Xia et al., “Genetic mapping of northern corn leaf blight-resistant quantitative trait loci in maize,” Medicine (Baltimore)., vol. 99, no. 31, p. e21326, 2020.

[14] D. S. Skibbe, G. Doehlemann, J. Fernandes, and V. Walbot, “Maize tumors caused by ustilago maydis require organ-specific genes in host and pathogen,” Science (80-.)., vol. 328, no. 5974, pp. 89–92, 2010.

[15] F. Li, X. Bai, and Y. Li, “A crop canopy localization method based on ultrasonic ranging and iterative self-organizing data analysis technique algorithm,” Sensors (Switzerland), vol. 20, no. 3, 2020.

[16] J. Pang, Z. Y. Bai, J. C. Lai, and S. K. Li, “Automatic segmentation of crop leaf spot disease images by integrating local threshold and seeded region growing,” Proc. 2011 Int. Conf. Image Anal. Signal Process. IASP 2011, pp. 590–594, 2011.

[17] S. Kai, Z. Liu, H. Su, and C. Guo, “A research of maize disease image recognition of corn based on BP networks,” Proc. - 3rd Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2011, vol. 1, no. 2009921090, pp. 246–249, 2011.

[18] K. Aurangzeb, F. Akmal, M. Attique Khan, M. Sharif, and M. Y. Javed, “Advanced Machine Learning Algorithm Based System for Crops Leaf Diseases Recognition,” Proc. - 2020 6th Conf. Data Sci. Mach. Learn. Appl. CDMA 2020, pp. 146–151, 2020.

[19] S. E. Sukmana and F. Z. Rahmanti, “Blight segmentation on corn crop leaf using connected component extraction and CIELAB color space transformation,” Proc. - 2017 Int. Semin. Appl. Technol. Inf. Commun. Empower. Technol. a Better Hum. Life, iSemantic 2017, vol. 2018-January, pp. 205–208, 2017.

[20] Z. Liu, Z. Du, Y. Peng, M. Tong, X. Liu, and W. Chen, “Study on Corn Disease Identification Based on PCA and SVM,” Proc. 2020 IEEE 4th Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2020, no. Itnec, pp. 661–664, 2020.

[21] J. Arun Pandian and G. Geetharamani, “Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network”, Mendeley Data, v1, 2019.

[22] R. Concepcion II, S. Lauguico, L. Alejandrino, E. Dadios, and E. Sybingco, “Lettuce Canopy Area Measurement Using Static Supervised Neural Networks Based on Numerical Image Textural Feature Analysis of Haralick and Gray Level Co-Occurrence Matrix,” AGRIVITA, Journal of Agricultural Science, 42(3), 472-486, 2020.

[23] L. dos S Coelho, “Gaussian quantum-behaved particle swarm optimization approaches for constrained engineering design problems,” Expert Systems with Applications, vol. 37, no. 2, pp. 1676-1683, 2010.

[24] J. Sun, W. Fang, V. Palade, X. Wu, X., and W. Xu, “Quantum-behaved particle swarm optimization with Gaussian distributed local attractor point, Applied Mathematics and Computation,” vol. 218, no. 7, pp. 3763-3775, 2011.

[25] R. Concepcion, S. Lauguico, R.R. Tobias, E. Dadios, A. Bandala and E. Sybingco, “Genetic Algorithm-Based Visible Band Tetrahedron Greenness Index Modeling for Lettuce Biophysical Signature Estimation,” IEEE Region 10 Conference, 2020.

A Buck Converter-based Battery Charging Controller for Electric Vehicles using Modified PI Control System

Md. Rezanul Haque and Md. Abdur Razzak
Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Aftabuddin Ahmed Road, Bashundhra R/A, Dhaka-1212, Bangladesh
Corresponding E-mail: hrezanul@gmail.com, razzak@iub.edu.bd

***Abstract*–A well functioned battery charger for electric vehicles can improve the life cycle of the battery. To control the overshoot, rise time and settling time with respect to the changing input voltage and load is the noticeable challenge in the field of battery charging. Usually, a PID controller has been used to overcome these challenges. However, PID controller is suitable for a specific configuration, and overshoot, oscillation happens when input voltage and load vary. In this paper, a buck converter-based battery charging controller is presented in which a modified proportional-integral (PI) controller is used. The modified PI controller is tuned for charging the battery with minimum overshoot and settling time. To verify the viability, the proposed system has been simulated in MATLAB. The simulation results show that the overshoot is reduced to 1.23% with the settling time of 0.01s for the 48V battery bank.**

***Keywords*–Battery charging controller, Asynchronous Buck Converter, Electric vehicle, Pulse Width Modulation, Proportional Integral**.

I. Introduction

In order to make the transportation system more efficient, environment friendly and robust the current noticeable interest has been increased to use the electric vehicle [1]. The electric vehicle (EV) can produce high speed and torque within a minimum time. Unlike conventional fuel-based vehicles the EV takes power from rechargeable battery [2]. To charge the battery with optimum efficiency ensuring the good battery life a power electronics controller is required which can convert the grid power to suitable DC for charging the battery [3].

For effective operation, it is necessary to charge the battery with lower overshoot and less settling time with respect to the variation of input voltage and load. To meet these demand, PID and other controller topologies have been proposed [4-7]. But they struggle to withstand the variation of input voltage and load. The PI or PID transient response is mostly S shaped and the delay in the initial point is noticeable as explained by Ziegler-Nichols First rule and it is very challenging to reduce the delay time, overshoot and settling time.

Many scholars avoid PI and proposed PID, forward PID, FOPID and other topologies which are complex, hard to implement, need high processing microcontroller and hence costly. While using PID the overshoot issue is still present.

Theerayod Wiangtong [8] avoided using PI and proposed PID with Flower Pollination Algorithm which makes the controller complex and still overshoot present at the output voltage although it was applied for small system. Adhul S V [9] proposed FOPID and avoid conventional PI and PID controller due to high overshoot. D. Ounnas, D. Guiza, Y. Soufi [10] also avoided PI controller due to overshoot and proposed digital PID controller for the Buck converter.

This paper presents a buck converter-based battery charging controller using modified proportional integral (MPI) control system to reduce the initial delay time, settling time and overshoot with changing input voltage and load.

II. Design of Proposed System

The block diagram of proposed buck converter controller using Modified Proportional Integral controller is depicted in Fig.1 The output voltage is sensed and compared with the processed reference value and the error goes to MPI controller and the PWM is generated based on that and thus the buck converter operates.

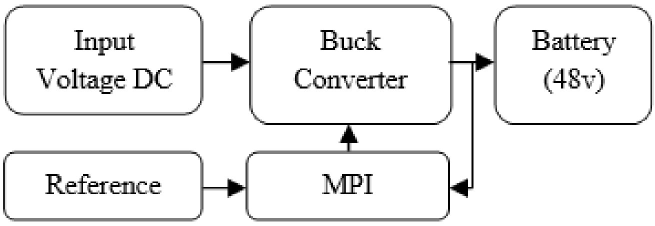


Fig. 1: Block diagram of proposed system

A. Modified PI Controller

The modified proportional integral (MPI) controller is shown in Fig.2. To find the difference between desired voltage $r'(t)$ and actual output voltage $y'(t)$, the MPI is required. The desired output voltage is sampled and further processed according to predefined tuned value. The desired output voltage which is instantaneous defined as

$$r'(t) = nu(t) \qquad (1)$$

In sample and process block the instantaneous desired voltage modified in the basis of equation (2).

$$r(t) = nr(t+0) - nr(t-m) \qquad (2)$$

where *n* is the given desired output voltage value and *m* is the desired settling time. The desired output voltage is then compared with actual output voltage and the difference between them is fed to the PI block defined as

$$y(t) = k_P e(t) + k_i \int e(t)\, dt \qquad (3)$$

where $e(t)$ is the difference between desired and actual output voltage.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

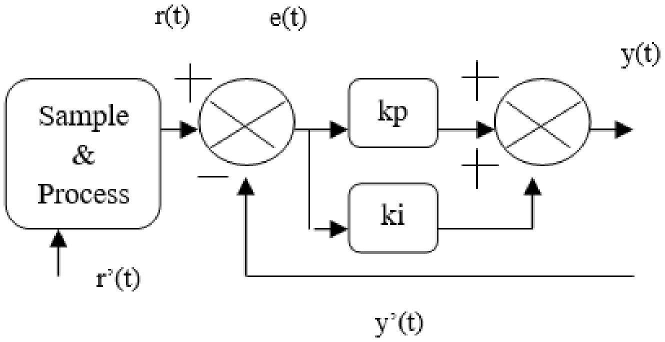


Fig. 2: Block diagram of proposed MPI system

B. PI Tuning Parameters using Zeigler-Nichols Method

The tuning parameters of PI controller have been derived using Ziegler-Nichol's frequency domain method as shown in Table I. The critical value of proportional constant k_{cr} has found from simulated results and hence the proportional constant k_p can be calculated as

$$k_p = 0.45k_{cr} = 0.45 \times \ 0.223 = 0.1$$

The period of oscillation p_{cr} has been found form simulated result and the T_i can be calculated as

$$T_i = \frac{p_{cr}}{1.2} = \frac{0.14}{1.2} = 0.117$$

Hance, the integral constant k_i can be found as

$$k_i = \frac{k_p}{T_i} = \frac{0.1}{0.117} = 0.854$$

These values of tuning parameters of PI controller have been used in the proposed system.

Table I. Ziegler-Nichols tuning parameters for PI controller

Tuning parameter	Value
k_p	$0.45k_{cr}$
T_i	$\frac{p_{cr}}{1.2}$

C. Buck Converter

To simplify the circuit design of the buck converter the following assumptions are considered:

1) all components are ideal, and
2) the inductor is operated in continuous conduction mode (CCM).

Maximum inductor current ripple can be calculated as

$$\Delta I_L = I_L \times \Delta I_{Lmax} = 10 \times 10\% = 1\text{A}$$

where the $I_L$is inductor current.

The maximum output voltage ripple can be calculated as

$$\Delta v_{out} = v_{out} \times 1\% = 48 \times 1\% = 0.48\text{v}$$

Inductor value is calculated as

$$L = \frac{v_{out}(v_{in} - v_{out})}{\Delta I_L \times f_s \times v_{in}} = 1.31017\text{mH}$$

Output capacitor value is found as

$$C = \frac{\Delta I_L}{8 \times f_s \times \Delta v_{out}} = 0.51243\text{uF}$$

The design parameters of DC-DC buck converter used in proposed charging system are listed in TABLE II.

TABLE II. Design Parameters of Proposed Charging System

Parameter	Values
Input DC Voltage	312v
Output DC Voltage	48v
Maximum Output Current	10A
Applied Switching Frequency	31kHz
Filter Inductance	1.31017mH
Filter Output Capacitance	0.51243uF
Max Inductor Ripple Current	1%
Output Voltage Ripple	1%

III. Simulation and Results Analysis

A complete circuit was developed in the MATLAB Simulink using the calculated and designed parameters to analyze the performance of the proposed system as shown in Fig.3. The MATLAB functional block consists of MPI equations as explained in section IIA. The desired voltage and actual voltage are then compared and based on that the PWM is generated at a frequency of 31kHz with the help of PI controller.

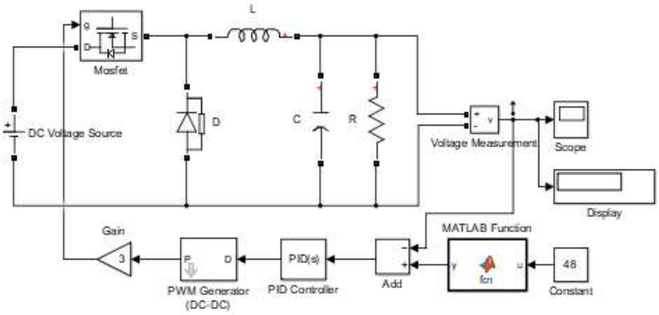


Fig. 3: Proposed MPI system

Figs. 4-8 show the direct comparison for different input voltages and loads for 48V output voltage. Both conventional PI and propose MPI used the same k_p and k_i values. The result shows that the conventional PI controller has high overshoot than that of the proposed modified PI where a settling time of 0.01s has been achieved. The conventional PI struggles for stabilizing the overshoot voltage when the input voltage and load changes whereas the proposed modified PI can handle this scenario easily and no overshoot happen during that case.

From Figs. 5-6 it is observed that the maximum overshoot of the output voltages of the buck converter increases 192.7% to a maximum of 206.25% for conventional PI controlled buck converter when the input voltage changes 290V-335V from its base voltage 312V for a fixed load of 30 ohm. On the other hand, the maximum overshoot of the output voltages of the buck converter increases 130.2% to a maximum of 209.3% for conventional PI controlled buck converter when the load changes from 10 ohms to 50 ohms for a fixed input voltage of 312V as shown in Figs. 7-8. When input voltage 312V and load 50 ohm the overshoot was maximum which is 209.3%. The results conclude when the load increases from base 30 ohm value the overshoot increases greatly as depicted in Fig. 9. This large percentage of overshoot is not suitable for EV's battery charging applications. Interestingly, a negligible overshoot of the output voltage of the buck converter exists while using the modified PI controller.

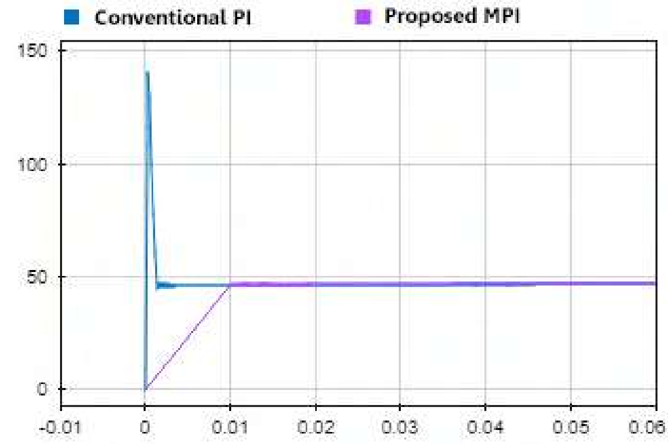


Fig. 4: Simulated Result for Vin 312 Vo 48 and load 30 ohm

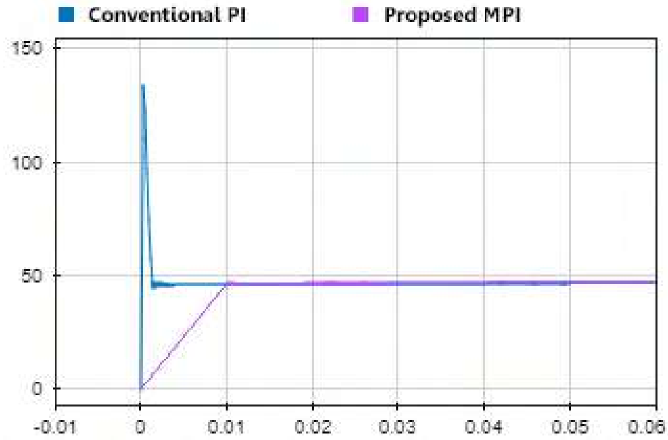


Fig. 5: Simulated Result for Vin 290 Vo 48 and load 30 ohm

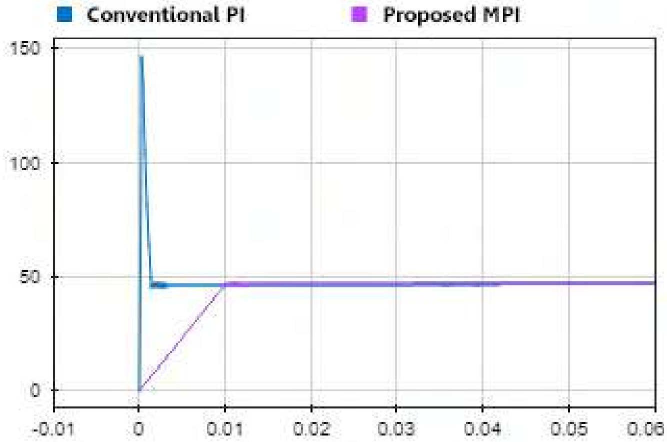


Fig. 6: Simulated Result for Vin 335 Vo 48 and load 30 ohm

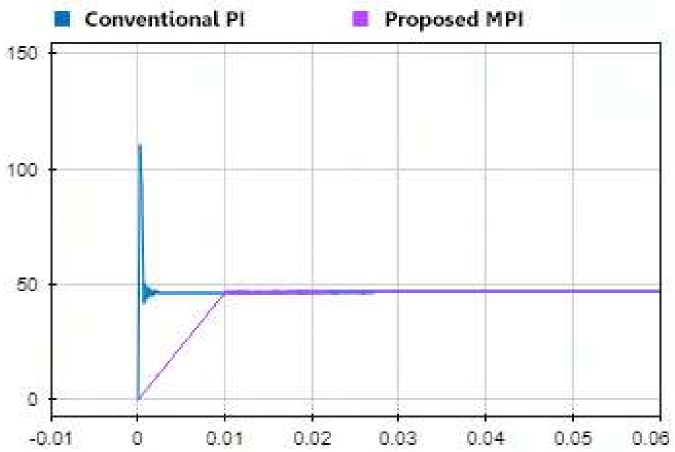


Fig. 7: Simulated Result for Vin 312 Vo 48 and load 10 ohm

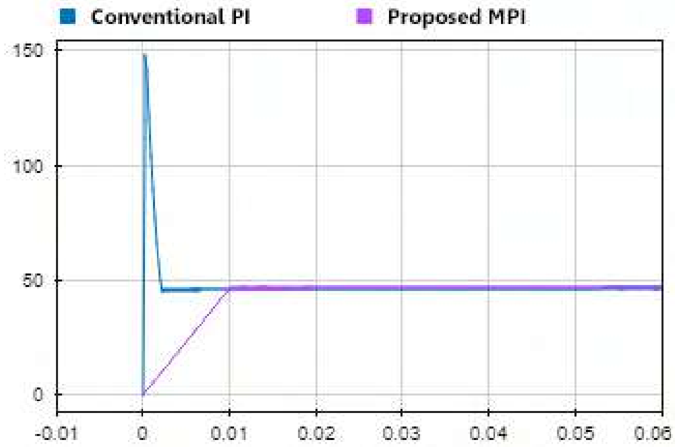


Fig. 8: Simulated Result for Vin 312 Vo 48 and load 50 ohm

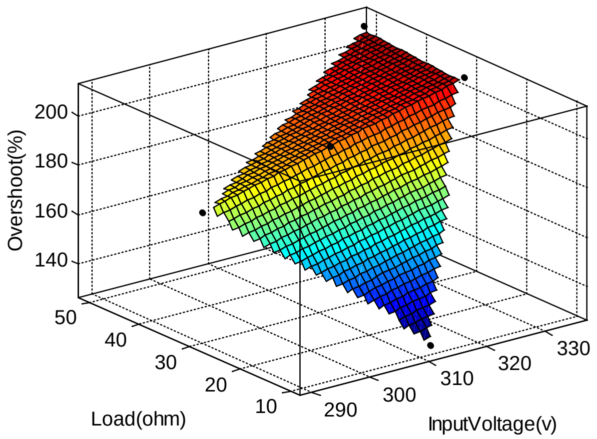


Fig. 9: Parameters for conventional PI

The voltage-gain transfer function for the proposed modified PI controlled buck converter calculated by MATLAB was found to be.

$$\frac{2.542e07}{S^2+6.933e07s+7.705e10} \tag{4}$$

The Bode diagram and pole-zero map for the proposed modified PI controlled buck converter as illustrated in Figs. 10 and 11, respectively, which confirms the stability of the system since all the poles of the voltage gain transfer function lies on the unit circle in the ***z***-plane.

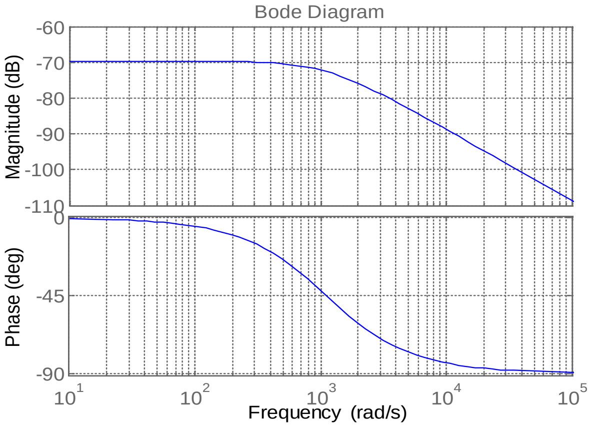


Fig. 10: Simulated Bode Diagram

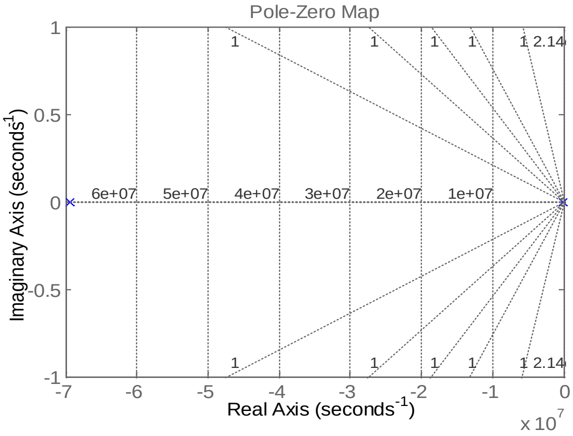


Fig. 11: Simulated Pole-Zero Map

IV. Conclusion

This paper emphasized on controlling the buck converters output voltage overshoot and settling time using a modified PI controller. A comparison of complete analysis of conventional PI and with the proposed modified PI has been investigated. The result shows that the maximum overshoot of the output voltage of the buck converter varies in the range 130.2% - 209.3% for the conventional PI controller-based buck converter when the input voltage and load changes, whereas the performance increases significantly with almost no overshoot while using the proposed modified PI controller based buck converter for a desired settling time of 0.01s. This modified PI controller-based buck converter, which has less settling time with negligible overshoot, will be very useful for EV's battery charging systems.

V. Future Work

A hardware prototype of the proposed system and its performance analysis will be presented in near future to confirm the viability of this simulation work.

References

[1] M. R. Haque, S. Das, M. R. Uddin, M. S. Islam Leon and M. A. Razzak, "Performance Evaluation of 1kW Asynchronous and Synchronous Buck Converter-based Solar-powered Battery Charging System for Electric Vehicles," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 770-773.

[2] A. K. Rajvanshi, “Electric and Improved Cycle Rickshaw as a Sustainable Transport System for India,” Current Science, vol. 83, no. 6, pp. 703–707, 2002.

[3] S. Chakraborty, S. I. Annie, and M. Razzak, “Design of Single-Stage Buck and Boost Converters for Photovoltaic Inverter Applications,” in International Conference on Informatics, Electronics & Vision (ICIEV). IEEE, 2014, pp. 1–6.

[4] S. Chakraborty, M. M. Hasan, I. Worighi, O. Hegazy, and M. A. Razzak, “Performance Evaluation of a PID-Controlled Synchronous Buck Converter Based Battery Charging Controller for Solar-Powered Lighting System in a Fishing Trawler,” Energies, vol. 11, no. 10, p. 2722, 2018.

[5] Mohammed Mahedi Hasan, Md. Mahamudul Hasan, Sajib Chakraborty, Mohamed El Baghdadi, M. A. Razzak and Omar Hegazy, “Evaluation of Failure Trends of a PID-controlled Synchronous Buck Converter Based Battery Charging Controller", 2020 IEEE International conference on Power Electronics, Smart Grid and Renewable Energy (PESGRE 2020) to be held in Cochin, Kerala, India from 2-4 January 2020.

[6] Sajib Chakraborty, Saurav Das, Mohammed Mahedi Hasan and M. A. Razzak, “A Novel MPPT-Based Synchronous Buck Converter for Solar Power System in Fishing Trawler", 7th IEEE Power India International Conference (PIICON 2016), 25-27 November, 2016, Rajastan, India.

[7] Sajib Chacraborty and M. A. Razzak, “Design of a Transformer-less Grid-Tie Inverter Using Dual-Stage Buck and Boost Converters”, International Journal of Renewable Energy Research, Vol.4, No.1. pp. 91-98 (2014).

[8] T. Wiangtong and J. Sirapatcharangkul, "PID design optimization using flower pollination algorithm for a buck converter," 2017 17th International Symposium on Communications and Information Technologies (ISCIT), Cairns, QLD, Australia, 2017, pp. 1-4.

[9] Adhul S V, T Ananthan, FOPID Controller for Buck Converter, Procidia Computer Science, Vol. 171, pp. 576-582, 2020.

[10] D. Ounnas, D. Guiza, Y. Soufi, R. Dhaouadi and A. Bouden, "Design and Implementation of a Digital PID Controller for DC–DC Buck Converter," 2019 1st International Conference on Sustainable Renewable Energy Systems and Applications (ICSRESA), Tebessa, Algeria, 2019, pp. 1-4.

Demand Analysis of Energy Consumption in a Residential Apartment using Machine Learning

Halima Haque, Adrish Kumar Chowdhury, M. Nasfikur Rahman Khan and Md. Abdur Razzak
Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Aftabuddin Ahmed Road, Bashundhara R/A, Dhaka-1212, Bangladesh
Corresponding E-mail: razzak@iub.edu.bd

***Abstract*—The challenge of keeping pace with the growth of population is synchronizing with the advancement of the technology. As a developing country it is difficult to sustain the balance between the increased inhabitants of Bangladesh and the total energy consumption. Machine learning (ML) can used to forecast the demand of energy consumption including production, organization and conservation of energy for the new buildings with respect to the citizens. This paper introduces a prediction analysis of energy consumption of a residential apartment using different machine learning models including multiple linear regression (MLR), random forest (RF), support vector machine (SVM), and k-nearest neighbors (KNN). While analyzing the models, the energy consumption data of twelve months from an apartment in Chittagong-district situated in Bangladesh was used. The analyses confirm the most effective model used for such energy demand criteria of residential buildings of that location. The outcome of the prediction method reveals that random forest (RF) model can reach to the best accuracy with the highest performance parameters.**

Keywords—energy consumption, machine learning, prediction, performance, learning models.

I. Introduction

Energy demand of a country renders crucial role to the ecosystem along with the development. Imbalance of energy consumption and demand can lead to the agony of the citizens in large scale. Around a decade, the emerge of machine learning models has taken part in almost every sector of global progress especially in energy consumption analysis. Moreover, forecasting of energy consumption building nowadays has highly subjected some prominent algorithms in machine learning to improve the conservation and modeling [1] [2]. In different area of application, the structure seems to represent different accuracy [2] [3]. In addition, support vector machine, gaussian mixture model, different neural networks and even new hybrid methods can be used in multiple platforms while acknowledging prediction analysis of energy demand or consumption [4] [5] [6]. Deep neural networking in high computational extent can sustain greater accuracy in different demanding proposals [7] [8]. In contrast, artificial neural network enables estimation to a certain level of interest where the goal focuses to the high precision in energy use [9] [10]. Besides depending on the feature extraction and the load pattern other machine learning models can signify the intention of conspicuous energy consumption. Reference [11] provides the most absolute outcome of the energy consumption forecast using support vector regression (SVR) machine based on cooling loads. Furthermore, improved results have been achieved while introducing k-shaped clustering before applying the SVR model [12]. On the other hand, hybrid models can attain the standard of accuracy while concerning the demand of short-term energy use [13]. Such classification approach highly enlarges the performance of the system while implementing on other algorithms in numerous feature analysis [14]. However, acknowledging temperature, humidity, solar parameters along with more factors, the gaussian mixture regression (GMR) even in heating, ventilation & air conditioning (HVAC) has conducted solid assessment of estimation [15] [16].

The major influences that can diversify the rate of demand in energy based on complex figures and uncertain features. In Bangladesh, around 50% of total households have domestic connections from 21.8 million grid connected consumers. Others are mostly commercial or industrial, where the total power generation is approximately 13,000MW. In favour of the growth, the intricate nature of preserving the energy demand is a challenge for the country. Among few of the implementation of machine learning in prediction models, this country has already accepted the peripherals [17] [18]. Besides, some machine learning algorithms have interestingly improved the building of energy consumption in non-residential sectors outside Bangladesh [19] [20]. Reference [21] and [22], applies such instances of promising accuracy in residential area holding the features with complication.

This paper emphasizes on different energy consumption prediction models of machine learning based on certain feature points like temperature, humidity, electric bill etc. The methods are evaluated by coefficient of determination (R^2), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), and coefficient of variance (CV). The energy demand assessment has been analyzed with different machine learning models including multiple linear regression (MLR), random forest (RF), support vector regression machine (SVR), and k-nearest neighbors (KNN).

The content of the paper is coordinated as follows: Section II upholds the dataset formal with detain; Section III presents the mathematical and understanding of different models; Section IV illustrates various performance parameters; Section V analyses the results of the predictive outcomes and finally draws conclusion according to the final remarks.

II. Dataset Formation

The dataset represents energy related factors of a six-story residential building including the ground floor having six types of flooring size in Chittagong district located in Bangladesh having the parameters of table 1. Various related features like area (square feet), occupancy, daytime

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

temperature (°C), nightly temperature (°C), average temperature (°C), daily sunshine (hours), average rainfall (mm), rainy days (days), sea temperature (°C), humidity (%), windspeed (kph), number of rooms, and price (BDT-excluding demand, vat, and other charges) are considered along with the 12 month of daily energy consumption from 1st January 2020 to 31st December 2020. Having six different types of flats each consisting of eight floors, the whole apartment contains 48 flats and 392 rooms in total which have been demonstrated in the table I. Likewise, the information of the consumption is being collected manually from the digital electric meter. The data-formation has 576 samples while observing all the obstacles that have limited the possible outcome. The entire dataset has been recorded as *.csv (comma separated values) file for simulation purpose.

TABLE I. BUILDING FORMAT

Type of flats	No. of floors	Size (square feet)	No of rooms
A	8	1415	9
B		1214	8
C		1206	8
D		1385	9
E		935	7
F		1047	8

III. MACHINE LEARNING MODELS

The machine learning methods include multiple linear regression (MLR), random forest (RF), support vector regression (SVR), and k-nearest neighbors (KNN) which are constituted below.

A. *Multiple Linear Regression (MLR)*

More than one input features can be related to the dependent feature that is the energy consumption while utilizing multiple linear regression model. Denoting x_0 as the constant or the intercept, p and n as the observations, β_p as the coefficient for each independent feature of p^{th} sample, $x_{e(p)}$ as the variable of p^{th} sample, and $\in$ as the error term of the fitted model; the energy consumption (y_e) of our dataset can be presented as,

$$y_e = x_0 + \sum_{p=1}^{n} \beta_p x_{e(p)} + \in \qquad (1)$$

B. *Random Forest (RF)*

Comprising of many adaptable decision trees with the ability of executing both regression and classification problems, the RF can compute with most effectiveness. Each tree contributes highly which predicting the best solution by voting. The ideology of acquiring highest accuracy in the sense of merging the training sets called by 'bagging' process, is used to ensemble the trees in this supervised algorithm. Hence, at the time of processing the model itself introduces certain randomness while equally spitting the subsets among the decision trees. The goal of such method is to recognize the best among all the features in each set. Consequently, this leads to better estimation process for any application.

C. *Support Vector Regression (SVR)*

The SVR can effectively handle outliers in any model by separating the target values and the provided data into high-dimensional hyperplanes. Different kernel functions are used to determine support vector classifier or regressors such as polynomial, linear, sigmoid, radial (radial basis function-RBF) etc. Moreover, without the transformation these functions can easily evaluate the correlation between each pair of points assuming them in greater dimensions. This phenomenon is known as the 'kernel trick' which eases the assessment of computation. The polynomial kernel uses the equation 2 for calculating the relations, where *x* and *y* are the observations of independent and dependent feature respectively, *r* is the polynomial-coefficient, and *d* is the polynomial-degree.

$$\{(x \times y) + r\}^d \qquad (2)$$

The equation 3 validates the radial function where γ weights the squared-distance and the impact. Unlike polynomial, RBF deals with infinite number of dimensions encouraging both taylor series expansion method and dot-multiplication. On the other hand, the linear kernel is associated with the equation 4, where n is the number of observations.

$$e^{-\gamma\,(x-y)^2} \qquad (3)$$

$$k(x, y) = \sum_{t=1}^{n} x_t y_t \qquad (4)$$

Most functions allow dot-multiplication for assessing the relationship while *r*, *d*, and *γ* are determined by cross-validation.

D. *k-Nearest Neighbors (KNN)*

Being able to use in both classification and regression model, KNN operates with '*k*' as the number of points closest to the required observation to find the distance using distance-measurement techniques. These techniques popularly include Manhattan, and Euclidian. A regressor would find the mean of the distance of total '*k*' points and assign that to the new one. The process stabilizes until the predictive point is in optimal state. In contrast, a classifier assigns the majority voted group-points as the predictive point where the '*k*' has to be an odd value. The corresponding equation 5 and 6 are allied with Euclidean and Manhattan distance measuring methods respectively, where x and y are the observations and the m is the total number of samples.

$$d\,(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \qquad (5)$$

$$d\,(x, y) = \sum_{i=1}^{m} |x_i - y_i| \qquad (6)$$

IV. PERFORMANCE PARAMETERS

For evaluating outcomes from all the models, the specifications as coefficient of variance (CV) for both test and train dataset, coefficient of determination (*R*^2), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) are being figured out and compared. The simulation environment used in the purpose of modelling is the Spyder IDE (integrated development environment). As the objective is to apply regressors, the *R*^2 holds the variance of targeted outcome figure which possess the equation 7. The more the value of *R*^2, the better the model performs. Additionally, the CV retains the total amount of distributed figures concerning with the computed mean that can be calculated through equation 8. The lesser the value of CV, the performance of the model gets better. Finally, the RMSE presents the overall term of error in a model regarding in the form of root-mean-squared which can be represented by the equation 9. Furthermore, the predictive inaccuracies of the model have to go through standard deviation in order to acquire the RMSE of the model.

$$R^2 = [\frac{\sum_{i=1}^{N}\left(y_{predict,i} - \bar{y}_{data}\right)^2}{\sum_{i=1}^{N}\left(y_{data,i} - \bar{y}_{data}\right)^2} \times 100]\ \% \qquad (7)$$

$$CV = [\frac{\sqrt{\frac{\sum_{i=1}^{N}\left(y_{predict,i} - y_{data,i}\right)^2}{N-1}}}{\bar{y}_{data}} \times 100]\ \% \qquad (8)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{predict,i} - y_{data,i}\right)^2}{N}} \qquad (9)$$

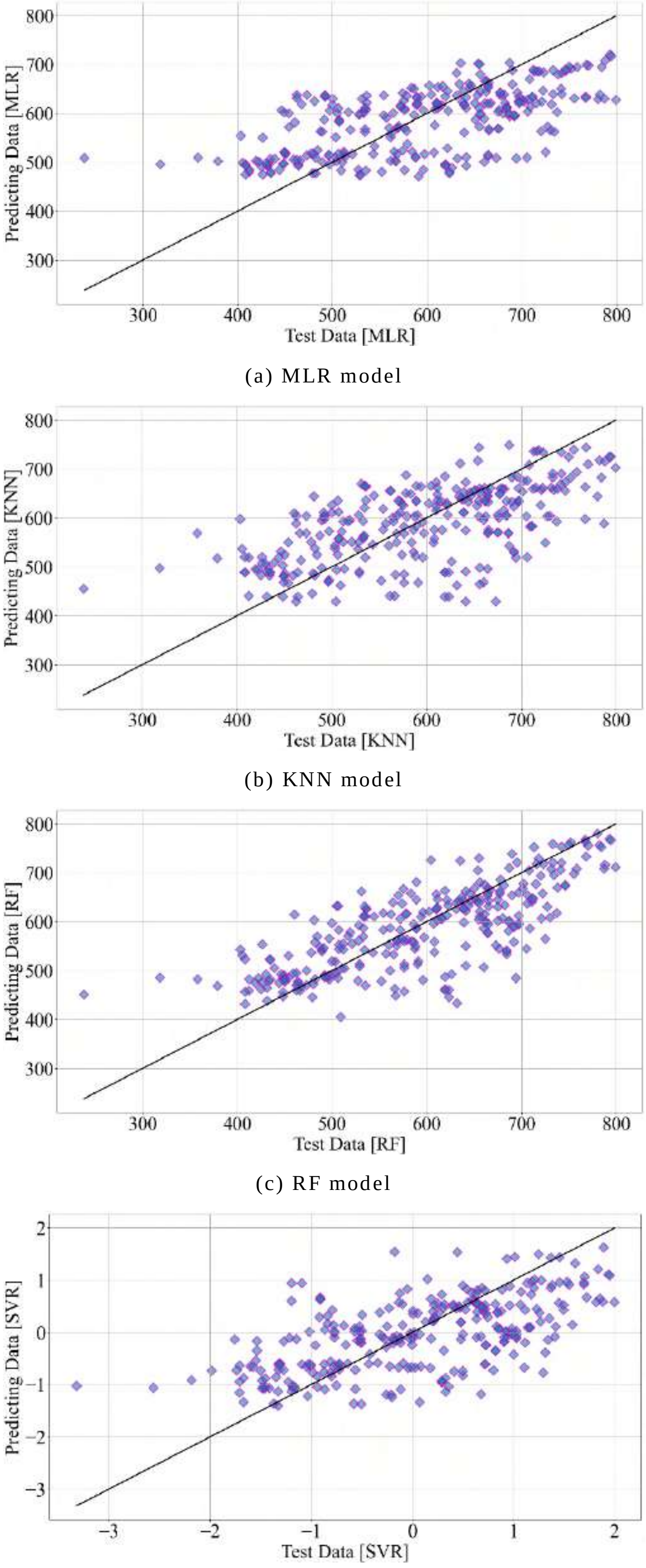
where the y_(predict,i) is the predicted energy consumption, y_(data,i) is the actual energy consumption or the test data, N is the sample size, and y ̅_data is the average of the energy consumption.

After evaluation, the performance parameters from the models in predictive energy consumption is listed in Table I. The main dataset has been split into 50:50 ratio (test:train) while computing the analysis.

TABLE II. PERFORMANCE PARAMETERS

Model Name	R^2 **(%)**	MAE	MSE	$RMSE$	CV_{test} **(%)**	$CV_{predict}$ **(%)**
MLR	38.06	67.87	6894	83.03	17.8	11.25
KNN [k=5]	36.22	67.44	7099	84.25	17.8	13.26
RF	56.88	52.72	4799	69.27	17.8	14.61
SVR (RBF)	32.32	0.65	0.68	0.82	2469	-874.92
SVR (Linear)	33.27	0.64	0.67	0.82	2469	-506.84
SVR (Polynomial : degree=5)	36.47	0.63	0.64	0.80	2469	-1771.90

The RF model responds well corresponding to the consumption rather than other models of the building. Securing 56.88% of coefficient of determination, RF has less coefficient of variance comparing with others. Fig. 1 exhibits the comparison among models where the scatter plotting represents the real and estimated points. Likewise, the black line is just the centre line where the test value is equal to estimated value. The RF can easily disperse along the line apart from other machine learning methods that have been used. The total energy consumption from the apartment is around 340MW per year. Such analysis can definitely amplify the process of managing energy demand throughout the country effectively. Besides, it will escalate the progress rate of the entire energy system along with scope of future analogical experiments.

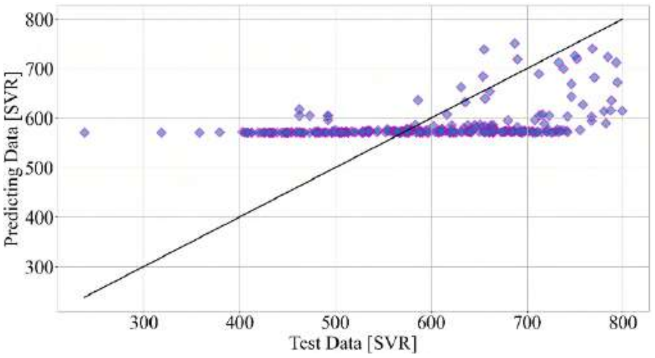


(a) MLR model

(b) KNN model

(c) RF model

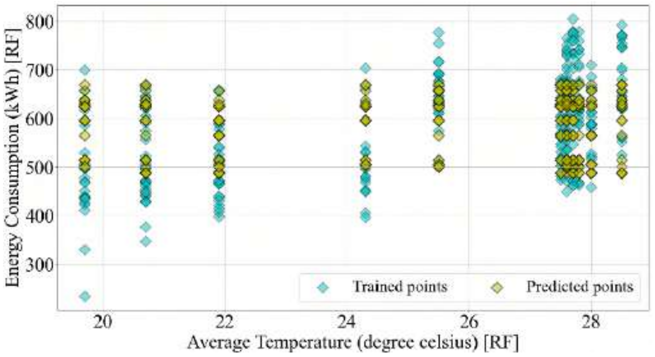
(d) SVR (polynomial kernel) model: with standardization



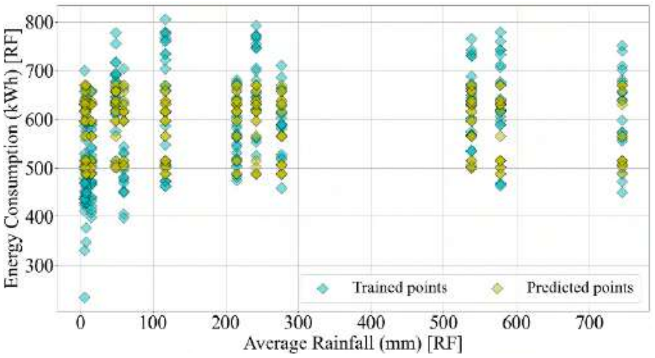
(e) SVR (polynomial kernel) model: without standardization

Fig. 1. Accuracy comparison among the models (a-e).

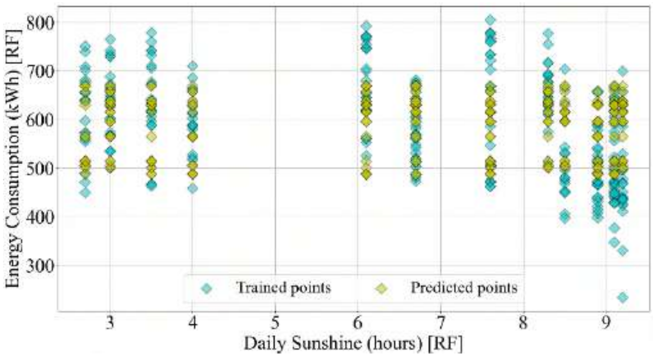
The RF method has been used to illustrate the correlation of individual variable responding with energy consumption as shown in Fig.2. Observing the change in each diagram appeared in Fig.2, dependent variables which have been considered in the dataframe according to our building parameter have steady transformation according to the energy consumption except the electricity bill (Price-BDT). Depending on the geographical features the outside weather is not so variant in nature, which leads to lower range of changing parameters. However, such patterns will cause prediction of energy consumption differently.



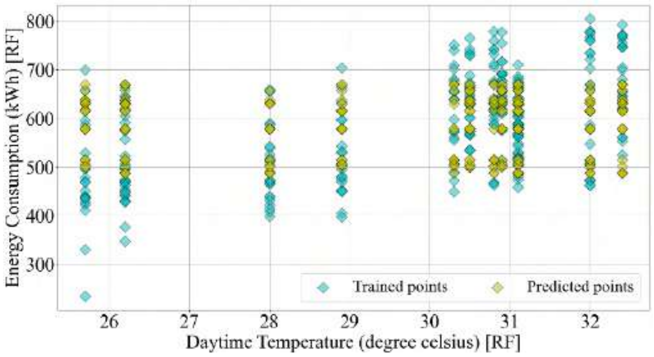
(a) Energy consumption vs average temperature



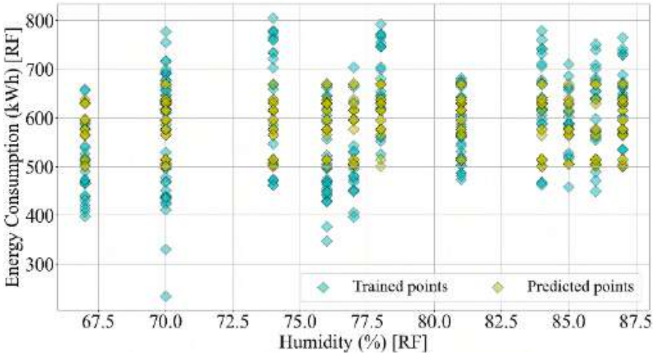
(b) Energy consumption vs average rainfall



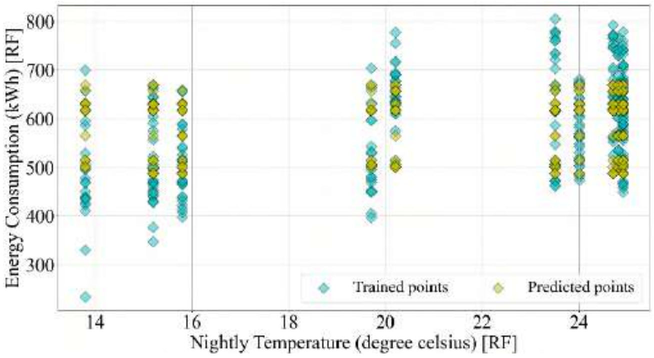
(c) Energy consumption vs daily sunshine



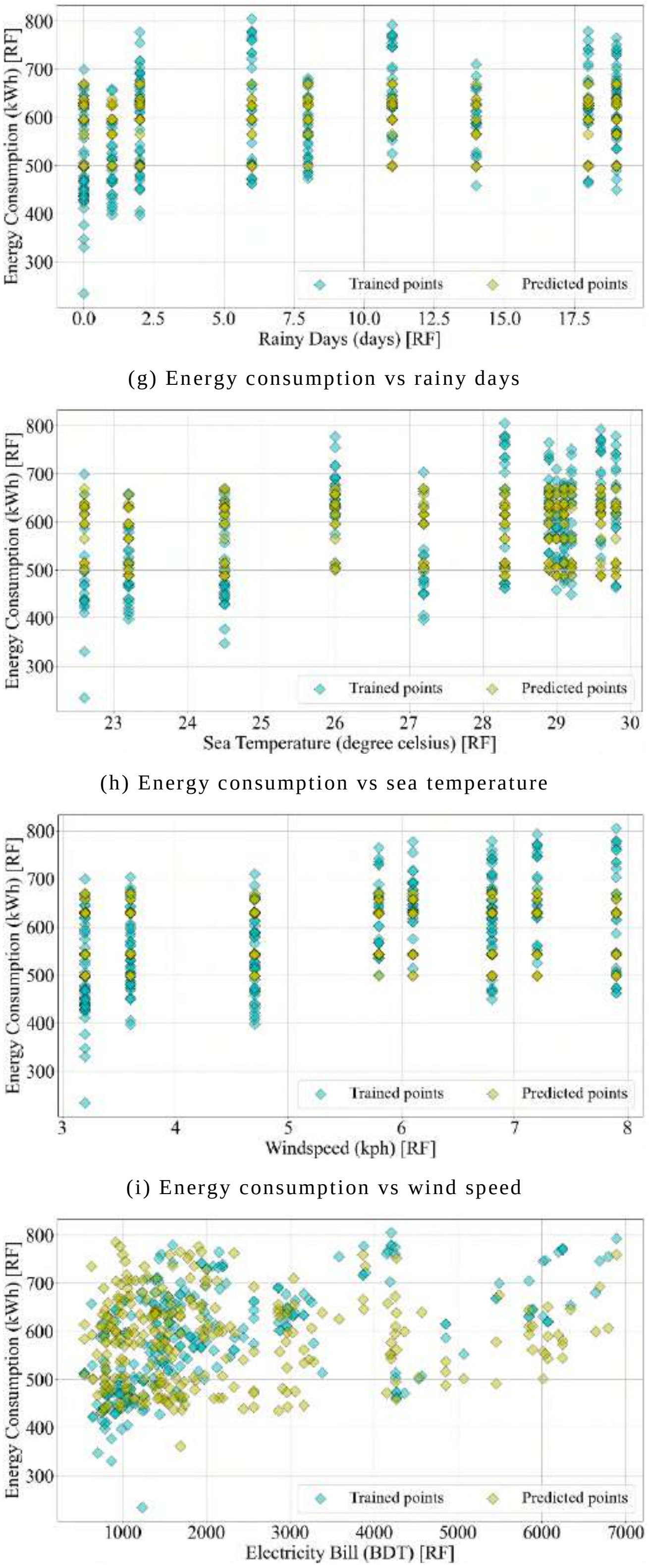
(d) Energy consumption vs daytime temperature



(e) Energy consumption vs humidity



(f) Energy consumption vs nightly temperature

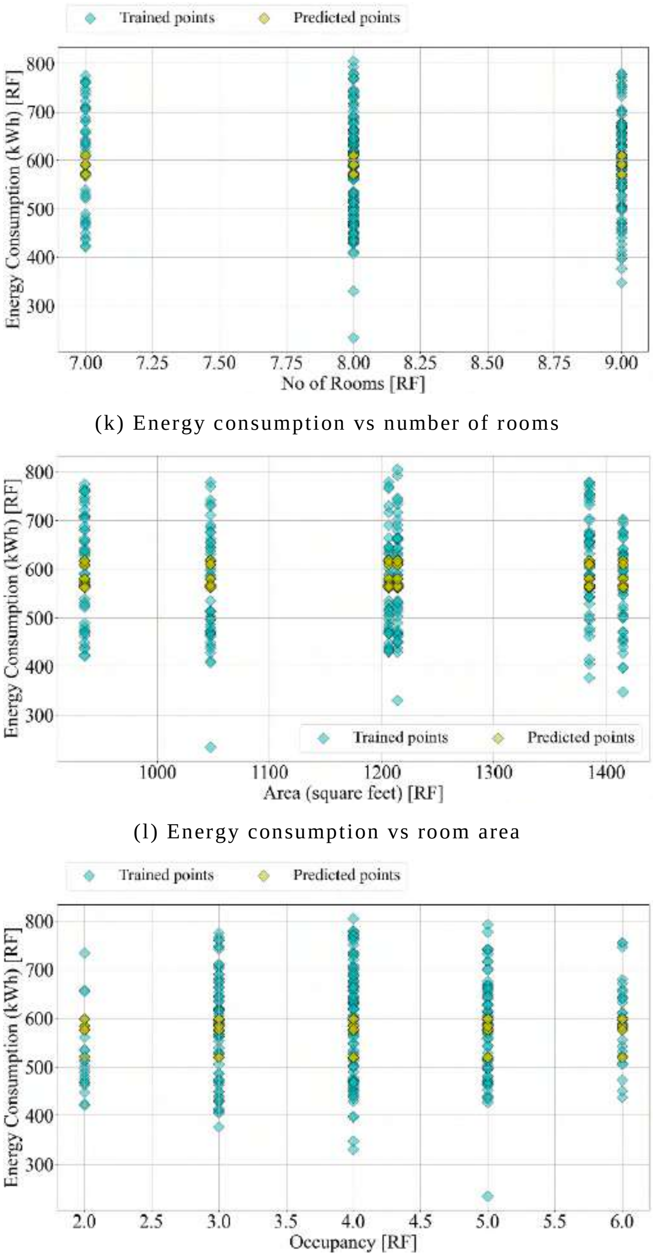


(g) Energy consumption vs rainy days

(h) Energy consumption vs sea temperature

(i) Energy consumption vs wind speed

(j) Energy consumption vs wind speed



(k) Energy consumption vs number of rooms

(l) Energy consumption vs room area

(m) Energy consumption vs occupancy

Fig. 2. Invidual correlation of input-output data (*a-m*) using RF model

V. Conclusion

The paper emphasizes on energy prediction analysis on demand purpose using different machine learning models: multiple linear regression (MLR), random forest (RF), support vector regression machine (SVR), and k-nearest neighbors (KNN). Table 2 identifies the correlations among the features that have been assembled systematically for presenting the features associated with each method. Among four models, RF can deal with the relationship among the provided feature categories. Yet other models like MLR, SVR_poly, and KNN has performed similarly for the peripherals of such energy data of building in Chittagong. Nevertheless, the performance in R^2 is not so good. All the models show poor figures which can be the reason of multi-variant nature among the independent variables corresponding to the dependent one. However, the models can be effectively applied for the prediction of demand analysis for the energy consumption.

VI. Future Work

Future study on this paper is required where more samples will be considered along with more in-depth survey of hourly data configuration for energy consumption. Other machine learning models like deep neural network, gaussian process regression, gaussian mixture regression model etc. will be taken into consideration for much extensive analysis. This paper will also acknowledge some profound features for further analyzation. Besides, the work method can also be adapted in other figures with different structures.

References

[1] Saleh Seyedzadeh, Farzad Pour Rahimian, Ivan Glesk and Marc Roper, "Machine learning for estimation of building energy consumption and performance: a review", Visualization in Engineering, 2018.

[2] Zhijian Liu, Di Wu, Yuanwei Liu, Zhonghe Han, Liyong Lun, Jun Gao, Guangya Jin and Guoqing Cao, "Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction", Energy Exploration & Exploitation, Vol. 34(4), 2019.

[3] Elena Mocanu, Phuong H. Nguyen, Madeleine Gibescu, and Wil L. Kling, "Comparison of Machine Learning Methods for Estimating Energy Consumption in Buildings", 13th International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2014.

[4] Zhengwei, Yanmin Han, Peng Xu, "Methods for benchmarking building energy consumption against its past or intended performance: An overview", Applied Energy, 2014.

[5] Safae Bourhnane, Mohamed Riduan Abid, Rachid Lghoul, Khalid Zine Dine, Najib Elkamoun, and Driss Benhaddu, "Machine learning for energy consumption prediction and scheduling in smart buildings", published online in SN Applied Sciences, 30 January 2020.

[6] Amir Mosavi, and Abdullah Bahmani, "Energy consumption prediction using machine learning; a review", Preprints, 11 March 2019.

[7] Amasyali, Kadir, El-Gohary, and Nora, "Deep Learning for Building Energy Consumption Prediction", published in Leadership in Sustainable Infrastructure, 2017.

[8] Chengdong Li, Zixiang Ding, Dongbin Zhao, Jianqiang Yi and Guiqing Zhang, "Building Energy Consumption Prediction: An Extreme Deep Learning Approach", published in Energies article, 7 October 2017.

[9] R. Mena, F. Rodríguez, M. Castilla, M.R. Arahal, "A prediction model based on neural networks for the energy consumption of a bioclimatic building", Energy and Building, 2014.

[10] Chirag Deb, Lee Siew Eang, Junjing Yang, Mattheos Santamouris, "Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks", Energy and Building, 2015.

[11] Hai Zhong, Jiajun Wang, Hongjie Jia, Yunfei Mu, and Shilei Lv, "Vector field-based support vector regression for building energy consumption prediction", published in Elsevier journal page & Applied Energy, 8 March 2019.

[12] Junjing Yang, Chao Ning, Chirag Deb, Fan Zhang, David Cheong, Siew Eang Lee, Chandra Sekhar, and Kwok Wai Tham, "k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement", Energy and Buildings, 31 March 2017.

[13] Mehrnoosh Torabi, Sattar Hashemi, Mahmoud Reza Saybani, Shahaboddin Shamshirband, and Amir Mosavi, "A Hybrid Clustering and Classification Technique for Forecasting Short-Term Energy Consumption", Environmental Progress & Sustainable Energy, 2018.

[14] Rigoberto Arambula Lara, Giovanni Pernigotto, Francesca Cappelletti, Piercarlo Romagnoni, Andrea Gasparella, "Energy audit of schools by means of cluster analysis", Energy and Buildings, 2015.

[15] Abhishek Srivastav, Ashutosh Tewari, Bing Dong, "Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models", Energy and Building, 2013.

[16] Yuna Zhang, Zheng O'Neill, Bing Dong, Godfried Augenbroe, "Comparisons of inverse modeling approaches for predicting building energy performance", Building and Environment, 2015.

Efficient acetone sensing by Pd nanoparticle loaded graphene Field Effect Transistor

Radha Bhardwaj and Arnab Hazra*

Dept. of Electrical & Electronics Engineering, Birla Institute of Technology and Science (BITS)-Pilani, Vidya Vihar, Rajasthan 333031, India

*Email: arnabhazra2013@gmail.com , arnab.hazra@pilani.bits-pilani.ac.in,
Tel: +91-1596-255724,

Abstract- In this work we have reported Pd/GO FET nanocomposite field effect transistor (FET) based acetone sensor. Pd nanoparticle loaded graphene oxide (GO) was prepared by one step spray coating technique at room temperature. The morphological and structural characterizations of developed pure GO and Pd/GO samples were performed with field emission scanning electron microscopy (FESEM), Raman spectroscopy and UV-Vis spectroscopy techniques. The effect of gate voltage on sensors at different temperature range (25- 75˚C) was investigated by I_{DS}-V_{GS} characteristic. GO and Pd/GO FET sensors showed optimum response at 50 ˚C temperature with and without applied gate voltage. The response of Pd/GO FET sensor was around 8 % under zero gate voltage (V_{GS}= 0 V) at operating temperature of 50 ˚C. Due to the application of gate voltage near Dirac point voltage ($V_{GS} \approx V_{dirac}$), both the sensors showed a significant increment in the response magnitude where pure GO exhibited 22 % and Pd/GO exhibited 45 % response in the exposure of 80 ppm acetone at 50 ˚C. The Pd/GO FET sensor showed ~6 times amplification in sensitivity as the consequence of applied gate voltage.

Keywords- Pd/GO hybrid, FET, acetone sensing, Amplified sensitivity

I. INTRODUCTION

Volatile organic compound (VOC) detection attracted a significant attention in sensing field due to their extensive application in detection of toxic, hazardous and flammable VOCs in pharmaceutical, water treatment, automotive industry, beverage industry, power plants, medical diagnosis, air quality treatment and food processing [1,2]. In a more specific and influential way, acetone is a breath marker for diabetes and halitosis disease and offer a valuable research perspective in detection of acetone in exhaled breath for disease monitoring and diagnosis [3]. In general, VOC sensor operates on a principle where surface interaction of sensing layer with target VOC results in modulation of electrical properties [4]. A variety of materials are involving as sensing layer like nanoscale semiconducting metal oxides [5,6], 2D materials, stable polymers [7] and thin metal films [8] etc. but among all of them 2D nanomaterials such as graphene, chalcogenides etc. and their composites attracted much attention due to their outstanding inherent properties such as good compatibility with environment, high specific surface area, high activation energy, room temperature stability, ballistic transport, fast response and low cost detection of VOC's[3,6, 9,10].

Since last two decades, graphene and their derivatives like graphene oxide (GO) and reduced graphene oxide (RGO) based gas/vapor sensors have been explored significantly due to their exciting properties like extremely high effective surface area, high mobility, electrical conductivity and easy functionalization properties [11]. However, research data also envisage that surface modified graphene has a greater number of defects and impurities which drastically improved the sensitivity and other sensing performances [3,6,9]. Incorporation of other sensing materials in GO improves its gas sensing properties and offers high sensitivity, fast response and recovery, long term stability etc. More significantly, GO based nanocomposites showed highly selective behaviour toward a target VOC. Bo Zhang et al reported RGO/α-Fe_2O_3 based acetone sensor and showed that mixing of RGO with metal oxide results in fast response and recovery with moderate sensitivity at 100 ppm but at high operating temperature (225˚C) [5]. Similarly, Pengpeng Wang et al reported ZnO/GO sensor for breath acetone detection with 35.8 % response for 100 ppm of concentration at relatively high operating temperature (240˚C) [9]. Noble metals (Ag, Au, Pd, Pt) nanoparticles functionalization on

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

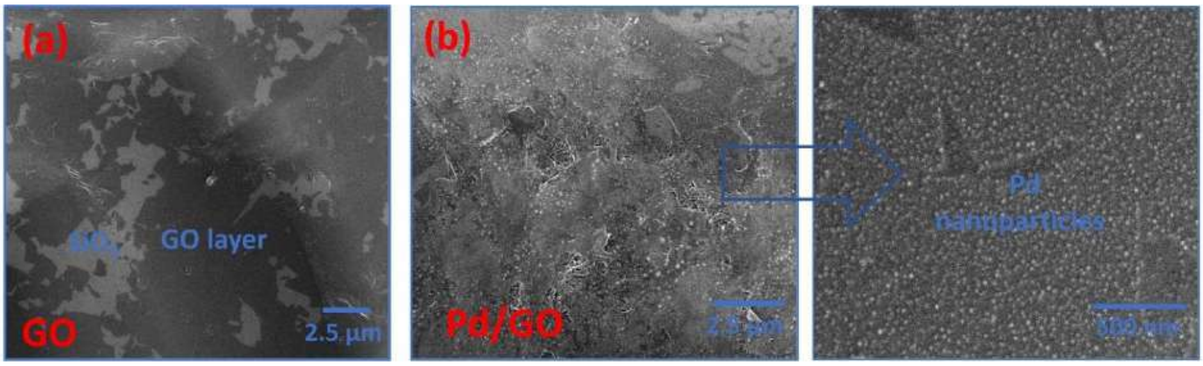


Fig. 1 FESEM images of pure GO (a), Pd/GO (b) and high-resolution image of Pd/GO samples on Si/SiO_2 substrate.

GO sensing layer influence sensing performance by their electrical and chemical sensitization mechanism [10, 12-15]. Ali Esfandiar and coworkers have shown the influence of Pd nanoparticles on WO_3/GO composite in hydrogen sensing performance. The sensor response was good enough with moderate operating temperature (100 ˚C) with outstanding response and recovery (less than 1 min) [14].

Besides the above mentioned advantages of GO composite, graphene and its derivatives also proved its potential in field effect transistors (FET) structure for efficient gas/VOC sensing.

Owing to the tuneable bandgap, limited majority carrier concentration, pure 2D structure, graphene and its derivatives exhibit field controlled current conduction in it [16-18]. Current in the graphene oxide channel can modulate significantly by applying suitable gate voltage in FET configuration that significantly improves the gas/vapor sensing performance of GO FET. A limited number of reports have been found on GO based FET system for gas sensing purpose [19,20].

In this work, we are reporting Pd/GO FET sensor system for the detection of acetone vapor at temperature range 25-75 ˚C. Here our main focus to study the effect of Pd nanoparticles loading on GO channel and the application of back gate bias voltage on acetone sensing performance of Pd/GO FET. The GO and Pd/GO FET sensors were fabricated on SiO_2/Si substrate by one step spray coating method. The attachment of Pd nanoparticles on GO surface was investigated by using scanning electron micrograph which demonstrated uniform distribution of GO sheets and as well as Pd nanoparticles with no aggregation. The effect of gate voltage on sensors at different temperature was investigated with I_{DS}-V_{GS} characteristics study and the effect of gate voltage on sensing performance of sensors was examined by exposing them to 80 ppm acetone at applied V_{GS} and different temperatures. In last, a possible sensing mechanism of proposed sensors was discussed.

II. EXPERIMENTAL

(a) Synthesis of nanocomposites

0.2 wt% (>80%, flake size: 0.5-5 µm; purchased from graphene supermarket) graphene oxide (GO) aqueous solution was prepared and further solution was ultrasonicated for 1 h for better dispersion. 1 M noble Pd nanoparticles solution was prepared by adding 0.17 gm palladium chloride ($PdCl_2$) into 1 L deionized water with continuous stirring at room temperature. Later, Pd nanoparticles solution was stabilized by adding diluted HCL dropwise with continuous sonication.

Boron-doped, ~500 µm thick <100> SiO_2/Si substrate having SiO_2 thickness of 90 nm was used as a substrate for the nanocomposite formation with resistivity of Si in the range of 0.001-0.005 Ω-cm. 10-20 µl GO solution was deposited on substrate by spray coating technique in air at the room temperature and sample is marked as S1. On the other hand, Pd/GO nanocomposite was synthesized by depositing 10-20 µl GO dispersion and then Pd nanoparticle solution in the same amount on substrate at room temperature and marked as S2.

After drying the samples, a thermal annealing was carried out on both the samples for 4 h at 250 ˚C temperature to achieve the thermal stability.

(b) Device fabrication

To fabricate the back gated field effect transistors, both top and bottom electrodes was deposited. Back gate contact was taken by selective removal of backside (1 mm×1 mm) SiO_2 layer of the substrate by HF etching with the help of negative photoresist. Top Au source and drain electrodes of thickness 150 nm was deposited by electron beam evaporation unit by using Cu physical mask.

(c) *Material characterizations*

Morphology of both the samples was characterized by field emission scanning electron microscopy (FESEM) technique. Structural properties and bandgap of samples was identified by Raman spectroscopy and UV-Vis spectroscopy techniques, respectively.

(d) Sensing characterization

GO and Pd/GO FET sensors were characterized by placing them inside a closed glass chamber with volume 450 ml. The glass chamber was then placed on a heating coil to vary the temperature from 25 to 75 ˚C. The sensors were characterized on a static mode gas sensing technique, where acetone was injected by micro syringe (Hamilton 705RN 50UL SYR) and sensors were recovered by continuous flowing of 450 SCCM air by mass flow controller (MFC). The reading of FET sensors was taken by the help of two source meters (keithley 6487). To measure the injected acetone Concentration (volume: V_1); C (ppm)=2.46 × (V_1D/VM) × 10^3 relation was used, where D (gm/mL), M (gm/mol) and V (Lit) represent density of liquid, molecular weight of liquid and volume of vaporization chamber respectively [21,22]. Response magnitude (RM) of the sensors was calculated by using the formula RM= $[(I_v- I_a)/I_v]\times100$ where I_a and I_v were the currents of the sensor in the air and acetone at operating temperature. The sensors were characterized at a constant (V_{DS}= 1 V) bias voltage throughout the study with varied gate voltages (V_{GS}).

III. RESULTS AND DISCUSSIONS

(a) Material Characterizations

Morphological study of both the samples done by FESEM as shown in the Fig. 1. Fig. 1(a) clearly demonstrated very thin layered GO deposition in a continuous manner. Pure graphene oxide image showing discontinuity into the ordered structure and random aggregation of layers and forming a uniform distribution (Fig. 1a). Fig. 1b is showing Pd nanoparticle distributed on GO layer which is placed on SiO_2/Si substrate. Uniform, compact and continuous distribution of nanoparticles was observed in S2 (Fig. 1b). High resolution image of Pd/GO represented by arrow, demonstrating average diameter of Pd nanoparticles around 20-30 nm.

Raman spectroscopy scan from 1000 to 2500 cm^{-1} is displaying in (Fig. 2a) to identify the defect density in samples and to compare the change in GO properties after addition of noble Pd nanoparticles. Two characteristic peaks of graphene were found in pure GO sample at around 1344 cm^{-1} (D band) and 1608 cm^{-1} (G band). Both the peaks were also found in Pd/GO sample with slight shift in peaks towards the lower wave number at 1337 (D band) and 1590 cm-1(G band) as shown in Fig. 2a. Shifting in peaks may be due the strong bonding and carrier transfer between Pd nanoparticle and GO. The D peak to G peak intensity ratio in pure GO was 0.92 and in Pd/GO was 1.39. The increase in relative intensity of D peak reveals increase in number of disorders due to the incorporation of Pd nanoparticles (Fig. 2a).

UV-Vis spectra of pure GO and Pd/GO samples is showing in Fig. 2b. Absorbance is high in Pd loaded sample compare to the pure GO as depicted in UV Vis spectra in Fig. 2b. Pd nanoparticle loaded sample has an absorption edge at 274.8 nm while pure graphene oxide sample showing its absorption edge at 330 nm. Pure GO has a bandgap of 3.7 eV and their mixing with Pd nanoparticles (noble metal)

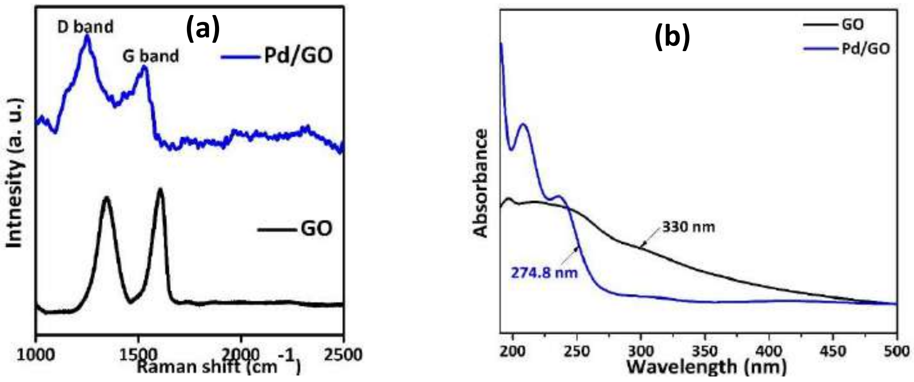


Fig. 2 (a) Raman spectra and (b) UV-vis spectra of pure GO, Pd/GO on Si/SiO_2 substrate.

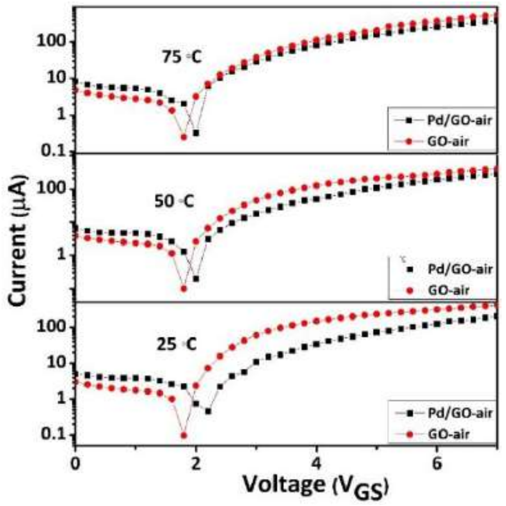


Fig. 3 I_D- V_{GS} characteristics of both FET sensors i.e GO and Pd/GO in air under V_{DS}= 1 V at a temperature range 25 °C - 75 °C.

show modification in bandgap and resulting bandgap was 4.5 eV in Pd/GO nanocomposite.

(b) I-V Characteristics of GO and Pd/GO FET sensors at different temperatures

Fig. 3 is showing I_{DS}-V_{GS} characteristic of both the GO and Pd/GO FET sensor in air for temperature range 25 ˚C- 75 ˚C. Modulation in the baseline current from GO to Pd/GO FET sensor was due to the conductivity difference between GO (*p*-type) and Pd nanoparticles (*n*-type). In Pd/GO FET sensor, accumulation/depletion of charge carriers between GO and Pd nanoparticles show shift in device current at (V_{GS}= 0V) and simultaneously in dirac point towards right as shown in the fig. 3. GO has very low majority charge carrier concentration and at applied V_{GS} the drain current start decreasing and at a particular voltage the drain current (I_{DS}) was found lowest with negligible carrier concentration. The value of V_{dirac} was related to the carrier concentration between source and drain terminal. As showing in the fig. 3, at each temperature range the value of the dirac point is contact for the both samples; for GO observed dirac point voltage was 1.8 V and for Pd/GO was 2.2 V. After dirac point the conductivity of the samples get changed as displayed in fig. 3.

(c) Acetone sensing performance with two terminals (V_{GS}= 0 V)

GO and Pd/GO FET sensors were tested for 80 ppm of acetone at (V_{GS}= 0 V) under the temperature range 25 to 75 ˚C as shown in Fig. 4. Response of the samples were increased linearly with increasing temperature and get saturated at 75 ˚C. At 25 ˚C, pure GO FET sensor was showing 4 % response with incomplete recovery and on the other side Pd/GO FET sensor was showing slightly high response around 6 % at V_{GS}= 0 V, as shown in Fig. 4a. However, 50 ˚C was considered as optimized temperature due to the favourable response for Pd/GO FET sensor around 8 % compare to the other temperature at V_{GS}= 0 V as shown in Fig. 4b. At higher temperature (75 ˚C), the response value of both the sensors gets decreased, from (fig. 4c). Moreover, Pd/GO showed high response magnitude for all the temperature range compare to pure GO at zero gate voltage possibly due to the electrical and chemical sensitization effect of Pd nanoparticles [23,12].

(d) Acetone sensing performance with three terminals (V_{GS}= Vdirac)

Fig. 5 shows the sensitivity modulation under the effect of applied gate voltage on GO and Pd/GO FET sensors for 80 ppm acetone and for the temperature range from 25 to 75 ˚C. Both the sensors were optimized under applied gate to source voltage which is very closed to the Dirac point voltage and have a different value for both the samples; for GO applied V_{GS}= 1.7 V and for Pd/GO applied V_{GS}= 1.9 V. The value of response magnitude significantly

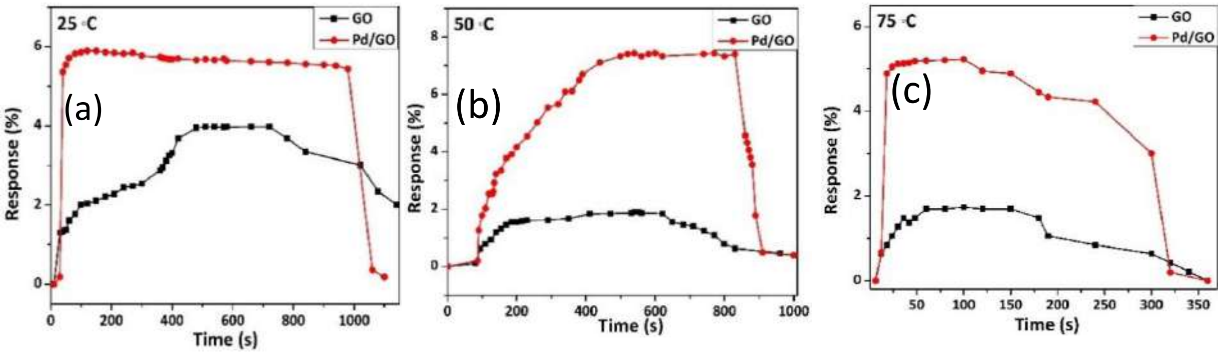


Fig. 4 The response variation curve of both FET sensors; GO and Pd/GO at 80 ppm acetone exposure with zero gate voltage at operating temperature (a) 25 °C, (b) 50°C and (c) 75 °C.

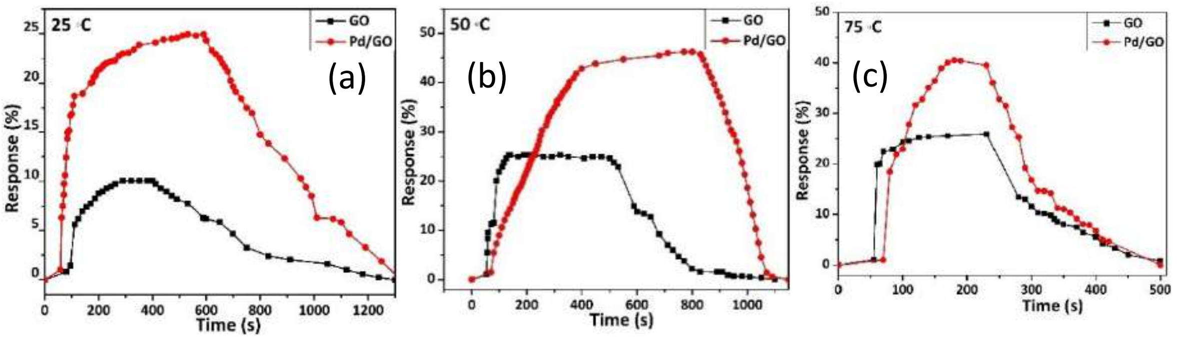


Fig. 5 (a) The response variation curve of both FET sensors; GO and Pd/GO at 80 ppm acetone exposure with corresponding applied gate voltage at operating temperature (a) 25 °C, (b) 50°C and (c) 75 °C.

get increased at applied V_{GS} and increase linearly with the applied temperature (Fig. 5). Here also the highest response value found at 50 ˚C applied temperature and showing response magnitude of 22 % and 45 % for GO and Pd/GO FET sensors respectively, at their corresponding V_{GS} as shown in (Fig. 5b). Moreover, response magnitude of sensors was comparably very high at applied V_{GS} at all the temperature range than zero gate bias. In contrast, Fig. 5 confirms that applied back gate voltage plays a key role in sensitivity modulation for both the sensors even at very low operating temperature (25 ˚C).

The general mechanism in acetone sensing involves chemisorption of oxygen groups (O^{2-}, O^{-}) on sensing material and also Pd/GO sensor form a junction between Pd nanoparticles and GO due to their work function difference and show a change in device current [5,9,12]. Introduction of reducing vapor, acetone molecules reacted with pre-absorbed oxygen species and the trapped electrons by oxygen molecules get released back to the conduction band of sensing layer and show modulation in device current. In the influence of gate voltage, the modulation in current value will be very high as we found in Fig. 5, due to the low carrier density in the channel at $V_{GS} \approx V_{dirac}$ and any change in carrier concentration by the analyte gas become very significant.

IV. CONCLUSION

Herein, we have prepared two sensors; GO and Pd/GO. Pd/GO FET sensor was showing efficient acetone sensing characteristics in the influence of applied gate voltage and Pd nanoparticles. FESEM image confirms the decoration of Pd nanoparticles on GO layer without an aggregation and discontinuity. Raman spectroscopy showed the characteristic peaks of GO in both the samples and Pd nanoparticles incorporation in GO results in peak shift. The ambipolar current-voltage behaviour of GO was visualized through the I_{DS}-V_{GS} characteristics at a temperature range 25-75 ˚C in air. An amplified sensitivity was found in both GO and Pd/GO under the effect of V_{GS}= 1.7 V and 1.9 V for GO and Pd/GO, respectively. Pd/GO FET sensor amplified response was ~6 times higher than the response obtained in two terminals configuration (V_{GS}= 0 V) for 80 ppm acetone at 50 ˚C operating temperature.

References

[1] N. Yamazoe, “Toward innovations of gas sensor technology,” *Sensors Actuators, B Chem.*, vol. 108, no. 1-2 SPEC. ISS., pp. 2–14, 2005, doi: 10.1016/j.snb.2004.12.075.

[2] K. D. Cleaver, “The analysis of process gases: A review,” *Accredit. Qual. Assur.*, vol. 6, no. 1, pp. 8–15, 2001, doi: 10.1007/s007690000179.

[3] D. Zhang, A. Liu, H. Chang, and B. Xia, “Room-temperature high-performance acetone gas sensor based on hydrothermal synthesized SnO2-reduced graphene oxide hybrid composite,” *RSC Adv.*, vol. 5, no. 4, pp. 3016–3022, 2015, doi: 10.1039/c4ra10942b.

[4] G. F. Fine, L. M. Cavanagh, A. Afonja, and R. Binions, “Metal oxide semi-conductor gas sensors in environmental monitoring,” *Sensors*, vol. 10, no. 6, pp. 5469–5502, 2010, doi: 10.3390/s100605469.

[5] B. Zhang *et al.*, “Enhanced gas sensing properties to acetone vapor achieved by α-Fe2O3 particles ameliorated with reduced graphene oxide sheets,” *Sensors Actuators, B Chem.*, vol. 241, pp. 904–914, 2017, doi: 10.1016/j.snb.2016.11.023.

[6] J. Kaur, K. Anand, A. Kaur, and R. C. Singh, *Sensitive and selective acetone sensor based on Gd doped WO3/reduced graphene oxide nanocomposite*, vol. 258. Elsevier B.V., 2018.

[7] S. Pirsa, “Chemiresistive gas sensors based on conducting polymers,” *Mater. Sci. Eng. Concepts, Methodol. Tools, Appl.*, vol. 1–3, pp. 543–574, 2017, doi: 10.4018/978-1-5225-1798-6.ch022.

[8] A. Pundt, “Hydrogen in nano-sized metals,” *Adv. Eng. Mater.*, vol. 6, no. 1–2, pp. 11–21, 2004, doi: 10.1002/adem.200300557.

[9] P. Wang *et al.*, “ZnO nanosheets/graphene oxide nanocomposites for highly effective acetone vapor detection,” *Sensors Actuators, B Chem.*, vol. 230, pp. 477–484, 2016, doi: 10.1016/j.snb.2016.02.056.

[10] S. Ge *et al.*, “Ag/SnO2/graphene ternary nanocomposites and their sensing properties to volatile organic compounds,” *J. Alloys Compd.*, vol. 659, pp. 127–131, 2016, doi: 10.1016/j.jallcom.2015.11.046.

[11] C. M. Yang, T. C. Chen, Y. C. Yang, M. Meyyappan, and C. S. Lai, “Enhanced acetone sensing properties of

monolayer graphene at room temperature by electrode spacing effect and UV illumination," *Sensors Actuators, B Chem.*, vol. 253, pp. 77–84, 2017, doi: 10.1016/j.snb.2017.06.116.

[12] R. Bhardwaj, V. Selamneni, U. N. Thakur, P. Sahatiya, and A. Hazra, "Detection and discrimination of volatile organic compounds by noble metal nanoparticle functionalized MoS2coated biodegradable paper sensors," *New J. Chem.*, vol. 44, no. 38, pp. 16613–16625, 2020, doi: 10.1039/d0nj03491f.

[13] P. Bindra and A. Hazra, " Electroless deposition of Pd/Pt nanoparticles on electrochemically grown TiO 2 nanotubes for ppb level sensing of ethanol at room temperature ," *Analyst,* no. 2, 2021, doi: 10.1039/d0an01757d.

[14] A. Esfandiar, A. Irajizad, O. Akhavan, S. Ghasemi, and M. R. Gholami, "Pd-WO3/reduced graphene oxide hierarchical nanostructures as efficient hydrogen gas sensors," *Int. J. Hydrogen Energy*, vol. 39, no. 15, pp. 8169–8179, 2014, doi: 10.1016/j.ijhydene.2014.03.117.

[15] L. Chen *et al.*, "Fully gravure-printed WO3/Pt-decorated rGO nanosheets composite film for detection of acetone," *Sensors Actuators, B Chem.*, vol. 255, pp. 1482–1490, 2018, doi: 10.1016/j.snb.2017.08.158.

[16] S. Su *et al.*, "Two-dimensional nanomaterials for biosensing applications," *TrAC - Trends Anal. Chem.*, vol. 119, p. 115610, 2019, doi: 10.1016/j.trac.2019.07.021.

[17] M. Jin, H. K. Jeong, W. J. Yu, D. J. Bae, B. R. Kang, and Y. H. Lee, "Graphene oxide thin film field effect transistors without reduction," *J. Phys. D. Appl. Phys.*, vol. 42, no. 13, 2009, doi: 10.1088/0022-3727/42/13/135109.

[18] A. Hazra and S. Basu, "Graphene Nanoribbon as Potential On-Chip Interconnect Material—A Review," *C*, vol. 4, no. 3, p. 49, 2018, doi: 10.3390/c4030049.

[19] R. Pearce, T. Iakimov, M. Andersson, L. Hultman, A. L. Spetz, and R. Yakimova, "Epitaxially grown graphene based gas sensors for ultra sensitive NO 2 detection," *Sensors Actuators, B Chem.*, vol. 155, no. 2, pp. 451–455, 2011, doi: 10.1016/j.snb.2010.12.046.

[20] A. Hazra, "Amplified Methanol Sensitivity in Reduced Graphene Oxide FET Using Appropriate Gate Electrostatic," *IEEE Trans. Electron Devices*, vol. 67, no. 11, pp. 5111–5118, 2020, doi: 10.1109/TED.2020.3025743.

[21] P. Bindra and A. Hazra, "Selective detection of organic vapors using TiO2 nanotubes based single sensor at room temperature," *Sensors Actuators, B Chem.*, vol. 290, no. January, pp. 684–690, 2019, doi: 10.1016/j.snb.2019.03.115.

[22] T. Gakhar and A. Hazra, "Oxygen vacancy modulation of titania nanotubes by cathodic polarization and chemical reduction routes for efficient detection of volatile organic compounds," *Nanoscale*, vol. 12, no. 16, pp. 9082–9093, 2020, doi: 10.1039/c9nr10795a.

[23] M. S. Barbosa, P. H. Suman, J. J. Kim, H. L. Tuller, and M. O. Orlandi, "Investigation of electronic and chemical sensitization effects promoted by Pt and Pd nanoparticles on single-crystalline SnO nanobelt-based gas sensors," *Sensors Actuators, B Chem.*, vol. 301, no. May, p. 127055, 2019, doi: 10.1016/j.snb.2019.127055.

Cyber Physical Systems, a New Challenge and Security Issues for the Aviation

Faisal Alrefaei
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, FL, USA
alrefaef@my.erau.edu

Abdullah Alzahrani
Department of Electrical and Computer Engineering
Oakland University
Rochester, MI, USA
alzahrani2@oakland.edu

Houbing Song
Department of Electrical Engineering and Computer Science
Embry-Riddle Aeronautical University
Daytona Beach, FL, USA
houbing.song@erau.edu

Mohamed Zohdy
Department of Electrical and Computer Engineering
Oakland University
Rochester, MI, USA
zohdyma@oakland.edu

Salma Alrefaei
Department of Business Administration
Taibah University
Yanbu, Medina, KSA
Rriiif09@gmail.com

***Abstract*—The aviation sector has been experiencing an increased revolution in the last century. It becomes a vital transporter for different purposes such as transport, cargo, delivery, and military purposes. The importance of aviation attracts governments and researchers to focus intensely on the aviation sector to improve aviation modernity and be robustness. Recently, Cyber-Physical Systems (CPS) have emerged as intelligent embedded systems to improving areas such as aviation, smart grid, power systems, and aerospace sectors. Therefore, CPS has become prominent and eligible for the aviation sector to improve productivity and efficiency to meet the stakeholder's satisfaction. Consequently, applying the CPS to the aviation sector opens new challenges related to system security caused by its operation neutrality. This paper innovative a review Aviation Cyber-Physical Systems (CPSs) framework in terms of creating secure systems. Moreover, it discusses what are the possible policies and solutions to make this system robust and secure.**

***Index Terms*—Aviation systems, cyber threats, cyber attacks,cyber-physical systems, wireless sensor network, cyber security, physical layer security.**

I. Introduction

The aviation sector developed in the last century and played a crucial role in improving aviation's different aspects. Aviation is used for various purposes, such as transporting people, cargo or military uses to defend or attack. These facts attracted the governments, organizations, enterprises, and researchers to improve the aviation sector to meet their purpose. Productivity, efficiency, and controllability are the characteristics that are required. Recently, emerging the cyber physical system (CPS) attracted their intention of using this new intelligent embedded system in the aviation and aerospace sector. The applying of the CPS met their requirements and drove the aviation sector to its goals. The nature of the CPS is the integration cyber world with the physical world. That integration provides benefits and opened up the challenges represented by the integration's security issues and vulnerability. Aviation uses to transport passengers, cargo, and military purposes. This revolution makes using aviation is an integral part of future intelligent plans, modernization of the organization and government in the upcoming century. Besides, the aviation management system helps in this revolution and makes the aviation sector easier to use with less cost. These days Aviation cyber-physical systems (ACPS) integrate the cyber layer such as computation, network, digital storage with the physical layer such as sensor, actuator, camera, etc. [1]. This integration makes ACPS become an intelligent embedded system. ACPS is acquiring valuable characteristics such as automation, efficiency, controllability, productivity, which must be presented in the intelligent systems. In contrast, integrating these systems and applying them in aviation opened up some challenges, such as security issues and vulnerability. Therefore, the security of the ACPS must be addressed and investigated.

Security in the aviation cyber-physical systems must be addressed in the earlier design. The importance of the security of the ACPS is crucial, where the potential threat attacks are presented as threats to the other organization. This paper focuses on the security issues raised in the physical layer to the application layer through the network layer. besides, it shows in detail the typical type of attacks that execute on the cyber and physical components with their impacts and the solution to mitigates the threats.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

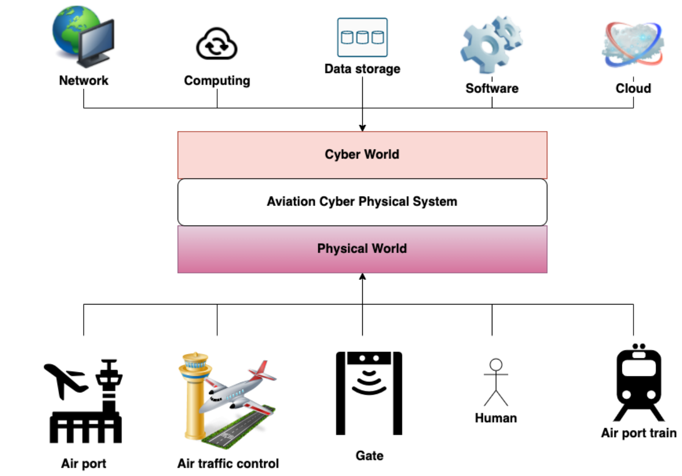


Fig. 1. Integrating cyber space with physical world

II. The importance of Security in the Aviation Cyber-Physical Systems.

The urgent needs and the nature of aviation's interoperation make the security considerations in the first line of the requirements. The ACPS is integrated into the two primary layers. Each one of them has a specific attack technique that occurs either active attack or passive. The cyber layer is represented by any intangible components such as a network, software, data storage, etc [1]. During transmission data, command signals, or measurement, this transmission expose to be damaged or data leak. On the other hand, the hacker intends to execute their malicious action on the sensor or actuator nodes in the physical layer. Therefore, Defense strategies and attack detection play a pivotal role in protecting commands and data transferred by the network.

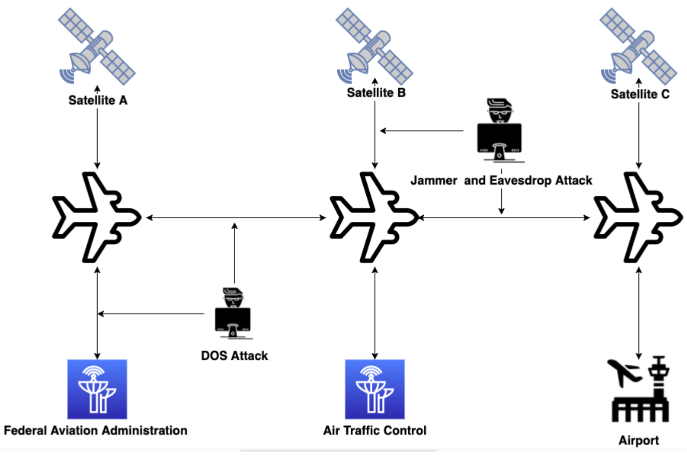


Fig. 2. Several potential paths of cyberattackss

The nature of the propagation of signal over the ACPS network is exposed to cyber-attacks. The security mechanisms are applied to the sensitive part in aviation such as cargo, passenger screen no-fly passengers, gates aircraft, etc., disclosing that information to the industry or airlines in the risk. The hackers can use some intelligent techniques to compromise the system.

III. Aviation cyber-physical systems Architecture

The basic concept of the ACPS is built on integrating the cyber layer with the physical world. This integration allows logically combining two layers to constitute the ACPS. The main three parts of the ACPS are the Application layer, Network Layer and Physical Layer.Each one of these layers its functionality relies on the one below [13]. Therefore, any compromising to any layer will affect all systems or at least one layer. Also, the operation way of ACPS is constrained, so; real-time plays a vital role in this performance. Securing each one of them is not a proper technique; the technique has to cover all security issues in all layers. The security technique must cover every single threat, bug, or vulnerability presented by the malicious attack. This section shows what the main layers that constitute the ACPS with their functionality . Also, the proper protection technique must be using to present stability to the system:

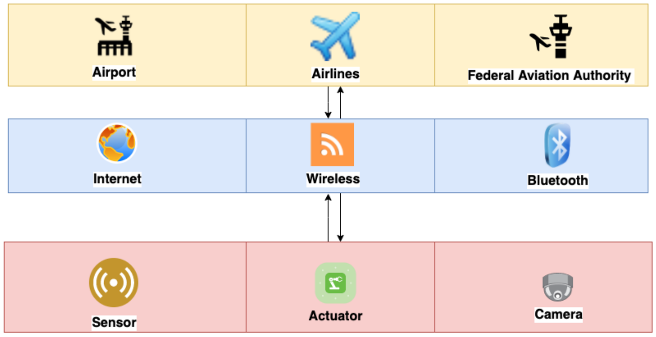


Fig. 3. ACPS architecture

1) Application Layer: The application layer is the layer that stores, analysis, and adjusts the physical world to the desired behavior. It consists of the controller of the system that decides and releases the commands based on the measurements that come from the Physical layer (source) [14]. The Application layer executed a complex algorithm on the aggregated data and transferred it back to the physical layer based on the system's status [2]. These layers also have sensitive data such as the airlines, passengers, airport, etc.

Attack on the application layer has an enormous impact on the entire system. In this layer, attacks are intended to damages the data either by violating privacy. It happens by leaking the data or unauthorized to this layer and manipulate the commands released to the physical layer[2]. Unauthorized access to this layer means that the system is driven by hackers who maliciously intended to damages the system for their purposes and goals. The type of attacks executes in this layer malicious code installed in the software deceptively or running buffer overflow against the resource to violate the availability and make the system unavailable [2]. Stack holders, organizations, and governments focus on this layer because it controls a physical layer's operation and service.

2) Transmission Layer: The transmission layer is the layer located between the application layer and the physical layer. The data in this layer fragment into small pieces calling as packets [2]. The packet is transferred and transmit, And the transmission layer rotates. This layer's data is transferred either via Bluetooth, local area network, wireless, or the Internet. The Internet is most commonly used to protect the transmitted data [2].

Security issues arise in this layer during the transmission, especially over wireless communication. The nature of the wireless communication propagating the data are suspicious to cyber attacks. In this layer integrity, confidentiality is violated as a result that the hacker can intercept the traffic and modify the data and inject commented data.

A. *Physical Layer*

The physical layer is the layer that has sensors to gathering information and acting on the physical world to adjust the system to work in the expected behavior. The aggregated data can be collected either by the sensor, camera light, etc [2] . the aggregated data is sent to the application layer to decide and release commands to be executed by the actuator.

The security issues in this layer are various, and the hacker can exploit some weaknesses in this layer. The limitation of the computation resources and memory are the main target of this equipment [15]. The hackers can use malicious action to consume the power or compromised the equipment, and sent an inaccurate measurement [16]. This equipment's nature is preferable for the hacker where the hacker can use the natural of wireless and damage the link between the sensor and the application layer.

IV. Cyber Attack on the Aviation Cyber-Physical System and it Suggested Solution

The importance of using aviation across the world makes it a primary goal for the hacker. They use an intelligent technique to disclose data and identify the vulnerabilities in the system components[17]. The hackers use a different type of attack to get their damage tasks completed.

According, to the architecture of the ACPS, it is integrated between the Cyber and Physical worlds. Each layer must be applied by an appropriate mechanism to protest their assessments, which can be, measurements, command signals, or equipment. The cyber layer presents vulnerabilities, such as malicious code, spoofing packets, and network protocol. In the physical world, the equipment's vulnerabilities or execution throw the jamming attack [10]. Therefore, the security engineer should address the security in all ACPS aspects where the influence compromising the cyber layer affects the physical layer directly. This section has shown the common cyber-attacks on the CPS.

A. *Evasdroping Attack*

This type of attack works on intercepting any data transferred over the system network. It does not need pre-knowledge about the system; it is just listing to the traffic. The eavesdropping attack does not do much to the system, such as a damaging or catastrophic consequence. It gathers valuable information about the aviation sector assessments either in the physical or cyber layers [18]. In the united state all planes mandatory to be equipped with Automatic Dependant-Broadcast (ADS-B). Its system is classified into two parts ADS-B in and ADS-B out [3]. This system broadcast unencrypted data of the airplane, such as location, altitude, speed, weather, etc [3]. This broadcasting allowing anyone can intercept these data and used them for a different purpose. Therefore, this kind of attack can gather data about the target to execute the attacks.

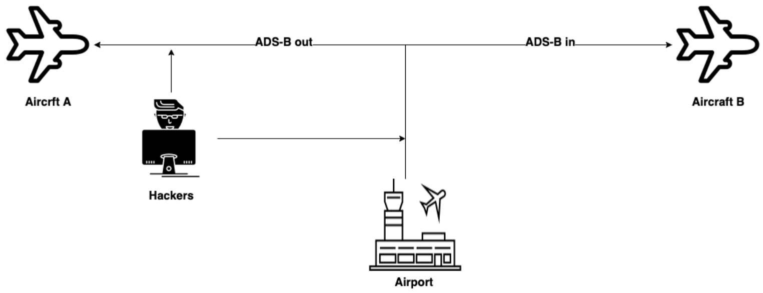


Fig. 4. Eavesdropping attack

B. *Denial of Service (DOS)*

Denial of Service is a typical attack that targets the network to make the system unavailable [2]. The hackers disrupt the network and service to make them inaccessible. The hacker to execute this attack has some techniques to achieve the attack [19]. For example, the hacker can target some airport resources, such as preventing the air traffic management system from sending their info to the airlines, planes, or stakeholders. Also, the hacker has another technique by filling the buffer of the airport systems to degrade the system performance to make them interact slowly until being shouting down useless.

C. *Jamming Attack*

A jamming attack is considered an active attack and works to disrupt the communication either by block the communication or interfered with it [8]. It has two main targets either modification the data stream or generating a false data stream [20]. The hackers block the links between the sensors node, so the measurement is blocked to be transmitted to the controller to operate services such as assigning gates, use baggage systems, and coordinate traffic on the runway.

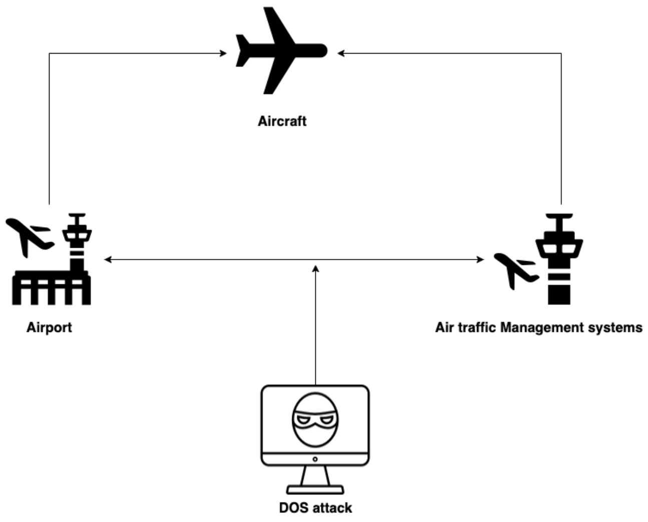


Fig. 5. DOS attack

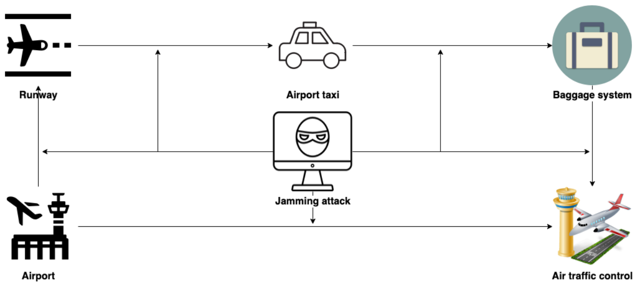


Fig. 6. Jamming Attack

D. Replay Attack

A replay attack is an attack that retransmits the data packet maliciously to repeat the measurements or signal commands to degrade the system [4]. Also, it aims to delay the valid measures from source to destination by repeating the packets [21]. These attacks compromise the sensor and actuator to execute their malicious action.Such as repeat the command to the aircraft to stay longer on the runway.

E. False Data Injection

A false data injection attack is an attack that violates the integrity of the packet data transmitted over the network; what hackers need in the malicious attack first unauthorized access to the network and start listening to the data traffic [5]. It does not require any knowledge about the system [5]. After that, the hacker injects its malicious data to drive the system to undesirable performance. False data injection can be used against the ground station or the aircraft [22]. The ground station packet injected directly affects the route and the arrival time by misleading the pilot and plane system control. This injection data drives aircraft to catastrophic consequences. The aircraft data packet injected broadcast fabricated ADS-B signals to the aircraft around or to the ground station [9]. Moreover, the connection link between the surrounding air crafts and the ground station might be injected. The destinations will not be able to define the valid data packet among the enormous received.

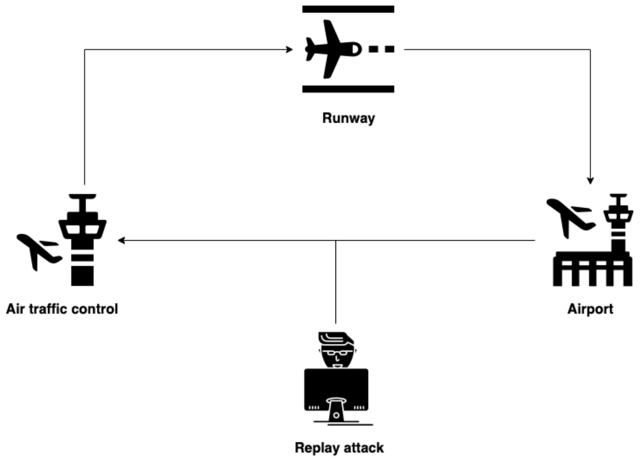


Fig. 7. Shows Replay Attack

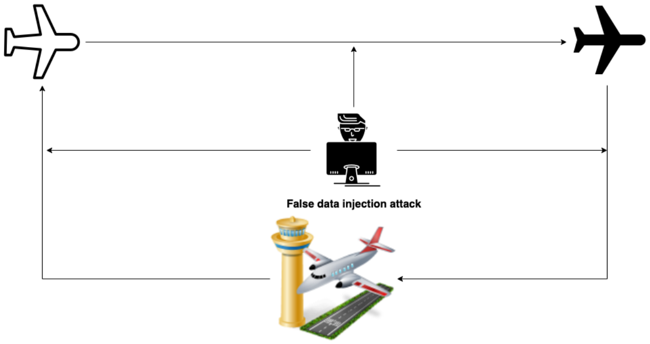


Fig. 8. False Data injection Attack

V. Threats in the Aviation Cyber-Physical System

Threats are actions used by a hacker to disrupt the system to reach undesirable behavior. The aviation sector is crucial, and addressing the threats at the beginning of designing the system is required [10]. The threats in aviation are presented in a different aspect [23]. It can be executed by installing malicious software, manipulating data transmitted over the network [26]. Therefore, The threats on aviation threats are presented in the cyber world, the physical world, or cyber-physical layers [1]. The cyber threats affect the physical world directly and vice versa. Therefore, this section, addressing the presenting threats in every aspect. In the cyber layer,

VI. Security of Wireless sensor network in the aviation sector

The wireless sensor network is one of the ACPS components. This technology plays a crucial role in enhancing the performance of the ACPS. It is intelligent, multi-function, and self-organizing technology [6]. It has limited the computation, battery, and memory resources. The sensors are the main components of the WSN and kinked each other through the link in short range. The WSN can increase the remotely level airport [24]. It can connect all operations such as assigning gate, taxi, and increasing interoperation in the airport. WSN uses widely in the aviation sector because of the low cost and high efficiency [11] .it can be used to sense and gather data from the physical world. Also, it is used to monitor and observe the behavior of the physical realm and adjusted it based on the erratic behavior. It is used in the terminals to sense the tempretautr on the airports, lounges, stores, etc. Also, it is used to drive the aircraft on the runway to the assigned gate.

The importance of securing the WSN in the aviation sector is an urgent request [25]. WSN is used to constitute in the airport to efficiently and effectively sense-data to keep the controller ready for any possible event. The security issues in the WSN such as damage to a data packet or drop when the congestion happened in the network [27]. Attacks cause to delay responded to the action where it would lead to a dangerous event. Also, the hacker containment the sensor channel to send the wrong measurement or data.

VII. Conclusion

The aviation sector has been evolving and becomes more efficient in the last century. The emerging of the CPS improves the aviation operation and makes that more efficient and effective. The CPS integrated a two-cyber and a physical layer. Therefore, the integration increases the performance and provide new services to airlines, stakeholders governments, and enterprise. In contrast, it opened up new challenges to the system where the security issues expanded and some vulnerabilities presented. The security engineer and researchers started to focus on the security issues to innovate new techniques to enhance the security and reliability in the ACPS.

References

[1] Sampigethaya, K. and Poovendran, R., 2013. Aviation cyber–physical systems: Foundations for future aircraft and air transport. Proceedings of the IEEE, 101(8), pp.1834-1855.

[2] Ashibani, Y. and Mahmoud, Q.H., 2017. Cyber physical systems security: Analysis, challenges and solutions. Computers & Security, 68, pp.81-97.

[3] Pollack, J. and Ranganathan, P., 2018. Aviation Navigation Systems Security: ADS-B, GPS, IFF. In Proceedings of the International Conference on Security and Management (SAM) (pp. 129-135). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[4] Ding, D., Han, Q.L., Xiang, Y., Ge, X. and Zhang, X.M., 2018. A survey on security control and attack detection for industrial cyber-physical systems. Neurocomputing, 275, pp.1674-1683.

[5] Alrefaei, F., Alzahrani, A., Song, H. and Zohdy, M., 2020, September. Security of Cyber Physical Systems: Vulnerabilities, Attacks and Countermeasure. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.

[6] Chelli, K., 2015, July. Security issues in wireless sensor networks: Attacks and countermeasures. In Proceedings of the world congress on engineering (Vol. 1, No. 20, pp. 876-3423).

[7] Pollack, J. and Ranganathan, P., 2018. Aviation Navigation Systems Security: ADS-B, GPS, IFF. In Proceedings of the International Conference on Security and Management (SAM) (pp. 129-135). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[8] Houbing Song, Danda Rawat, Sabina Jeschke, and Christian Brecher. Cyber-Physical Systems: Foundations, Principles and Applications. ISBN: 978-0-12-803801-7. Boston, MA: Academic Press, 2016, pp. 1-514.

[9] Houbing Song, Glenn A. Fink, and Sabina Jeschke, Security and Privacy in Cyber-Physical Systems: Foundations, Principles and Applications. ISBN: 978-1-119-22604-8, Chichester, UK: Wiley-IEEE Press, 2017, pp. 1-472.

[10] J. Wang, Y. Liu and H. Song, ”Counter-Unmanned Aircraft System(s) (C-UAS): State of the Art, Challenges, and Future Trends,” in IEEE Aerospace and Electronic Systems Magazine, doi: 10.1109/MAES.2020.3015537.

[11] I. Butun, P. Österberg and H. Song, ”Security of the Internet of Things: Vulnerabilities, Attacks, and Countermeasures,” in IEEE Communications Surveys & Tutorials, vol. 22, no. 1, pp. 616-644, Firstquarter 2020, doi: 10.1109/COMST.2019.2953364.

[12] Yu Jiang, Houbing Song, Yixiao Yang, Han Liu, Ming Gu, Yong Guan, Jiaguang Sun, and Lui Sha. 2018. Dependable Model-driven Development of CPS: From Stateflow Simulation to Verified Implementation. ACM Trans. Cyber-Phys. Syst. 3, 1, Article 12 (January 2019), 31 pages. DOI:https://doi.org/10.1145/3078623

[13] Ekedebe, Nnanna, Wei Yu, Chao Lu, Houbing Song, and Yan Wan. ”Securing transportation cyber-physical systems.” In Securing Cyber-Physical Systems, pp. 163-196. CRC Press, 2015.

[14] Yunchuan Sun, Houbing Song (Eds.) Secure and Trustworthy Transportation Cyber-Physical Systems. ISBN: 978-981-10-3891-4. Singapore: Springer, 2017.

[15] Guido Dartmann, Houbing Song, and Anke Schmeink. Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things. ISBN: 9780128166376. Elsevier, 2019, pp. 1-360.

[16] F. alrefaei, ”The Importance Of Security In Cyber-Physical System,” 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-3, doi: 10.1109/WF-IoT48130.2020.9221155.

[17] Jiang, Y., Wang, M., Jiao, X., Song, H., Kong, H., Wang, R., Liu, Y., Wang, J. and Sun, J., 2019, July. Uncertainty theory based reliability-centric cyber-physical system design. In 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 208-215). IEEE.

[18] Sun, Y., Song, H., Jara, A.J. and Bie, R., 2016. Internet of things and big data analytics for smart and connected communities. IEEE access, 4, pp.766-773.

[19] Jeschke, S., Brecher, C., Song, H., Rawat, D. (2017). Industrial internet of things: cybermanufacturing systems . Springer. https://doi.org/10.1007/978-3-319-42559-7

[20] Song, H., Srinivasan, R., Sookoor, T. and Jeschke, S., 2017. Smart cities: foundations, principles, and applications. John Wiley Sons.

[21] Liu, Y., Weng, X., Wan, J., Yue, X., Song, H. and Vasilakos, A.V., 2017. Exploring data validity in transportation systems for smart cities. IEEE Communications Magazine, 55(5), pp.26-33.

[22] W. Li and H. Song, ”ART: An Attack-Resistant Trust Management Scheme for Securing Vehicular Ad Hoc Networks,” in IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 4, pp. 960-969, April 2016, doi: 10.1109/TITS.2015.2494017.

[23] Z. Lv, H. Song, P. Basanta-Val, A. Steed and M. Jo, ”Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics,” in IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1891-1899, Aug. 2017, doi: 10.1109/TII.2017.2650204.

[24] Mehmood, A., Mukherjee, M., Ahmed, S.H., Song, H. and Malik, K.M., 2018. NBC-MAIDS: Naïve Bayesian classification technique in multi-agent system-enriched IDS for securing IoT against DDoS attacks. The Journal of Supercomputing, 74(10), pp.5156-5170.

[25] M. Albalawi and H. Song, ”Data Security and Privacy Issues in Swarms of Drones,” 2019 Integrated Communications, Navigation and Surveillance Conference (ICNS), 2019, pp. 1-11, doi: 10.1109/ICNSURV.2019.8735133.

[26] Y. Liu, J. Wang, H. Song, J. Li and J. Yuan, ”Blockchain-based Secure Routing Strategy for Airborne Mesh Networks,” 2019 IEEE

International Conference on Industrial Internet (ICII), 2019, pp. 56-61, doi: 10.1109/ICII.2019.00021.

[27] X. Yue, Y. Liu, J. Wang, H. Song and H. Cao, ”Software Defined Radio and Wireless Acoustic Networking for Amateur Drone Surveillance,” in IEEE Communications Magazine, vol. 56, no. 4, pp. 90-97, April 2018, doi: 10.1109/MCOM.2018.1700423.

IEMTRONICS

International Conference

Toronto, Canada