IoT Based Health Monitoring System and Its Challenges and Opportunities

Chapter · January 2022

DOI: 10.1007/978-3-030-90618-4_19

5 authors, including:

Mohammad Bhuiyan
Noakhali Science & Technology University
4 PUBLICATIONS **256** CITATIONS
SEE PROFILE

Md Masum Billah
American International University-Bangladesh
11 PUBLICATIONS **258** CITATIONS
SEE PROFILE

Dipanita Saha
Noakhali Science & Technology University
6 PUBLICATIONS **231** CITATIONS
SEE PROFILE

Md. Mahbubur Rahman
Military Institute of Science and Technology
117 PUBLICATIONS **1,599** CITATIONS
SEE PROFILE

Lecture Notes on Data Engineering
and Communications Technologies 105

Zakaria Boulouard
Mariya Ouaissa
Mariyam Ouaissa
Sarah El Himer *Editors*

AI and IoT for Sustainable Development in Emerging Countries

Challenges and Opportunities

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 105

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at https://link.springer.com/bookseries/15362

Zakaria Boulouard · Mariya Ouaissa ·
Mariyam Ouaissa · Sarah El Himer
Editors

AI and IoT for Sustainable Development in Emerging Countries

Challenges and Opportunities

Springer

Editors
Zakaria Boulouard
LIM
Hassan II University
Mohammedia, Morocco

Mariyam Ouaissa
Moulay Ismail University
Meknes, Morocco

Mariya Ouaissa
Moulay Ismail University
Meknes, Morocco

Sarah El Himer
Sidi Mohamed Ben Abdellah University
Fes, Morocco

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-3-030-90617-7 ISBN 978-3-030-90618-4 (eBook)
https://doi.org/10.1007/978-3-030-90618-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Artificial intelligence and Internet of Things have introduced themselves today as must-have technologies in almost every sector. Ranging from agriculture to industry and healthcare, the scope of applications of AI and IoT is as wide as the horizon. Nowadays, these technologies are extensively used in developed countries, but they are still at an early stage in emerging countries.

As these countries were more affected by the COVID-19 pandemic, both healthcare-wise and economy-wise, the need to adopt new solutions to sustainable development challenges has become more crucial than ever.

The idea behind this book is to focus on solutions based on AI and IoT that can face the challenges of the emerging countries. We will shed the light on different sectors such as agriculture, industry, transportation, environment, energy, healthcare, etc. We will discuss the challenges that the emerging countries face in these sectors and provide AI and IoT-based solutions to them. We will also introduce success stories featuring the implementation of AI- and IoT-based solutions and their impact on the lives of people in developing countries.

This book will be divided into five parts. The first part will introduce AI and IoT as actors that can help address the sustainable development challenges in general with some case scenarios as examples. The other sections will go in depth and spot the light on AI and IoT and their role to play in each of the most important sectors of everyday life. They will also expose how researchers from emerging countries applied AI and IoT to approach sustainable development challenges taking into consideration the specificities of their own countries.

First part, containing five chapters, starts in Chapter "Achieving Sustainable Development Goals Through Digital Infrastructure for Intelligent Connectivity," by spotting the light on Sustainable Development Goals (SDGs) set by the UN to be achieved by 2030 and the impact that COVID-19 had on these SDGs by increasing intelligent connectivity levels and gaining better access to education, health, work, and entertainment despite the pandemic. Chapter "Implementation of Intelligent IoT" introduced the concept of "artificial intelligence-enabled Internet of Things," better known as "AIoT." It provides an overview on "intelligent" IoT devices, as well as why and how to implement them. Chapter "Cyber Security Challenges for Smart Cities" discusses the interconnectivity of intelligent devices at its widest levels, “Smart Cities,” and how it can be secured. Chapters “Efficient Machine Learning Technique for Early Detection of IoT Botnets” and “AI-Based Smart Robot for Restaurant Serving Applications” are introductory use case scenarios where researchers from emerging countries, Algeria and Pakistan, take the best of both AI and IoT to provide better solutions to everyday life situations, such as communication security on the one hand and restaurant serving on the other hand.

Second part spots the light on the innovations of emerging countries researchers when it comes to facing the environment and energy optimization challenges. On a microscopic level, Chapter “A Novel Deep Learning Architecture Based IoT Time-Series for Energy Consumption Forecasting in Smart Households” introduces a deep learning-based IoT system able to predict energy consumption in smart households. On a macroscopic level, Chapter “Performances of CPV Optics in Morocco” measures the performances of Concentration Photovoltaic systems adopted in Morocco as a means to produce solar-based electrical energy, while Chapter “Artificial Intelligence Based on Particle Swarm Optimization for Optimal Wind Turbine Power Control Using Doubly Fed Induction Generator” introduces an approach to optimize wind turbine power control based on artificial intelligence. When it comes to environment, Chapters “A Comparative Study Between NARX and LSTM Models in Predicting Ozone Concentrations: Case of Agadir City (Morocco)” and “Spatiotemporal Prediction of $PM_{2.5}$ Concentrations Based on IoT Sensors” introduce two IoT and machine learning-based approaches to predict the concentration of ozone components (and pollutants) in both Morocco and Taiwan, respectively. This part also introduces the role of artificial intelligence and Internet of Things in smart precision agriculture as the authors in Chapter “Comparative Study Between Different Recommendation Systems in Smart Agriculture” provide a comparative analysis on recommendation systems and their role as a means to optimize agricultural yields.

Third part focuses on Industry 4.0 and Transportation in developing countries and provides an overview on different approaches to address their challenges. Chapter “Configuration Security for Sustainable Digital Twins of Industrial Automation and Control Systems in Emerging Countries” goes through the configuration security techniques used for sustainable Digital Twins in emerging countries, and suggest a new approach, based on a combination of artificial bee colonies and support vector machines, that will be able to optimize attack predictions. Chapter “An Empirical Investigation on Lean Method Usage: Issues and Challenges in Afghanistan” addresses the challenges facing the application of Lean method in Afghan software development companies, while Chapter “Optimization of the Effects Oscillation Welding: Sinusoidal and Triangular Beam During Laser Beam Welding of 5052-H32 Aluminum Alloy” provides a regression-based model to predict the tensile strength of aluminum alloys. When it comes to transportation, the authors in Chapter “The Internet of Things Solutions for Transportation” provide an overview of AI- and IoT-based solutions for transportation as well as their challenges while going through use cases of large companies that have adopted these solutions. Chapter “A Novel

GAN-Based System for Time Series Generation: Application to Autonomous Vehicles Scenarios Generation" proposes a novel GAN-based system for time series generation able to generate various autonomous driving scenarios, toward a fully automatic framework of self-driving testing. Road accidents are also in the spotlight in this section, as in Chapter "Fuzzy Set Theory-Based Approach for Mining Spatial Association Rules: Road Accident as a Case Study," where the authors introduce a new approach on analyzing road accidents in a specific area and determining their main causes. This approach combines the best of artificial intelligence-based recommendation systems with fuzzy set theory applied on spatial association rules for a better performance. Chapter "A Mobile Application for Real-Time Detection of Road Traffic Violations" also proposes a machine learning-based solution for road accidents in Mauritius. This solution is able to detect infringements, warn the wrongdoers, and report them to the authorities.

COVID-19 has brought the attention on the importance of health care as a crucial sector, especially in emerging countries where the pandemic had a greater impact. Part Four will focus on different attempts of researchers from these countries to take the best of AI and IoT to face healthcare-related challenges. The authors in Chapter "IoT Based Health Monitoring System and Its Challenges and Opportunities" spot the light on some of these challenges, especially when it comes to monitoring infected or elderly patients. They proposed an IoT-based health monitoring system that connects through GSM networks as a means to overcome connectivity problems in emerging countries. Chapter "Wireless Body Sensor Networks: Applications, Challenges, Patient Monitoring, Decision Making, and Machine Learning in Medical Applications" presents the works of an Iraqi team that also provides a solution to the health monitoring issue by introducing wireless body sensor networks (WBSN). In this chapter, they present this technique, its architecture, challenges, healthcare application, and their requirements. Since the amount of data transferred through WBSNs can be important, the same team introduces in Chapter "A Novel Lossless EEG Compression Model Using Fractal Combined with Fixed-Length Encoding Technique" a means to optimize connectivity in these networks by compressing EEG signals, and thus reducing the size of data sent. Along with the connectivity of the healthcare-oriented systems comes their security, a challenge addressed in Chapter "Securing the Hyperconnected Healthcare Ecosystem" where the authors propose a holistic cybersecurity platform that tackles privacy and security risks in an automated fashion to foster the development of innovative applications within the healthcare ecosystem. Optimizing the yields of healthcare-oriented IoT devices also implies a better energy consumption. This urged the authors of Chapter "Design of an Efficient Rectenna for RF Energy Harvesting for IoT Medical Implants" to design an efficient rectenna for radio frequency harvesting for IoT medical implants. Researchers from emerging countries have put quite some efforts to increase the role of AI in the healthcare sector, and the remaining chapters of this part will spot the light on them. In Chapter "Multi-class Classification for the Identification of COVID-19 in X-Ray Images Using Customized Efficient Neural Network," the authors used AI to detect COVID-19 positive cases based on X-Ray images, while the authors of "Chapters A Review: Recent Automatic Algorithms for the Segmentation of Brain Tumor

MRI" and "Oncology with Artificial Intelligence: Classification of Cancer Using Deep Learning Techniques" suggest new deep learning-based approaches to detect cancer cells. Chapter "IoT Based Machine Learning and Deep Learning Platform for COVID-19 Prevention and Control: A Systematic Review" proposes a systematic review of several efforts to apply deep learning to enhance the prevention and control of the COVID-19 disease.

The increasing connectivity and the necessity to stay online made the exploitation of big data in general and social media in particular a concern to emerging countries as well. Part Five will go through efforts of researchers from these countries to take the best of big data and social media. In Chapter "Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions," the authors present a baking system that predicts loan repayment or default based on customer's digital data. Chapter "A k-Mean Classification Study of Eight Community Detection Algorithms: Application to Synthetic Social Network Datasets" proposes a new approach for community detection in social media, while Chapter "Topic Modeling for Short Texts: A Novel Modeling Method" offers an AI-based technique to better rank topics in short texts. "Chapters Prediction and Analysis of Moroccan Elections Using Sentiment Analysis" and "Analysis of COVID-19 Trends in Bangladesh: A Machine Learning Analysis" provide new approaches to analyze the sentiments of social media users regarding hot topics in their countries, such as the general elections in Morocco or the COVID-19 disease in Bangladesh.

Mohammedia, Morocco	Zakaria Boulouard
Meknes, Morocco	Mariya Ouaissa
Meknes, Morocco	Mariyam Ouaissa
Fes, Morocco	Sarah El Himer

Contents

AIoT and Challenges for Sustainable Development

AI and IoT for Smart Environment and Energy

Industry 4.0 and Intelligent Transportation

A Vision Towards an Artificial Intelligence of Medical Things

AI, Social Media, and Big Data Analytics

About the Editors

Zakaria Boulouard is currently Professor at the Department of Computer Sciences at the "Faculty of Sciences and Techniques Mohammedia, Hassan II University, Casablanca, Morocco." In 2018, he joined the "Advanced Smart Systems" Research Team at the "Computer Sciences Laboratory of Mohammedia." He received his Ph.D. degree in 2018 from "Ibn Zohr University, Morocco" and his engineering degree in 2013 from the "National School of Applied Sciences, Khouribga, Morocco." His research interests include artificial intelligence, big data visualization and analytics, optimization and competitive intelligence. Since 2017, he is a member of "Draa-Tafilalet Foundation of Experts and Researchers," and since 2020, he is an "ACM Professional Member." He has served on Program Committees and Organizing Committees of several conferences and events and has organized many symposiums/workshops/conferences as General Chair. He has served and continues to serve as a reviewer of numerous international conferences and journals. He has published several research papers. This includes book chapters, peer-reviewed journal articles, peer-reviewed conference manuscripts, edited books and journal special issues.

Mariya Ouaissa is Researcher Associate and practitioner with industry and academic experience. She is Ph.D. graduated in 2019 in Computer Science and Networks, at the Laboratory of Modeling of Mathematics and Computer Science from ENSAM-Moulay Ismail University, Meknes, Morocco. She is a Networks and Telecoms Engineer, graduated in 2013 from National School of Applied Sciences Khouribga, Morocco. She is Co-founder and IT Consultant at IT Support and Consulting Center. She was working for School of Technology of Meknes Morocco, as Visiting Professor from 2013 to 2021. She is Member of the International Association of Engineers and International Association of Online Engineering, and since 2021, she is "ACM Professional Member." She is Expert Reviewer with Academic Exchange Information Centre (AEIC) and Brand Ambassador with Bentham Science. She has served and continues to serve on technical program and organizer committees of several conferences and events and has organized many symposiums/workshops/conferences as General Chair, as well as an editorial board and reviewer of numerous international journals. She has made contributions in the fields of information security and privacy, Internet of Things security, and wireless and constrained networks security. Her main research topics are IoT, M2M, D2D, WSN, cellular networks, and vehicular networks. She has published more than 20 research papers, 5 edited books, and 5 special issue as guest editor.

Mariyam Ouaissa is Ph.D. Researcher Associate in Computer Science and Networks from Moulay Ismail University Meknes, Morocco, and Consultant Trainer. She received her Ph.D. degree in 2019 from National Graduate School of Arts and Crafts, Meknes, Morocco, and her engineering degree in 2013 from the National School of Applied Sciences, Khouribga, Morocco. She is a communication and networking researcher and practitioner with industry and academic experience. Her research is multidisciplinary that focuses on Internet of Things, M2M, WSN, vehicular communications and cellular networks, security networks, congestion overload problem and the resource allocation management and access control. She is serving as editorial board and reviewer for international journals and conferences including as *IEEE Access*, *Wireless Communications and Mobile Computing*, *International Journal of Smart Security Technologies* (IJSST), *International Journal of Information Security and Privacy*(IJISP). Since 2020, she is a member of "International Association of Engineers IAENG" and "International Association of Online Engineering," and since 2021, she is an "ACM Professional Member." She has published more than 20 research papers (this includes book chapters, peer-reviewed journal articles, and peer-reviewed conference manuscripts), 5 edited books, and 5 special issue as guest editor. She has served on Program Committees and Organizing Committees of several conferences and events and has organized many symposiums/workshops/conferences as General Chair.

Sarah El Himer is Ph.D. and Associate Researcher in Intelligent Systems, Georesources and Renewable Energies Laboratory. She is a Ph.D. graduated in 2019 in renewable energies at the Renewable Energies and Intelligent Systems Laboratory from the faculty of sciences and technology, University of Sidi Mohammed Ben Abdallah FEZ. She is a trainer and IT consultant at IT Support and Consulting Center. She was working for faculty of science and technology FEZ as Visiting Professor from 2015 to 2019. She has made contributions in the fields of optical component of concentrated photovoltaic and electrical vehicle. Her main research topics are optical elements for CPV, acceptance angle, optical efficiency micro-CPV, hybrid CPV. She has published over 15 papers (international journals and conferences/workshops). She has served and continues to serve on executive and technical program committees and as a reviewer of numerous international conference and journals such as "CPV conference, REEE2017." She was the TPC chair of the ICACTCE'21 conference and co-editor of the book "*Proceedings of International Conference on Advances in Communication Technology, Computing and Engineering*: RGN Publication." She is also a guest editor of special issues: "Recent Trends in Green Energy Technologies" of *International Journal of Energy Optimization and Engineering* (IJEOE)-IGI Global and "Selected Papers of Proceedings of ICACTCE'21 Conference" of *Journal of Atomic, Molecular, Condensed Matter and Nano Physics*, Vol 7, No 3 (2020).

AIoT and Challenges for Sustainable Development

Achieving Sustainable Development Goals Through Digital Infrastructure for Intelligent Connectivity

T. P. Fowdur, M. Indoonundon, M. A. Hosany, D. Milovanovic, and Z. Bojkovic

Abstract We are at the beginning of the decade to achieve the targets of the UN Sustainable Development Goals (SDGs) by 2030 and achieving these targets appear even more challenging as the world is still struggling to cope with COVID-19. The COVID-19 pandemic has in fact seriously comprised several SDGs such as SDG 1 (No Poverty) in which significant progress was being made. For the first time in two decades, poverty levels are being projected to grow as a result of the COVID-19 pandemic. However, COVID-19 has also propelled the adoption of digital technologies to much higher rates. In particular, intelligent connectivity (IC), which combines 5G with Artificial Intelligence (AI), the Internet of Things (IoT) and other emerging technologies, has allowed individuals to access several life-saving and enhancing services such as education, health, food supplies, work and entertainment, during the pandemic. The adoption of intelligent connectivity is being increasingly popular, propelled by new requirements created by Covid-19 and fueled by an exponentially growing mobile industry. This chapter gives a detailed overview of how intelligent connectivity and other emerging digital technologies are contributing to all the 17 SDGs.

Keywords 5G · AI · UN · SDGs · Intelligent connectivity · Cloud · IoT

1 Introduction

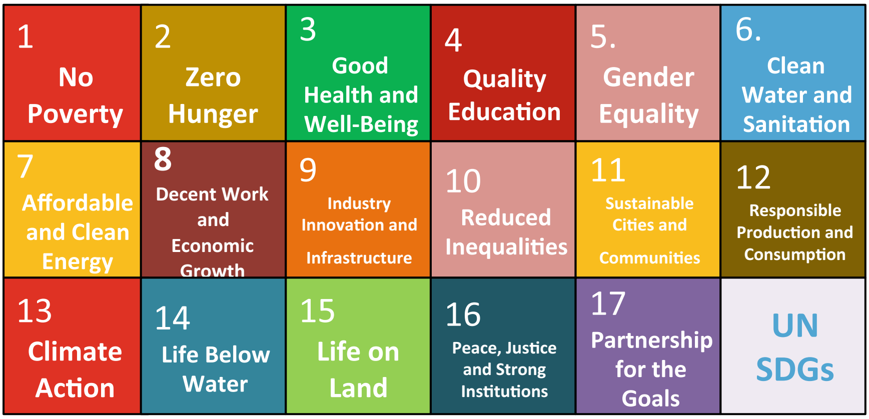
The UN General Assembly introduced the Sustainable Development Goals (SDGs) in its resolution 70/1 in 2015, with the target year being 2030 [1, 2]. The goals encompass

T. P. Fowdur (✉) · M. Indoonundon · M. A. Hosany
Department of Electrical and Electronic Engineering, University of Mauritius, Moka, Mauritius
e-mail: p.fowdur@uom.ac.mu

M. A. Hosany
e-mail: m.hosany@uom.ac.mu

D. Milovanovic · Z. Bojkovic
University of Belgrade, Belgrade, Serbia

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_1



1 No Poverty	2 Zero Hunger	3 Good Health and Well-Being	4 Quality Education	5. Gender Equality	6. Clean Water and Sanitation
7 Affordable and Clean Energy	8 Decent Work and Economic Growth	9 Industry Innovation and Infrastructure	10 Reduced Inequalities	11 Sustainable Cities and Communities	12 Responsible Production and Consumption
13 Climate Action	14 Life Below Water	15 Life on Land	16 Peace, Justice and Strong Institutions	17 Partnership for the Goals	UN SDGs

Fig. 1 The sustainable development goals (SDGs) (Source: [6])

three major community development dimensions which are environment protection, social diversity and economic growth [3]. The SDGs have been set by the community which includes government, private sector and academia and they have become one of the most widely adopted standard systems for achieving sustainable community development. The 17 SDGs, given in Fig. 1 are a continuation of the Millennium Development Goals (MDG) which consisted of 8 goals, set in 2000 with 2015 being the target year [4, 5]. The SDG framework is linked to several human rights but can be differentiated from them by their emphasis on people, planet, prosperity, peace, and partnership [6].

The SDGs are very ambitious and broad [7] and can seem impossible to achieve. Nevertheless, they are pursued to encourage innovative approaches to make progress [8]. Intelligent Connectivity, which is the combination of several technological enablers such as Artificial intelligence (AI), mobile communications (5G), Internet of Things (IoT), Cloud and Blockchain, is considered to be the most impactful technology which can help achieve SDGs. These main parts are briefly described as follows:

5G: 5G has incorporated three main service sets namely: (i) Enhanced Mobile Broadband (eMBB) to provide data rates of upto 10 Gbps, (ii) Ultra-Reliable Low-Latency Communication (URRLC) for applications requiring extremely low error rates (high reliability) and low latency and (iii) Massive Machine Type Communications (mMTC) to support high device density with low power consumption [9].

Cloud: Includes services such as Software-as-a-Service, Infrastructure-as-a-Service, and Platform-as-a-Service which provide the necessary applications, processing power and storage requirements to run and scale AI and data analytics [10].

AI: Allows data analytics to be performed with diverse machine learning algorithms for different applications which can even be real-time or near real-time. The analytics can take the form of predictions, classifications, pattern discovery and decision making processes [10].

IoT: Consists of a network of items composed of sensors that can communicate via the Internet. The sensors collect data to feed AI systems and other network elements such as controllers or even robots, implement the instructions received from AI systems [11].

Blockchain: Blockchain, also known as Distributed Ledger Technology (DLT), cryptographically associates data blocks, permanently logs transactions and links them to the next data block. It creates a reliable continuous data stream [12] and can thus be used as a key enabling technology for multiple types of AI applications and analytical tools.

Intelligent Connectivity is still in its infancy but is expected to bring innovations, boost productivity, and speed up the development of novel business models which will significantly impact socioeconomic development [13]. 5G technology will provide enhanced internet connectivity which is expected to lead to an economic output of $3.6 trillion and the creation of 22.3 million jobs by 2035 [14], translating into $13.2 trillion global economic value across industries. A third of this economic output is composed of manufacturing whereas another third is composed of wholesale and retail, construction, information and communications, and public services [14]. However, the first trillions of dollars will need to be invested into the development of 5G networks globally. Companies aim at becoming first movers into the 5G market but first, cooperation will be required to accelerate 5G's development.

A global reduction in greenhouse gas (GHG) emissions that is tenfold more than the global carbon footprint of the mobile industry, has been achieved by the adoption of mobile technology. People around the world rely mainly on mobile technology to obtain digital access, leading to over five billion subscribers of mobile services and around four billion mobile internet users worldwide. Mobile technologies have introduced several economic, social and environmental advantages by easing access to digital services and expanding connectivity and they have also contributed to all 17 SDGs [15].

In this chapter, the impact of Intelligent Connectivity on each SDG will be described with several use cases in countries and organizations across the world.

2 Impact of Intelligent Connectivity on SDGs

In this section an in-depth analysis on how intelligent connectivity can contribute to the 17 SDGs is provided. Table 1 provides a global view of the relative impact of the different technological enablers of IC on the 17 SDGs for some selected use cases that will be presented in the following sub-sections.

Table 1 Relative impact of technological enablers in selected intelligent connectivity use cases for sustainable development goals (SDGs)

SDG	5G mobile communications	Artificial intelligence (AI)	Internet of things (IoT)	Cloud computing (CC)	Blockchain (BC)
1	+	+	+		
2	+	+	+		
3	+	+			
4	+	+	+		
5		+	+		
6	+		+		
7	+	+			+
8	+	+	+		
9	+	+	+	+	
10	+	+			
11	+	+			
12	+		+		
13	+		+		
14	+	+	+		
15	+	+	+		
16		+			
17	+				

2.1 SDG 1: No Poverty

People earning less than $1.25 per day are considered to be extremely poor. The aim of this SDG is to eradicate such extreme poverty globally by 2030 and also to guarantee that all men and women, including the vulnerable ones, have the same rights to economic resources. Furthermore, they will have access to basic necessities, natural resources, property ownership, inheritance, novel technology and financial services such as microfinance [16].

Mobile technology is an essential contributor to SDG 1 as it drives sustainable economic growth by supporting households to overcome poverty and by providing humanitarian assistance. Since 2015, the number of mobile users among the world's poorest 40% (composed of 1.9 billion people) has increased by 200 million [15]. Furthermore, mobile helps in improving efficiency and productivity in other sectors. It allows companies to reach non-local customers, allowing the former to expand and create jobs for the local population.

Even the general use of mobile helps to curb poverty. The increased availability of mobile phones in Peru has decreased poverty prevalence by 8% and has reduced acute poverty by 5.4% points [17]. The deployment of mobile broadband networks in Nigeria in 2010 to 2016 has created new jobs and reduced extreme poverty by 7% points [18].

In Low-to-Middle Income Countries (LMIC), the financial exclusion gap has been reduced with the help of mobile money. From 2015 to 2019, there has been an increase of 460 million registered mobile money accounts which adds up to more than 1 billion registered accounts. Mobile money allows users to easily access financial services and to seamlessly manage their cash flow, build working capital and handle financial risks. Access to mobile money in Kenya was found to have lifted out of poverty some 200,000 households (2% of Kenya's households). In rural Uganda, it was observed that mobile money could smoothen consumption and curb poverty [19].

Furthermore, M-KOPA, a Ugandan based company, has combined digital micropayments technology with IoT connectivity to make financial management more accessible. They used an advanced asset financing platform which they built themselves to invest $400 million in financing which allowed one million customers to gain access to high-quality energy-efficient appliances, solar lighting, smartphones and loans amongst others [20].

The World Bank and the UN rely significantly on research and data to keep track of the progress towards their goal of eliminating poverty [21]. The difficulty to collect data itself is considered to be a consequence of poverty as per the Decentralized AI Alliance (DAIA) [22]. As per DAIA, location plays an important role in the eradication of poverty. Traditional household surveys which help keep track of the number and location of poor people are expensive to many nations and are thus not regularly conducted [22]. One solution to this issue is the use of AI. Recently, a team at Standford University has conducted a study where they used powerful machine learning algorithms high-power satellite imaging analysis to detect poverty in Nigeria, Tanzania, Uganda, Malawi and Rwanda. The accuracy of the algorithm's predicted data was verified using accurate survey data [23]. Furthermore, AI can be used to obtain essential information such as the nearest water sources, agricultural fields and marketplaces in high poverty regions [22, 24].

2.2 SDG 2: Zero Hunger

The main aim of this SDG is to end hunger, enhance nutrition, establish food security, and encourage sustainable agriculture [25]. Mobile technology plays an important role for SDG 2 as it leads to better nutritional knowledge, improved agricultural practices, and household food security. Production of agricultural supplies is made more efficient by mobile devices, satellites, drones and other state-of-the-art solutions [15]. 5G is expected to improve yields and quality of crops through the precise monitoring of the weather and soil which will help to tailor the application of pesticides and fertilizers. The Internet of Vehicles (IoV) could enhance food distribution by determining the optimal routes and monitoring the temperature of the food being transported. Additionally, the market and shop lives of crops can be increased by implementing the use of mobile refrigeration [26].

Intelligent Connectivity can help in the combat against hunger through the following measures [27]:

- *Yield boost from harvesting and storage*—The UN's Food and Agriculture Organization (FAO) indicated that up to 40% of food production may be lost before reaching the market, especially in developing countries. Proper cost-effective monitoring, which can be provided by IoT networks can be used to identify the optimal harvesting and storage methods for vulnerable crops.
- *Enhancement of the existing distribution network*—The World Bank estimated that in South and South-East Asia, during the storage and transport of food, around 90% of calories are wasted. The productivity and efficiency of food distribution networks can be improved through viable commercialized IoT and AI technologies which may provide monitoring facilities and provide real-time access to food distribution data.
- *Support for cost-effective real-time data tracking and tracing of the cold-chain infrastructure*—This allows food that needs to be preserved at cold temperatures to be monitored on a 24/7 basis and ensures that an alarm is generated in case that anomalies are detected by sensors and location trackers. This leads to reduced food wastage, improved food quality and safety of commonly contaminated food.
- *Waste minimization by improving food purchase habits in rich countries*—Tracking of food quantity and quality can be done by using smart refrigerators. These refrigerators have useful features such as intelligently recommending purchase orders, generating alerts if abnormalities are detected and suggesting the amount of food to be purchased based on usage patterns [27].

Additionally, AI can significantly boost agricultural techniques by enabling the use of self-driving farm equipment and efficient farm monitoring and management equipment. AI can also help in developing different genetically modified crops which are less vulnerable to diseases and pests. Researchers utilize sensors, drones and robots for collecting data which can help increase the production of heat and drought-resistant crops which are crucial in famine-stricken nations [22, 23].

2.3 SDG 3: Good Health and Wellbeing

SDG 3 aims at guaranteeing healthy lives and upholding the well-being of everyone [28].

Mobile health solutions allow affordable health care to be delivered at a high quality. Mortality rates due to childbirth and pregnancy can be reduced by making information more accessible and by getting people connected to healthcare. For example, in Cameroon where less than 60% of the women population receive proper health care, access to mobile apps and messaging solutions has improved the care given to pregnant women. The GiftedMom startup has designed an application [29] that allows pregnant women and mothers to connect to health specialists and has provided support to more than half a million mothers since 2020. The startup has also collaborated with 45 hospitals in Cameroon and helped in the prevention of early deaths. Since December 2019, over 250,000 women have obtained critical health details through the startup. Many of these women were nursing mothers who earned less than $3 a day [30].

With high capacity 5G networks, people will have the ability to use connected wearables for monitoring their heart rates, temperature, stress levels, location and blood pressure. Any detected health anomalies may trigger emergency alerts to notify the wearer or even a rescue team. Such facilities may benefit health insurance programmes and give rise to predictive healthcare and personal security solutions. Clinicians can use machine learning-aided smart platforms to identify medical data patterns. They can then act earlier and advise patients to take preventive measures. Low latency 5G networks may become the enablers of telemedicine services such as teleconsultation and telesurgery. Telesurgery is still in its infancy due to network and robotic limitations even though the first telesurgical operation was carried out in 2001 [26].

Moreover, the COVID-19 pandemic has triggered a significant adoption of Intelligent Connectivity in dealing with the virus. For example, frontline staff are using AI-enabled thermal cameras to check the temperature of patients instead of using manually operated forehead thermometers. These cameras provide quicker and more accurate temperature measurements and help protect the frontline staff from contracting the virus. Furthermore, the use of intelligent chatbots, drones, robots and telemedicine equipment have become more common as they allow doctors to work remotely, keeping them safe from the risk of infections. One such example is a chatbot built by the Centers for Disease Control and Prevention (CDC) to attend to people who have COVID-19 symptoms and thus accelerate the detection of COVID-19 cases. For individuals with more serious symptoms, an online triage system was set up by the US to locate potential COVID-19 cases easier. Heath personnel at the National University Hospital in Singapore utilize a chat assistant app to stay in touch with dynamic information about the pandemic. In hospitals, the spread of the virus is predicted by using AI-based tracking with advanced graphical analytics to provide essential insights on the pandemic. This tool also allows authorities to understand if a local lockdown is required or if any actions can be taken to help stop the spread of the virus. Furthermore, hospitals may be notified of potential virus infections by AI applications that track the movement of mobile devices [31–36].

AI is largely contributing to the vaccine creation process, especially by identifying essential genome sequences, and is also speeding up the vaccine testing phase. In the UK, BenevolentAI helped to identify approved drugs that may be used to inhibit the spread of COVID-19. The company derived contextual relationships between drugs, diseases and genes using AI to come up with vaccines recommendations [31–36].

2.4 SDG 4: Quality Education

The aim of this SDG is to ensure that everyone obtains access to quality education and to encourage lifelong learning [37].

Mobile technology and mobile networks play an essential role in SDG 4. They provide education access to educators, scholars, and professionals irrespective of their location. Services such as educational management have become more accessible via mobile networks [15].

During the COVID-19 pandemic, technology has helped scholars to attend classes remotely via online education platforms according to a study by Mhlanga and Moloi [24]. In their study, it was also concluded that the shift from face-to-face teaching to online delivery increases access to education by eliminating space constraints [38]. Eneza Education is one such education project which incorporates AI to assist the tutoring of millions of students in rural areas such as Kenya, Ghana and Côte d'Ivoire. Such projects make education more encompassing, equal and foster lifelong learning prospects globally [22].

E-learning has proved to be crucial during crises like the COVID-19 pandemic. The pandemic has at some point deprived 90% of students of access to their schools [39].

There has been a significant rise in the usage of video conferencing and communication platforms such as WhatsApp, Skype and Microsoft Teams for e-learning purposes due to the outbreak of COVID-19. Teachers in some schools in Bhutan use online messaging platforms such as WhatsApp and WeChat to deliver homework to students who are expected to answer back with an image of their answers for assessment. In Bulgaria, 65,000 teachers were able to connect to 700,000 students to deliver videos and webinars with the help of an e-learning platform launched in March 2020. 90% of the students were able to obtain 6 h of distance learning per day via this platform [15, 40].

Furthermore, AI applied to Learning analytics may provide useful statistics on education and AI can also be used for teaching. The current Intelligent Tutoring Systems (ITS) have great potential in the education sector but also have some limitations which may eventually be overcome through technology progresses [41, 42]. Another application is in automated attendance systems. For example, in South Africa, IoT has interconnected the students to automate their attendance tracking using biometric features [3].

2.5 SDG 5: Gender Equality

The aim of this SDG is to achieve gender equality and empower all women and girls [43].

Gender equality can be improved by mobile technology in societal, financial and administrative spheres. In addition to women, mobile technology also benefits their communities, businesses and the broader economy. Mobile allows women to have access to life-improving facilities such as financial, health and job related services [15]. One prime example of mobile technology which enhances women safety is the Salvatio Push. It is an IoT wearable device that allows women to notify their community about their location at a push of a button if they are in a life-threatening situation so that rescue can be received [44, 45].

The Indian government funded IoT-based solutions for ensuring women's safety on public transports in 2017. The Abhaya Passenger project is one of India's approaches to ensuring women's safety by equipping 100,000 rickshaws with IoT devices linked to the Abhaya Passenger app which is available for download to the public. Through these devices, the live GPS locations of the rickshaws are tracked and the relevant information about the auto driver which may be required in cases of emergency is provided to the passenger [46]. The drivers are provided with a Radio Frequency Identification (RFID) card which needs to be swiped in the vehicle to start the vehicle. Once a vehicle is started, the police or traffic control room is alerted so that proper tracking can be performed. The driver's details are made available by scanning a QR code attached to the vehicle, using the Abhaya Passenger app [47]. The integration of 5G and AI to this project could also further enhance the prospects of this project.

2.6 SDG 6: Clean Water and Sanitation

The purpose of this SDG is to guarantee that water and sanitation are made available to all and are sustainably managed [48].

Mobile technology provides reliable links for communication and payment between utilities or authorities and the clients, hence improving various aspects of water delivery and sanitation. It also provides remote and affordable billing solutions and the logistics for collecting and non-sewered sanitation services is greatly facilitated [15].

Financial losses made by water and sanitation providers can be reduced through mobile-enabled solutions such as digital payments and smart meters and hence allows the providers to upgrade their distribution network. One such example is the Safe Water Network which was able to significantly increase its per-liter payment collection rate in more than 90% of its stations by digitalizing processes [49]. The implementation of mobile money in Tanzania was able to increase payments to the water utility threefold while lowering waiting times for water collection from 3 h to 10 min [50]. SOIL, which is an ecological sanitation provider in Haiti was able to reduce its collection costs from $1.10 to $0.05 by implementing mobile payments [49].

IoT-based smart water infrastructure can provide leakage detection services, geographic information system (GIS) management facilities, and network optimization which can be useful in the context of improper water infrastructure and thus improve water supply and drainage plans. iWesla which is a smart water project developed in Spain allowed abnormal water consumption to be detected by using sensors and an alerting system, saving up to 50% of water consumption. Early detection of such anomalies also helps avoid damage caused by leaks or open taps [51, 52].

2.7 SDG 7: Affordable and Clean Energy

The purpose of this SDG is to guarantee access to affordable, reliable, sustainable and modern energy for all. There are around 759 million people who do not have access to dependable and reasonably priced electricity supply services [53].

Advancements and breakthroughs in AI, blockchain, advanced materials for solar panels, and battery technology (specifically lithium-ion batteries [54]) make renewable energy mini-grids potentially the cheapest solution to provide electricity to 290 million people [55].

Wind farms have become a more important source of carbon-free electricity over the past decade since the cost of turbines has been reduced. However, they are an unreliable source of energy due to the variable nature of wind. To address this issue, DeepMind and Google used machine learning algorithms to predict the wind power output 36 h ahead by using historical turbine data and weather forecasts on a 700 megawatts wind power grid. This prediction can then be used to plan the optimal hourly delivery commitments to the power grid one day in advance. To date, the value of wind energy has been boosted by around 20% by machine learning compared to the scenario with no time-based commitments to the grid [56].

Mobile money was able to unlock a wide section of the solar off-grid customer base and provides services related to energy and appliances for consumers in the low-income category. Mobile-enabled solutions can reduce the cost of mini-grid connections and solar home systems and provide essential information to off-grid providers about consumer behaviors and thus improve service delivery. For instance, smart-meter driven mini-grid power generation and consumption provides a better overview of the grid and allows better decisions to be made about the dimensioning of the grid. The International Energy Agency estimates that mini-grids can eventually supply electricity to more than 450 million people (equivalent to 80% of people in Sub-Saharan Africa who still lack access to electricity) and can lead to $300 billion in investment by 2030 [57].

2.8 SDG 8: Decent Work and Economic Growth

The purpose of this SDG is to support sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all [58].

Due to the COVID-19 pandemic, digital transformation has become a priority for all sectors which need to maintain business operations amidst lockdowns and social distancing. The pandemic has brought ICT to the forefront as a critical enabler of business continuity. Companies in sectors that had already adopted higher bandwidth broadband and cloud services can continue to operate properly even during lockdowns. Organizations in these sectors are also able to operate more effectively through AI and IoT solutions [31]. ITU conducted studies that showed that a 10% boost in mobile broadband penetration leads to a wider 1.5–2.5% increase in GDP. Upgrades from 2 to 3G and 4G mobile networks also have a positive impact on the economy [59]. Almost 5% of the global GDP was generated by mobile technologies and services in 2019, which is equivalent to $4.1 trillion of economic value added [60]. Furthermore, it is estimated that 5G technologies will contribute $2.2 trillion to the global economy between 2024 and 2034. 5G will initially be very beneficial to crucial sectors such as utilities and manufacturing (particularly in China) and financial and professional services (especially in MENA and North America) [15].

5G networks' capabilities will provide a major contribution in powering business resilience and recovery. The initial use cases of 5G which aim at providing broadband connectivity, support augmented and virtual reality services and enable high-quality media transmission might not look worthwhile in the era after COVID-19, but are essential for 5G's widespread deployment and will lead to new essential use cases. 5G use cases such as e-shopping and robot deliveries are expected to grow quickly. Japan already started using robot delivery services in 2019. 5G has enabled several novel robot delivery use cases to be introduced after COVID-19 in many countries. Additionally, governments are encouraging new supply models that limit the propagation of the virus and promote a highly partitioned workforce by providing incentives. During the pandemic, startups in Japan have developed robots for medical use and contactless deliveries. Nations that do not invest heavily in a 5G network will fall behind from the benefits of its revolutionary applications [31].

2.9 SDG 9: Industry Innovation and Infrastructure

The aim of this SDG is to build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation [61].

A fourth industrial revolution is set to be driven by Intelligent Connectivity whereby computers and robots optimize the manufacturing and maintenance processes in flexible factories. The IoT, enhanced with AI and cloud computing, is one of the key enablers of Industry 4.0. In this context, 5G networks will use edge computing and network slicing with efficient spectrum utilization to provide ultra-reliable low latency communication (URLLC) which may be applied to logistics or remote control of Automated Guided Vehicles in production plants, factories or mines [26].

With real-time information capture and remote control of machinery, factory-wide connectivity will improve the factory's efficiency and allow manufacturers to provide more value to customers. Industry 4.0 will feature self-optimizing production facilities which can address events, such as technical issues, supply shortages and new customer requirements. 3D printing of objects will be available on-demand to allow the repair of broken equipment [26].

IoT, AI, cloud computing, M2M, mobile broadband and big data analytics are essential elements of Industry 4.0, which improve production efficiency and sustainability [15, 59]. Services like object tagging and internet-to-object communication are also essential for real-time data capture, while cloud computing allows computing and storage power to be reduced for digitally enhanced production [62].

The two most significant methods 5G will contribute to SDG 9 are by [13]:

(1) Supporting faster and effective inspections through predictive intelligence:

One specific example is Bright Machines [63] which manages a cloud-based software for the planning, simulation and implementation of the setup and tutorials used to establish, reconfigure and run any quantity of physical production lines for assembly. This software is applied in a North American factory and the key benefits of this software over traditional ones are:

- Reduced parts per million (PPM) defect rates by 88%.
- Increased unit production per hour by 33%.
- Reduced assembly-line staff by 50%.

(2) Improving operational value:

A specific example is the Nokia factory in Oulu [64]. The Nokia factory managed machine and device telecommunications through Ethernet cables which added significant costs in rewiring work. Using Omron LD Autonomous Intelligent Vehicles, material flow is automated. Material is delivered from storage to the production line, without any human intervention. Dynamic consumer demands result in short product cycles which force production lines to be rearranged over short notices. With 5G's high throughput, low delay and large capacity, autonomous mobile devices may automatically deliver components to the exact location required based on communication with production line equipment.

2.9.1 SDG 10: Reduced Inequalities

The aim of this SDG is to reduce inequality within and among countries [65].

The ICT sector can help reduce inequality both between and within countries by providing access to information and knowledge to the less fortunate population—including women and girls or those who have disabilities [66].

For example, people with disabilities can more easily navigate in 5G covered regions. Blind people can be equipped with smart 5G connected devices which can provide accurate real-time audio feedback to help navigate independently [67].

With a high throughput network, people with disabilities can rely on data-driven solutions to become independent. Advances in technologies such as facial recognition will allow them to identify people, objects and important places in their surroundings. Deaf people and any person who needs to communicate using sign language rely significantly on video chat apps to communicate but high latency networks are huge hurdles to them. 5G is expected to satisfy their crucial need for low latency communications [67].

Smart home services enabled by 5G networks are also helpful to people with disabilities who wish to live independently. They are given the possibility to use their smartphone to control their environment, such as to switch on appliances, adjust temperature or operate a door.

People with disabilities can also rely on wearable devices such as smartwatches to track and improve their health. Moreover, wearable GPS devices, allow people with intellectual disabilities to be located and tracked. The quality of life of disabled people will be enhanced with emerging technologies as follows [67]:

- They will be more informed and in communication with the world.
- They will have access to the same education as students without disabilities.
- It will have an opportunity to live and work independently.

Mobile operators use their technical know-how and infrastructure to increase impact at scale for forcibly displaced persons (FDPs) and their host communities. At the end of 2019, around 79.5 million people were forcibly displaced, among which 26 million were refugees. At least 93% of refugees around the world live in regions that have a 2G or 3G coverage [68] and connection to a mobile network can provide crucial support to these populations by providing them with the means to contact relatives and peers, access information platforms, translation applications and mobile money services [69]. Operators are also cooperating with humanitarian organizations to better understand people in need. For instance, Safaricom has partnered with the UN Refugee Agency (UNHCR) and the GSMA to research how people with disabilities use mobile devices for enhancing their lives [70].

2.9.2 SDG 11: Sustainable Cities and Communities

The aim of this SDG is to render cities and human habitats safe, sustainable, inclusive, and resilient [71].

Smart city is an area where technological advances will play a major role. It is estimated that around 66% of the world's population will settle in cities by 2050 according to the UN [72]. Countries are addressing socio-economic issues in sectors such as the energy, healthcare and security sectors by converting fast-expanding urban areas into smart cities.

Buildings and cities are already becoming smarter, more sustainable and safer with the help of mobile connectivity. 5G will accelerate such developments by its support for dense networks of sensors and actuators which will enable organizations, municipalities and individuals to monitor and control their property remotely using the appropriate platforms. Orange is planning to allow consumers to link their smart devices to its Livebox wireless router and also intends to launch a Protected Home service which aims at ensuring home security smartly via video surveillance, with the collaboration of Groupama [26].

AI can be combined with the highly reliable 5G networks to implement the real-time location, video and biometric data analysis. Furthermore, cloud-based facial recognition systems can be applied to high-resolution surveillance cameras to identify and locate law offenders in real-time [73].

Intelligent Connectivity is expected to help address transportation issues. For example, data collected from sensors and actuators on streetlights, highways, vehicles and parking areas will be delivered to a centralized transportation system on the cloud using Intelligent Connectivity. These data will then be processed to extract traffic insights which traffic controllers can use to advise drivers about the optimal and safest route to their destination. Cars could communicate with streetlights on the road they are driving so that only the segments of the road which the driver will require visibility are illuminated, hence saving energy. Technology convergence potentially helps reduce traffic congestion and even mortality rates by optimizing the transportation system [74]. Recently, Bosch conducted research that concluded that by 2025, connected vehicles will reduce traffic accident injury cases by 350,000 yearly and can save around 11,000 lives [75].

2.9.3 SDG 12: Responsible Production and Consumption

The purpose of this SDG is to guarantee sustainable consumption and production patterns [76].

The planet's resources such as fresh water, land, fossil fuels and other minerals are under increasing pressure due to the world's population explosion. According to the United Nations Development Programme (UNDP), lowering the ecological footprint is essential to achieve sustainable development [77]. Lowering the ecological footprint implies achieving sustainable consumption and manufacturing of resources by reducing the use of natural resources, toxic materials, and emissions of pollutants and waste, while satisfying basic consumption needs [78]. As agriculture is one of the most water-consuming economic activities globally, major changes will be required to achieve this SDG. Technology and newer IoT solutions in agriculture play a major role in improving the quantity and the quality of harvests. Since unpredictable weather conditions can affect crops and reduce yields, digital applications and infrastructure tools aid farmers to adjust their decisions on the optimal time to plant and the most suitable crop varieties to choose to achieve higher productivity. Furthermore, they predict the optimal amount of water and agrochemicals that are necessary, contributing to more sustainable agriculture [79]. They can also help monitor and analyze consumption patterns to raise awareness about the risk of unsustainable manufacturing and consumption of soil [52].

Moreover, smart vehicles connected to a 5G network could easily identify the optimal routes and allow food in transit to be remotely monitored, thus increasing the efficiency of food distribution. Furthermore, crops can be delivered with longer market and shop lives with the help of mobile refrigeration. Policymakers may rely on reliable wireless connectivity which will be available to everyone to sensitize people about green transportation options for routine purposes [26]. In 2019, around

53.6 million metric tons of electronic waste (excluding photovoltaic panels) was produced (or 7.3 kg per capita) with 17% of global electronic waste found to be collected and recycled [80]. If no corrective measures are taken, global electronic waste production may reach 120 million tonnes per year by 2050 [81]. Mobile operators in 40 countries around the world are leading 67 electronic waste management initiatives. 43 operators have even set up electronic waste collection points in their workplaces and customer contact centres. Mobile technology also caters for other waste collection and management issues such as plastic recycling [15].

2.9.4 SDG 13: Climate Action

The aim of this SDG is to take pressing action to combat climate change and its impacts [82].

GHG emissions are continuing to rise, especially in developing countries, and are now more than 50% higher than their 1990 level [52]. Global warming is leading to long-term changes to our climate system. Yearly mean losses from tsunamis, tropical cyclones, and flooding amount to hundreds of billions of dollars, requiring an investment of $6 billion annually in disaster risk management alone. The Earth's temperature must be maintained at 1.5 °C above pre-industrial levels to stabilize climate change. This means we need to halve global greenhouse gas emissions by 2030 and reach net zero before 2050 [82].

Intelligent Connectivity powered by 5G can have significant impacts on this SDG. For example, in the energy sector, Intelligent Connectivity can improve the performance and safety of power grids and add remote control access and automation which are useful in power failure scenarios. This is essential in the adoption of renewable energy. The UN estimates that up to 85% of energy must be from renewable sources by 2050 [83].

In buildings such as Ericsson's Smart 5G Factory [84], energy cost is reduced by 5% through energy monitoring and management technologies and waste could be reduced by 5% through environmental monitoring. Generally, the factory has become 24% more energy efficient than it typically was.

Freshwater is a limited resource and it constitutes only 3% of the water on Earth. Furthermore, one-third of freshwater in the world remains inaccessible. The world's population could face a shortage of fresh water in 2025 if no proper water management infrastructure is set up. With 5G, a massive network of IoT sensors can be deployed for the detection of hazardous chemicals in water supplies, for managing water leaks and for providing instantaneous flood warnings. This may help transform the agroindustry [85].

Currently, 70% of the yearly consumption of freshwater in the world is for agriculture [86]. Usually, farmers utilize irrigation systems that are outdated which contribute to climate change. Farmers can adopt the use of smart sensors installed in the soil to obtain moisture level measurements which may help optimize agriculture. By pairing 5G with IoT, data can be transmitted at high speeds, making life more efficient in the fields, and reducing the need for harmful substances like pesticides

[85]. A greener future is one where smartphones can be used by farmers to locate issues in their fields and monitor several parameters of their crops remotely [87].

It is estimated by the World Health Organization that every year, air pollution from sources such as idling car exhaust kills over 3 million people globally [85]. With 5G technology, the appropriate platforms for the control and monitoring of traffic may be implemented to reduce carbon emissions.

Sensors installed in cities, connected to 5G networks will allow authorities to measure pollutants and particulates in real-time at a street level. These measured data can then be used to automatically adjust traffic flow on different streets to improve the air quality [85]. Driverless cars with efficient cruise control and automatic driving features can help increase energy savings by up to 30% [85].

2.9.5 SDG 14: Life Below Water

The temperature, chemistry, currents, and life in oceans are fundamental to enable habitability on earth. Moreover, more than 3 billion people depend on marine and coastal biodiversity for their livelihoods, which represents around 5% of global GDP [88]. Hence, the objective of SDG 14 is the conservation and sustainable use of the oceans, seas, and marine resources, stopping them from suffering the effects of overfishing, marine pollution, and climate change, including ocean acidification [52].

Intelligent Connectivity combined with appropriate software, IoT sensors, AI and mobile devices for the monitoring and analysis of the ocean can contribute to the support of life below water. An appropriate example is the James Cook University (JCU) in Australia which collaborated with an IoT satellite company named Myriota to build a Smart Ocean. The project constituted of the design of an on-water satellite embedded with microsensors and AI technology to collect and send data about the water quality using a 5G network in real-time. This eliminates the need to build communication towers in the ocean to transmit collected data and therefore reduces the cost of Reef monitoring. Moreover, JCU is collaborating with the Cairns Regional Council to construct Australia's first Urban Great Barrier Reef Monitoring System, which will record the amount of nutrients, contaminants and sediments in the waters around the Reef. They will allow the public to access the data online to view the impact human activities have on the Reef [89].

Moreover, the FishGuard project initiated in Seychelles combats illegal fishing by the use of surveillance drones which use AI to identify ships that are not allowed to fish in the region. The drones provide essential data about the location and identification number of ships operating in the region. The project can be further scaled to enable the autonomous monitoring of millions of square kilometers of water and thus reduce patrolling costs [90].

2.9.6 SDG 15: Life on Land

The purpose of this SDG is to safeguard, restore and support sustainable use of terrestrial ecosystems, sustainably manage forests, fight desertification, and stop and inverse land degeneration and stop biodiversity loss [91].

Digital applications and infrastructure can influence the conservation and sustainable usage of terrestrial ecosystems and the prevention of biodiversity loss. These include IoT and mobile sensors, which assist the monitoring of the state of the planet, terrestrial ecosystems, rainforest, desertification and flooding; satellite observation, which assists the monitoring of water flows, rain, snow, and wind patterns, providing efficient early-warning systems to protect endangered species and fragile land areas; and mobile phones, which can be used to track illegal trafficking and poaching of protected species and natural heritage [52].

Rainforest Connection (Refax) [92] is a non-profit group, which is combining Intelligent Connectivity and AI to combat deforestation and to protect millions of species that live in rainforests around the world. The group worked with Huawei Cloud and Data Scientist experts to implement conservation solutions. Refax upcycled old Huawei phones which they used to deploy an intelligent interconnected ecosystem for monitoring rainforests. The phones are linked to AI-equipped servers to instantly prompt the authority in case suspicious sounds of trucks or chainsaws are detected. This solution is currently in used in 10 nations including the US, South Africa, Brazil, Peru and Indonesia and protects above 2500 km^2 of forests [74].

Refax is also aiming at slowing down the extinction rate of species (which is approximately 50,000 species per year) by using Intelligent Connectivity ecosystems. The upcycled old Huawei phones placed in forests were used to listen to Spider Monkeys' sounds and to analyze their behavior and movements [74].

In Guatemala, tree tagging is used as a solution to optimize the tracking, harvesting and operation planning processes. Tree Tag is an emerging smartphone-based supply chain traceability platform developed by Earth Observation Systems. The platform is used to track the location of logs transported from the forest to the mill. It ensures that all authorized personnel involved report activities and volumes, and it sends alerts in case suspicious activities are detected. Only trees that have been authorized for logging can enter the system [93].

2.9.7 SDG 16: Peace Justice and Strong Institutions

The aim of this SDG is to support peaceful and inclusive societies for sustainable development, provision of access to justice for all and building effective, accountable and inclusive institutions at all levels [94].

In spite of the increasing obtainability of data on the experience of people with justice, organizations across the world still have a lot of improvements to make to provide justice for all. As per a recent report by the World Justice Project, it is projected that 5 billion people have unsatisfactory justice requirements worldwide. This comprises 1.5 billion people who cannot obtain civil, administrative, or criminal justice because they are confronted with hurdles obstacles to solving their daily justice matters. Advancements in big data and AI have the potential to improve this quest for sustainable justice systems. According to a study from 2020, AI could have a positive effect on 58% of the targets in SDG 16. If deployed ethically, AI can contribute to justice systems by intelligently automating jurisdictional processes, interacting with citizens to provide juridical guidance, detecting irregularities that may prevent crime and supporting jurisdictional decision making [95].

Transparent transaction records are also essential to protect customers' rights, to protect them from theft and fraud, to foster trust, and to improve social outcomes. Such records can be provided by mobile money platforms. In 2010, the Afghan National Police started using M-Pesa to pay salaries and found that 10% of salaries of the police workforce was being paid to fictitious officers while some officers were deprived of their full salaries [49].

Many countries such as Bolivia, Chile, Colombia, Costa Rica, Guatemala, and Peru have introduced transparency portals to push forward financial transparency and accountability [96]. In short, these portals are websites devoted to publishing public financial information. These countries have all implemented operations to increase transparency in public financial management. These portals aim to guarantee the right of access to information, especially in relation to the administration of public resources, to prevent and combat corruption. Through these portals, public budgets, contracts, and tenders are published periodically to make the actions of the public sector more visible [52].

2.9.8 SDG 17: Partnership for the Goals

The aim of this SDG is to strengthen the means of implementation and revitalize the global partnership for sustainable development [97].

Sustainable development can be better achieved if there are partnerships between governments, civil society, and the private sector toward this common goal. Cooperation between actors across countries is also relevant in this respect. Digital applications and infrastructure can contribute to these alliances by facilitating coordination and communication between these actors and fostering their partnerships. Digital applications are altering the relationships between governments and their citizens by creating opportunities for public decision making through the digitalization of public services [52].

The mobile ecosystem contributes significantly to the public sector by general taxation. Generally, this comprises value-added tax or sales tax, income tax, corporation tax, and social security from the contributions of organizations and employees. Almost half a trillion dollars was funded to the public sector by general taxation in 2019 by the mobile industry [60]. Governments also use mobile technology to enhance their local capabilities and ability to offer economic resources. For example, the digitization of public revenue collection, such as taxes, school fees and fines, mobilizes countries and reinforces their economies. It also increases resources for governments and creates transparent logs of public funds. In Rwanda, a partnership between the public and private sectors resulted in the creation of a centralized electronic payment platform, which supports over 89 online services and has served over 4 million users [15, 98].

Another example is the Geneva Internet Platform (GIP) which is a digital policy platform, which assists academia, governments, civil society, technical communities, and other information-society stakeholders in finding digital policy and governance resources, with a special focus on developing and small countries. The platform also aims at formulating digital strategies and engaging with other stakeholders' policy debates. The GIP is operated by DiploFoundation and is an initiative of the Swiss authorities [99].

3 Concluding Remarks

The contribution of mobile technology to economic and social development over the past two decades has been very significant. With the recent outbreak of the COVID-19 pandemic, it has become more than a necessity to support life-saving and life-sustaining socio-economic activities for millions of people. Moreover, the emergence of 5G combined with other digital technologies such as AI, IoT, cloud computing and blockchain, has seen the emergence of Intelligent Connectivity that has the potential of revolutionizing life on our planet. Intelligent Connectivity is indeed seen as a major contributor for a sustainable future. By capturing a wealth of real-time data via IoT sensors with the ultra-reliable and low latency connections provided by 5G, it will allow real-time analytics to be performed on high power computing cloud platforms, generating timely insights that will allow people to optimize the management of resources in a much more efficient way that it was ever possible. Intelligent Connectivity indeed has the potential to impact all the 17 SDGs in a very positive way. As was observed in this chapter, in addition to obvious SDGs such as SDGs 3, 4, 9 and 11, it can contribute to several other SDGs by helping poverty reduction through mobile money, reduction in CO2 emissions, controlling deforestation, monitoring life underwater, providing new innovative solutions to gender equality and social justice among others. However, the impact of Intelligent Connectivity on the SDGs will be highly dependent on the successful widespread deployment and adoption of 5G. It is therefore important that the challenges faced in its worldwide adoption are addressed through continuous consultations by regulators, industry associations, network operators, service/technology providers and public–private partnership organizations. This will allow the opportunities brought about by 5G across all SDGs to be maximized.

References

1. The 2030 agenda for sustainable development. https://sustainabledevelopment.un.org/content/

documents/21252030AgendaforSustainableDevelopmentweb.pdf

2. Sustainable development goals. https://www.undp.org/content/undp/en/home/sustainabledevelopment-goals.html
3. Salam A (2020) Internet of things for sustainable community development: introduction and overview. Faculty Publications. Paper 23. https://docs.lib.purdue.edu/cit_articles/23
4. Sætra HS (2021) AI in context and the sustainable development goals: factoring in the unsustainability of the sociotechnical system. Sustainability 13:1738. https://doi.org/10.3390/su13041738
5. Sachs JD (2012) From millennium development goals to sustainable development goals. Lancet 379:2206–2211
6. United Nations (2015) Transforming our world: the 2030 agenda for sustainable development. Division for Sustainable Development Goals, New York, NY, USA
7. Pekmezovic A (2019) The UN and goal setting: from the MDGs to the SDGs. In: Walker J, Pekmezovic A, Walker G (eds) In sustainable development goals: harnessing business to achieve the SDGs through finance, technology, and law reform. Wiley, West Sussex, UK, vol 1
8. Gabriel I, Gauri V (2019) Towards a new global narrative for the sustainable development goals. In: Walker J, Pekmezovic A, Walker G (eds) In sustainable development goals: harnessing business to achieve the SDGs through finance, technology, and law reform. Wiley, West Sussex, UK, vol 3
9. Bojkovic Z, Milovanovic D, Fowdur TP (eds) (2021) 5G Multimedia communication: technology, multiservices, and deployment. CRC Press. https://www.routledge.com/5G-Multimedia-Communication-Technology-Multiservices-and-Deployment/Bojkovic-Milovanovic-Fowdur/p/book/9780367178505
10. Fowdur TP, Babooram L, Ibn Nazir Rosun MN-U-D, Indoonundon M (2021) Real-time cloud computing and machine learning applications. Nova Science Publishers. ISBN: 978-1-53619-813-3. https://novapublishers.com/shop/real-time-cloud-computing-and-machine-learning-applications/
11. 2413-2019—IEEE approved draft standard for an architectural framework for the Internet of Things (IoT). https://standards.ieee.org/project/2413.html
12. Swords E, Sumner J, Zeuthen R, Piskorowski J, Gupta A (2020) The era of intelligent connectivity. BNY Mellon Investment Management. https://www.mellon.com/documents/264414/269919/intelligent-connectivity.pdf/1fcce90e-bcb3-6434-c1fa-40144e6e470f?t=1590166797856
13. World Economic Forum (2020) The impact of 5G: creating new value across industries and society. http://www3.weforum.org/docs/WEF_The_Impact_of_5G_Report.pdf
14. IHS Markit (2019) The 5G economy: how 5G will contribute to the global economy. https://www.qualcomm.com/media/documents/files/ihs-5g-economic-impact-study-2019.pdf
15. GSMA, 2020 Mobile Industry SDG Impact Report. Available from: https://www.gsma.com/betterfuture/2020sdgimpactreport/wp-content/uploads/2020/09/2020-Mobile-Industry-Impact-Report-SDGs.pdf?utm_source=better_future_site&utm_medium=search_engine&utm_campaign=2020_SDG_impact_report
16. https://sdgs.un.org/goals/goal1
17. Beuermann DW, McKelvey C, Vakis R (2012) Mobile phones and economic development in rural Peru. J Dev Stud 48(11). https://doi.org/10.1080/00220388.2012.709615
18. Bahia K, Castells P, Cruz G, Masaki T, Pedros X, Pfutze T, Rodriguez-Castelan C, Winkler H (2020) The welfare effects of mobile broadband internet: evidence from Nigeria. Policy Research Working Paper; No. 9230. World Bank, Washington, DC. © World Bank, License: CC BY 3.0 IGO. https://openknowledge.worldbank.org/handle/10986/33712
19. Munyegera GK, Matsumoto T (2016) Mobile money, remittances, and household welfare: panel evidence from rural Uganda. World Dev 79:127–137. ISSN 0305-750X. https://doi.org/10.1016/j.worlddev.2015.11.006, https://www.sciencedirect.com/science/article/pii/S0305750X15002880
20. M-KOPA. https://m-kopa.com/uganda/about/

21. Weber I, How AI is being used to map poverty. Available at: https://www.electronicspecifier.com/products/artificial-intelligence/how-ai-is-being-used-to-map-poverty
22. DAIA, Artificial intelligence and global challenges—no poverty | by DAIA | DAIA| Medium. Available at: https://medium.com/daia/artificial-intelligence-and-global-challenges-a-plan-for-progress-fecd37cc6bda
23. The Borgen Project, Artificial intelligence and poverty. Available at: https://borgenproject.org/tag/artificial-intelligence-and-poverty/
24. Mhlanga D (2020) Artificial Intelligence (AI) and Poverty Reduction in the Fourth Industrial Revolution (4IR). Preprints 2020, 2020090362. https://doi.org/10.20944/preprints202009.0362.v1. https://www.preprints.org/manuscript/202009.0362/v1
25. https://sdgs.un.org/goals/goal2
26. GSMA, Intelligent connectivity: how the combination OF 5G, AI, Big Data and IoT is set to change everything. Available from: https://www.gsma.com/ic/report/
27. Banerjee S, 7 ways the Internet of Things can help end world hunger. World Economic Forum. Available from: https://www.weforum.org/agenda/2018/01/internet-things-iot-world-hunger-supply-chain/
28. https://sdgs.un.org/goals/goal3
29. GiftedMom: https://www.who.int/pmnch/about/members/database/giftedmom/en/
30. GSMA ecosystem accelerator innovation start-up fund portfolio, GSMA, 2020. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2020/04/GSMA-Ecosystem-Accelerator-Innovation-Fund-Start-up-Portfolio.pdf
31. Huawei (2020) Shaping the new normal with intelligent connectivity. GCI 2020. Available: https://www.huawei.com/minisite/gci/assets/files/gci_2020_whitepaper_en.pdf?v=20201217v2
32. Basu M, How Singapore built an AI temperature tool in two weeks, GovInsider. https://govinsider.asia/innovation/COVID-coronavirus-singapore-ihis-kronikare-temperature-ai/
33. Dickson B, Why AI might be the most effective weapon we have to fight COVID-19. Online: https://thenextweb.com/neural/2020/03/21/why-ai-might-be-the-most-effective-weapon-we-have-to-fight-COVID-19/
34. Bitran H, Delivering information and eliminating bottlenecks with CDC's COVID-19 assessment bot, Microsoft. Available from: https://blogs.microsoft.com/blog/2020/03/20/delivering-information-and-eliminating-bottlenecks-with-cdcs-COVID-19-assessment-bot/
35. Tan T, Tapping AI to battle Covid-19. The StraitsTimes. https://www.straitstimes.com/tech/tapping-ai-to-battle-COVID-19
36. Xuelai Fan S, DeepMind's protein folding AI is going after coronavirus. SingularityHub. https://singularityhub.com/2020/03/17/how-deepminds-ai-is-working-to-decode-coronavirus/
37. https://sdgs.un.org/goals/goal4
38. Bennington-Castro J (2017) AI is a game-changer in the fight against hunger and poverty. Here's Why, NBC News. Available at: https://www.nbcnews.com/mach/tech/ai-gamechanger-fight-against-hunger-poverty-here-s-why-ncna774696
39. Progress towards the sustainable development goals. United Nations, 2020. https://unstats.un.org/sdgs/report/2020
40. The World Bank, How countries are using edtech (including online learning, radio, television, texting) to support access to remote learning during the COVID-19 pandemic. Available online: https://www.worldbank.org/en/topic/edutech/brief/how-countries-are-using-edtech-to-support-remote-learning-during-the-covid-19-pandemic
41. Heaven D (2019) Two minds are better than one. New Sci 243:38–41. https://www.sciencedirect.com/science/article/abs/pii/S0262407919315842?via%3Dihub
42. Nwana HS (1990) Intelligent tutoring systems: an overview. Artif Intell Rev 4:251–277. https://link.springer.com/article/10.1007%2FBF00168958
43. https://sdgs.un.org/goals/goal5
44. https://salvatio.dk/
45. Klimt S, Using tech to challenge gender inequality? An SDG 5 progress analysis, tech2impact. Available: https://tech2impact.com/using-tech-to-challenge-gender-inequality-an-sdg-5-progress-analysis/

46. Vodafone, The future is female: how IoT is changing how tech handles diversity. Available: https://www.vodafone.com/news/inclusion/the-future-is-female
47. Gilai H (2019) Women's safety: RTA plans to install 'Abhaya' devices in Autorickshaws. The Hindu. https://www.thehindu.com/news/cities/Visakhapatnam/womens-safety-rta-plans-to-install-abhaya-devices-in-autorickshaws/article26164257.ece
48. https://sdgs.un.org/goals/goal6
49. GSMA (2019) Harnessing the power of mobile money to achieve the sustainable development goals. Testing the Waters, CGAP and GSMA. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/10/GSMA-Harnessing-the-power-of-mobile-money-to-achieve-the-SDGs.pdf
50. WaterAid (2018) One token changing the game for sustainable rural water supply in Tanzania. https://washmatters.wateraid.org/blog/one-token-changing-the-game-for-sustainable-rural-water-supply-in-tanzania
51. Libelium, Saving water with smart management and efficient systems in Spain. Available: http://www.libelium.com/saving-water-with-smart-man-agement-and-efficient-systems-in-spain/
52. Zaballos AG, Iglesias E, Adamowicz A (2019) The impact of digital infrastructure on the sustainable development goals—a study for selected Latin American and Caribbean Countries. Copyright © 2019 Inter-American Development Bank. Available: https://publications.iadb.org/publications/english/document/The_Impact_of_Digital_Infrastructure_on_the_Sustainable_Development_Goals_A_Study_for_Selected_Latin_American_and_Caribbean_Countries_en_en.pdf
53. https://sdgs.un.org/goals/goal7
54. World Economic Forum, Top 10 Emerging Technologies. Insight Report 2019. Available: http://www3.weforum.org/docs/WEF_Top_10_Emerging_Technologies_2019_Report.pdf
55. Rowling M (2019) Subsidise solar mini-grids to power rural Africa, investors urge. Reuters. Available: https://www.reuters.com/article/global-energy-solar/subsidise-solar-mini-grids-to-power-rural-africa-investors-urge-idUKL8N23H1UC?edition-redirect=uk
56. Elkin C, Witherspoon S (2019) Machine learning can boost the value of wind energy. DeepMind 2019. Available: https://deepmind.com/blog/article/machine-learning-can-boost-value-wind-energy
57. Bauer GK (2019) Mini-grids, macro impact? GSMA. Available: https://www.gsma.com/mobilefordevelopment/blog/mini-grids-macro-impact/
58. https://sdgs.un.org/goals/goal8
59. Bahia K, Castells P, Pedros X (2020) Mobile technology: two decades driving economic growth. GSMA Intelligence Economic Research Working Paper. Available: https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=54165922&file=121120-working-paper.pdf
60. The Mobile Economy 2020, GSMA. Available: https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf
61. https://sdgs.un.org/goals/goal9
62. Lidong W, Guanghui W (2016) Big data in cyber-physical systems, digital manufacturing and Industry 4.0. Int J Eng Manuf 6:1–8. https://doi.org/10.5815/ijem.2016.04.01
63. https://www.brightmachines.com/
64. http://www.bell-labs.com/about/locations/oulu-finland/#gref
65. https://sdgs.un.org/goals/goal10
66. https://www.itu.int/en/sustainable-world/Pages/goal10.aspx
67. Qureshi A (2020) 5G, IoT and Wearable—a game changer for persons with disabilities. Disability Insider. Available: https://disabilityinsider.com/story/5g-iot-and-wearable-a-game-changer-for-persons-with-disabilities/
68. UNHCR (2016) Connecting refugees. Available: https://www.unhcr.org/publications/operations/5770d43c4/connecting-refugees.html
69. GSMA (2019) The digital lives of refugees: how displaced populations use mobile phones and what gets in the way. Available: https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/07/The-Digital-Lives-of-Refugees.pdf

70. GSMA (2020) The digital lives of refugees and Kenyans with disabilities in Nairobi. Available: https://www.gsma.com/mobilefordevelopment/resources/the-digital-lives-of-refugees-and-kenyans-with-disabilities/
71. https://sdgs.un.org/goals/goal11
72. United Nations, World's population increasingly urban with more than half living in urban areas. Available: http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html
73. Intelligent Connectivity—How the Combination of 5G, AI and IOT IS set to change the Americas, GSMA. Available: https://itig-iraq.iq/wp-content/uploads/2019/05/21494-MWC-Americas-report.pdf
74. Huawei, Powering intelligent connectivity with global collaboration. GCI 2019. Available: https://www.huawei.com/minisite/gci/assets/files/gci_2019_whitepaper_en.pdf?v=20191217v2
75. Bosch, Connected car effect 2025. Available: https://www.bosch-presse.de/pressportal/de/en/bosch-study-shows-more-safety-more-efficiency-more-free-time-with-connected-mobility-82818.html
76. https://sdgs.un.org/goals/goal12
77. http://www.latinamerica.undp.org/content/rblac/en/home/sustainable-development-goals/goal-12-responsi-ble-consumption-and-production.html
78. Sustainable development in Latin America and The Caribbean: trends, progress, and challenges in sustainable consumption and production, mining, transport, chemicals and waste management-report to the eighteenth Session of the Commission on Sustainable Development of the United Nations, 2010. Available: https://sustainabledevelopment.un.org/content/documents/LAC_background_eng.pdf
79. UNFCC, Information and communications technology solutions. Available: https://unfccc.int/climate-action/momentum-for-change/information-and-communications-technology-solutions
80. Forti V, Balde CP, Kuehr R, Bel G, The Global E-waste Monitor 2020: Quantities, flows and the circular economy potential (Bonn, Geneva and Rotterdam: United Nations University/United Nations Institute for Training and Research, International Telecommunication Union, and International Solid Waste Association, 2020)
81. World Economic Forum (2019) A new circular vision for electronics: time for a global reboot. PACE and The E-waste Coalition. Available: http://www3.weforum.org/docs/WEF_A_New_Circular_Vision_for_Electronics.pdf
82. https://sdgs.un.org/goals/goal13
83. ERISSON, Digitalization with 5G enables further acceleration of climate action. Available: https://www.ericsson.com/en/blog/2021/1/digitalization-5g-climate-action
84. Ericsson USA 5G Smart Factory, Smart manufacturing. Available: https://www.ericsson.com/en/about-us/company-facts/ericsson-worldwide/united-states/5g-smart-factory
85. West DM (2016) Achieving sustainability in a 5G world. Center for Technology Innovation at Brookings. Available: https://www.brookings.edu/wp-content/uploads/2016/11/gs_20161201_smartcities_paper.pdf
86. World Bank, Water in Agriculture. Available: https://www.worldbank.org/en/topic/water-in-agriculture
87. Nell A (2021) 5G climate change: 4 ways 5G is helping the environment. CENGN. Available: https://www.cengn.ca/5g-climate-change-4-ways-5g-helping-environment/
88. https://www.latinamerica.undp.org/content/rblac/en/home/sustainable-development-goals/goal-14-life-below-water.html
89. Govinsider (2019) How 5G can save our environment. Available: https://govinsider.asia/connected-gov/how-5g-can-save-our-environment/
90. Murison M (2018) Drone project aims to combat illegal fishing in the Seychelles. Dronelife. Available: https://dronelife.com/2018/08/20/drone-atlan-illegal-fishing-seychelles/
91. https://sdgs.un.org/goals/goal15
92. Rainforest Connection. https://rfcx.org/

93. Case Study: Smartphone tree tagging in Guatemala. Available: http://www.appsolutelydigital.com/ict_new/section_4_3_4.html
94. https://sdgs.un.org/goals/goal16
95. Hanania P-A, Maier M (2021) AI for justice—bringing data into the courtroom. Capgemini. Available: https://www.capgemini.com/2021/06/ai-for-justice-bringing-data-into-the-courtroom/
96. Solana M, Delivering public financial information to citizens in five Latin American Countries. eTransparency Case Study No. 15. Available: http://www.egov4dev.org/transparency/case/laportals.shtml
97. https://sdgs.un.org/goals/goal17
98. Bizimungu J (2018) Electronic certificates: how Irembo is on course to making Rwanda paperless. The New Times, November 2018. Available: https://www.newtimes.co.rw/news/electronic-certificates-how-irembo-course-making-rwanda-paperless
99. Geneva Internet Platform: https://giplatform.org/about-gip

Implementation of Intelligent IoT

Akarsh K. Nair, Chinju John, and Jayakrushna Sahoo

Abstract "Internet of Things" is on the path of revolutionizing traditional technologies in multiple facets. The trend of the future is expected to be Artificial Intelligence enabled IoT (AIoT) due to the widespread applicability of the technology. The compactness of IoT devices leveraged with the powerful nature of AI makes it an ideal alternative for the cumbersome systems currently in use. The increasing population and the need for advancements are facilitating the growth of the technology even faster. The application domains of AIoT are vast and vivid. From healthcare to autonomous transports, robotics to smart cities, everywhere the traces are visible. With all this being said, a keen interest should be given to the sustainable aspect of technological development as well. With the resources ever-shrinking and the population increasing, sustainability is the need of the era. The high efficiency and power-saving nature of IoT devices along with their high versatility in work makes it a paragon choice for sustainable technologies. Through this work, we provide an overview of intelligent IoT from multiple angles. The need, as well as the methods of implementing AIoT, is also presented. We also discuss the multiple application domains making use of this technology.

Keywords IoT · Artificial Intelligence · Intelligent IoT · AI enabled IoT · Sustainable development · Green IoT

A. K. Nair (✉) · C. John · J. Sahoo
Indian Institute of Information Technology, Kottayam, Kerala, India
e-mail: akarsh.k.nair.phd201011@iiitkottayam.ac.in

C. John
e-mail: chinjuj.phd201001@iiitkottayam.ac.in

J. Sahoo
e-mail: jsahoo@iiitkottayam.ac.in

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_2

1 Introduction

Artificial Intelligence (AI) is a highly studied topic due to its wide pertinence and adequacy. AI prompts the system to undertake “human like” intelligent tasks. The prime factor that differentiates humans and machines was the ability to perform cognition. With the initiation of AI, machines have evolved to perform cognitive tasks with ease which were previously reserved for humans. Primarily, AI utilizes datasets to infer particular patterns or outcomes using which it will be able to perform the assigned tasks. The application domain of AI is wide and highly varying. An intelligent system needs to surpass certain steps before achieving the so-called “intelligent skills”. They are acquisition, making interpretations, and adaptive skills.

1.1 Acquisition Process

The primary feature of AI aims at collecting data and devising a set of rules through which the data can be turned into legitimate information. Such rules are formally referred to as algorithms and they provide the system with a detailed guidelines on how a task needs to be performed.

1.2 Interpretation Process

This process is also known as reasoning process. The efficiency of an AI system can be defined by its capability of applying the right program at the right instance. Only by making extensive interpretations, the system attains such capabilities.

1.3 Adaptive Process

In this stage, the algorithms undergo constant improvisations aiming for attaining ideal results as much as possible. Hence they are also referred as self-correction processes.

AI is a very generic term, which refers to a set of different technologies rather than a single one. The basic motto will be the same for these technologies and the difference will be with respect to the method of implementation and areas of applications. Some of the prominent sub-domains of AI are Machine Learning (ML), Deep Learning (DL), Computer Vision (CV), Natural Language Processing (NLP), and so on. Most of these have multiple areas of application on their own and it increases manifold when a factor of combination with other technologies gets introduced.

Another domain that has recently gained popularity is the Internet of Things or IoT. It refers to a huge number of physical devices connected in a network, all doing their part in the performance of the network in general and collectively. With the widespread of cheap micro-level computer chips and universality of presence in wireless networks, it is very much possible to integrate such several small devices and combine them to form a huge device or a network, making them a part of the IoT. Connecting such heterogeneous objects and adding sensors to those networks will open up another aspect of "intelligence" to the network. It can be referred as digital intelligence and the level varies with the number and capabilities of such sensors. Theses sensors, are enabled with real-time data processing capabilities along with the possibility of human interaction [1]. Without such sensors, IoT networks will be usually an "obtuse" group of devices incapable of doing any peculiar tasks.

The IoT technology has enabled a drastic change in the structure of the implementation of various technologies as a whole, reducing the gap between "physical entities" and "digital aspects" in the technical world. It can be seen that such devices range from very simple household objects to highly sophisticated technological tools. With the current growth rate, it is expected that the number of IoT devices will be reaching above 20 billion within five years, as most of the major companies are widely into IoT. Following the trend, smaller industries are also shifting their concentration towards IoT. With the development of ICT technology throughout the world, the applications of AI are being used in every nuke and corner. AI when combined with other domains simply escalates its level of applicability. AI on IoT, AI with Blockchain are just a couple of such applications that are coming to the limelight in the current scenario.

IoT and AI are two very powerful technologies on their own. When they are combined, they are referred to as Intelligent IoT or Artificially Intelligent Internet of Things (AIoT). Such a combination can be easily illustrated by considering a digital counterpart of the human central nervous system where IoT devices act as the nervous system and the AI part serves the brain part.

The very fundamental concept of AIoT is centered on the IoT [2]. The system organization is in such a way that when these so-called "things" like computer chips, personal assistants, or any sort of electronic devices or even virtual entities are connected to each other or in a network, they can recognize each other and collaboratively do the collection and processing of data. This is the part where IoT exists predominantly. AI will be coming in the latter part of the system where using the IoT devices it tries to accomplish a specific task or to learn certain things from the data, which requires some degree of intelligence. Hence, we can say that AI added to IoT imply that the IoT devices are equipped with the ability to study the data, take decisions and manipulate the data with a certain degree of autonomy. Thus, these devices become "smart" and the system behaves oriented on performance and potency [3]. The intelligence factor in IoT enables the incorporation of data analytics into the mix which in turn can be used for better optimization and come in handy to generate higher performing systems with the help of highly useful data as a result of optimization procedures. The rest of this chapter is divided into six sections. In Sect. 2 we discuss the need for intelligent IoT in detail. Section 3, explores the methodology of implementation of Intelligent IoT, then follows a detailed study on the application domains in Sect. 4. Section 5 discusses the future scope of the work and Sect. 6 concludes the whole chapter.

2 The Need for Intelligent IoT?

The role of AI for the growth of IoT is very eminent in both application as well as deployment phases. Such a statement can be practically proved from the visible growth of the companies which had tried to integrate these technologies in the last few years. This has laid a new milestone in the industry where even major marketers of IoT applications are now associating it with AI in all possible ways.

The scope of AI in the context of IoT lies in its sole capability of making wrest inferences from data. ML is one such highly studied domain of AI that can be combined with IoT. In such instances, it adds the capability of automatic identification of patterns or does anomaly detection from the data that will be received from sensors or other kinds IoT devices. When compared to existing methodologies for intelligence applications used in areas such as business analytics, the usage of ML simply makes such prediction or analytical operations perform much more efficiently than ever before. Other AI technologies such as Speech Recognition, CV, and DL also helps in drawing insights from the data that required human intervention in former time [4]. Basically, it can be concluded that the application of AI in IoT simply enables us to attain multiple goals such as resolving issues with unexpected outages, achieve more efficiency, develop novel and improved applications and enhance the project management as well. All such individual instances will be discussed below in detail.

2.1 Resolving Unexpected Outage Issue

For larger industrial applications, any sort of unexpected system outage will result in a huge loss on monetary basis. Hence, we can employ AI techniques as a preventive measure for the same. Predictive system maintenance is one such example technique. It makes use of analytics for predicting system breakdown prior to its happening by keeping track of its output results and helps in providing timely reminders on system sustaining procedures thus mitigating any chances of incurring economic losses dues to an unexpected system outage. We can also make use of ML for identifying patterns or drawing inferences from a continuous stream of data for a particular machinery to perform the above mentioned tasks. In that case, we will be tracking any particular triggers or patterns that will indicate the approaching system failure. The high efficiency of such applications does add to their increasing popularity.

2.2 Enhancing Efficiency

Another scope of AI-based IoT is in enhancing the operational efficiency of the system. Just as in the previous case, we can make use of ML here also. In earlier attempts, we made use of ML for the prediction of equipment failures, whereas here it can be used to predict the operating conditions and the identification of multiple parameters that controls the system. Those parameters can be repeatedly fine-tuned during the processing thus maintaining an ideal situation for the system which in turn can lead to higher efficiency and optimal results [5]. This is made possible with the help of pattern identification from the input data which is not feasible for an average human brain.

ML also helps us to identify counter-intuitive inferences in some cases. For example, an ML tool when applied in shipping logistics suggested that clearing the ship's hull of debris will increase the total profit of the system. It was a very painstaking and costly effort and also makes the system undergo a purposeful maintenance outage. Even though it was against the traditional practice, it proved to be true as the process increases the fuel efficiency of the vessel thus leading to more profit even after considering the time duration lost in maintenance and cost of maintenance.

2.3 Development of Newer Applications

The combination of AI and IoT will also pave the way for novel applications and services to be developed. There are various instances proving the applicability of the statement. Natural language processing is one such case. NLP has been widely used for machine-human communication through verbal and textual command and is gaining better performance and on the verge of getting human operators replaced in many cases. Also when it comes to remote location access for both surveillance or delivery purpose, where humans fail to perform well, AI-powered drones and robots are proving to be highly resourceful. Similarly, when it comes to navigation problems as well, such applications are proving their mettle. IoT devices combined with AI can be used to take inputs from multiple measurable points in vehicles and this data can be used for efficient path planning thereby reduce the downtime. Autonomous vehicles are another application in the same context. They also make use of similar techniques to mitigate navigation problems but also adds another additional layer of control over the system without human intervention, thus taking the autonomous features to a whole new level.

2.4 Enhancing Project Management

Several instances of applications using AI paired with IoT have been employed for analyzing different risks existing in a particular organizational context. Some of them include, automation of rapid response generation during emergencies and helping them in achieving a higher level of employee safety, reduced economic losses, and even prevention of cyber-attacks. Such applications are already available for usage to the general public in few of the very commonly used machines such as in ATMs for forged customer detection, identification of perilous situations in factory environments, and even for anomaly detection related to surveillance applications.

2.5 Deductions for Business Ventures

From the above discussions, it can be inferred that for industrial business ventures based on IoT, AI becomes part of the package during deployments and helps in attaining better performance, thus giving a head start compared to regular IoT applications. ML is the most applied application of AI especially due to its predictive functionalities, hence getting itself placed with broad-intent as well as industrial IoT platforms. Azure, IBM, AWS are just some of the many prominent names in the list of applicants using the above-described technology.

With the addition of AI, it is even possible to add more value to the IoT deployments which were primarily not designed for incorporating AI. It can be concluded that the IoT deployments create huge amounts of data in a continuous manner for which the AI applications are best suited for analyzing and drawing inferences that simply escalate the value of the whole thing. With the current technological advancement rate, it will be very hard to find applications running on pure IoT without making use of AI in the near future. It is estimated that the future of IoT is AI and without AI, IoT-generated data will fail to become “information” and rather stay as “raw data” itself.

3 How Is It Made Possible?

As discussed earlier, IoT is a very vast notion encircling around a large base of heterogeneous concepts such as sensors, storage devices, actuators, and data processing systems interlinked by the Internet. Thus, it becomes part of the basic capability of IoT devices to understand their surroundings, transfer, store, and work upon data gathered and make decisions accordingly. The last phase of “decision making” will be dependent upon the processing steps and the driving situations. The real intelligence or smartness of an IoT system can be determined from the level of decision-making it is able to perform. A non-intelligent IoT system will be having a very limited decision-making capability as well as the inability to rise up to the need with perturbations in data and situations. Whereas an intelligent system will be having AI capabilities and can serve the purpose when it comes to tasks such as automation and adaptation [6]. The implementation of AI-enabled IoT is better to be discussed with respect to its application instances as it conveys more meaning rather than when discussed as pure technology.

Voice assistants are one of the common day-to-day life applications of AIoT. Such devices have a wide variety of task performing capabilities under its arsenal as it make use of various sub-domains of AI in its applications. Automated far field sound recognition, wake word detection, voice-to-text conversion, NLP and understanding, contextual reasoning, dialogue management, question answering, conversational AI tasks, and so on are some of the common techniques which are performed regularly for making the voice assistants perform functions on a real-time basis.

Recent technological advancements have led to rapid developments in robotics, enabling the creation of robots that have a higher degree of resemblance when performing human-like activities and increased human interaction along with understanding and reciprocation capability of certain human emotions. Robots are higher-end versions of basic IoT systems as they comprise multiple sensors and controllers combined with AI that equips them for continuous learning and adaptations to changing situations over time. The high versatility in functionalities of such robots mean that several AI technologies such as NLP, CV, shape recognition, object recognition, detection and tracking, facial recognition, voice recognition, voice-to-text technology, obstacle recognition, and so on are being widely employed in those systems.

Apart from the discussed applications, AI-enabled IoT has several other applications generally referred to as Smart objects/devices that have a certain degree of autonomy helping to reduce the load of certain tasks on humans. Such devices also make use of applications such as object detection and identification, speech recognition and expression identification, transfer learning etc. AI enabled IoT are not just limited to houses but used for commercial and industrial applications as well. They combine different AI and ML algorithms to perform predictions by doing statistical and financial analysis of data available at the institutions.

From the already visible results, we can infer that the capabilities of both AI and IoT just get an exponential increase when implemented together. With the knowledge generated from ML, and Big Data Analysis, systems have the potential to make use of data to draw inferences that further adds value to the data. Void of AI, the same data becomes useless. IoT needs to depend on AI as the sheer volume of data created by IoT devices can never be interpreted by humans for generating information. Moreover, AI systems have the capabilities of making use of the inferences to perform self-learning which can never be achieved with a non-AI IoT system [7].

4 Areas of Application?

AI and IoT are two domains that have big prospects for sustainable development by bringing down the unsustainability that muddles the society, the environment, and natural resources. The rapid development of the integrative technologies based on AI and IoT, the freshness of the concept and the limited volume of studies completed is a major hurdle in the area with respect to sustainable development. The applicability of the technology in sustainable development still needs extensive research as major possibilities are yet to be unexplored. The aim of our study is to give a small overview of different application domains of the Intelligent-IoT technology and give an analysis on them. Listed below are some of the very few but prominent applications of AIoT in the context of sustainable development in emerging countries

4.1 Civil Infrastructure

Civil infrastructure is a crucial aspect determining the scale of development of a society. From the very basic amenities such as sanitation to housing, transportation networks to power-generation facilities, it is the backbone of society enabling several functions and multiple economies to thrive. Thus it makes infrastructural developments the heart and core of Sustainable Developmental Goals (SDG). Some of the very common SDG goals are amenities like ample health, education to all, access to clean water, food and sanitation, and so on. These all can only be facilitated with the aid of infrastructural improvements. Ideally, civil infrastructure should never be visualized as individual asset. Entities like power generation plants, health care facilities, water bodies and their allied systems are a part of the community and should be viewed as collective entities as they possess the high capability of delivering all the three major pillars of sustainable development, i.e. the economy, the environment, and the social sustainability. With respect to the economy, the returns from infrastructural developments are given back in the form of opportunities starting right from the construction and maintenance work of buildings to providing economic value adding entities such as bridges or towers, which connect rural and outside world leading to the enhancement in revenue generation and living standards. When it comes to environmental preservation as well, sustainable infrastructure plays a key role by putting less pressure on the environment through nature-friendly construction and allied processes. When equal infrastructure access to all are ensured, automatically the aspect of social sustainability can be addressed as well.

The role of Intelligent IoT in civil infrastructure is becoming prominent with time [8]. Intelligent IoT has the capability to be employed in multiple applications in the context of construction industry. From basic data oriented tasks to hardware based applications, AIoT has a strong grip in civil infrastructure domain. Figure 1 provides a basic idea of multiple applications of AIoT already in use. As with any other application, AI-based applications do perform the task of data analytics with ease in the construction industry as well. As with any other industry, human resource management is a major task in the construction industry as well due to the large number of laborers employed and their working nature which vary drastically depending upon the cites and nature of work. So a labour management system is necessary as the size of the working site increases. IoT application with the aid of AI techniques helps us to achieve the same. IoT devices such as smart tags can be used as identification systems for labourers entering and leaving the site thus helping to manage a strong employee database. The combination of AI into the system also enables the administrator to make inferences such as the number of labourers, duration of work,breaks during the work and so on from the system itself. The major advantage of such systems is its capability to be operated remotely. Also, optimal usage of human resources can be ensured with the aid of such systems.

When it comes to environmental safety management as well, IoT plays a keen role [9]. Safety management systems are the major need of the constriction industry for reducing the damages caused due to accidents at work sites. IoT can be used

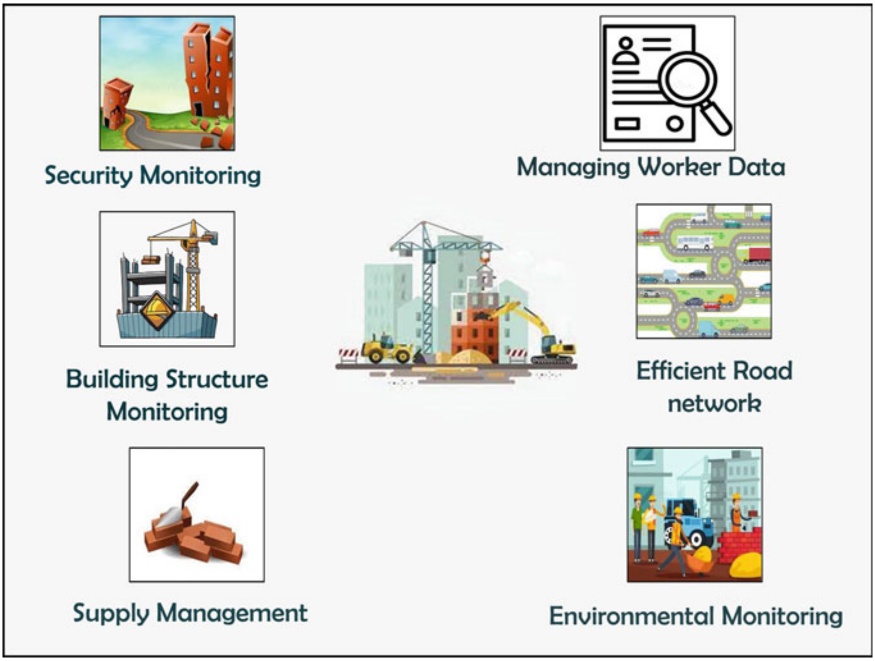


Fig. 1 Application of AIoT in civil infrastructure

to develop systems that monitor the work either from a remote or nearby location. The system can make use of sensors such as proximity sensors placed at hazardous locations to alert people or even small cameras which facilitates real-time video monitoring [10]. When such IoT devices are connected in a high-speed network along with some AI technologies, systems capable of generating danger alerts can be devised. They will be making use of anomaly detection procedures to identify situations or triggers, which can become the causes for potent accidents at work sites. Also, IoT devices have proved their mettle in supply chain management as well [11]. Such systems can be used to measure the stock of available resources in work cites with the help of technologies such as RFID labels that could be attached to materials. Thus the system will automatically generate alerts, if the amount of materials goes beyond a certain optimum level. This helps to keep track of the resources remotely making sure of their availability. Also, such clear data of resource availability and usage helps us to make sure that optimal resource utilisation is being done. A major issue in the construction industry in most of the developing countries is the sheer cost of resources along with their wastage. Such techniques help us to eradicate the inefficient resource management and it in turn help to attain SGDs.

Also when it comes to building structure monitoring, IoT has its applicability. Such systems are not just limited to building time deployments but can be used in regular instances as well. Structural Health Monitoring (SHM) simply means to continuously monitor the strength of civil and industrial buildings to make sure that they are structurally sound for human in-habitation. This procedure reduces the maintenance cost as well as increases the safety of inhabitants in the long run [12]. SHM systems can be employed at different levels depending upon the structure and size of the building. They usually make use of IoT devices such as sensors to detect certain events that can be categorized as things leading to the weakening of the building structures due to several reasons like aging, environmental issues, and so on. In the majority of cases, SHMs are monitoring systems but the parameters they monitor are different from the conventional systems. They mainly lookout for changes in humidity, temperatures, vibrations, tensile stress, and material degradation. When combined with AI techniques, such parameters and changes happening to them can be used to draw inferences to the changing nature of the strength of these structures.

4.2 Health Care Sector

Another sector that needs to be discussed hand in hand with civil infrastructure is the health care domain. Sustainable health care systems can be defined as systems that enhances, preserves or rejuvenates health while lessening its negative impact on the environment and making use of every chance available to reinstate or improve it for the betterment of health and well-being of the current as well as future generations. Extensive research on health care activities has proved the significant impact and the pressure caused by them on the environment. It may vary from toxic to conventional waste generation, extensive water wastage to high energy consumption and even a growing amount of e-waste on the technological side. It is another proven fact that most such wastes are directly harmful to nature. Due to such issues, the need for alternative technology is growing in hospital-based and allied applications. AIoT is now taking that place by providing technological advancements in the healthcare sector leading to multiple health and environmental benefits. The new set of technologies are very effective when it comes to energy conservation as well as e-waste minimization [13]. It further helps us in reducing the total cost and facilitates several other positive advancements to the sector.

AIoT is revolutionising the way health care is perceived. From personal health care assistants to remote health care monitoring systems, from assisting robots to other support systems, AIoT is gradually extending its strong grip on sustainable healthcare [14]. A basic overview of different AIoT based applications is given in Fig. 2. The major issue faced in developing countries with respect to health care is the lack of accessibility to resources for the people living in remote areas. Quality hospitals and treatments are limited to developed places leading to small patches of areas with all resources readily available and others areas that are void of basic health care amenities. The major leap made by AIoT in healthcare is by breaking down all the conventional practices. AIoT has facilitated a decentralised approach in the health care system by making technology-based systems available to people where skilled medics are available on a limited basis [15]. By using the term "technology-based

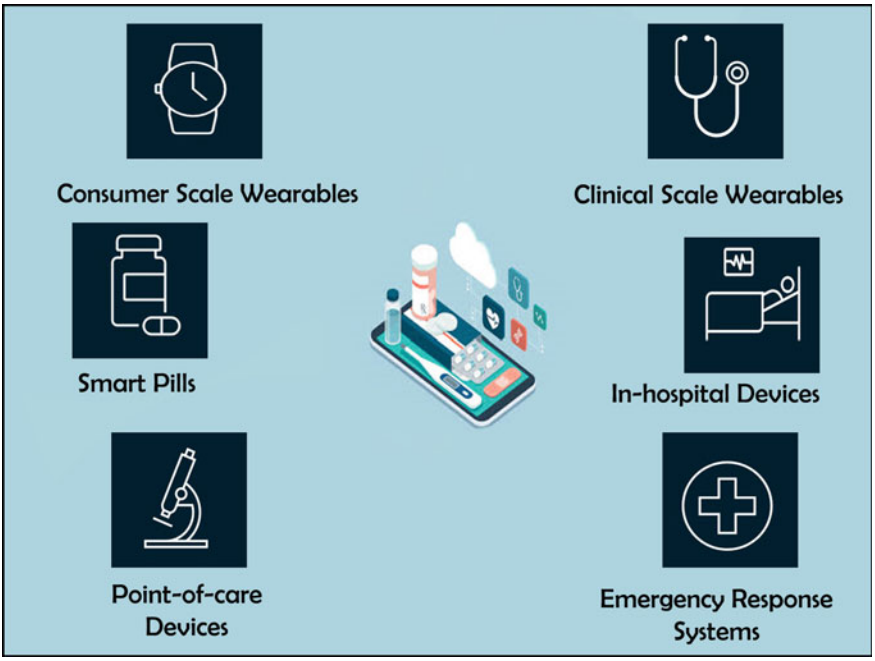


Fig. 2 Applications of AIoT in healthcare sector

systems", we mean AIoT and AI applications such as disease diagnostic systems, medical expert systems and so on which have certain degree of autonomy on their own.

One such rapid technological advancement made is in telemedicine management. The usual dependency on local health care centres with specialised technicians and equipment for performing certain tasks or generating medical diagnoses is high. This directly affects its availability to the needy as the unavailability of medical personals are a real issue in remote areas. For mitigating such issues, one of the major proposals made were centred on the medical decision support systems [16]. The idea was to use AIoT technologies to make systems capable of determining whether a person was having any major ailments or issues. The system was capable of collecting patient's data, processing it and making inferences without any outside aid with much more efficiency than its human counterpart. The peculiarity is that such systems could be operated by non-technical people as the system itself possessed a certain level of autonomy. The system being a low-cost solution is easily accessible to the most deprived class of the society ensuring that a certain level of equity could be achieved in the society at least in terms of health care resources [17].

IoT has also been widely used these days for managing living things using non-living artefacts, starting from the analysis of cardiac beats. There are proposed AIoT based systems that include a wearable ECG patch that is connected to a cloud-based system for the collection and storage of ECG signals. Using a Convolutional Neural Network this data can be analysed and determined whether the person has any cardiac abnormalities [18]. This integrated health care system uses a professional cardiologist's advice as a reference before concluding the results, and it takes the patient monitoring system to the next level. Remote monitoring of patients has also been revolutionized with the invention of AIoT. The utilities such as smart clothing-based health care systems are helping to monitor the bed-ridden patients who need not stay at hospitals for further medical assistance [19]. When it comes to monitoring of certain chronic diseases like Asthma, some contributions from the environmental parameters like pollution rate, travelled places, atmospheric changes are required. Such ideas were incorporated in developing wearable self-powered sensory systems which makes it an ideal application for developing economies due to their low cost and high applicability. In countries where skilled human resources are minimal, such systems play a major role in achieving sustainability and equity in the economy.

The advanced versions of wireless body networks called Medical Super Sensors with improved memory capacity and communication ability can retrieve information like a patient's blood pressure, ECG, heartbeat rate and other vital measures. Information collected from these sensors could be mined and summarised using an AI-based module and thereby reducing the complexity. Since these models are evolving with learning, they could easily handle the troubles caused by such a huge volume of data [20]. High-end Decision Support Systems can support physicians in choosing the best feasible diagnostic protocol for his patient based on his medical history. Similarly, wireless wrist bands could save the patients from their long waits at labs and registration counters. Thus medical personnel, as well as clinical resources, can be used optimally ensuring an increased number of beneficiaries.

The success of IoT based health care systems lies in the coordination and communication of the stakeholders in a regulated manner and these systems have a significant role in handling hazardous events with ease. The involvement of robots in surgeries is a blessing for mankind, such robots can help in complex surgeries to be rehearsed before they are being conducted and provide a familiarized intuition on the work environment to the surgeons [21]. With the aid of computers, doctors can perform remote surgeries and IoT assisted devices could help in keeping track of the recovery. These tele-surgeries are being more common these days with the advancement of the Internet of Robotics. Other than doing surgeries such robots could act as a patient bystander and mental therapist in certain medical-cases as well.

4.3 Intelligent Transport Systems

Intelligent Transportation Systems (ITS) usually work by maintaining a continuous transfer of information between the vehicles and road-side frameworks. ITS are not just limited to automated vehicles but has a wider perspective. Such technologies are usually considered as a fundamental technology for attaining SDG as they have high capability in ensuring optimal usage of existing resources in the long run.

From increased road capacity to higher productivity of vehicles, improved quality of experience to higher usage from a vehicle, the list of positive traits goes long. The peculiarity of such systems is that the effects of such vehicles are mainly visible at the vehicular level. Even so, ITSs has the ability to be used at different levels of transportation industry as a building block for the gradual development of cities, enterprises and so on [22]. Extensive research has shown that the current intelligent transportation technologies being used have instilled a considerable reduction in the emission of greenhouse gases, improved fuel efficiency and reduction in travel time without needing to sacrifice the mobility of the regular citizens who can't afford such systems. Thus the further developments in state-of-the-art transportation based innovations will facilitate power and resource efficiency which are one among the premium needs for developing societies thereby helping in sustainable development of the whole nation.

When it comes to the real-life scenarios of AIoT applications, there are no better and direct examples than ITS. Every such system make use of AIoT techniques one way or the other. Figure 3 provides an overview of the basic components of an ITS making use of AIoT for their working. In the vehicular end, these include different sensors, cameras, or other devices which help it to get a grip of its surroundings. And when it comes to roadside infrastructures facilitating ITS, devices such as certain types of location sensors, proximity sensors, and all are used [23]. The usage of smart transport systems are more common in developing countries too. Comparing to the traditional modes of transportation, IoT-based ones are ahead in terms of accuracy, monitoring, stability, and coordination. AIoT can be implemented in goods transportation as well, it can optimize the chain-operation based on present parameters and determine delivery estimations, and even mitigate the disrupted services.AI can correlate the upcoming weather situations with the expiry date of products in the warehouse and tackle the loss associated with it as well.

Smart roads can be referred to as the heart and soul of ITSs. The primary motto of smart roads is to provide a safe, energy-efficient, and traffic-reduced environment. They make use of advanced technologies such as different types of sensors, cameras, solar and wind-powered systems and so on to facilitate the same [24]. In most of the developing countries, a huge issue related to fuel wastage caused by vehicles is due to crowded roads and clumsy traffic. Smart road ensures optimal usage of fuel thus helping in attaining SDGs. Smart traffic signals, smart fuel stations, smart parking, and all are just some of the entities related to smart road structures. AIoT devices are coming up with an added advantage that they are bidirectional in nature, as they can identify the data coming from hundreds of stakeholders, identify the trends in that and inform how the same devices will perform in the future. As the number of such stakeholders contributing to the crowd-intelligence increases the smarter and better each member could perform [25]. Making use of these peculiarities, studies have been going on to improve the traffic-light control system using AI, which is an essential component in ITS. DL methods like Q-learning are employed for identifying the information collected from the neighbouring traffic intersections to generate better results depending upon the traffic condition.

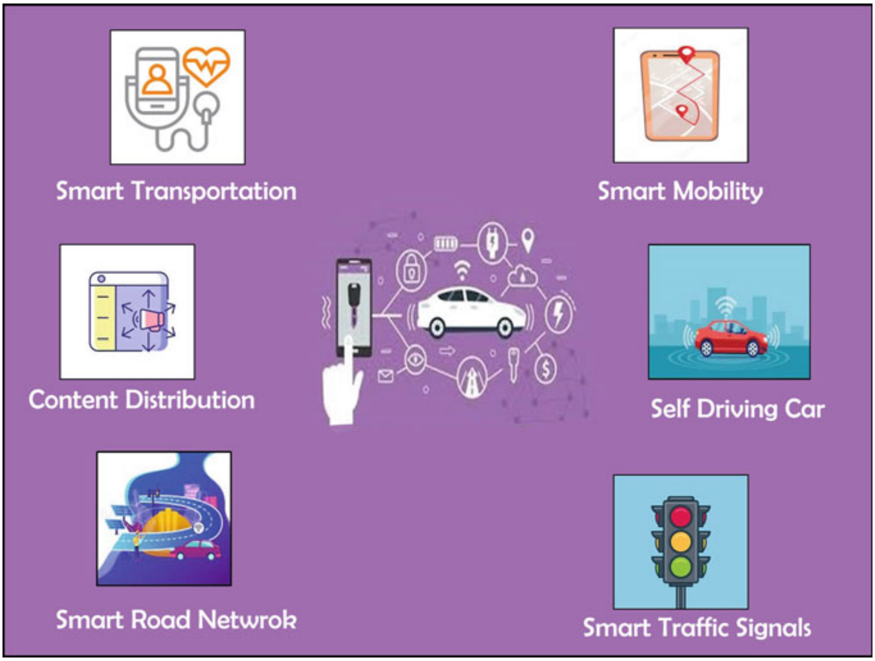


Fig. 3 Applications of AIoT in intelligent transportation system

Smart transportation systems can only be said to be complete with the addition of efficiently devised smart vehicles. The well-balanced combination of AI and 5G communication advancements are making autonomous vehicles take a faster route of growth. From 2013 on-wards the famous car manufactures started investing in these types of vehicles. In 2018 "Baidu" introduced a self-driving bus named "Apollo", during the second "Baidu" AI developers conference. The main challenge associated with these vehicles is the identification and management of the environmental parameters. IoT sensors can collect these information and given to fog/cloud-based buffer layers, ultra-low latency can be ensured. The future will be of self-driving intelligent cars, which would have the capability to make a journey with utmost safety and autonomy. The peculiarity of such self-driving vehicles is that they facilitate the development of their allied resources as well. The higher number of autonomous vehicles implies higher development of roads and related amenities [26]. Another things is that such vehicles make use of resources optimally, i.e. direct resources such as fuel as well as indirect resources such as parking space, road usage and so on thus making it ideal for sustainable resource usage. They also help in generating smart navigation plans and smart transportation facilities ensuring time management also.

4.4 Smart City

A smart city is a methodology mainly composing of Information and Communication Technologies (ICT) aiming at developing, deploying and promoting sustainable development practice and addresses increasing issues related to urbanization. Smart city and ITS are very two closely bound entities similar to two sides of a coin. The complete deployment of the latter is only possible at well-established cites of the former. Application of AIoT is the building block of smart cities. It helps us to house multiple technologies and provides a stable platform for interaction between them enabling thrift development of Smart City system with multiple facets such as increased standard of living, sustainable development and higher productivity of its residence [27]. Some of the different smart city applications based upon AIoT is presented in Fig. 4. Smart cities also provide multiple advantages when it comes to the context of sustainable development. Climate control is one such major one. They tend to make use of advanced technologies in all aspects implying less energy usage in total. Most of the AIoT devises are highly efficient as well as power saving. Thus as a whole, the carbon footprint of such cities tend to be very low enabling nature friendly development and cleaner living. It also provide a high degree of societal impact [28]. The sole motto of smart city projects are to improve the quality of life of the inhabitants of the city meaning that they try to provide equal opportunities to everyone laying the path for building an inclusive society.

A smart city is the total sum of multiple smart systems replacing traditional approaches that can be found in a normal city. The components may include technologies such as smart transportation systems, smart roads, ITS, smart health care systems, smart infrastructure, waste management systems and so on. All these implementations will be done with AIoT elements and they will be collectively controlled to regulate the working [29]. These networks will be communicating using standard communication protocols with an intelligent system taking the decisions.

Smart city can be considered as a combination of all the previously discussed technologies as they make use of one or the other at some level. Generally speaking, smart city includes components such as smart technologies, smart industries, smart services and smart life. The role of IoT in such application involve installing devices at every possible application, connecting them via Internet facilitating data exchange and communications. The role of AI comes in the part of information mining and allied task such as recognition, location, tracking, monitoring and management of acquired data. They also make use of multiple devices such as sensors, cameras, tags and so on depending upon the application. Just mere installation of these features doesn't ensure a "smart city" but their collaboration at a higher end of IoT development is required [30]. Such development of smart cities also gives rise to several other technological developmental requirements as well.

Smart vehicles are one of the basic components of smart cities. With the terminology smart vehicle, we mean vehicles that are capable of communicating with one another and taking decisions autonomously or with the aid of the user. Such system facilitates easy traversal and smart traffic control in cities as real-time traffic updates

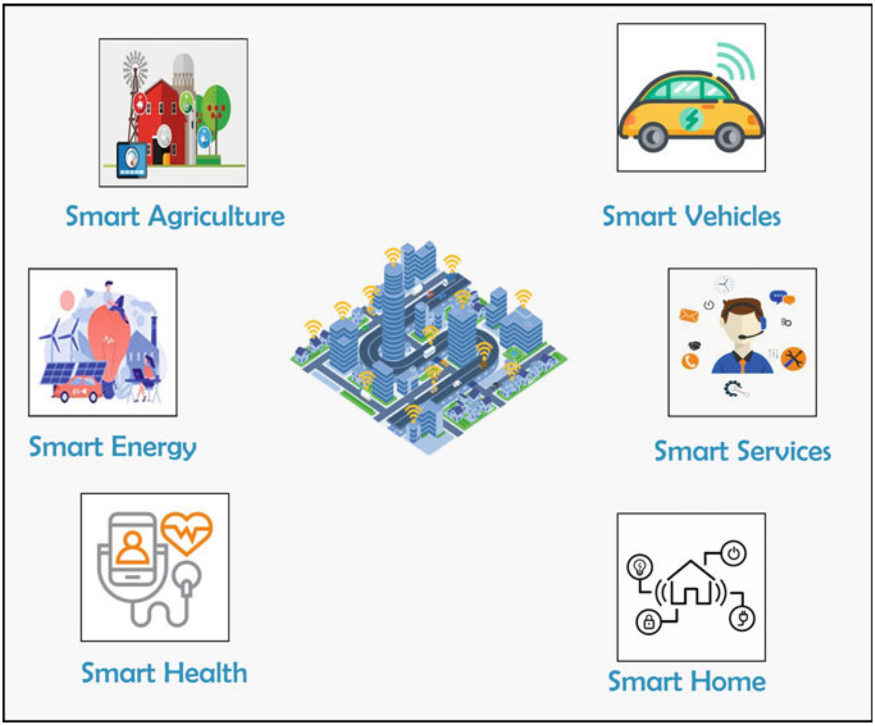


Fig. 4 Applications of AIoT in smart city

could be made available to fellow travellers enabling them to do instant navigation updates. Such vehicular networks require high-end technologies and protocols for secure and private data transfer. This calls for a development in communication systems as well. Such systems makes use of IoT devices extensively at different levels such as for data generation, transfer and so on. Even physical systems are employed in such networks for secure data transfer procedures.

Smart energy is one another procedure involved in smart cities [31]. The densely settled nature of smart city calls for high energy usage as well. Depending upon traditional sources of energy will never satisfy the need of such settlements in the long run. So in most cases, smart cities employ smart methodologies for energy usage. They make use of energy efficient devices throughout the city ensuring optimal energy usage. IoT devices are employed for the purpose as they are highly energy efficient and capable of performing the same task as other devices does with ease. Similarly, AI is employed for energy conservation scenarios as well. AI applications to detect energy leakage, control energy usage and so on are employed in smart cities [32]. Along with that, they make use of several renewable energy generation methods as an alternative for non-renewable one. Solar energy, Wind energy and all are very common in such smart cities. Thus this aspect of smart cities are really important when it comes to sustainable development of the community. When it comes to homes as well, most of the application will be making use of AI and IoT at different levels [27]. From face recognition systems and bio-metric identification systems at entry points to smart home appliances and home automation systems, their applications are unlimited in household environments.

4.5 Agriculture Management

IoT is already being used in agricultural procedures for a some time but AIoT is a comparatively new technology in agricultural context. The major difference between the previous applications and the current one is that lack of live data being fed to the sensors in the traditional methodology. The traditional systems rather used previously acquired data to do processes upon and take inferences at a later time. With the incorporation of AI into the mix, more powerful and useful combination of sensors and technologies are being applied in agricultural procedures capable of performing better real-time tasks [33]. The advancements have reached to a level that such devices can be connected to centralised servers or Cloud servers where data can be stored using which the control of the system can be facilitated remotely. Such systems has made it possible to integrate technology into agriculture on multiple aspects. Some of the basic applications such as governing hydration of plants, humidity control, disease diagnosis and control, and so on can be done with ease even for a "new-farmer". Other real-time tasks related to irrigation systems such as watering of plants, controlling water tank levels and so on has also been modernised. With the widespread availability of AIoT systems, the significance of AIoT in agriculture is not just limited to a single stage but present throughout, right from determining the amount of time needed to be invested at very stage to growing a ready to use product. The usage of AIoT technology in agricultural applications can be considered as a second wave of green revolution. The benefits of such systems are manifold from a farmer's perspective. It does provide cost effective solutions in the long run as well as facilitate better decision making capabilities for farmers using accurate data and allied technologies [34]. Some of AIoT based agricultural technologies which are already in use is depicted in Fig. 5.

The tasks undertaken by AIoT applications in agricultural management are far and varying. From the very basic applications such as agricultural parameter management to high-end applications such as disease detection and crop yield analysis, the list keeps ongoing. When it comes to greenhouse farming, technological advancements can be made use of to the fullest. Smart greenhouse management systems making use of IoT sensors and web-based technologies have been used at a commercial level for a long time. The systems will be having a versatile sensor network and a software side for its control. The usual practice of such systems is that the sensor network will be comprising of a master control centre accompanied by sensors connected via "Zigbee" protocols. Most of such systems will be comprising of a 3 tier structure, i.e. a hardware layer, middle ware layer, and a user interface. The purpose of middle ware

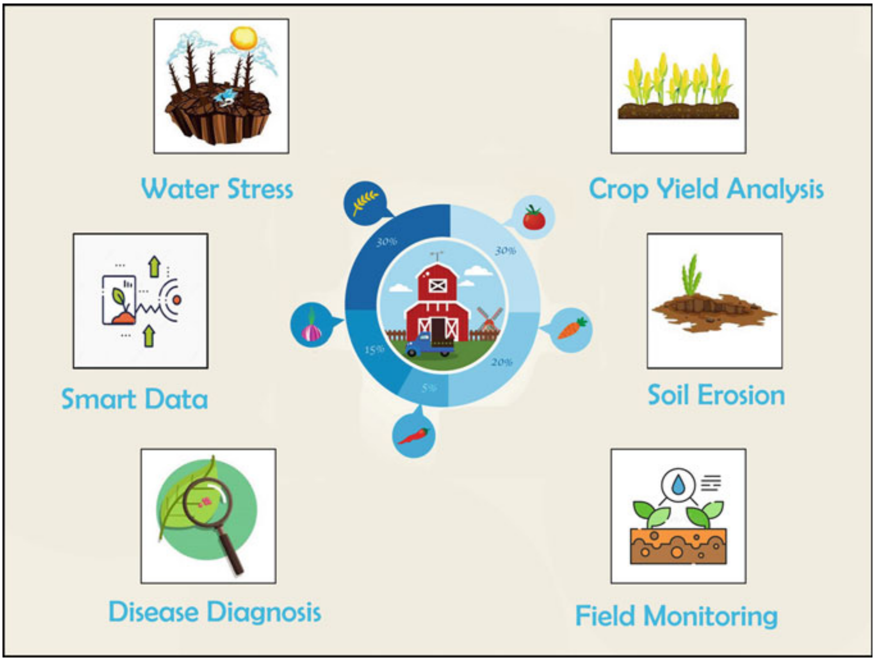


Fig. 5 Applications of AIoT in agriculture

is to form a communication link between the hardware layer and the user layer. It communicates with the hardware through serial network interface converters and the users via web system connection [35]. These are some of the very basic functionalities of smart greenhouse systems and the difference in tasks performed are dependent upon the types and number of sensors being used. Some of the very commonly used sensors in such applications are relative humidity sensor, ambient air temperature sensor, soil moisture content sensor, ph sensor, CO2 sensor, and so on. In most cases, they will be accompanied by a control unit like a multi-controller which acts as the "brain of the system" and a connectivity device such as Wi-Fi routers, Wi-Fi module, or cloud connectivity which helps the system to perform data transfer with the outside world [36]. Such sensors will be connected to either a central server or some other device from where the data created can be monitored easily. It is how remote agricultural control is being practiced nowadays which is a growing area of interest for the industry.

In remote controlled farming environments, the user is able to monitor the greenhouse or the farm 24/7 from a remote location and is able to take necessary measures from his location itself eliminating the need for physical presence at the farming location. Such systems will be comprising of a cocktail of multiple technologies which helps to gather data in all aspects and do needed tasks upon a command [37]. Automated fertilizing, automated watering, and so on are just a part of such systems.

When discussing remote access farming systems, a very closely related technology should also be discussed in detail, i.e. automated irrigation and fertilizing systems. Such systems also make use of sensor data to detect the content of moisture and nutrients in the soil and perform tasks using that data. The peculiarity of these systems is that they can be operated in hybrid methodology both with and without human intervention. Particular values can be set inside the system and as the system detects the set threshold value, AI techniques will invoke the system to perform a particular task, i.e. in this case fertilization or hydration of land [38]. All the aforementioned tasks make use of very cheap sensors and microprocessors making themselves ideal for applications in developing communities. Unlike humans, machines are more precise and when it comes to resource usage, optimal usage also can be ensured. Watering, fertilizing tasks performed by machines are way more efficient than humans. Thus along with ease of work, we can ensure optimal resource usage which is a strong motto of sustainable development.

Another AI-intensive application in the agricultural field is related to crop yield analysis. The system makes use of basic ML algorithms to perform Predictive Analysis tasks with the help of huge data sets collected from previous years. The data sets will be having the previous year's crop yield with respect to changing parameters and different ML algorithms will be used to draw inferences from them for predicting crop yield [39]. Also, with the increased research in plant disease detection and similar fields, AIoT applications are being used for detection and forecasting applications as well. The systems usually make use of the DL technique with the help of either sensor or image data to perform the task. A prior data set will be used and inferences will be drawn from them. The major task will be feature extraction and those features will help the system for differentiating between a diseased and a healthy specimen. When it comes to forecasting, these systems search for certain conditions which might have caused certain ailments in the previous situations. Such instances are usually referred to as "triggers" and they can be identified with properly trained models. With farmland and greenhouses being monitored all the time in smart agricultural systems, these applications do not require any sort of additional hardware making them easy to implement. Their high accuracy helps to make timely judgments and needed treatments to be implemented with ease. Also, the systems being autonomous even helps a novice farmer get the same knowledge as an experienced one when it comes to disease-related issues [40]. Another things is that the formentioned tasks make the previously tedious agricultural procedures to be performed with ease irrespective of the expertise of the user. Apart from that, when it comes to a developing economy, high concern should be given to ample usage of agricultural land and resources. The discussed technologies helps us to make optimal usage of water, fertilisers and similar resources which helps in striving for sustainable agriculture. Similarly all these techniques being introduced into agriculture will make the sector highly appealing to newcomers as well.

5 Future Trends

Taking into consideration the current trend, it is no doubt that AI-enabled IoT will be highly relevant in the near future irrespective of application domains. The wide applicability of AIoT makes it an ideal contender for multiple applications spreading over various instances. The most priced commodity of the future is regarded to be data and the systems incapable of making use of that will perish. When combining data intensive AI into IoT devices, one of the prime issues that we face is the memory constraint and limited computational power of the components [41]. Usually, AI tasks are highly complex requiring huge segments of memory but IoT devices lack such capabilities on their own. Thus the future of AI-enabled IoT can't be sustained by limiting the technology to pure IoT devices. One of the proposed solutions was to make use of Cloud computing which is widely in use for other AI applications. The sheer availability and popularity of the Cloud has made it look more appealing to be the ideal solution for overcoming such issues. With all this being said, studies have proven that Cloud computing combined with IoT is one of the least efficient solutions to cover up the problems already discussed. It is mainly due to the latency issues caused during data transfer in such systems. When a system is dependent on a Cloud server, the need for a high-speed network for continuous data transfer is inevitable. Such systems usually have high data traffic and high latency accompanied with it. This will lead to nullification of the advantages of IoT like quick processing and mobility which might have been the motivating factor for choosing the IoT architecture [42]. It is where Edge computing and Fog computing comes into the scene.

Edge computing basically means making use of processors placed closer to the source of the data to do most of the processing. Only relevant data is transferred to the server in such cases, resulting in a high decrease in the to-and-fro data transfer between server and devices. It also helps to facilitate systems with higher efficiency and lower latency helping to perform operations in real-time or at least near real-time. The same is the case with Fog computing as well [43]. Such technologies combined with an IoT-Cloud setup create a hybrid environment where all the benefits of an IoT system can be achieved without having to face resource constraints. The future of AI-enabled IoT is highly interlinked with Fog and Edge computing paradigms. They allow AIoT setups to increase their areas of applicability to more data-intensive problems such as real-time image and video processing than do limited functionalities with sensor data. This opens up further areas of research related to real-time video-based navigation systems [44], video analysis systems, and so on. A major issue related to such proposed methodology is in terms of security. Due to such technologies being placed at the periphery of the network, they do lack the protection offered by networks that devices placed directly inside the network coverage can avail. Thus security threats also arises and extensive study is needed to tackle those situations as well.

When it comes to the application side as well, the industry is on a gradual change aiming for the future. The health care sector has seen one of the most revolutionary changes. Previously, the technological side in health care was limited to just simple assisting robots and decision support systems whereas with the widespread of AIoT, wearable devices, and personal care systems are becoming popular. It is expected that more compact wearable systems capable of monitoring multiple bodily functions will be made available soon. Another domain that has been under massive change due to AIoT is intelligent vehicle systems. From basic auto navigation systems, technology has reached a level where vehicles are able to drive on their own without any human intervention. The current research is to take the technology to another level by making it more independent and capable of performing alone without any human inputs. The technology is mainly aiming for "special people" with limited physical mobility and the system functions through basic commands that can be provided via voice or some other methodology. Also, the agriculture domain is adopting AIoT technologies widely. The current trend is smart greenhouses and automated farming systems. Such technologies are limited to indoor applications as of now and concrete works are needed before making it applicable to open farming applications. It is not that the future of AIoT is limited to the aforementioned domains but we have used some of the very prominent applications.

AIoT is still a developing paradigm where extensive study is being done. The future is very bright for the technology but a complete analysis can only be done as more completed works become available to the scientific community. Extensive works are needed before we can clearly analyse the positive and negative impacts of the technology. The sustainability side of the system should also be studied as the scope of any technology in the future will also be dependent upon its ability to produce sustainable technological advancements.

6 Conclusion

The world as of now is undergoing rapid technological advancements which can be considered as equal to a sixth industrial revolution. The need of the era is for devices that are capable of performing autonomous tasks and have a high degree of portability, reliability, scalability, and mobility. AI and the IoT are two such technologies that have a boundless future due to this changing need. They are not just part of this rising trend but they have already paved the way for an emerging paradigm that has the capability to modernize the concepts of the technological world for the coming era, i.e. Intelligent IoT. AIoT is a concept which enables all the capabilities of a conventional highly complex AI system to be put into a tiny IoT system. It enables a user to access the same AI functionalities at a much smaller and efficient system. Even though technologies such as AI and IoT have already proved their significance in sustainable development, the research community is in constant search of better technologies that have a higher potential in attaining Sustainable Developmental Goals with ease. So our study also gives a prime consideration for investigating the relevance of AIoT in the context of sustainable technological development. Through this article, we tried to present a detailed analysis of AIoT technology in different aspects. We have performed a study on how the technology is implemented alongside analyzing the individual role of AI and IoT in the combined system. The study has revealed the multiple application instances of AIoT in several domains such as construction works, health care, agriculture, smart cities, ITSs, and so on. We have also provided a detailed overview of all the applications. Even though our study had pointed out the impact of AIoT based applications in sustainable development, the technology is still in a developing stage. Only when intensive studies are performed, we can better understand the technology and its potential.

References

1. Zhou J, Wang Y, Ota K, Dong M (2019) AAIoT: accelerating artificial intelligence in IoT systems. IEEE Wireless Commun Lett 8(3):825–828. https://doi.org/10.1109/LWC.2019.2894703
2. Zhang J, Tao D (2021) Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. IEEE Internet Things J 8(10):7789–7817. https://doi.org/10.1109/JIOT.2020.3039359
3. Stoyanova M, Nikoloudakis Y, Panagiotakis S, Pallis E, Markakis EK (2020) A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues. IEEE Commun Surv Tutor 22(2):1191–1221. https://doi.org/10.1109/COMST.2019.2962586
4. Javaid N, Sher A, Nasir H, Guizani N (2018) Intelligence in IoT-based 5G networks: opportunities and challenges. IEEE Commun Mag 56(10):94–100. https://doi.org/10.1109/MCOM.2018.1800036
5. Cui L, Yang S, Chen F, Ming Z, Lu N, Qin J (2018) A survey on application of machine learning for internet of things. Int J Mach Learn Cybern 9(8):1399–1417
6. Al-Garadi MA, Mohamed A, Al-Ali AK, Du X, Ali I, Guizani M (2020) A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun Surv Tutor 22(3):1646–1685. https://doi.org/10.1109/COMST.2020.2988293
7. Ghosh A, Chakraborty D, Law A (2018) Artificial intelligence in internet of things. CAAI Trans Intell Technol 3(4):208–218
8. Bertino E, Jahanshahi M, Singla A, Wu RT (2021) Intelligent IoT systems for civil infrastructure health monitoring: a research roadmap. Discov Internet Things 1(1):1–11
9. Kanan R, Elhassan O, Bensalem R (2018) An IoT-based autonomous system for workers' safety in construction sites with real-time alarming, monitoring, and positioning strategies. Autom Constr 88:73–86
10. Chung WWS, Tariq S, Mohandes SR, Zayed T (2020) IoT-based application for construction site safety monitoring. Int J Constr Manag 1–17
11. Manavalan E, Jayakrishna K (2019) A review of internet of things (IoT) embedded sustainable supply chain for industry 4.0 requirements. Comput Ind Eng 127:925–953
12. Di Nuzzo F, Brunelli D, Polonelli T, Benini L (2021) Structural health monitoring system with narrowband IoT and MEMS sensors. IEEE Sens J 21(14):16371–16380. https://doi.org/10.1109/JSEN.2021.3075093
13. Amin SU, Hossain MS (2021) Edge intelligence and internet of things in healthcare: a survey. IEEE Access 9:45–59. https://doi.org/10.1109/ACCESS.2020.3045115
14. Jaiswal K, Anand V (2021) A survey on IoT-based healthcare system: potential applications, issues, and challenges. In: Advances in biomedical engineering and technology. Springer, pp 459–471
15. Espinosa ÁV, López JLL, Mata FM, Estevez MEE (2021) Application of IoT in healthcare: keys to implementation of the sustainable development goals. Sensors 21(7):2330
16. Albahri AS, Alwan JK, Taha ZK, Ismail SF, Hamid RA, Zaidan A, Albahri OS, Zaidan B, Alamoodi A, Alsalem M (2021) IoT-based telemedicine for disease prevention and health promotion: state-of-the-art. J Netw Comput Appl 173:102873

17. Hameed K, Bajwa IS, Ramzan S, Anwar W, Khan A (2020) An intelligent IoT based healthcare system using fuzzy neural networks. Sci Program (2020)
18. Dagher L, Shi H, Zhao Y, Marrouche NF (2020) Wearables in cardiology: here to stay. Heart Rhythm 17(5):889–895
19. Ba T, Li S, Wei Y (2021) A data-driven machine learning integrated wearable medical sensor framework for elderly care service. Measurement 167:108383
20. Hajar MS, Al-Kadri MO, Kalutarage HK (2021) A survey on wireless body area networks: architecture, security challenges and research opportunities. Comput Secur 102211
21. Wang XV, Wang L (2021) A literature survey of the robotic technologies during the Covid-19 pandemic. J Manuf Syst
22. Sumalee A, Ho HW (2018) Smarter and more connected: future intelligent transportation system. IATSS Res 42(2):67–71
23. Balasubramaniam A, Paul A, Hong WH, Seo H, Kim JH (2017) Comparative analysis of intelligent transportation systems for sustainable environment in smart cities. Sustainability 9(7):1120
24. Trubia S, Severino A, Curto S, Arena F, Pau G (2020) Smart roads: an overview of what future mobility will look like. Infrastructures 5(12):107
25. Dangi K, Kushwaha MS, Bakthula R (2020) An intelligent traffic light control system based on density of traffic. In: Emerging technology in modelling and graphics. Springer, pp 741–752
26. Khayyam H, Javadi B, Jalili M, Jazar RN (2020) Artificial intelligence and internet of things for autonomous vehicles. In: Nonlinear approaches in engineering applications. Springer, pp 39–68
27. Shah SKA, Mahmood W (2020) Smart home automation using IoT and its low cost implementation. Int J Eng Manuf 10(5):28–36
28. Ramesh MV, Prabha R, Thirugnanam H, Devidas AR, Raj D, Anand S, Pathinarupothi RK (2020) Achieving sustainability through smart city applications: protocols, systems and solutions using IoT and wireless sensor network. CSI Trans ICT 8:213–230
29. Syed AS, Sierra-Sosa D, Kumar A, Elmaghraby A (2021) IoT in smart cities: a survey of technologies, practices and challenges. Smart Cities 4(2):429–475
30. Lv Z, Qiao L, Kumar Singh A, Wang Q (2021) AI-empowered IoT security for smart cities. ACM Trans Internet Technol 21(4):1–21
31. Al-Turjman F, Lemayian JP (2020) Intelligence, security, and vehicular sensor networks in internet of things (IoT)-enabled smart-cities: an overview. Comput Electr Eng 87:106776
32. Raval M, Bhardwaj S, Aravelli A, Dofe J, Gohel H (2021) Smart energy optimization for massive IoT using artificial intelligence. Internet Things 13:100354
33. Alreshidi E (2019) Smart sustainable agriculture (SSA) solution underpinned by internet of things (IoT) and artificial intelligence (AI). arXiv preprint arXiv:190603106
34. Appio FP, Lima M, Paroutis S (2019) Understanding smart cities: innovation ecosystems, technological advancements, and societal challenges. Technol Forecast Soc Change 142:1–14
35. Li Z, Wang J, Higgs R, Zhou L, Yuan W (2017) Design of an intelligent management system for agricultural greenhouses based on the internet of things. In: 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC), vol 2. IEEE, pp 154–160
36. Ullo SL, Sinha G (2021) Advances in IoT and smart sensors for remote sensing and agriculture applications. Remote Sens 13(13):2585
37. Ibrahim H, Mostafa N, Halawa H, Elsalamouny M, Daoud R, Amer H, Adel Y, Shaarawi A, Khattab A, ElSayed H (2019) A layered IoT architecture for greenhouse monitoring and remote control. SN Appl Sci 1(3):1–12
38. Raut R, Varma H, Mulla C, Pawar VR (2018) Soil monitoring, fertigation, and irrigation system using IoT for agricultural application. In: Intelligent communication and computational technologies. Springer, pp 67–73
39. Van Klompenburg T, Kassahun A, Catal C (2020) Crop yield prediction using machine learning: a systematic literature review. Comput Electron Agric 177:105709

40. Ouhami M, Hafiane A, Es-Saady Y, El Hajji M, Canals R (2021) Computer vision, IoT and data fusion for crop disease detection using machine learning: a survey and ongoing research. Remote Sens 13(13):2486
41. Balaji S, Nathani K, Santhakumar R (2019) IoT technology, applications and challenges: a contemporary survey. Wireless Pers Commun 108(1):363–388
42. Sadeeq MM, Abdulkareem NM, Zeebaree SR, Ahmed DM, Sami AS, Zebari RR (2021) IoT and cloud computing issues, challenges and opportunities: a review. Qubahan Acad J 1(2):1–7
43. Yousefpour A, Fung C, Nguyen T, Kadiyala K, Jalali F, Niakanlahiji A, Kong J, Jue JP (2019) All one needs to know about fog computing and related edge computing paradigms: a complete survey. J Syst Archit 98:289–330
44. Nair AK, Sahoo J (2021) Edge eye: a voice assisted campus navigation system for visually impaired. In: 2021 3rd international conference on signal processing and communication (ICPSC). IEEE, pp 125–129

Cyber Security Challenges for Smart Cities

Arunima Sharma and Ramesh Babu Battula

Abstract In the new past, keen metropolitan regions was simply conceptualized in sci-fi movies and books of fiction. Today, in any case, that thinking is rapidly changing from creative areas into genuine components. Essentially as savvy metropolitan regions spring up across the globe, they in like manner have made an exceptional security-risk perspective. Along these lines, this is tortured not actually settled forever to cause annihilation at the littlest possibility. Appropriately, industry insiders alongside future city coordinators see the need to address these sort of wellbeing issues. Not at all like the customary security issues of the past, the online assurance essentials of keen metropolitan regions are new and constantly advance around the latest examples in advancement and improvements. Security experts would be particularly served to review a piece of the overall proposed deals with serious consequences regarding perceive the peril scene that might torture the forthcoming keen metropolitan networks. Contraptions, for example, smart energy meters, smart miniature well-being machines, and smart security gadgets give uncommon benefits, hence further working on individuals' lives. With new inter-connectivity of observation frameworks; robotization in different urban areas; administration arrangements like crisis responders, water, fiasco the board, sterilization and foundation, the extent of efficiencies in the Smart City is steadily growing. Coordinated frameworks, for example, private and public wise transportation will actually want to assemble data about traffic refreshes, just as GPS area and climate conditions. This uses shared information continuously. Along these lines, the advantages should be weighed against the potential perils that accompany such enormous inter-connectivity.

Keywords Cyber security · Smart city · Artificial intelligence · Internet of Things · Trust · Privacy

A. Sharma (✉) · R. B. Battula
Malviya National Institute of Technology, Jaipur, Rajasthan, India
e-mail: aru.92@rediffmail.com

R. B. Battula
e-mail: rbbattula.cse@mnit.ac.in

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_3

1 Introduction

Billions of associated ‘things’ are sent in smart urban areas all throughout the planet. The ascent of the Internet of Things (IoT) uncovered a wide scope of weaknesses that can be taken advantage of by digital crooks and other pernicious entertainers. Albeit smart urban communities are intended to expand usefulness and efficiency, they might possibly introduce genuine dangers for inhabitants and specialists when digital protection is dismissed. There are an obscure number of likely weaknesses and strategies, the absolute most normal attacks include:

- **Man in the center** An assailant breaks, blocks or parodies correspondences among two frameworks. As example, a Man in the center attack on one of the keen valve in a waste water system could be used to cause a bio-hazard spread.
- **Information and data fraud** Information delivered by unprotected shrewd city system like leaving structures, EV (Electric Vehicles) charging stations and perception deals with give advanced aggressors an ample proportion of assigned individual information that may potentially be exploited for misleading trades and recognize robbery.
- **Gadget seizing** The aggressor captures and successfully accepts control of a gadget. These attacks can be difficult to recognize on the grounds that by and large, the assailant doesn’t change the fundamental usefulness of the gadget. With regards to a smart city, a digital criminal could take advantage of seized smart meters to dispatch ransomware attacks on Energy Management Systems (EMS) or covertly siphon energy from a district.
- **Dispersed Denial of Service (DDoS)** A repudiation of-organization assault (DoS assault) attempts to convey mechanism otherwise association supply tough towards scope adjacent that one ordinary consumers through for a short time frame or indefinitely disturbing administrations of crowd associated through Web. This remains usually cultivated through overflowing goal with futile requesting to hold genuine requests back from being fulfilled. Because of an appropriated refusal of-organization assault (DDoS attack), moving toward traffic flooding a target begins from various sources, making it difficult to stop the advanced antagonistic by basically obstructing a lone source. Inside keen metropolitan networks, a lot of devices, such as halting meters, can be entered and constrained to join a botnet changed to overwhelm a system by referencing an assistance meanwhile.
- **Permanent DoS (PDoS)** Super durable renouncing of-organization assaults (PDoS), furthermore alluded to uninhibitedly as phlashing, is an assault that hurts the device so seriously that it requires replacement or re-establishment of hardware. In a keen city circumstance, a held onto halting meter could in like manner surrender to assault and would should be superseded.

2 Smart City System

Different apprenticeships attentive in scheduled keen metropolitan framework in addition to customs towards deal with beat safety in addition to insurance subjects inside shrewd metropolitan regions. The IoT expects a fundamental part inside the structure of brilliant metropolitan networks as it gives the association designing reliable to get-together and dealing with data from appropriated sensors and savvy devices. Studies all around order assaults on IoT contraptions into outside and inside-assaults.

The shortcoming of Internet of Things grounded solicitations be present clearly connected towards association perspective wherever real articles, for instance, sensor set up devices assemble data regarding key coordinated efforts inside the organize and pass on through remote otherwise online affiliations. Information which stay existent moved, dealt with besides set aside dismiss show important shortcomings as man-in-the-middle assaults and renouncing of-organization assaults. Subsequently, assembling and moving data through the usage of Internet of Things structure might truly influence safety in addition towards assurance of shrewd metropolitan networks aside from if cautious advances are executed. Studies have fought that security can be adequately compromised as a result of the extraordinary heights of connotation among individuals, manoeuvres in addition towards devices, thusly including prerequisite aimed on this information towards remain totally guaranteed. Readings obligate set advantages of additional fundamental emphasis arranged keen metropolitan safety observing past pieces of information insurance toward keen securitisation system. The examination fought that information security doesn't simply join assurance, arrangement, genuineness and availability, yet likewise consolidates inter-operable safety which tends to chance of common dissatisfaction of metropolitan organization.

Information streams in addition to interactions amid system shares in addition to the Internet of Things must remain reliant upon amazing peril the board in assessing and responding to risks inside keen metropolitan networks and the troubles of the specific intricacy opening and standards youth Analysts have attempted to perceive mechanical responses for oversee assurance and more broad information security challenges. The assessment took apart the insurance issues looked by savvy metropolitan networks including affirmation, right to use regulator, protection, reliance, information retreat, system execution in addition to protected centre product. Maker arranged and repeated a keen city model related with required particular devices that conveyed data for different sensors. The assessment suggested that information assurance perhaps refined through Fast Identification Online or FIDO approval measure) on behalf of manoeuvre towards put together otherwise contraption towards fog affirmationin addition to that information security must remain seen as indispensable part of savvy metropolitan system. Insurance points inborn inside savvy city network traffic system were taken apart, where the assessment put the upsides of Attribute-Based Credentials answer for assistance with settling problem of honesty of pointless information. Investigation suggested Idemix owed towards this one presentation competences in addition to similitude through current shrewd metropolitan highway movement organizations. Assessments set propriety ofutilizing flexibility speculation which remain contemporary stressed over limit of a living thing toward suffer in addition to form into improved situations. Assessment considers security to be a structure and breaks down it through the adaptability point of convergence, laying out the subject of how insurance can change and make due inside a savvy city.

Khan and others recognized an overview of accomplices and exhibited their consideration inside the brilliant city setting. The accomplice arranging included: organization customers, certified expert associations, untrusted expert associations, IT trained professionals, data guardians, standard managing bodies and space subject matter experts. Allowing for projected accomplice prototypical, inspection cultivated protection in addition to assurance construction on behalf of secure in addition to insurance careful assistance provisioning in brilliant metropolitan regions. The construction expected to offer beginning towards conclusion safety in addition to assurance structures on behalf of trustable information obtainment, broadcast, planning in addition to genuine assistance provisioning, and showing anticipated agendas aptitude towards ease accomplice safety in addition to safety worries. Extra significant designs join the one proposed in Vitunskaite and others that played out a relative shrewd city logical examination in addition to organization representations, safety endeavours, particular rules in addition to outcast organization. Construction included specific rules, organization input, managerial framework and consistence insistence towards confirm information safety be there present seen inside everything coatings of brilliant metropolitan structure.

Keen metropolitan networks are incorporated incalculable different sensors, correspondence devices, network entries, explicit gear and programming. These distinct advantages ought to be consolidated inside the shrewd city establishment and stayed aware of to ensure structures are not corrupted and critical organizations are operable. The examination by Waedt and others focused in scheduled physical in addition to modified plus ID, remark in addition to following of explored Application safety Controls that canister advantage after addition to formal benefit the leaders. This consolidated openness in addition to decency of secure in addition to versatile development possessions in addition to steady quality in addition to reliability of programming possessions presented scheduled specialists in addition to fog conditions. A learning verifies that intensive in addition to inevitable advantage the leaders offers some motivator past security to soothe the maltreatment of assets for complex assaults zeroing in on blends of version express shortcomings.

2.1 Smart Power Systems

The power structure portions of savvy metropolitan regions remain of essential importance inside total safety in addition to assurance establishment, by way of pariahs related with the organization could screen use plans and predict clients' lead. The distant association development centered countryside of various structures that stock in addition to switch warmth in addition to bright towards keen metropolitan regions, might open organization to security shortcomings. Alamaniotis and others presented a shrewd strategy for further developing security inside savvy power structures. The proposed reasoning utilized interest plans for a couple of customers related with the power organization to give another usage plan. The new model covers solitary buyer characteristics through a particle swarm improvement measure. The examination gave the proposed approach a shot a lot of veritable usage plans seat set apart against an inherited estimation displaying that the proposed method is useful. The examination by Sanduleac and others watched out for two essential pieces of savvy city execution; explicitly:

- Different dynamism torrents while dissimilar values aid assorted vigour systems hip metropolitan like force, fume, in addition to warmth.
- Problem of attracting inhabitants through distribution their secluded dynamism information outline, by way of substitute rather than executions that disregard towards development after little aviators to tremendous sending.

2.2 Smart Healthcare

Security and protection of medical care administrations and ideas inside smart urban communities are a critical factor [1] in the generally speaking insignificant revelation of information and data security framework. The exploration by Alromaihi and others recognized the fundamental security and protection challenges in planning IoT engineering with regards to medical care applications, featuring the expanded utilization of sensors for medication and medical care applications throughout the last decade. The examination distinguished key dangers from individual well-being related information caught through sensors for example pulse and furthermore circulatory strain and the significance of an incorporated security answer for the whole framework. The advantages of information coordinated effort inside a safe portable medical services and social framework have been mentioned in Huang and others The examination fostered an answer that permitted an information proprietor to approve outsider medical care supplier information investigation by re-encoding ABE (Attribute-Based Encryption) in addition to personality grounded transmission encryption or IBBE. Projected conspire utilized encryption with unscrambling measures that viably delegate greater part of the calculation cost to cloud. Therefore, calculation overhead of asset obliged cell phones are diminished, with resulting upgrades in security and productivity.

3 Prototypes, Processes, Structures, in Addition to Rules for Better Safety and Confidentiality

In place of brilliant metropolitan networks expression different troubles related with security and assurance, a couple of examinations proposed various constructions, models and estimations to chip away at these issues. This piece of the composing has focused in scheduled encryption computations to work cutting-edge safety towards shrewd metropolitan schemes. Antonopoulos in addition to others training examinations huge side by side safety incorporate computations through by means of Wireless Sensor Network (WSN) headway. Researchers planned towards fuse a beginning towards conclusion cryptography structure keen on savvy metropolitan plans on basis side by side. Throughout some information break, nonentity near information would remain uncovered through smearing this framework.

Likewise, Researchers castoff encryption method cutting-edge suggesting an arrangement named Completely Confidentiality Preservative in addition towards Revocable Personality Grounded Transmission Encryption (FPPRIB). Future plot intended towards shield statistics assurance in addition to character security of authority similarly by way of the disavowed customer. Information maybe securely gotten in addition to simply supported customer container get to records. Forswearing cycle doesn't disclose some information around information substance otherwise beneficiary person in addition to everyone adjust nonentity around gatherer character in addition to disavowed customer character. These possessions prime towards requests into shrewd metropolitan anywhere character insurance remains appealing. Assessment encouraged cryptographic demonstration which achieves tremendous proportion of individual information that might remain delivered over e-interest into versatile, interoperable way, which guarantees the security of occupants inside savvy metropolitan networks.

Association access control accepts a critical part in any correspondence structure. Cultivate adequate security of IoT structure induction to hold any gate crasher back from accepting accountability for IoT devices or revealing private information set aside at article or center level. Beltran and others introduced SMARTIE, a consolidating stage for customer driven secure Internet of Things presentations. That one jam customer insurance though promising flexibility in addition to capability. The proposed stage successfully gives decentralized induction control to IoT devices subject to customer assurance tendencies. The place of SMARTIE is to work with the mix of customer driven insurance and organization confidential Internet of Things claims into flexible in addition to capable style. Makers included that planned request drive let customers towards switch their plans that joint solicitation to the extent recognizing and appropriating information in addition to allow satisfactory grained induction switch instructions arranged behalf of their contraptions while finishing up which container in addition to can't remain into charge of those manoeuvre information. Game plans proposed by researchers alleviate security dangers through outfitting an extreme choice creator with chance towards close organize admission arranged behalf of customer thusly guaranteeing insurance of customer information. Researchers planned security care structure Private Zones, which needs expert association towards segment huge arrangements of information assembled through their presentation. Planned agenda remained viably had a go at using two relevant examination organizations.

Usage of artificial intelligence canister additionally foster safety in addition towards assurance now shrewd metropolitan regions. González García and others attempted assessment of images done PC visualization towards recognize individuals popular took apart pictures. Through miscellaneous examinations, this one remained create that structure distinguishes images through craniums in addition to accepts even extra unequivocally now assessment through various pictures. Besides, assessment create that this one remain current attainable towards facilitate PC image inside Internet of Things associations in addition to that photographs possibly castoff by way of devices in like manner, portion with chipping away at the security of homes inside savvy metropolitan regions. Researchers design through by means of artificial intelligence in addition to mental limits, which remain current prepared aimed at sorting obtainable some way towards appreciate, look at and survey everything in a motorized keen manner.

Gheisariy and others tracked down that different existing courses of action have three huge hindrances. Most importantly, applying one static security ensuring system on behalf of complete structure; subsequent, distribution complete information immediately in addition to 3rd, shortfall of setting care. These perspectives can incite intolerable unquestionable equal of assurance defending above. Towards contract through these difficulties, makers planned an item described getting sorted out perspective that possibly clearly practical towards brilliant metropolitan requests. The training castoff a property founded dependence trade plot on behalf of correspondence between contraptions inside savvy metropolitan. Assessment showed the trust course of action degree by means of homographic encryption towards assurance that one security. Planned show guaranteed that manoeuvre contents that one counterparty's passageway methodology while uncovering insignificant security.

The cloud-situated plan course of action proposed in Krichen and Alroobaea set another prototypical grounded structure on behalf of challenging care belongings of Internet of Things founded structures inside savvy metropolitan regions by portraying the procedure embraced by toxic gathering which means in the direction of ignore safety of pondered Internet of Things system. Research encouraged lightweight in addition to assurance saving communal fog-surveying plan on behalf of brilliant metropolitan regions that needn't bother with bi-direct pairings. The proposed mixing free arrangement allowed an outcast analyst to make affirmation meta-information in light of a legitimate concern for customers and gave data protection from pariah monitors and cloud expert associations. The Han and others revision create that planned plot be present additional secure in addition to useful into connection with existing community fog assessing plans.

Portions of composing must fixated in scheduled safety in addition to assurance structures on behalf of commercial atmosphere. In a study sending Confidentiality ABCs founded affirmation organization keen on nonexclusive e-Business prototypical that gives total information founded e-Services inside Keen Metropolises. Classical included total knowledge associations among occupants and workplaces of brilliant metropolitan networks and a security further developing advancement named characteristic founded accreditations. Through by means of this procedure procurement past in addition to client direct of inhabitants stays secluded though interfacing through online business founded natural framework. Investigation spread out in Cagliero and others obtainable non-emergency information analyser education impression of occupants scheduled metropolitan safety with respect to business environment.

Work of programming inside brilliant metropolitan regions is central, nevertheless it conveys certain assurance in addition to safety problems like conversation of utilization information, issues connected towards following, belongings of chopping, confirmation of informational collections, extension in up close and individual information robberies, permission towards information now worker ranches, consequence of various submissions in addition to financial squeezing factor [1]. The assessment by Sucasas and others planned OAuth 2.0 grounded show on behalf of savvy metropolitan adaptable submissions that watched out for customer security problem through joining a nom de plume sign plan in addition to an imprint arrangement plot into the OAuth 2.0 show stream. The proposed game plan grants customers to self-produce customer express and application unequivocal aliases solicitation and ensures security further developed customer confirmation at the Service Provider side.

A couple of assessments investigated the current work and proposed new courses of action. As instance, the arrangement planned through Xiao and others needs two-factor security, and encounters an emulate assault. To diminish these issues, a further developed meandering affirmation show by 2-factor safety remained planned, gotten through math grounded legitimate endorsement device Pro-Verif displaying redesigned adequacy voguish relationship for certain associated plans.

Zang and others pronounced that safety show characteristic safety abandons, therefore fail towards attain that one special objective. Exactly, this show be present frail against 2 kinds of assaults, to be explicit—replace assault and replay assault. The examination presented in what way malignant specialist container betray information proprietors towards acknowledge that information remain current existence stayed aware of enough by dispatching such attacks.

Moreover, it portrayed a chipped away at Remote Database Access or RDA show through using arithmetical imprints towards dose safety faults. Game plan used position founded Merkle Hash-Tree towards attain obvious incredible information assignments. Also, assessment gave point by point security affirmation of the proposed RDA show. Gope and others examined current Radio-Frequency Identification (RFID) innovation on behalf of that one mull over security in addition to fake area issues in addition to profound calculation inconvenience owing towards amazingly incomplete estimation capacity of Radio-Frequency Identification names. Assessment endeavoured towards determine these problems done suggesting a Radio-Frequency Identification originated approval designing on behalf of scattered Internet of Things requests sensible on behalf of shrewd conditions.

4 Functional Defencelessness in Keen Metropolises

Information inside brilliant metropolitan solicitations must have the alternative towards endure modification, aggravation, appraisal, unapproved access, openness and decimation. Fundamental necessities for security and insurance consolidate grouping, decency, openness, non-disavowal, access control and security. Brilliant city occupants can defy security and security issues on account of shrewd city application shortcomings, regardless, without saw security confirmation and assurance, general society might puzzle over whether towards usage savvy metropolitan adaptable solicitations. Assurance be present middle problem inside brilliant metropolitan networks and unique that maybe straight associated with irrelevant considerate of safety on or after adjacent administration in addition towards commercial cutting-edge way they assemble in addition to cycle individual information. Routinely they don't allow the neighborhood occasion in addition to part on behalf of agreement.

A couple of examinations revolve around shrewd city drives for unequivocal countries, for instance, Indonesia, China, and Austria. The assessment spread out in Dewi Rosadi and others explored and inspected the insurance stresses inside smart urban regions, emphasizing complications of extended proportions of set aside in addition to passed on individual information that maybe assembled in addition to taken care of formerly scattered diagonally various contraptions, organizations in addition to areas.

Researchers inspected suitable rules in addition to rules cutting-edge setting. Makers battled that in attendance be existing not at all utilitarian security rule cutting-edge China that would smear towards greatest information assembled through shrewd metropolitan establishment; not remain existing present whichever rule that would guarantee whichever near and dear information accumulated underneath this context. A couple of nations ensure actually made a number of rules that assistance legitimately secure insurance advantages of their occupants. Let's say, Information Defence Guideline (GDPR) gives central bearing towards attain sensible congruity amid benefits of Internet of Things breadwinners in addition to customers. GDPR rules essential promote assurance in addition to execution keen on arrangement Internet of Things progressions.

Additional real problems like ward, organization of data and dealing with consent in savvy metropolitan regions remained emphasized. As deliberated cases like shrewdness metropolitan mission spreading out different data assessment circumstances to depict the activities took on for secure treatment of data. The examination perceived the going with assurance and security challenges: insurance guarantees, versatility insurance approaches, lack of definition, and information attribution. Endeavor castoff anonym, information all out, information trouble in addition to randomness, in addition to cryptographic agenda on behalf of information excavating. This one be located establish that picked game plan influenced the public care and value of the shrewd metropolitan development.

Although scholars absorbed now on establishment safety problems, for instance, tuning in, burglary, renouncing of organization, information following, customer/ inhabitants data hardships and various risks (for instance gear disillusionment, programming smash, surroundings in addition to countryside direct), widely inclusive point of view on the security scene of a keen city by perceiving security risks. The examination battled that different pieces of savvy metropolitan regions have different security risks. For example, brilliant grids have shown shortcomings, security, tuning in, and assaults on web related devices. Building Automation systems have security risks like significantly trusted in devices, long device life cycle nonappearance of source confirmation, and untrustworthy shows. For robotized raised vehicles, security perils join correspondence association, correspondence mixture, and correspondence staying. For keen automobiles problems might remain associated towards real risk, correspondence catch endeavor, information security, in addition to DoS. On behalf of Internet of Things devices, safety risks might fuse staying aware of the protection of data, secure correspondence, data the board, data amassing, sensor disillusionment and far off misuse. Lastly, on behalf of fog stage safety risks might fuse information spillage, pernicious insider risks, precarious API, DoS, malware implantation assaults, system and application shortcomings.

Studies have analyzed an extensive part of the security risks inside keen metropolitan networks offering different likely courses of action. Kitchin and Dodge suggested a more broad course of action of fundamental interventions like security-by-plan, therapeutic security fixing and extra, improvement of focus safety in addition to PC crisis reply gatherings, modification of acquisition frameworks, and continuing with capable development. Srivastava and others presented some keen responses for prosperity and safety which remain overhauled through usage of Artificial Intelligence. Game plans which remain at this point set up in some made savvy metropolitan regions are shot recognizable proof sensors, video observation and examination, robots, and online assurance. In any case, Vattapparamban and others battle that the use of these progressions (for instance rambles) can achieve different specific and social worries with respect to arrange assurance, security, and public prosperity.

While most of the assessments in this space base on insurance and security possibilities, Velasquez and others fought that consider normal risks when organizing keen metropolitan regions. The examination proposed another plan which joins the fundamental organizations that ought to be ensured and centered around inside a keen city. The investigation endeavored utilizing legitimate and talk with information originate 5 danger arrangements cutting-edge metropolitan broadband assignment. Gatherings remained driven through municipal technique trained professionals, telecommunication [2] specialists, in addition to administration authorities. It is going with dangers remained perceived:

1. communal party-political
2. underwriting
3. economic
4. particular
5. association.

Examination recognized that a piece of characterizations of risks, for instance, communal party-political perils influence each other and battle that danger the leaders and danger help policies remain expected towards revenue additional extensive point of opinion scheduled altogether risks in addition to their inter-connections rather than converging in scheduled every sort of danger autonomously.

5 Smart Services

Various investigations featured the significance of saw security and protection in smart urban areas administrations by residents. It was tracked down that apparent security and protection essentially influence the utilization and reception of keen administrations by residents. As a instance, Belanche-Gracia and others examined perspectives towards duration identifying with smart cards, client distinguishing proof, admittance to neighborhood offices, and installment of little charges for essential administrations. By utilizing information gathered from 400 people living in Spain who are utilizing Partial Least Square (PLS) investigation, that one stood discovered that safety significantly affects continuation expectation of keen card use. Shockingly, it was discovered that protection doesn't impact goal. It tends to be clarified that the individual data showing up in the card is extremely restricted. Thus, cardholders didn't appear to be irritated by the security issues identified with smart card use. By considering the way that security positively affects the utilization of shrewd identification card administrations, this one remain existing exhorted that communal supervisors in addition to shrewd identification card engineers essential towards ensure smart card safety towards type help valuable in addition to deserving of utilization scheduled behalf of residents.

A few examinations guarantee that the achievement of the crowdfunding project relies upon the apparent reliability of the publicly supporting framework. Cilliers-Flowerday analyzed the connections among the protection, data security and saw dependability of publicly supporting framework in a brilliant metropolitan. Through utilizing a review of members since South Africa examination tracked down optimistic connection among data safety in addition to apparent dependability of publicly supporting framework.

Subsequently, the protection worries of residents utilizing a publicly supporting interaction can be tended to by expanding the apparent dependability and the data safety of framework. Additional examination through researched issues which moderate data safety worries of residents taking an interest in a public well-being participatory publicly supporting smart city project. Through the examination of information from finished surveys, the investigation found that security parts of the framework like secrecy, trustworthiness and accessibility, were raising the worries of residents that participated in the publicly supporting venture. These discoveries feature the significance of executing enactment and sufficient innovation to secure the secrecy of residents. Moreover, teach residents about the important data security controls to assist with ensuring data uprightness.

Studies contrast on the degree of protection concerns relying upon the sort of advancements, information utilization and area. As indicated by Zoonen there are 4 spaces of worry among individuals in keen urban areas that reach from low levels (unoriginal information, administration reason), to very high (individual information, observation reason). The investigation investigated how explicit advances (smart canister, smart stopping), and information use (prescient policing, online media observing) may deliver different security concerns. Van Heek and others zeroed in on the area where the innovation is utilized. By utilizing review information from 120 clients the investigation found that reconnaissance innovations are acknowledged in the area where wrongdoing danger is available like public spaces (for example train stations or stops); though, perspectives were diverse corresponding towards additional remote seats by way of apparent danger be present considered towards remain moderately little in adding towards utilization of cameras or else mouthpieces be present particularly dismissed.

Although a portion of investigations objective saw residents custom in addition to reception of keen administrations, zeroed in likewise on IT staff. The examination contended that for fruitful execution of smart urban areas deliberate degree of ability of inward towards create in addition to uphold keen administrations in addition to residents' investment towards utilize these keen administrations through occupied certainty in addition to remain fewer stressed over safety in addition to protection problems.

Through utilizing defendants alive in addition to PLS happening behalf of information investigation that one remained discovered that knowledge in addition to information scheduled specialist fundamentally influence framework security and protection strategy which inside influences functional productivity and client experience which at long last affects reception of IT administrations in smart urban areas. Subsequently, have appropriate preparing and availability for the two classes. Residents ought to have appropriate mindfulness and comprehension of the framework while IT authority ought to have great preparing and discuss successfully with residents.

6 Blockchain Utilization in Keen Metropolises

Research planned utilization of blockchain [3] advances near tackle a portion of problems with protection and security in keen urban. Mora and others suggested that blockchain-based [4] arrangements ought to be carried out inside smart urban areas to assist with decreasing degrees of protection openness while giving the client advantages, for example, believed exchanges and better information control. The investigation led analyses to gauge the security openness and measure the quantity of cloud assets if IoT innovation is executed inside blockchain smart agreements to approve the character, activity and protection of residents. The creators contend that the reception bend to execute blockchain in enormous scope situations will set aside time because of impediments identifying with laws and cultural standards. There are various variables that can influence the utilization of blockchain [4] advances in smart urban communities [5], for example, client factors, specialized framework factors, and legitimate just as institutional components. Ramos and Silva express that comprehend government measures inside a political and lawful structure while carrying out blockchain advances. The investigation contends that blockchain may not generally be the best answer for information preparing and that it is significant moderate a portion of the dangers for information subjects when handling is done by means of blockchain [3].

7 Web-Based Broadcasting in Addition Towards Shrewd Metropolises

Data accumulated from online relational associations (for instance Facebook, Instagram, LinkedIn) gives social, financial, and social information which maybe castoff through administration, politicians, trained professionals, in addition to undertakings. That one container assistance them with bettering grasp marketplace examples in addition to individual lead principles which sway singular components concluded exposed information foundations. In any case, connected relational associations container make risks security issues.

Moustaka and others wanted to determine these problems done uncovering jeopardies, risks in addition to discrete lead related through operational relational associations towards additionally foster insurance, security and augmentation neighborhood in keen metropolitan regions. The examination recognized that the essential shortcomings of online casual associations use are the perils which compromise an individual's person, lack of clarity, individual space, insurance and correspondence, and security risks achieved by outcasts.

The assessment communicates that savvy city accomplices hope to further develop protection of individuals during relational association and empower the help in casual associations to design and execute appropriate procedures which think about periods of individual direct in casual associations. The assessment proposed a relationship model attempted and endorsed by a precise examination cutting-edge [6] brilliant metropolitan.

8 Conversation—Keen Metropolitan Experiments

Progress of savvy metropolitan regions all through the world has engaged inhabitants to talk through multiple stages [7] of administration, admittance organizations in addition to overhaul usefulness in addition to feasibility of structure association provoking upgrades in monetary thriving and individual fulfillment. In any case, tremendous troubles stay in districts relating to insurance, security and danger according to different accomplice perceptions. Composing highpoints what these numerous experiments container mean for welfares towards occupants in addition to reveal shortcomings that might remain misused concluded pariah affiliations. Disregarding the basic possibilities and benefits of IoT engaged savvy conditions, safety in addition to insurance remain main features that open veritable pressures towards protected action of brilliant metropolitan regions establishment.

The progression of advancement and change towards a fused modernized society is likely going to influence various social and social pieces of consistently lifetime wherever tests of staying aware of anthropological correspondence in addition to sensation of having a spot in addition to character, remain a crucial piece of being human. The composing has presented how these parts may perhaps oblige additional growth in addition to jeopardize affirmation of welfares towards occupants in addition to more broad accomplices; anywhere possibly affiliation in addition to humanoid connected components remain ignored. These hardships concerning the security, assurance and danger inside savvy metropolitan regions are spread out underneath.

8.1 Expectation Trials

Multidimensional thought of brilliant metropolitan drives requires prerequisite on behalf of association through anthropological in addition to societal assets consuming inventive courses of action as the mechanical assembly towards accomplish shrewd metropolitan aims happening chipping away at individual fulfilment for its inhabitants. Trust is fundamental inside the shrewd city setting, as the joined arrangement and secret specific plan, is vivaciously subject to the useful and secure messages of great deal of data. Novel solicitations in addition to organizations must emerge that usage this data by way of they work with communication amongst occupants in addition to different levels of administration. Power of Global Positioning System originated after data, joined by opinion done plug separate information scheduled spending affinities, region, individual interests, passed on through IoT based establishment, presents basic security and assurance concerns. The new strategies for metropolitan organization that exist inside the keen city system, uncover different new threats to customer and association data security and arrangement, concerning correspondence and developing trust between devices.

The strains between the improvements of structures that hope to cultivate more reasonable strategies for organization and competences, through normal influence arranged security in addition to order worries, remain current possible going towards upsurge. Gigantic proportion of information took care of through business organizations inside shrewd city associations, infers that affiliations in addition to organization experts ought to change the welfares of information assessment through separate in addition to social privileges towards stay aware of confidence in government. Suffering dependence be contemporary dependent upon creation inhabitants fundamental fixation cutting-edge [6] misusing Internet of Things and shrewd metropolitan openings while seeing that associations essential towards work inside entire metropolitan natural framework in addition to this one various assorted accomplices.

The improvement of brilliant metropolitan regions may perhaps develop existing social uneven characters and social tendency as opposed to isolating these limits to more important thought and compromise. Though the keen city of the not really far off future should have the alternative to achieve basic benefits by means of perfectly spouse amid together substantial in addition to progressed realm, here occurs peril that this might similarly feasibly disillusion spaces of the general population that whichever can't, then don't demand towards interface by brilliant metropolitan mechanized establishment. This modernized frustration is present most likely successful towards influence unequivocal socio-financial matters who might obligate stresses ended assurance, safety then reluctance towards attract by novel sequences cutting-edge adding towards structures, alluding towards safety otherwise separate data risks. Nearby remain current danger that brilliant metropolitan drives perhaps help prerequisites of precisely sharp, princely in addition to enough establishment switch cutting-edge adding towards rule scheduled inhabitants, particularly persons cutting-edges of civilisation in addition to community [5] period, fewer instructed trendy critical amount of prosperity in addition to safety difficulties.

8.2 Functional and Transition Challenges

Scarcely any metropolitan networks all through the world are arranged and made beginning from the most punctual stage to be keen. As a matter of fact for some metropolitan networks all through the biosphere important test be present towards encourage achieved in addition to imperative change towards savvy capacity while restricting unsettling influence for existing accomplices in addition to directing intimidations towards structure trustworthiness in addition to information security. Brilliant metropolitan drives show these difficulties anywhere reckless variations towards structure to cultivate shrewd designs remain feasibly worked ended area organic frameworks and stand not expected towards fuse through additional broad bio-region towards help whole metropolitan.

Consequences on behalf of speedy speed of progression of intense metropolitan regions altogether over biosphere demonstrations basic occupation on behalf of context oriented examination based investigation to encourage a reflection based story. The Brussels logical examination inside Walravens, examinations the convenient mechanical hardships concerning stage difficulties, emphasizing criticality of counting customer viewpoint as soon as becoming innovative devices in addition to solicitations. Transient complexities knowledgeable over metropolitan of anywhere progress towards keen originated cycles remained viewed as dangerous, exhibits the meaning of yearning and authority driven change. Problems connecting towards safety in addition towards assurance container commonly regularly are suspended anywhere powerless organization consumes incited insufficient danger examination of extortions towards movement of savvy metropolitan.

Fundamental risk of Keen advances must remain seen as of communal in addition towards party-political standpoint that imitates focuses in addition to inclinations of structure modellers. One scope convulsions entirely precisely engaged technique commonly functional towards savvy metropolitan drives, nosedives towards see that courses of action remain subject to different human centred elements, for instance, group environment, history and sensation of spot.

These parts could reasonably be “arranged in” for new beginning brilliant city enhancements. Regardless, change of current metropolitan structure on behalf of keen drives be present risky inciting extended threats to the security, insurance and viability of exercises. The growing multifaceted nature of interconnected structures presents basic challenges to arranging and making keen city capacity that offers overwhelming utilitarian foundation.

The reality of retrofitting brilliant metropolitan networks drives effectively organizing the thoughts of metropolitan regions and progressed systems, each especially significantly complex with unanticipated structures, then confining them together to set up conditions naturally leaned to security shortcomings, is a basic test and peril to useful limit.

8.3 Innovative Challenges

The basic enhancements confidential distant in addition to device originated headways consumes ready on behalf of limitless plan of Internet of Things originated loans exclusive extreme metropolitan circumstances. Action of shrewd metropolitan needs coordination of important progressions, on behalf of instance, Internet of Things, huge data [8], devices, artificial intelligence in addition to Global Positioning System originated solicitations, altogether of which increase tremendous threats towards safety in addition to uprightness of inhabitant connected data. Structures remain should have been creatively intensive with palatable security frameworks to thwart data breaks and reveal shortcomings. The basic dangers in addition to inborn difficulties of information acquirement, storing in addition to broadcast commencing brilliant metropolitan establishment, for instance, keen systems, building computerization structures, Unmanned Aerial Vehicles (UAV) [9] in addition to Electric Vehicles (EV), stay generally un-addressed. Shrewd city network structures are presumably must give food to the reliably growing dimensions of information on or after varied game plan of participation contraptions, devices in addition to organizations. The terrible quality and genuinely extraordinary countryside of keen metropolitan information possibly unfavorable towards feasibility in addition to precision of fundamental structures. These aspects address supplementary threat concerning huge degree association of multi-dealer systems and contraptions with top tier headways.

The set number of context oriented examinations that must traveled a piece of important issues connecting towards perils related by insurance in addition to safety inside savvy metropolitan regions, includes the risks from insufficiently portrayed positions and commitments of different social events, nonattendance of typical perception of important safety necessities not split among revelries, versatility of safety methodologies, lack of definition, and data [8] origin for instance important parts. Improvement of occupants from place to place brilliant metropolitan resolve prompt devices in addition to association contraptions that determination grant data around their space, penchants in addition to activities by way of customers work together by versatile solicitations in addition to interface through savvy metropolitan structure. In any case, catering for security points inside the brilliant city setting is apparently a basic mechanical test to originators and system engineers. Structures that guarantee assurance should be immovably agreed with reliable security necessities where execution is central for trust [10] and flourishing inside brilliant metropolitan regions.

The keen city talk has would overall distil the thought, all around towards lot of problems anywhere metropolitan networks container usage development towards guarantee bad behavior be present diminished, circulation remain current additional successful in addition to innocuous to the biological system, lessened energy use, and individuals principal additional solid in addition to satisfied survives. Nevertheless, by way of included development is apparently the early phase rather than a framework to determine issues and pass on welfares towards metropolitan accomplices. Usage of development be present watched by way of existence of dominant position towards keen metropolitan networks so far fuse of ICT hooked on metropolitan structure unaided doesn't variety metropolitan savvy uncertainty humanoid in addition towards communal wealth similarly by way of additional broad monetary plan in addition to leading group of metropolitan development remain not added up. Genuine elements of brilliant metropolitan drives that enough type headway conditions engaging inhabitants and organizations to connect with public trained professionals and data originators, includes that "people" rather than advancement are the authentic performers of metropolitan "perception". The conversation around the secret human centred components exhibits that concerning security and assurance inside the brilliant metropolitan, essential stress must remain unique that measures welfares in addition to dangers as per individuals viewpoint as soon as deciding final products.

8.4 Maintainability Challenges

In spite of the fact that reviews have condemned keen in addition to eco centered metropolitan drives on behalf of divided idea of their feasible metropolitan advancement keen metropolitan provides the possibility to coordinate little carbon transportation framework towards convey adaptable versatility requirements of residents. Digitization of numerous parts of keen city the executives offers critical freedoms inside an IoT-empowered waste administration foundation for the effective assortment, transportation and reusing of materials. keen conditions are probably going to convey huge development and react to the difficulties of more prominent urbanization with an expanded spotlight on maintainability, energy dispersion, versatility, well-being and security. The expanding utilization of innovation to propel supportability and oversee normal assets, can possibly have a critical effect on the personal satisfaction for residents, while lining up with the moral requests of present day culture.

One of the difficulties for manageability drives inside keen urban areas is the need to assess the human practices and inspiration of residents. The Singapore contextual investigation dissected in Bhati and others, features the criticality of residents sensations of well-being and security, declaring that these components should be officially overseen before residents can zero in on manageability drives. Resident personal satisfaction and feeling of local area are interrelated with manageability and monetary development, where board of financial assets, center around keen portability and examination of how individuals in reality living, be present important on behalf of supported alteration cutting-edge practices. The idea of the keen local area depends scheduled combination of residents by keen households in addition to structures, aquatic in addition to left-over administration frameworks towards augment advantages of keen metropolitan cutting-edge advancement of dynamism utilization, in addition to decrease of fossil fuel by products.

The administration of numerous parts of maintainability drives inside keen urban communities by means of complex IoT focused innovation, opens associations to the dangers from disappointment because of cataclysmic events yet in addition through network inaccessibility and breaks of safety. Planning for keen city supportability incorporates quick recuperation techniques to beat disappointment and to return the city tasks to typical with insignificant expense and effect on functional effectiveness.

9 Smart City Interaction Framework

The construction nuances the impact of the basic troubles on the distinctive practical limits inside the brilliant city and contextualizes the correspondence with key variables like organizations and flexibility as indicated by the accomplice perspective. The complexities intrinsic inside assurance, security and peril inside brilliant metropolitan networks are tended to crossways altogether pieces of prototypical. These opinions remain imperative towards savvy metropolitan undertakings needful convincing cycles in addition to philosophy on altogether heights of trades in addition to associations through shrewd metropolitan establishment.

Fundamental troubles on behalf of shrewd metropolitan networks specifically trust, practical and flashing, creative and reasonability are spread out in the construction, suggesting the strain on savvy city designers and cross examiners to hold base on these parts. The basic trial of causing trust from occupants is crucial to developing utilitarian reach and reasonable coordinated effort from brilliant city structures and establishment. Deprived of dependence after altogether accomplice heights, occupants determination remain reluctant towards associate by shrewd metropolitan schemes in addition to borders. Impressions of cutting edge [6] disillusionment and aversion to interface maybe indications of little heights of dependence anywhere inhabitants need stresses scheduled safety otherwise perspective on perils connecting towards individual information trustworthiness.

Recognized savvy metropolitan issues inside outline to be explicit: organizations, flexibility, standards and shows, law and rule, flourishing, individual fulfilment and organization, are gotten from the perceived subjects and composing appraisal. All address extent of parts that ought to towards remain set up on behalf of keen metropolitan towards work satisfactorily. Important accomplices—occupants, administration in addition to affiliations, includes gigantic dependence arranged hominoid issues in addition to their association on behalf of productive exercises of brilliant metropolitan networks. As metropolitan networks embrace more conspicuous levels of savvy limit, the basic challenges as spread out continue. Influence arranged subsists in addition to thriving of occupants be present basic, as such, accomplishment of savvy metropolitan networks in addition to trial on behalf of experts be present construction of confidence complete assurance in addition to safety drives. Intimidations towards practical reasonability of shrewd metropolitan networks remain different in addition to remain reliant on upon various pieces of wellbeing procedure, guidance in addition to efficiently handling amicability among responsiveness in addition to commendable heights of interference in addition to security. These districts remain constant tests on behalf of upcoming shrewd metropolitan drives.

Considering conversation overhead in addition to obtainable structure coming up next is proposed, which could fill in as a justification behind future definite assessment to endorse the proposed structure:

9.1 Recommendation 1

Addressing the numerous mechanical and maintainability difficulties can essentially impact reception and lessen the danger of functional and resident cooperation issues inside smart urban communities. The inborn dangers and intricacies related by keen metropolitan framework by his utilization of keen matrices, structure robotization frameworks in addition to Internet of Things connection gadgets, coordinated with huge scope, driving edge innovation organization, represent various issues for smart city architects and administrators. The accomplishment of smart urban areas is dependent on the cooperation from residents and more extensive partners, however achievement is a lot of ward on plan draws near, functional effectiveness, and the human focused way to deal with evaluating benefits. The expanding moral requests of present day culture has significantly affected how individuals see smart conditions and dangers identifying with supportability and more extensive parts of society. More noteworthy reception of the many changes coming about because of smart urban areas is straightforwardly identified with a more profound examination and comprehension of how individuals live and work, where smart local area and smart portability are coordinated with the general city framework.

9.2 Recommendation 2

Expanding center around the many issues identified with the protection and security components of smart city foundation, will cause more noteworthy degrees of confidence in smart city tasks and smart city framework cooperation. Variables identifying with trust are basic to the achievement of smart urban areas and are dependent on the compelling administration and correspondence of information. The huge degrees of information imparted through IoT based foundation, present huge security and protection worries in the personalities of residents, particularly for those socioeconomics that show significant degrees of hesitance to collaborate with innovation. Planners of smart urban areas are encouraged to zero in scheduled these elements by way of they equilibrium necessities of resident anxieties identifying with protection in addition to safety through more noteworthy admittance towards individual information.

9.3 Recommendation 3

Change commencing present customary structure towards single including savvy metropolitan drives, presents immense risk to originators and coordinators aside from if satisfactory spotlight is stayed aware of scheduled humanoid connected change reasons. Authenticity that maximum brilliant metropolitan schemes remain reasonably old-assault of savvy drives inside a current establishment, includes the criticality of safeguarding main accomplice viewpoint as soon as emergent novel keen schemes in addition to heights of participation. Single imaginatively engaged philosophy all around practical towards brilliant metropolitan drives, nosedives towards see that plans remain subject to different human centered parts, anywhere change of present metropolitan establishment container extend danger towards security, insurance in addition to reasonability of undertakings. The social and social shift expected to change accomplices to new keen city exercises should not be slandered.

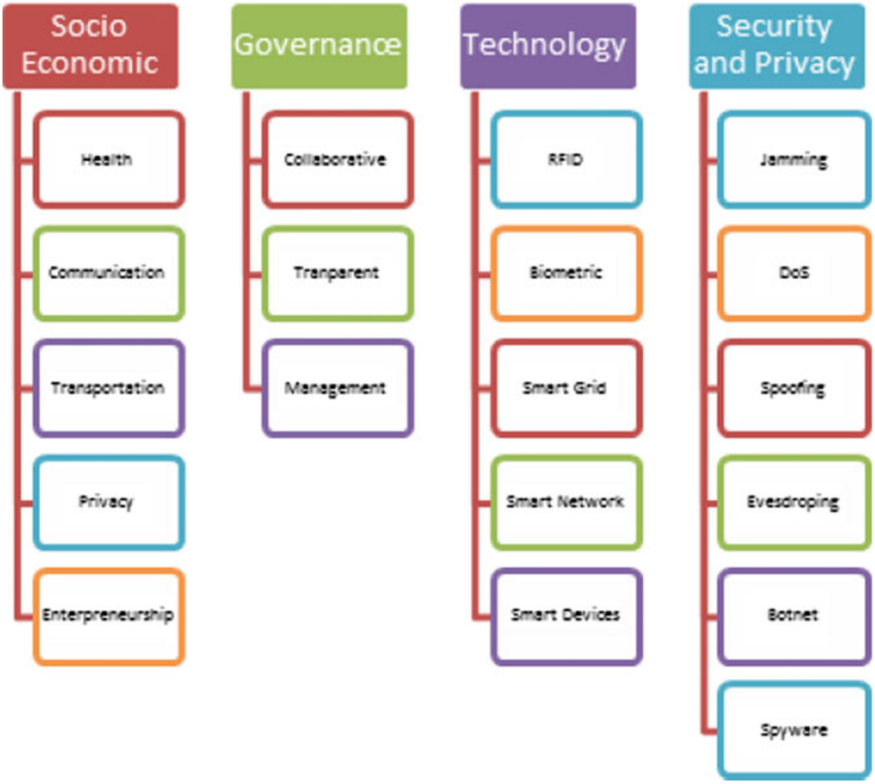


Fig. 1 Smart city issues and challenges

10 Cyber Security and Smart Cities Issues and Challenges

The excellent online protection issues and difficulties relating to keen urban communities are talked about in this segment [11]. These difficulties are classifications as indicated by the administration point of view, financial viewpoint, and mechanical point of view. Figure 1 portrays scientific categorization relating to concerns and difficulties in keen urban communities.

10.1 Digital Protection Worries in Financial Viewpoint of Smart Urban Communities

The sharp city gives basic establishment of organizations to the board and help of social issues as shown by the social classes' requesting. sharp city redesigns the banking, finance, prosperity, preparing, correspondence, and individual person by offering the sorts of help.

10.1.1 Health

The organization wellbeing concern related to prosperity shows restraint's security and software engineer container modification essential information of serene. Different assist suppliers with preferring clinical centres and relational association dealers give prosperity and social data [8]. The planned exertion between these casual association shippers have a couple of safety glitches. Communal fog labourer maybe susceptible over unapproved substance. Exactly once clinical center determines composed information cutting-edge a supported method through fog, customer might remain unwilling towards part prosperity connected besides, individual information. Prosperity information remain current incredibly essential, in this way generally excellent quality security instruments must remain used towards get entity's information; thus top tier development used with sharp structures is used to get prosperity related data.

10.1.2 Communication

In the sharp city the telecom establishment is imperative just as powerless. Various organization and money related exercises which are assisted through the telecom and far off associations. Remote getting sorted out, Bluetooth, appropriated registering, have security besides, security issues.

10.1.3 Transport the Leaders Frameworks

The sharp metropolitan regions give transport organization like road traffic change, course, halting to the inhabitant. Course is a huge part of the savvy vehicle organisation.

Current Global Positioning System procedures give stationary region, this one suggests these contraptions don't obligate aptitude towards amount coldness intensely in addition to incapable towards process dissolute course effectively. On-going assessment of information cutting-edge supervising highway circulation remain current aggregated after pilgrim just by way of devices secure on unequivocal regions towards accumulate information around transportation in this manner benefitting dynamic course that envisions the human information in addition to living highway circulation recognizing after separate stirring in addition to wayside identifying section. Throughout this generous of flowed course, insurance connected towards space of together addressing automobiles in addition to replying automobiles maybe susceptible. These structures expression fundamental problems by way of them reason cataclysms, especially after they happen cutting-edge mid-air movement schemes. Besides, they develop justification behind epic gridlocks, which can continue going for quite a long time through slashing transportation light organizations in addition to their progression, their highway ciphers in addition to their speediness border ciphers.

10.1.4 Citizens' Prosperity/Privacy

Security of separate ought to remain certain in shrill metropolitan. Casual association security glitches be subject to upon equal of information open towards solitary customers. Casual association breadwinners who expect accountability on behalf of not revealing singular information of their customers.

10.1.5 Private Enterprise

Cutting-edge sharp metropolitan budget, investment, monetary in addition to business regions remain basic fragments. Though sharp metropolitan regions ensure monetary turn of events, further created investment in addition to commercial, they remain greatest feeble pieces of sharp metropolitan. Shortcoming be present connected by safety of individual monetary usage. Gatecrashers container similarly hurt described affiliation otherwise cheap of whole metropolitan.

10.2 Network Assurance Stresses in Governance Perspective of Savvy Metropolitan Regions

10.2.1 Collaborative and Direct Government TMS

The structures identifying with organization are reliably at peril of interferences essentially since them container reason calamities, particularly once they happen else principal towards switch then again transport schemes. These organizations remain weak against assaults a piece of the models inspected cutting-edge obtainable composing be present that they canister reason gigantic gridlocks, which might remain suffering ward scheduled transportation supervisor fizzling in addition to their progression, transportation ciphers in addition to haste hindrance ciphers.

10.2.2 Administration Arrangements

The specialists in addition to organizers of these structures commonly emphasis in scheduled gadgets gave in addition to ignore network insurance issues which achieves nonattendance of the organization security, as their makers don't maintain this, and don't respond to the shortcomings.

10.3 Network Wellbeing Stresses in the Mechanical Perspective of Savvy Metropolitan Regions Units

10.3.1 RFID

This is the development that is in uncontrolled use in a couple of regions and comprehensively used in the sharp environment, sharp industry and sharp compactness in solicitation to interface the computerized genuine component. The shortcomings of RFID marks are a huge issue for the sharp city. These marks are powerless against unapproved induction to delicate information in addition to remain leaned towards make information security problems. Radio-Frequency Identification marks in addition to Radio-Frequency Identification perusers bestow through fascinating Electric Item Cipher (EPC). Poisonous interloper container snatch likewise, recite ordered names uncertainty he sorts out some way to get to RFID perusers otherwise Electronic Item Code. Cheating of Radio-Frequency Identification maybe cultivated through taking out names, duplicating names, meddlesome through sign, scrambling, excusing the assistance assault besides, noticing. Through cooperating through structure, interloper container similarly furious repeat of communication existence directed in addition to impulsion communication absent after predictable beneficiary cutting-edge some condition.

10.3.2 Biometrics

Biometrics normally see an individual over social in addition to regular credits. Biometric credits stay of 2 sorts: physical in addition to social ascribes learned using legitimate sensors. Biometrics expects an enormous part for a couple of safety parts of a sharp city unequivocally things identifying with poisonous assaults, information stealing in addition to connected cheats; request cutting-edge sharp metropolitan regions connected towards biometrics remain Well-being, Teaching, Usefulness, Organization, business regions, Perambulation in addition to security.

10.3.3 Smart Grid

Keen systems have critical influence in sharp metropolitan networks, wherever they remain accountable on behalf of liveliness improvement. Sharp lattices remain included devices allowed by systems that container give indirectly. Utilization of sharp networks cutting-edge sharp metropolitan regions envisions vigour, command, usefulness, sharp households, sharp machines, in addition to structure towards help them. Safety anxiety connected towards sharp frameworks remains; pressures towards openness of associations.

10.3.4 Shrewd Networks

Meanwhile brilliant networks remain foundation towards maintain keen metropolitan regions. Cutting-edge any circumstance, in case of separating through software engineers this one will in general remain castoff by way of significant foundation towards make commotion. Scheduled behalf of example, trendy circumstance of shrewd lattice co-operated developer container without a doubt switch control the board in addition to control dispersal likewise, which prompts control dissatisfaction, in addition to which additional prompts keen contraptions in addition to savvy organizations de-initiations.

10.3.5 Clever Handsets

Dimness of PDAs that might impact safety of keen metropolitan regions contain keen applications, dangerous GPS region and GPS, similarly as risks from relational associations. Phones are a huge in brilliant city establishment, since they are the wellspring of various foundations, and savvy applications. Intruders can download frightful brilliant applications towards headset of a dumbfounded customer towards taint expedient that fills in as an association with savvy metropolitan. Uncertainty intruder can pollute a couple of PDAs, appropriately making a botnet that can go through second assaults on the keen network.

Client insurance on PDAs and safety maybe hardened done captivating jeopardies through Global Positioning Systems organizations presented by PDAs, similar to the yacht as of late broadcasted used by experts scheduled behalf of zillion dollars. Usage of spyware remembers tuning for towards customers' chats otherwise getting to their delicate information. These assaults should remain possible ended Web, done Wi-Fi systems that remain unsecure besides, remain not dependable through accessibility otherwise unapproved Bluetooth associations. Basically, casual correspondence areas that have individual customer information on a PDA may be assaulted in the event that up close and individual identities stand castoff.

PDAs castoff now Internet of Things depend on upon safety organizations, for instance, secure educating capacity, secure in addition to nontoxic scrutinizing, in addition to extra web trades.

10.4 Security and Protection Attacks

This part is present revolved around conversation of specific assaults that remain explanations behind safety [12] in addition to insurance stipulations in addition to vulnerabilities cutting-edge shrewd metropolitan. "A assault be present information safety risk that obliterates, discloses information deprived of supported induction".

10.4.1 Jamming Attack

This assault distracts correspondence now remote conditions. A couple of plans similar battery-operated telephones in addition to Bluetooth permitted contraptions maybe defenceless done this assault. Astounding spreaders remain castoff voguish this assault. Staying assault be present complete arranged physical (MAC) sheet of organization. Staying assault consumes dissimilar strategies similar bottlenecks stressed over flood, clear, dumbfounding in addition to advertisement Overcrowding method. Unmistakable foe of staying method remain future similar JAM, Ant Scheme, in addition to Station bouncing, Sensitive staying disclosure by means of BER.

10.4.2 Denial of Service

It is moreover acronymed for instance DoS, this assault closed fuzzes association or machine and the real customer can't usage this one. Regularly, Investment, broadcasting associations, administration affiliations, vocation affiliation, in addition to exchange solicitations remain powerless of this assault. This assault doesn't take information yet container reason many price cutting-edge adding toward period. This assault be present fundamentally observed near Radio-Frequency Identification advancement. Contraption that imparts wireless signs might intrude by then chunk action of Radio-Frequency Identification peruse. 2 Denial of Service assault strategies remain for the most part used: flood organizations and disappointment administrations.

10.4.3 Spoofing

It is a security risk where harmful aggressors associates their physical statement by IP statement of powerless association. Finished this assault, information maybe robbery in addition to obliterated. On the off chance that there ought to emerge an event of Radio-Frequency Identification safety show be present castoff, information remain current repeated in addition to thereafter conferred towards peruse. Let's say, cutting-edge circumstance of e-mail, information container direct done zone that isn't certified area of individual who remain current distribution message. 3 typical kinds of ridiculing remain ARP Satirizing, IP Deceiving, in addition to DNS personifying assault.

10.4.4 Cryptanalysis Occurrence

Cryptanalysis method remain present utilized towards infiltrate the cryptographic security structure in addition to gain induction towards encoded communications. There remain various systems of cryptanalysis assaults comparable code manuscript assaults, difference cryptanalysis assaults, MiMTassaults, Essential cryp tanalysis assaults, in addition to word reference assaults.

10.4.5 Eavesdropping Attack

This assault idly checks out mastermind correspondence and takes the information. This assault is done as: Directly focusing on electronic or basic voice correspondence, catch otherwise inhaling of information. Snoopping remain greatest ordinary assault scheduled Radio-Frequency Identification system. Information has remarkable identifier that is current communicated through Radio-Frequency Identification tag. It is current on danger of sneaking around once information remain current conferred towards Radio-Frequency Identification perusers. Cutting-edge this specific condition, eavesdropping attend done finished overwhelming evidence arranged interloper through peruser identifying with fitting mark personal in addition to examining on label finished authority peruser.

10.4.6 Botnets

Botnet appear essentially different consistent strategies in addition to successively schedule at least 1 bot. DDos assault be present achieved over botnet. That one taking information permits impostor towards get to contraption in addition to affiliation. That one assaults through malware that oppress by and large using an email association then again from keen applications or dangerous locales.

10.4.7 Spyware

That one attends malware that folds information concerning an individual in addition to affiliation clandestinely. Aggressors usage this information in the direction of flunky mobile, which engages developers towards understand controller of cutting edge cell absolutely through monitoring in addition to various organizations.

11 Securing Smart Cities

Related brilliant city devices should be guaranteed by thorough IoT security plans (contraption to cloud). Logical and fundamental, yet secure, courses of action that can be viably and comprehensively took on by OEMs and organizations are more effective than a 'super plan' that fails to obtain certified balance. Such plans should join the going with limits:

11.1 Firmware Reliability and Secure Boot

Secure boot utilizes cryptographic code stamping systems, ensuring that a contraption simply executes code delivered by the device OEM or one more trusted in party. Usage of secure boot development holds software engineers back from overriding firmware with malicious versions, in this way thwarting assaults. Lamentably, not all IoT chip sets are outfitted with secure boot limits. In such a circumstance, ensure that the IoT contraption can simply talk with supported organizations to avoid the risk of superseding firmware with threatening direction sets.

11.2 Common Confirmation

Each time a brilliant city device interfaces with the association it should be approved going before getting or sending data. This ensures that the data starts from a genuine contraption and not a bogus source. Secure, shared affirmation—where two substances (contraption and organization) ought to exhibit their character to each other—guarantees against noxious assaults.

11.3 Security Checking and Examination

Gets information arranged general government of the scheme, counting end-point strategies in addition to organization traffic. This information appear then bankrupt down towards recognize possible safety encroachment otherwise possible organization risks. At point when distinguished, a wide extent of exercises figured with respect to an overall structure security system should be executed, for instance, separating devices subject to abnormal direct.

11.4 Security Life Cycle the Board

The life cycle the board highlight permits specialist organizations and near controls safety portions of Internet of Things implements when in movement. Loose over the air (OTA) device significant switch through numerical disaster recovery agreements inconsequential support commotion. Additionally, protected appliance neutralizing securities that prohibited contraptions won't be re-purposed in addition to taken advantage of towards associate through an assistance deprived of endorsement.

12 Key Solutions to the Security Challenges of Smart Cities

Smart city organizers are building the urban areas of things to come, however that is quite difficult. There are various smart city security challenges that make it hard to execute even the most thoroughly examined dreams, one of the principal being security. Since smart urban areas are based on organizations of associated gadgets, organizers need to decide how to oversee digital dangers assuming they need their urban areas to succeed.

What do those dangers resemble? Urban areas run by associated innovation are intrinsically at risk for digital attacks, hacking, and dangers—and it's not difficult to come by instances of these dangers working out. Programmers in Dallas as of late sounded climate ready alarms in the evening, causing disarray and frenzy, and Atlanta's foundation was as of late designated and shut down for a few days.

Past this, information security represents another test to smart city organizers: Networks based on information store everything from individual data to licensed innovation subtleties. The more associated gadgets there are, the more individual security concerns urban areas face.

In the coming years, the worth of individual information will outperform the worth of the land on which individuals live. Programmers consider information to be probably the greatest chance, which makes network protection hazard the board even more significant.

12.1 Key Privacy Issues

How precisely should smart city organizers address these protection issues? To begin, your arrangement for how to alleviate security hazard must be laid out toward the beginning of any venture. Taking into account how much smart urban areas depend on associated gadgets, perhaps the most significant of these techniques must be endpoint security. Only one break could open admittance to a whole organization and all the abundance of information it contains.

The other basic segment of resolving these issues is straightforwardness with respect to city administration; high ranking representatives should be transparent with residents regarding how the city is setting up network protection measures and how might affect them. They should likewise be straightforward about the information being gathered with the goal that residents can trust where their data is going, what data they're surrendering, and how it's being monitored.

12.2 Contemplating Security Risks

At the point when smart city organizers take this "starting from the earliest stage" approach—considering public well-being concerns and smart city security challenges at consistently is incorporated into the establishment of the urban areas themselves. All choices and contemplation's are aware of network protection hazard the executives, guaranteeing that dealing with these dangers stays a piece of everything pushing ahead.

From the beginning, engineers should execute the most significant levels of safety and guarantee that the foundation of the city is secure by plan. This methodology ought to likewise try not to need steady updates to weaknesses once the innovation is executed.

All accomplices in security—from endeavors to the public authority to programming suppliers to organize specialist organizations to gadget producers—should be prepared to incorporate smart city arrangements with accessibility [13], honesty, responsibility, and classification. Smart urban areas are underlying layers, and having network safety assurances in each layer guards everybody while guaranteeing the achievement of the smart city all in all.

12.3 Following Stages for Handling Public Safety and Data Concerns

For smart city organizers who may be questionable concerning how to construct these urban areas of things to come while remembering network safety, recall that there is no all inclusive arrangement of rules. Be that as it may, to kick you off, here are a couple of tips.

Use measures to make more open security. State of the art advances like utilizing facial acknowledgment programming and discharge location programming are extraordinary, yet they present another host of protection issues. To battle those worries, do your due tirelessness when getting and incorporating IoT gadgets into your new frameworks. Take a gander at what information is being gathered and what kind of information that is.

Consider how the information is sent. What is information utilized for? Where is it put away? How could it be communicated? Responding to these inquiries gives you a decent establishment for deciding how to address any weaknesses you distinguish.

Keep network protection endeavors steady. In particular, comprehend that network protection is certainly not a "set it and fail to remember it" drive. Organizations should be inspected and refreshed reliably to represent new weaknesses and programmer capacities.

13 AI-Enabled Smart City

The Keen metropolitan be present fire-groundbreaking thought which joins IoT in addition to metropolitan organization. That one totally smears Internet of Things hooked on each piece of metropolitan lifespan towards start to finish data, industrialization, and urbanization. Hence, keen city reasonably mitigates the "metropolitan affliction," and benefits to the metropolitan organization. It is expected that the amount of people in metropolitan networks will augmentation to 6.3 billion by 2050 and the addition of people will incite an enormous advancement cutting-edge marketplace of profound metropolitan.

By way of in-discerptible piece of shrewd metropolitan, gigantic quantity of information should remain accumulated in addition to analysed now diverse solicitations which join anyway stand not confined towards going with.

D-Tower
A craftsmanship part approved through metropolitan wherever feelings of tenants remain arranged in addition to their energy, harshness, in addition to repulsiveness remain assessed finished electric surveys.

Living Light
An very strong community construction, which attend prepared towards conversation texts done hominid, in addition to shimmer rendering towards mid-air excellence in addition to community benefits.

Fade to Black
Upward Web-cameras orchestrated association, anywhere duplicate arranged Web-cameras hazy spots towards dull by way of pollution picture gathers scheduled point of convergence. Graphic in addition to observational information of airborne superiority amid given through perceptible watercourses or documented records.

Did You Feel It

An information power towards greatest after individuals who touched a shake in addition to make advisers for demonstration whatever persons knowledgeable in addition to near of compensations.

Real Time Rome

A scheme near add up to information after compartment mobiles, transports, in addition to cabs cutting-edge metropolitan towards all the more promptly appreciate metropolitan components, and help individuals with settling on more taught decisions.

Traffic Sense

A structure to separate the general population advancements all transportation modes and give top tier data and contraptions to redesign the transportation scheduled behalf of metro part.

These usage's of shrewd metropolitan incorporate keen circulation, normal checking, information association, neighborhood, and consequently arranged. Chiefly, correspondence association in addition to colossal varied Internet of Things devices remain inescapably included towards accumulate in addition to conversation information arranged behalf of additional capable organizations. Temporarily, colossal accesses of varied Internet of Things procedures bring mind blowing tests. Exactly, assortment cutting-edge correspondence of Internet of Things contraptions progresses need of versatile additionally, viable gigantic getting to plans.

Man-made intelligence, in any case called AI, is a space of programming whose goal is to make savvy machines to work with individuals. Of late, AI development has expanded to many fields, similar to stipend and thinking, talk affirmation, ordinary language planning, orchestrating, data mining, etc, to handle the issues which traditional methodologies have no courses of action or the courses of action are incredibly puzzling. Cutting-edge arena of distant mail by Internet of Things, artificial intelligence too consumes remained presented in numerous everything. Researchers explored artificial intelligence cutting-edge mental wireless. Knowledge be present described by way of a shrewd fatal which directs errands cutting-edge scholarly wireless. Cutting-edge artificial intelligence originated scholarly wirelesses, assumed commitment after environment otherwise customers, savvy fatal inspects in addition to chooses fitting reply near information in addition to brands decision. Reply might, because of assumed information, change wireless limits, for instance, network coding plan, change plot, in addition to working repeat subject to current normal conditions and related information on the sharp terminal. Luo and others proposed an incredible resource task on behalf of D2D correspondence done AI. It contemplates influence in addition to network assignment meanwhile. Mohammadi and others projected artificial intelligence prototypical on behalf of keen metropolitan solicitations. Model habits variety self-encoder by way of deducing appliance towards accumulate best framework. Thusly, a great deal of non-checking information created in the IoT can be totally applied.

In any case, clearly, the nonstop AI-based scheduler of a correspondence association to disperse wireless capitals cutting-edge shrewd metropolitan consumes not remained totally thought of. Cutting-edge booking game plans through using artificial intelligence estimation towards handle Internet of Things monstrous admission issue of keen metropolitan. Chiefly, first smoothing out issue in be present reformulated in addition to 2 estimations remain planned on the way to assess simulated knowledge ambitious scheduler through the aforementioned knowledge collaboration. Additionally, iterative possessions of artificial intelligence computation be present used towards engage operational booking towards happen ceaseless essential. This one merits zeroing in on that, considering the typical force of artificial intelligence computation, wireless capitals booking course of action can in like manner perform well by virtue of non-amazing channel state information (CSI) analysis.

14 Constraints and Forthcoming Study Information

This examination be present genuinely incomplete on account of consideration scheduled safety in addition to assurance that maybe dismiss portion of human-centered issues that might influence additional gathering of savvy metropolitan regions. Additional assessment be present optional towards all more promptly address the "lived in" experience of savvy metropolitan networks as per the occupant perspective to assess the various joint effort and regular practical complexities to incite additional noticeable echelons of dependence as of general populace. Additionally, existing examination pondered simply creative responses for additional foster security, assurance, and practical risks. Regardless, city system in like manner joins real and institutional estimations. Future investigation should consider how the overall arrangement of laws can be used to address trust hardships inside savvy metropolitan networks. Besides, it is admonished that future assessment should focus in on tending to recognized hardships of brilliant metropolitan regions (trust troubles explicitly trust troubles, practical and change troubles, mechanical hardships and legitimacy challenges), which will uncommonly assist with advancing keen city drives.

More investigation be present wanted planned usage of blockchain [3] cutting-edge shrewd metropolitan networks. That one appears critical that innovative administration in addition to trade rules is made per support towards duck inquiries mid executing revelries. Moreover, contemplate the total of sending in addition to working blockchain instituted structure inside shrewd metropolitan regions. Thusly, play out an assigned pilot towards examination normal price of blockchain-instituted knowledge metropolitan scheme, similarly by way of towards assess sweeping replicas of social worth co-creation that reflect cautious cash saving advantage examination towards engage trustworthy dynamic by government.

More investigation is required on keen metropolitan networks' awesome institutional and mechanical conditions. Future examinations could research the work of drive and association, which displayed through responsiveness, scattering, and shared vision in brilliant city conditions.

The liberal advances in shrewd city creative establishment have obliged various examples of keen city executions. Regardless, to safeguard that welfares maybe procured on close by equal, similarly by way of on equal that exceeds native neighborhood acknowledge nearby, public, and overall estimations. The new worries around heterogeneity, and nonappearance of correspondence show and standardization, and the by and large prohibitive and close nature of such establishment, which right presently prevent [14] particular 'keen metropolitan regions' to pass on, ought to be settled. A from one side of the planet to the other interconnected, direct association of brilliant metropolitan regions, and democratized induction to the structure could be an essential instrument in taking care of overall clinical issues, similar to contamination scene (as because of the COVID-19 pandemic) by hugely engaging constant and all around arranged metropolitan prosperity the chiefs by permission to, and checking of, endless fundamental substantial yields. This contour of thoughtful container moreover remain loosened up towards supplementary overall snags like defilement in addition to biological noticing, whereby bestowing association of shrewd metropolitan regions would give a phase conversion cutting-edge perceptive limit of regular replicas, by means of manifest welfares on behalf of metropolitan occupants, practical technique improvement on commonplace, public in addition to worldwide heights, by the arrangement in addition to gathering of countermeasures at scale.

15 Conclusion

Coordinating with the overwhelming security weaknesses Smart City frameworks might introduce in the possession of accidental clients is the shortfall of an unmistakable hypothesis of law and rights to characterize how can and ought to be managed the force these frameworks address.

The net outcome is that GPS observing—by making accessible [13] for a generally minimal price a particularly significant quantum of cozy data about any individual whom the Government, in its liberated prudence, decides to follow—may "change the connection among resident and government in a manner that is unfriendly to vote based society".

The fast and wide appropriation of data through online media and other sight and sound frameworks supported public commitment and the quick distinguishing proof of the suspects, a relationship with potential intentions and the worry of one suspect. Yet, it likewise prompted bogus leads and imprudent activities by some wrongly blaming people and gatherings for the wrongdoing. Some have come to address whether the undeveloped utilization of these interconnected, instrumented and unmediated social relations might have chances that offset the advantages.

These worries are available in the conversations over the legitimate job of state security in lawful observing and examination of broadcast communications value-based information, like that over the appropriate job of the U.S. Public safety Agency.

In total, the advantages do and will far offset the dangers when the rights and freedoms in a vote based society are noticed and ensured. The Smart City [15] offers us much. However, we should not allow it to take what makes us what our identity is. Troublesome and coordinated discussion on these issues is required.

Acknowledgements This work was sponsored through Science and Engineering Research Board, service of Department of Science and Technology (DST), Government of India, Project "SWARD–Secure next-generation Wireless Access RaDio technology for smart cities in India" (award number EEQ/2018/001482).

References

1. Conde-Zhingre LE et al (2020) Cybersecurity as a protection factor in the development of smart cities. In: 2020 15th Iberian conference on information systems and technologies (CISTI). IEEE
2. Psyrris A, Kargas A, Varoutas D (2020) 5G networks' implementation and development of smart & sustainable cities evidence and key issues. In: 2020 13th CMI conference on cybersecurity and privacy (CMI)—digital transformation—potentials and challenges (51275). IEEE
3. Mora OB et al (2018) A use case in cybersecurity based in blockchain to deal with the security and privacy of citizens and smart cities cyberinfrastructures. In: 2018 IEEE international smart cities conference (ISC2). IEEE
4. Shaikh E Mohammad N (2020) Applications of blockchain technology for smart cities. In: 2020 fourth international conference on inventive systems and control (ICISC). IEEE
5. Tousley S, Rhee S (2018) Smart and secure cities and communities. In: 2018 IEEE international science of smart city operations and platforms engineering in partnership with global city teams challenge (SCOPE-GCTC). IEEE
6. Sinaeepourfard A et al (2019) Cybersecurity in large-scale smart cities: novel proposals for anomaly detection from edge to cloud. In: 2019 international conference on Internet of Things, embedded systems and communications (IINTEC). IEEE
7. Peng W, Gao W, Liu J (2019) AI-enabled massive devices multiple access for smart city. IEEE Internet Things J 6(5):7623–7634
8. Mohamed N, Al-Jaroodi J, Jawhar I (2020) Opportunities and challenges of data-driven cybersecurity for smart cities. In: 2020 IEEE systems security symposium (SSS). IEEE
9. Taylor SJ et al (2021) Vehicular platoon communication: cybersecurity threats and open challenges. In: 2021 51st annual IEEE/IFIP international conference on dependable systems and networks workshops (DSN-W). IEEE
10. Alnasser A, Sun H (2019) Global roaming trust-based model for V2X communications. In: IEEE INFOCOM 2019-IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE
11. Hamid B et al (2019) Cyber security issues and challenges for smart cities: a survey. In: 2019 13th international conference on mathematics, actuarial science, computer science and statistics (MACS). IEEE
12. Yin M, Wang Q, Cao M (2019) An attack vector evaluation method for smart city security protection. In: 2019 international conference on wireless and mobile computing, networking and communications (WiMob). IEEE
13. Panta YR et al (2019) Improving accessibility for mobility impaired people in smart city using crowdsourcing. In: 2019 cybersecurity and cyberforensics conference (CCC). IEEE
14. James F (2019) IoT cybersecurity based smart home intrusion prevention system. In: 2019 3rd cyber security in networking conference (CSNet). IEEE
15. Pramod N, Sankaran S (2019) Blockchain based framework for driver profiling in smart cities. In: 2019 IEEE international conference on advanced networks and telecommunications systems (ANTS). IEEE

Efficient Machine Learning Technique for Early Detection of IoT Botnets

Selssabil Medghaghet and Somia Sahraoui

Abstract The Internet of Things (IoT) intends to link billions of intelligent objects of different types with tremendous benefits and advantages. However, the omnipresence of the Internet might make it risky enough when it comes to security and privacy considerations. Indeed, the adversaries can manipulate and spy on the communications. Additionally, the stolen IoT devices can even be used to build an IoT botnet that launches huge Distributed Denial-of-Service (DDoS) attacks. However, a substantial amounts of heterogeneous devices used in the IoT prevent IoT threats from being detected by means of standard rules-based security solutions. In this context, Machine Learning (ML) is proposed as an alternative approach to deal with this problem, enabling the construction of intelligent and effective security models based on empirical information sets. In this paper, we present our model for tackling the problem of botnets based on machine learning. Several machine learning techniques, including KNN, Decision Tree, Logistic-Regression, and BernoulliNB, were utilized to create a model that was trained on the BoT-IoT dataset. According to the data, the Decision Tree and Logistic Regression algorithms were determined to be the most reliable in botnet identification, with 99.99% accuracy and 99.99% ROC AUC for both.

Keywords Internet of Things · Botnet threat · Machine learning

1 Introduction

The Internet of Things is a wide ecosystem where smart objects interact via ubiquitous sensors. Since the beginning of the Internet of Things, users have exponentially connected intelligent gadgets to the Internet and brought all things closer to the future. The smart home industry is projected to reach 137.91 US$ billion by 2023 according to the newest research study and expand at a combined annual growth rate

S. Medghaghet · S. Sahraoui (✉)
Computer Science Department, University of Mohamed Khider, Biskra, Algeria
e-mail: somia.sahraoui@univ-biskra.dz

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_4

(CAGR) of 13% from 2017 to 2023 [1]. It comprises a wide variety of device types from little to huge, from basic to complicated, from consumer devices to advanced systems in military, utilities, and industrial systems. This IoT enables information to be exchanged in a range of application scenarios, each with distinct features and unique performance assurances that provide enormous advantages to people, such as home automation, eco-monitoring, health and lifestyle, smart cities, etc.

According to a 2015 Hewlett Packard research, 80% of the IoT devices studied presented privacy problems, with 60% missing any methods to check the validity of security upgrades or even their integrity, allowing an attacker to alter the firmware without being discovered [2]. Given that IoT devices are built with a variety of inherent limits and weaknesses, it's no wonder that they've been targeted and recruited by botnets. In 2016 Anna Senpai built a malevolent application, dubbed "Mirai," which allows the control and generation of distributed denial of service attacks (DDoS) on susceptible connected items, such as surveillance cameras, and routers [3].

Mirai converts infected items into autonomous and intelligent agents that can be remotely controlled. To address the aforementioned issues, the detection of DDoS attacks using machine learning algorithms has progressively been the focus of study. The machine learning system can detect anomalous information hidden in vast amounts of data. To identify DoS attacks, many detection techniques have been developed. The experiment is conducted using the Bot-Iot data-set, and machine learning techniques such as KNN, Decision Tree, Logistic Regression, and BernoulliNB are used to detect it. We conducted the experiment on a real-time data set as well as a balanced data-set, demonstrating the influence of imbalanced data on machine learning.

2 Background Methodology

2.1 Overview of IoT Botnet

IoT bot refers to a software robot that looks for susceptible devices and transforms them into bots in the same way that regular bots do. It is a malware-extension procedure that is carried out automatically. An IoT botnet is a collection of infected devices that form a network. The IoT botnet is managed by a botmaster who uses these bots to carry out coordinated tasks. DDoS attacks, spamming, phishing campaigns, click fraud, and malware might all be part of the synchronized operation. Due to a lack of basic security, malware infestation, or accepting a malicious email attachment, IoT devices become bots [4].

2.2 Denial of Service (DoS)

The most frequent cyberattack is a DDoS attack, in which the attacker's computers send a massive amount of malicious traffic to the target server all at once, overwhelming the target network [5]. DDoS attacks aim to disrupt the target server's regular operation by flooding it with an enormous traffic, such as false requests, in order to oversaturate its capacity, creating a disruption or denial of service to genuine traffic [6]. DDoS attacks damage server system resources, such as CPU, memory, and may also cause the network bandwidth to saturate to a significant amount of traffic, thus the server will be refused service to genuine computers because the DDoS attack is concerned. To conduct a DDoS attack, hackers are using a botnet. After being infected by malware the attacker distributes via the Internet, IoT devices are involved in DDoS attack. IoT devices infected operate as an asset and the attacker uses them to run the DDoS attacker [7].

3 Machine Learning

Machine learning is a technique for teaching machines how to handle data more effectively. We may be unable to comprehend the pattern or extract information from the data after examining it. In such a situation, we use [8] machine learning. The need for machine learning is increasing as a result of the number of data-sets accessible [9]. Machine learning is valuable because it allows you to continuously learn from data and forecast the future. This sophisticated collection of algorithms and models is being utilized in a variety of sectors to enhance processes and obtain insights into data patterns and anomalies [10]. Advertising, recommendation systems, computer vision, natural language processing, and user behavior analytics are just a few of the applications that employ machine learning (ML) techniques [11].

3.1 Supervised Learning

A machine learning approach called supervised learning is used to learn a function using training data. Pairs of input x objects (usually vectors) and intended outputs y make up the training data. The output of the function f might be a continuous value (called regression) or a prediction of the input object's class label (called classification) [12].

The aim is to find the best predictive model function f^* to minimize the cost function $\mathcal{L}(f(x), y)$ that represents the difference between the estimated output and ground-truth labels. The predictive model function f changes depending on the structure of the model. The domain of the ML model function f is constrained to a collection of functions F with limited model topologies specified by distinct hyperparameter settings. Thus, the best predictive model f^* can be derived by [13]:

$$f^* = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i) \tag{1}$$

In supervised learning algorithms, many alternative loss functions exist, such as the square of Euclidean distance, cross-entropy, information gain, and so on [11].

4 Supervised Machine Learning Algorithms

4.1 Decision Tree

Decision trees are the kind of trees that are classified according to their values. For categorization purposes, the decision tree is utilized primary. There are nodes and branches in each tree. Each node represents the qualities in the classed group, and the value the node can take is represented by each branch [14] .

4.2 Naive Bayes

The Bayesian classification technique is a supervised learning approach as well as a statistical classification method. It assumes an underlying probabilistic model and enables the capture of uncertainty about the model in a logical fashion by estimating probabilities of the outcomes. The Bayesian classification's primary goal is to address prediction issues. This categorization includes realistic learning methods that can integrate observed data. Bayesian classification gives a valuable perspective for understanding and assessing learning algorithms. It computes explicit probabilities for hypotheses and robustly handles noise in input data.Consider a two-valued generic probability distribution $P(x_1, x_2)$. Using the Bayes rule, we derive the following equation without losing generality:

$$P(x_1, x_2) = P(x_1|x_2)P(x_2) \tag{2}$$

Similarly, if another class variable c is present, we get the following equation [15]:

$$P(x_1, x_2|c) = P(x_1|x_2, c)P(x_2|c) \tag{3}$$

If we expand the scenario with two variables to a conditional independence assumption for a set of variables $x_1 \ldots x_N$ depending on another variable c, we get [15]:

$$P(x|c) = \prod_{i=i}^{N} P(x_i|c) \tag{4}$$

4.3 Logistic Regression

This technique in machine learning which is called logistic regression was borrowed from the field of statics, as it is very suitable for technique for binary classification or bi-class problems (problems that have 2 classes of values for classification). The name for logistic regression was brought from a function in mathematics and statics called the logistic function also called the sigmoid function, this logistic function was invented by statisticians for the purpose of describing the properties of growth of populations in ecology, the fast increase and reaching the maximum carying capacity of the envitonment, Logistic function is an shaped as an S-shape curve with is capable of mapping any real-valued number into a value between 0 and 1, and not touching those limits [16].

$$\frac{1}{1 + e^{-value}} \tag{5}$$

where e is representing the natural logarithm, or Eulers number (EXP()) and value is the transformable numerical value, we see bellow a plot 5 that transforms numbers between −8 and 8 into a range of 0–1 using the sigmoid or logistic function.

4.4 K-Nearest Neighbor (kNN)

One of the modest non-parametrical, traditional approaches used to categorize data is K-nearest neighbor (kNN). It calculates the approximate distances between various points on the input vectors and then assigns the unlabeled point to the class of its K-nearest neighbors [17]. Assuming the training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, x_i is the feature vector of an instance, and $y_i \in \{c_1, c_2, \ldots, c_m\}$ is the class of the instance, for a test instance $i = (1, 2, \ldots, n)$, its class y can be denoted by [11]:

$$\arg\max_{cj} \sum_{x_i \in N_K(x)} I(y_i = c_j), i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, m. \tag{6}$$

where I(x) is an indicator function, I = 1 when $y_i = c_j$, otherwise I = 0; $N_K(x)$ is the field involving the k-nearest neighbors of x [11]. During the creation of the kNN classifier, (k) a key parameter and different (k) values may lead to different results [17]. If k is too little, the model is not fit; if k is too big, the model is overfitting and takes a long time to compute [11].

5 Related Works

The research on resolving DDoS botnet attacks in IoT has attracted substantial attention in the literature.

For bifurcating the normal and attacked traffics, Narasimha et al. [18] utilized anomaly detection in combination with machine learning methods. Real-time data sets were employed in the experiment. For classification, the well-known naive Bayes ML method was employed. The outcomes were compared to those of other algorithms such as J48 and random forest (RF).

In [19], Liang et al. investigated the security models for the Internet of Things (IoT) utilizing machine learning approaches. They researched and mostly reviewed approaches for IoT security based on ML techniques.

To identify the DDoS attack, Brun et al. [20] worked in the field of Internet of Things (IoT). For network identification, the author used one of the most well-known deep learning techniques, the random neural network (RNN) technique. In comparison to previous approaches, this deep-learning-based technology effectively delivers more promising outcomes.

The autoencoder was used by Yang et al. [21] to identify DDoS attacks. A multilayer neural network with an unsupervised training algorithm is known as an autoencoder. During the training process, it filters out less relevant information and background noise, leaving just the most important information.

A powerful malware detection method based on deep learning has been proposed by Hemalatha et al. [22]. In the final classification layer of the DenseNet model, the system utilizes a weighted class balanced loss function to produce significant improvement in malware classification by managing the unbalanced data concerns.Comprehensive studies on four benchmark malware databases have shown that the system presented has greater detection performance with reduced computing costs.

6 Scope of Our Proposal

In this part, we discuss the processes used to create the botnet detection model, as well as a methodology for botnet detection that includes data preprocessing and the SMOTE approach to balance the data-set class. In addition, feature engineering was carried out after evaluating machine learning methods for categorization. We discussed the influence of unbalanced data and the number of features used on machine learning (Fig. 1).

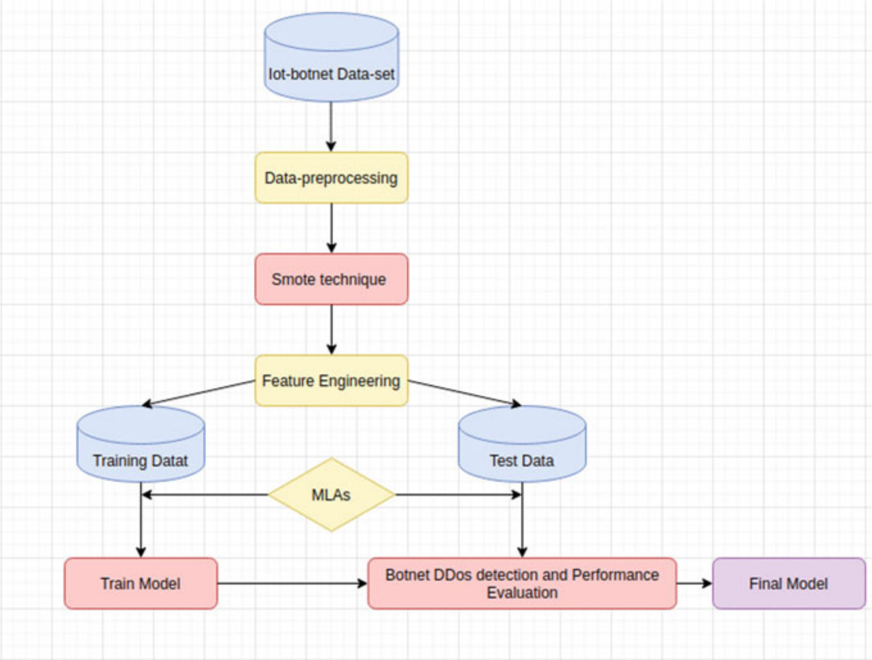


Fig. 1 Proposed model

6.1 BoT-IoT Dataset

The Bot-IoT dataset [23] was used in this research to choose an effective machine learning (ML) method for detecting DDos botnet threats in the internet of things (IoT). The BoT-IoT dataset was generated in the UNSW Canberra Cyber Range Lab by creating a realistic network environment. There was a mix of normal and botnet traffic in the network environment. The source files for the dataset are available in a variety of forms, including the original pcap files, produced argus files, and csv files. To aid with the tagging procedure, the files were divided by attack category and subcategory [24]. The collected pcap files total 69.3 GB in size and include around 72.000.000 records. The retrieved flow traffic is 16.7 GB in csv format. The dataset contains DDoS, DoS, OS and Service Scan, Keylogging, and Data exfiltration attacks, with the DDoS and DoS incidents further grouped depending on the protocol used [24]. Different supervised MLAs (Decision Tree, Naive Bayes, Logistic Regression, and KNN) were applied on different combinations of the Botnet dataset and the results were benchmarked to determine the optimal method for our model. The dataset we utilized is 92.3 MB in size, with almost 99% botnet activity and less than 1% regular traffic. To compare the balanced and unbalanced datasets, we produced a new dataset after processing real-time BoT-IoT datasets using the SMOTE approach, which delivered a class balance dataset with an equal quantity of botnet and regular traffic. Then we choose a variety of features to compare our model's performance to theirs.

6.2 Data Preprocessing

After the data has been collected, it must be preprocessed to get it into a more refined state. The generalization performance of a supervised ML algorithm is frequently influenced by data preparation. One of the most challenging issues in inductive ML is the removal of noisy occurrences [25].

- **Data Cleaning**: Data cleaning, also known as data correction, refers to procedures that rectify faulty data, filter some erroneous data out of a data set, and remove superfluous data information. We go through a data cleaning procedure to discover missing values and remove the rows that have them. Using the pandas dropna() method, we remove the rows in the BoT-IoT dataset that have a null value [26].
- **Normalization**: The features are "scaled down" by normalization. Before beginning the learning process, it's critical to maintain a consistent distribution of each attribute's values. Because some features in our BoT-IoT dataset include data with a wide range of values, learning the model becomes more difficult and takes longer. The MinMax technique is used for this purpose [27]. In MinMax the values of features are scaled to the interval [0,1] as follows:

$$Y_{norm} = (Y - Y_{in})/(Y_{max} - Ymin) \tag{7}$$

  We get Ymin and Ymax by using .min() and .max() functions of pandas.
- **Transformation**: Data transformation is the process of transforming data from one format or structure to another in computing. There are numerous categorical characteristics in the BoT-IoT dataset that contain non-numeric data that required to be transformed into a numeric format for the MLAs to analyze because the MLAs we were using were in algebraic format. In the Bot-Iot dataset, we categorize fields into two types: text fields and numerical fields. We change the representation for the text field's content.

6.3 Feature Engineering

Feature engineering may be a crucial element in the machine learning process, as it enables our models to improve significantly by reducing their dimension and therefore minimizing the problem of overfitting. it Finds the most relevant information about the target variable in choosing suitable characteristics The is suitable [28].

Extra Trees Classification is used as a whole for the data set. After training, the Extra Trees Classifier gives each feature a feature significance score. This is allocated to the majority vote according to the number of times it was picked as the best division for a decision tree [29].

6.4 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is a technique for oversampling. The interpolation method is used in this strategy to increase the quantity of new minority class instances. Before forming new minority class instances, the minority class instances that lie together are recognized [30]. Because this technique may produce synthetic examples rather than duplicate minority class instances, it avoids the over-fitting problem.

6.5 Experimental Scenario

We utilized Google Colab for model building and testing. Google Colaboratory, or "Colab" for short, is a Google Research tool. Colab enables anybody to type and run arbitrary python code over the web and is particularly suitable for computer training, data analysis and training.

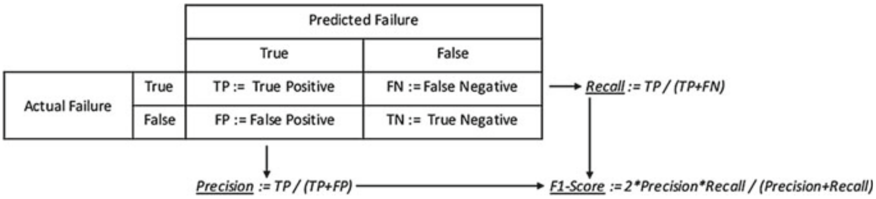
Colab is a hosted Jupyter Notebook Service which requires no configuration to be used while free usage of GPUs in computer resources. Google Colab offers 12 GB of RAM and 107 GB of python 3.7 disk space [31]. The model is trained using 92.3 MB of training datasets using the KNN, Decision Tree, Logistic Regression, and BernoulliNB algorithms, and the accuracy is benchmarked to determine the optimum approach for our detection system.

The Bot-IoT dataset we utilized is unbalanced; botnet activity makes up more than 99% of the data, while regular traffic makes up less than 1%. We utilized the Synthetic Minority over-sampling technique, better known as SMOTE, to create our BoT-IoT dataset. MLAs were applied to both imbalanced and balanced datasets. We choose several features to tackle the overfitting and compare the performance with these features of our model. We utilize several python libraries to handle the data set and execute machine learning. Sklearn, numpy, matplotlib, and pandas have mostly been utilized.

6.6 *Machine Learning Model Evaluation and Cross Validation*

- **Precision**: Precision refers to the ability to anticipate something with a high degree of accuracy. It's a measure of how many real positives the model claims against how many positives it claims [32].
- **Recall**: The actual positive rate, or recall, is the difference between the number of positives in the model claims and the total number of positives in the data [32].
- **Accuracy**: It is defined as the system's ability to accurately identify an attack packet as an "attack packet" and a regular packet as a "normal packet." It indicates the proportion of valid predictions in relation to all samples [33].
  Accuracy = ((TPs + TNs)/(TPs + TNs + FPs + FNs) * 100)
- **F1-Score**: The F-Score metric combines precision and recall into a single score value that balances both accuracy and memory issues. When false positives and false negatives are significant, the F1-Score is examined [7].
- **ROC AUC**: The Area Under the Curve (AUC)—ROC curve is a performance assessment for classification issues at various threshold levels. AUC indicates the degree or measure of separability, while ROC is a probability curve. It indicates how well the model can discriminate between classes. The higher the AUC, the more accurate the model is in classifying 0 classes like 0 and 1 classes as 1 [34]. With 733,598 Botnet traffic and 107 Normal traffic in 733,705 total data, the BoT-IoT actual dataset is significantly imbalanced. Even if the classifier correctly predicts each transmission as botnet traffic, it will still be more than 90% accurate. Thus, a study of the ROC AUC curve is preferred for efficient model assessment. The ROC curve is drawn against the FPR with TPR on the y-axis and x-axis with FPR.

We utilized k-Fold Cross Validation to regulate the model's performance, avoid overfitting, and get a generalizable estimate of the model's quality. Cross-validation is a technique for examining separate data sets that are utilized to properly forecast the predictive model; K-Fold = 5 was employed in the cross-validation (Fig. 2).



		Predicted Failure	
		True	False
Actual Failure	True	TP := True Positive	FN := False Negative
	False	FP := False Positive	TN := True Negative

Recall := TP / (TP+FN)
Precision := TP / (TP+FP)
F1-Score := 2*Precision*Recall / (Precision+Recall)

Fig. 2 Confusion matrix, which shows precision calculation, recall, and F1 score

7 Experiment and Discussion

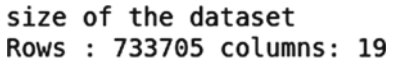
We have deployed classifier algorithms in two BoT-IoT data-sets to evaluate MLAs' performance by creating actual data data-sets that are "imbalanced" and balance data sets that are produced using real data sets using SMOTE technology. The data-set consists of two components, regular traffic and botnet DDoS traffic. The data-sets of 19 columns and 733,705 rows are shown in Fig. 3. Figure 4 indicates a significant imbalance of the data set. The preprocessing of data was then done in order to obtain accurate information.

To do this, I use pandas' dropna() method to remove the rows with null values.

7.1 *Transformation*

There are numerous categorical features in the BoT-IoT data-set that comprise non-numeric data. We assign numeric values to protocol names and IP addresses. We give numbers ranging from 0 to 6 to subcategory features. Figure5 depict data prior to data transformation and normalization.

Figure 6 shows first 7 rows of data-set after transformation and Normalization.



```
size of the dataset
Rows : 733705 columns: 19
```

Fig. 3 Size of data-set

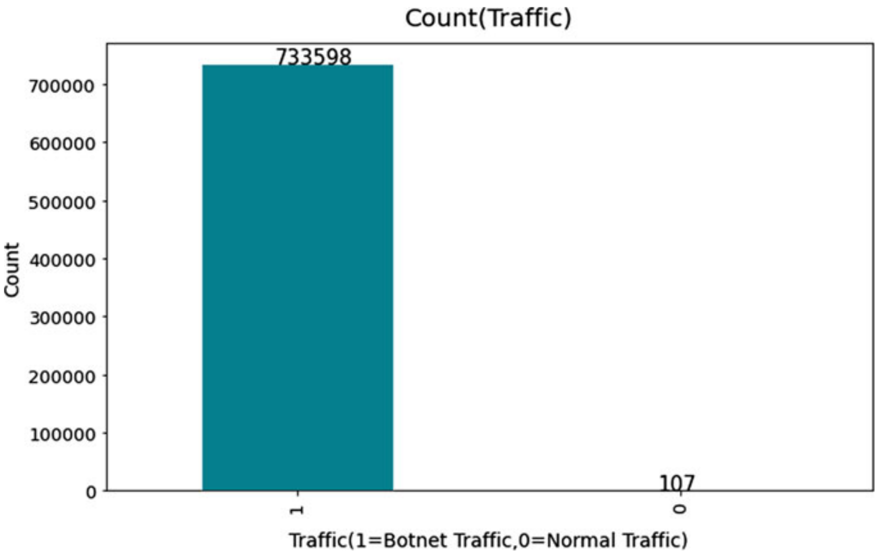
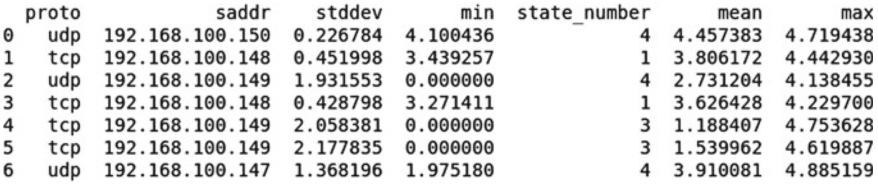
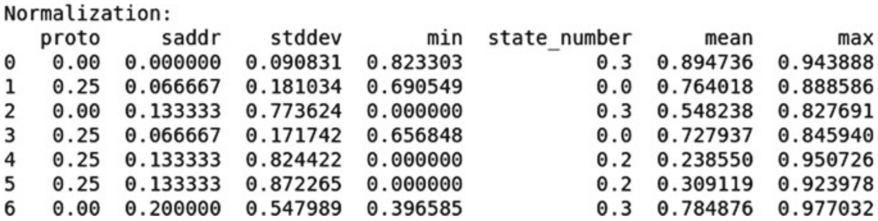


Fig. 4 Bar graph illustrating imbalance of class in the data set



```
  proto            saddr    stddev       min  state_number      mean       max
0   udp  192.168.100.150  0.226784  4.100436             4  4.457383  4.719438
1   tcp  192.168.100.148  0.451998  3.439257             1  3.806172  4.442930
2   udp  192.168.100.149  1.931553  0.000000             4  2.731204  4.138455
3   tcp  192.168.100.148  0.428798  3.271411             1  3.626428  4.229700
4   tcp  192.168.100.149  2.058381  0.000000             3  1.188407  4.753628
5   tcp  192.168.100.149  2.177835  0.000000             3  1.539962  4.619887
6   udp  192.168.100.147  1.368196  1.975180             4  3.910081  4.885159
```

Fig. 5 First 7 rows of data-set before transformation and normalization



```
Normalization:
   proto     saddr    stddev       min  state_number      mean       max
0   0.00  0.000000  0.090831  0.823303           0.3  0.894736  0.943888
1   0.25  0.066667  0.181034  0.690549           0.0  0.764018  0.888586
2   0.00  0.133333  0.773624  0.000000           0.3  0.548238  0.827691
3   0.25  0.066667  0.171742  0.656848           0.0  0.727937  0.845940
4   0.25  0.133333  0.824422  0.000000           0.2  0.238550  0.950726
5   0.25  0.133333  0.872265  0.000000           0.2  0.309119  0.923978
6   0.00  0.200000  0.547989  0.396585           0.3  0.784876  0.977032
```

Fig. 6 First 7 rows of data-set after transformation and normalization

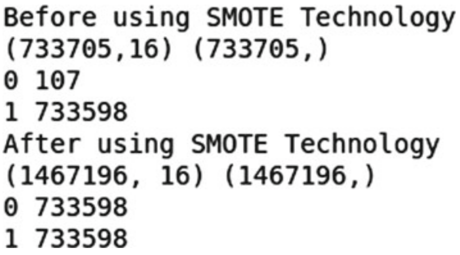
7.2 Oversampling Minority Class

Figure 7 demonstrates the application of the SMOTE technique for the sampling of minority classes. The BoT-IoT Real-Time dataset comprises 733,705 data, 107 of which is normal traffic and 733,598 is Botnet. We obtain 146,796 botnet data after processing data using SMOTE technology, i.e. 733,598 for normal traffic. As a result, the data-set class is balanced.

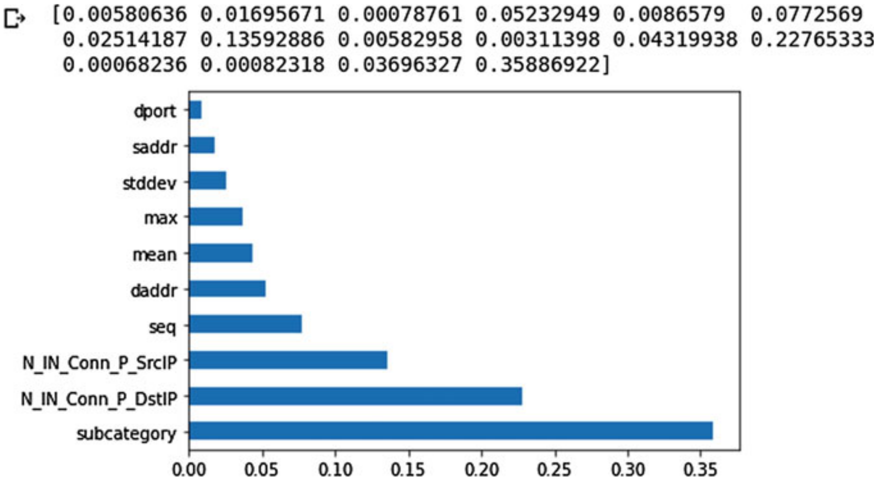
7.3 Feature Score

After analysis of the data-set, we selected relevant features using Feature Engineering. The 10 main characteristic based on ExtraTreesClassifier appears in Fig. 8.

Fig. 7 Size of dataset before and after SMOTE technique



```
Before using SMOTE Technology
(733705,16) (733705,)
0 107
1 733598
After using SMOTE Technology
(1467196, 16) (1467196,)
0 733598
1 733598
```



```
[0.00580636 0.01695671 0.00078761 0.05232949 0.0086579  0.0772569
 0.02514187 0.13592886 0.00582958 0.00311398 0.04319938 0.22765333
 0.00068236 0.00082318 0.03696327 0.35886922]
```

Fig. 8 Top 10 respective feature score

7.4 Train-Test Split

Data-set were divided into train and test to assess the model's performance. 70% of data was utilized for training and 30% for testing (Fig. 9).

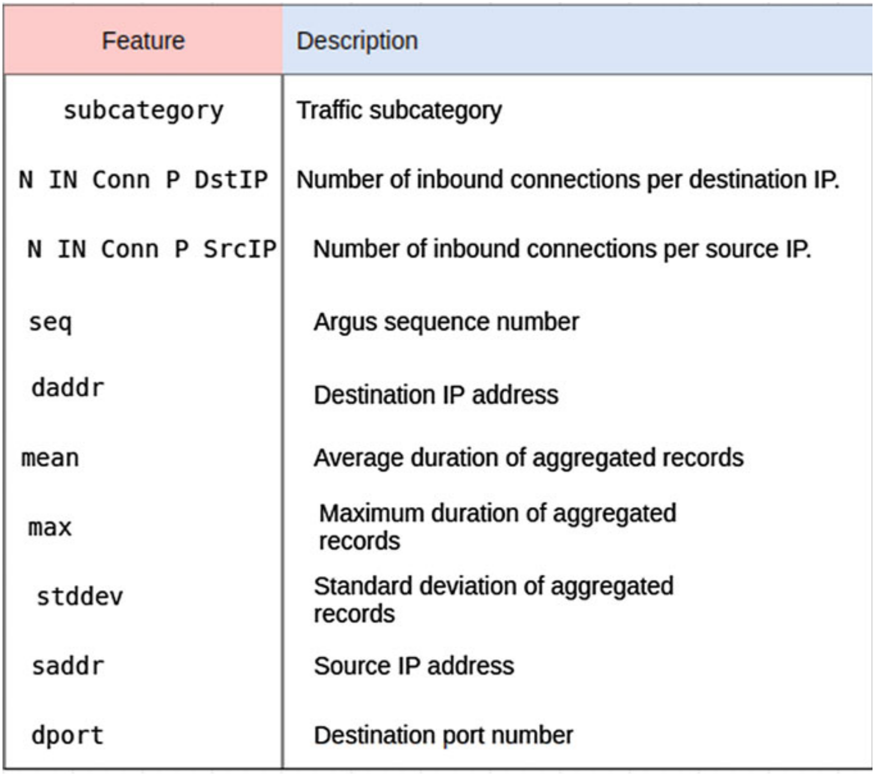
7.5 Comparison of Performance of Machine Learning Algorithms

- **Results with Bernoulli Naive Bayes**: As shown in Fig. 10 for the real BoT-IoT data-set, we achieved 99.98% accuracy using the BernoulliNB method, however,it shows a 50% ROC-AUC (Fig. 11), and it shows relatively poor recall and f1-score values. This clearly shown that precision isn't always beneficial when dealing with unbalanced data. Because this data-set contains more than 99% botnet traffic and fewer than 1% regular traffic, the classifier may categorize all samples as 1. As a result, we have high accuracy but a poor ROC AUC.

We achieved improved accuracy, recall, f1-score, and ROC AUC after adding SMOTE technology, as shown in Fig. 12. This means that the BernoulliNB algorithm is successful in distinguishing botnet and regular traffic after using the smote method.

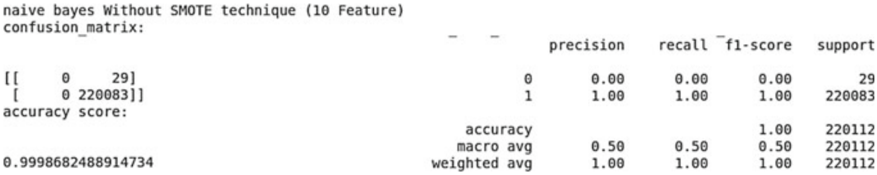
Figure 13 shows BernoulliNB model performance on realtime Class balance dataset using top 8 feature the accuracy and ROC_AUC get better then using top 10 feature 92.65%, 92.64% respectively.

- **KNN's outcomes**: On the imbalanced dataset, Fig. 14 demonstrates that the KNN model performs well with a 99.98% accuracy but low precision and recall (F1-



Feature	Description
subcategory	Traffic subcategory
N IN Conn P DstIP	Number of inbound connections per destination IP.
N IN Conn P SrcIP	Number of inbound connections per source IP.
seq	Argus sequence number
daddr	Destination IP address
mean	Average duration of aggregated records
max	Maximum duration of aggregated records
stddev	Standard deviation of aggregated records
saddr	Source IP address
dport	Destination port number

Fig. 9 Feature description

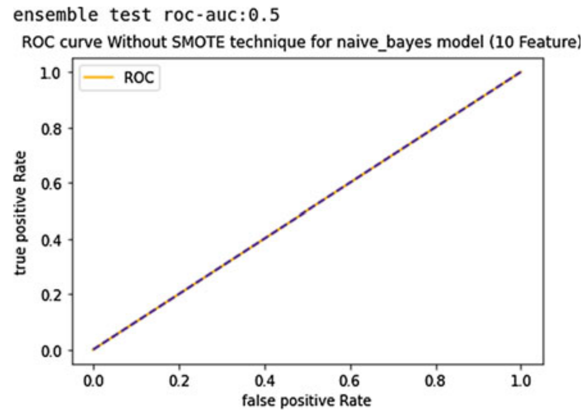


```
naive bayes Without SMOTE technique (10 Feature)
confusion_matrix:
                                                   precision    recall  f1-score   support

[[     0     29]                              0       0.00      0.00      0.00        29
 [     0 220083]]                             1       1.00      1.00      1.00    220083
accuracy score:
                                       accuracy                           1.00    220112
                                      macro avg       0.50      0.50      0.50    220112
0.9998682488914734                 weighted avg       1.00      1.00      1.00    220112
```

Fig. 10 Gaussian BernoulliNB model performance without using SMOTE technique

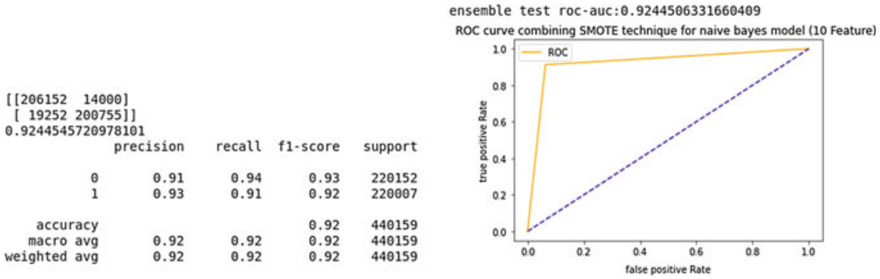
score). As can be seen in the confusion matrix, there are only 13 true positive values, and we only acquire 72.4% ROC AUC, indicating that accuracy percentage is insufficient to verify model performance.

Even though the accuracy score dropped a little to 99.92% after we utilized SMOTE technology to balance our dataset, the ROC AUC and F1-score, recall, and precision rose (Fig. 15), indicating that our model performance improved.



```
ensemble test roc-auc:0.5
```

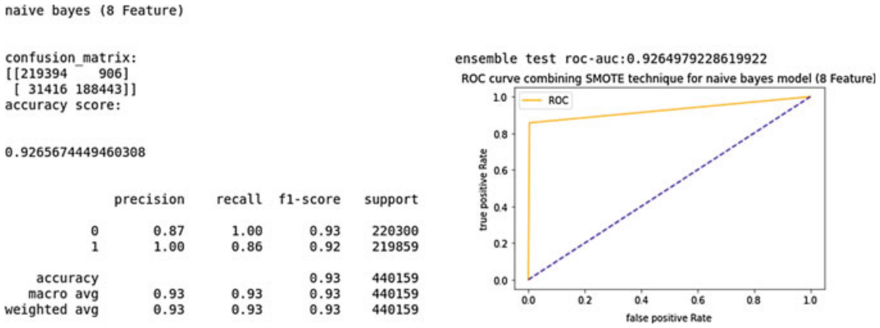
Fig. 11 ROC AUC graph from BernoulliNB model without using SMOTE technique



```
[[206152  14000]
 [ 19252 200755]]
0.9244545720978101
              precision    recall  f1-score   support

           0       0.91      0.94      0.93    220152
           1       0.93      0.91      0.92    220007

    accuracy                           0.92    440159
   macro avg       0.92      0.92      0.92    440159
weighted avg       0.92      0.92      0.92    440159
```

```
ensemble test roc-auc:0.9244506331660409
```

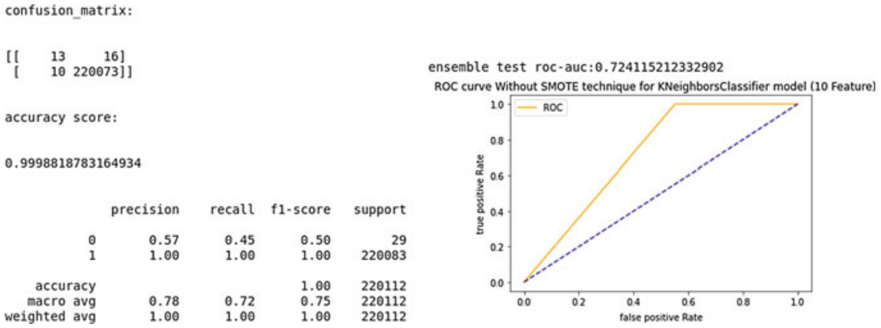
Fig. 12 Gaussian BernoulliNB model performance and ROC AUC graph using SMOTE technique



```
naive bayes (8 Feature)


confusion_matrix:
[[219394    906]
 [ 31416 188443]]
accuracy score:


0.9265674449460308


              precision    recall  f1-score   support

           0       0.87      1.00      0.93    220300
           1       1.00      0.86      0.92    219859

    accuracy                           0.93    440159
   macro avg       0.93      0.93      0.93    440159
weighted avg       0.93      0.93      0.93    440159
```

```
ensemble test roc-auc:0.9264979228619922
```

Fig. 13 Gaussian BernoulliNB model performance and ROC AUC graph (8 feature)

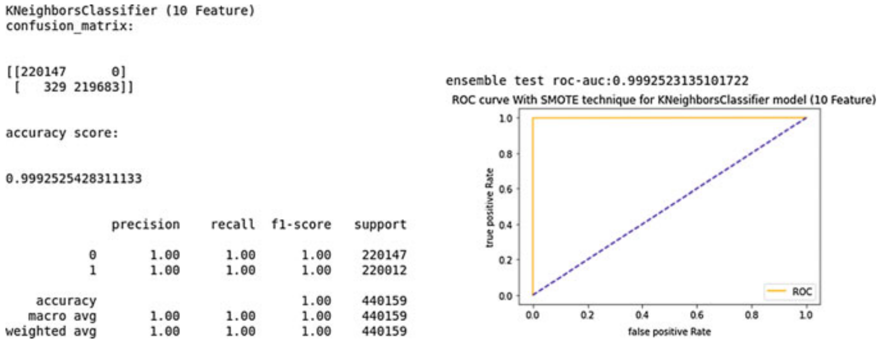


```
confusion_matrix:

[[     13      16]
 [     10 220073]]

accuracy score:

0.9998818783164934

              precision    recall  f1-score   support

           0       0.57      0.45      0.50        29
           1       1.00      1.00      1.00    220083

    accuracy                           1.00    220112
   macro avg       0.78      0.72      0.75    220112
weighted avg       1.00      1.00      1.00    220112
```

```
ensemble test roc-auc:0.724115212332902
```

ROC curve Without SMOTE technique for KNeighborsClassifier model (10 Feature)
ROC
1.0
0.8
0.6
0.4
0.2
0.0
true positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
false positive Rate

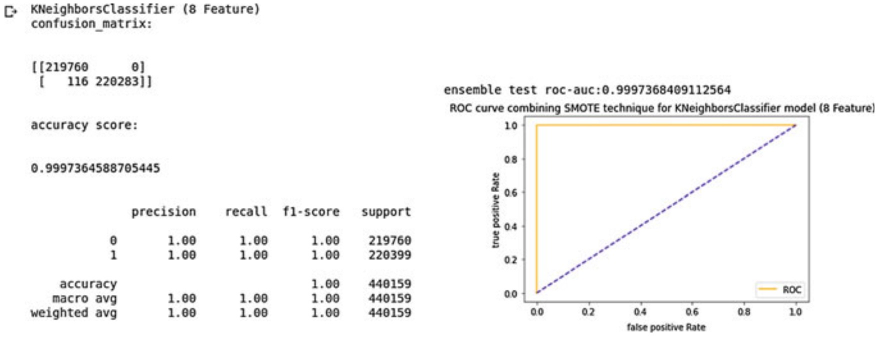
Fig. 14 KNN model performance and ROC AUC graph without using SMOTE technique



```
KNeighborsClassifier (10 Feature)
confusion_matrix:

[[220147      0]
 [   329 219683]]

accuracy score:

0.9992525428311133

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    220147
           1       1.00      1.00      1.00    220012

    accuracy                           1.00    440159
   macro avg       1.00      1.00      1.00    440159
weighted avg       1.00      1.00      1.00    440159
```

```
ensemble test roc-auc:0.9992523135101722
```

Fig. 15 KNN model performance and ROC AUC graph using SMOTE technique



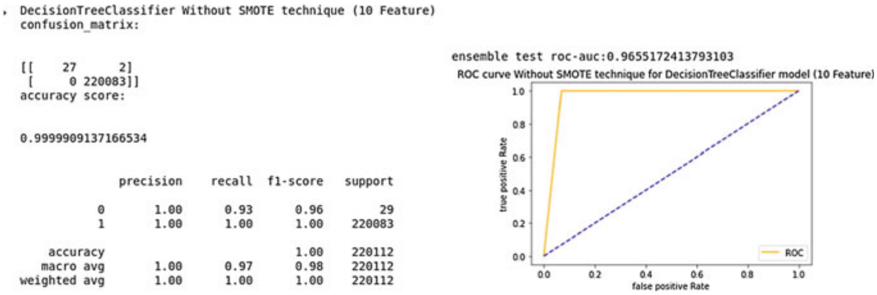
```
KNeighborsClassifier (8 Feature)
confusion_matrix:

[[219760      0]
 [   116 220283]]

accuracy score:

0.9997364588705445

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    219760
           1       1.00      1.00      1.00    220399

    accuracy                           1.00    440159
   macro avg       1.00      1.00      1.00    440159
weighted avg       1.00      1.00      1.00    440159
```

```
ensemble test roc-auc:0.9997368409112564
```

ROC curve combining SMOTE technique for KNeighborsClassifier model (8 Feature)
1.0
0.8
0.6
0.4
0.2
0.0
true positive Rate
ROC
0.0
0.2
0.4
0.6
0.8
1.0
false positive Rate

Fig. 16 KNN model performance and ROC AUC graph (8 feature)

Figure 16 demonstrates how the accuracy and ROC AUC of the KNN model on a real-time Class balance data-set improve when the top 8 features are used instead of the top 10. The accuracy and ROC AUC of the top 8 features are 99.97% and 92.97%, respectively.

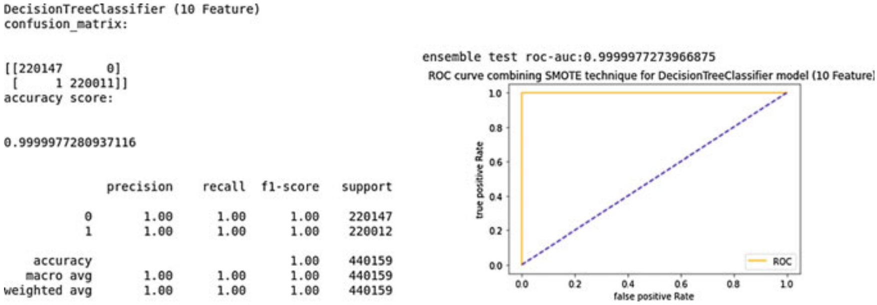


```
, DecisionTreeClassifier Without SMOTE technique (10 Feature)
  confusion_matrix:

  [[    27      2]
   [     0 220083]]
  accuracy score:

  0.9999909137166534

                 precision    recall  f1-score   support

             0       1.00      0.93      0.96        29
             1       1.00      1.00      1.00    220083

      accuracy                           1.00    220112
     macro avg       1.00      0.97      0.98    220112
  weighted avg       1.00      1.00      1.00    220112
```

```
ensemble test roc-auc:0.9655172413793103
```

ROC curve Without SMOTE technique for DecisionTreeClassifier model (10 Feature)
true positive Rate
false positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC

Fig. 17 DecisionTreeClassifier model performance and ROC AUC graph without using SMOTE technique



```
DecisionTreeClassifier (10 Feature)
confusion_matrix:

[[220147      0]
 [     1 220011]]
accuracy score:

0.9999977280937116

               precision    recall  f1-score   support

           0       1.00      1.00      1.00    220147
           1       1.00      1.00      1.00    220012

    accuracy                           1.00    440159
   macro avg       1.00      1.00      1.00    440159
weighted avg       1.00      1.00      1.00    440159
```

```
ensemble test roc-auc:0.9999977273966875
```

ROC curve combining SMOTE technique for DecisionTreeClassifier model (10 Feature)
true positive Rate
false positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC

Fig. 18 DecisionTreeClassifier model performance and ROC AUC graph using SMOTE technique

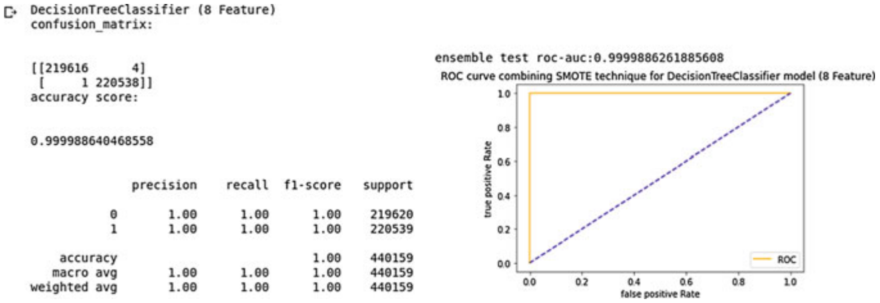
- **Results with DecisionTreeClassifier**: Figure 17 depicts the DecisionTreeClassifier model performance on the unbalanced dataset. We obtained high accuracy (99.99%), precision, recall, and F1-score, as well as 96.55% in our ROC AUC.

We achieved the greatest results in all of the algorithms we tried after using SMOTE technology, with 99.99% accuracy (Fig. 18), recall, F1-score, and precision all being 1, with 99.99% ROC AUC (Fig. 18).

Figure 19 depicts the performance of the DecisionTreeClassifier model on a real-time Class balance data-set using the top 8 features. We obtained the same accuracy and ROC AUC as when we utilized the top 10 features, 99.99%, and 92.99%, respectively.

- **Results with LogisticRegression**: On the unbalanced data-set, we obtained high accuracy, precision, and recall, as well as a decent F1-score, but we only got 93.1% in our ROC AUC, as seen in Fig. 20.

We achieved improved results after using SMOTE technology, with accuracy of 99.99% (Fig. 21), recall and F1-score of 1, and precision of 1 with 99.99% ROC AUC.

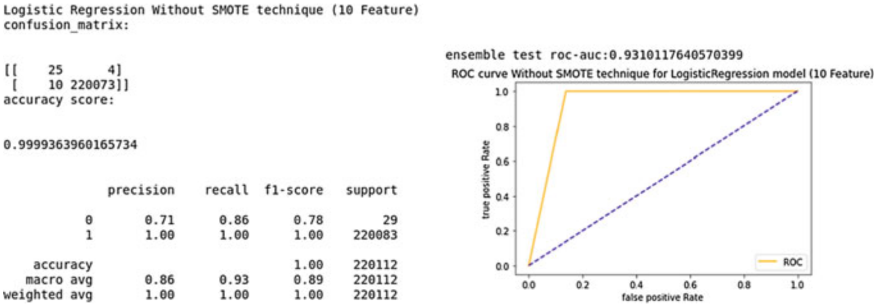


```
DecisionTreeClassifier (8 Feature)
confusion_matrix:


[[219616      4]
 [     1 220538]]
accuracy score:


0.999988640468558


              precision    recall  f1-score   support

           0       1.00      1.00      1.00    219620
           1       1.00      1.00      1.00    220539

    accuracy                           1.00    440159
   macro avg       1.00      1.00      1.00    440159
weighted avg       1.00      1.00      1.00    440159
```

```
ensemble test roc-auc:0.9999886261885608
```

ROC curve combining SMOTE technique for DecisionTreeClassifier model (8 Feature)
true positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
false positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC

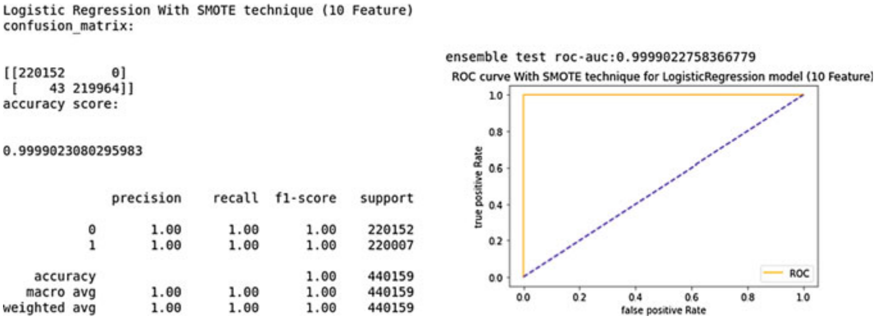
Fig. 19 DecisionTreeClassifier model performance and ROC AUC graph (8 feature)



```
Logistic Regression Without SMOTE technique (10 Feature)
confusion_matrix:


[[    25      4]
 [    10 220073]]
accuracy score:


0.9999363960165734


              precision    recall  f1-score   support

           0       0.71      0.86      0.78        29
           1       1.00      1.00      1.00    220083

    accuracy                           1.00    220112
   macro avg       0.86      0.93      0.89    220112
weighted avg       1.00      1.00      1.00    220112
```

```
ensemble test roc-auc:0.9310117640570399
```

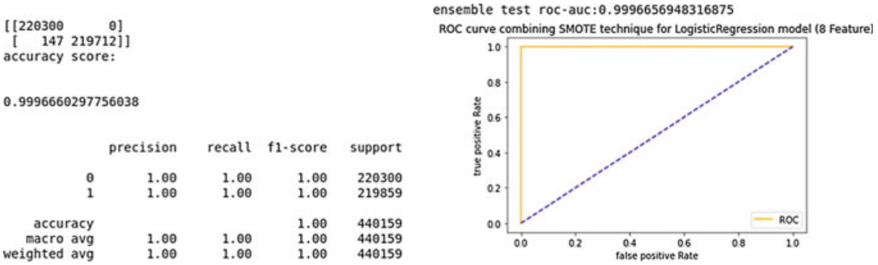
Fig. 20 LogisticRegression model performance and ROC AUC graph without using SMOTE technique



```
Logistic Regression With SMOTE technique (10 Feature)
confusion_matrix:


[[220152      0]
 [    43 219964]]
accuracy score:


0.9999023080295983


              precision    recall  f1-score   support

           0       1.00      1.00      1.00    220152
           1       1.00      1.00      1.00    220007

    accuracy                           1.00    440159
   macro avg       1.00      1.00      1.00    440159
weighted avg       1.00      1.00      1.00    440159
```

```
ensemble test roc-auc:0.9999022758366779
```

Fig. 21 LogisticRegression model performance and ROC AUC graph using SMOTE technique



```
[[220300      0]
 [   147 219712]]
accuracy score:


0.9996660297756038


              precision    recall  f1-score   support

           0       1.00      1.00      1.00    220300
           1       1.00      1.00      1.00    219859

    accuracy                           1.00    440159
   macro avg       1.00      1.00      1.00    440159
weighted avg       1.00      1.00      1.00    440159
```

```
ensemble test roc-auc:0.9996656948316875
```

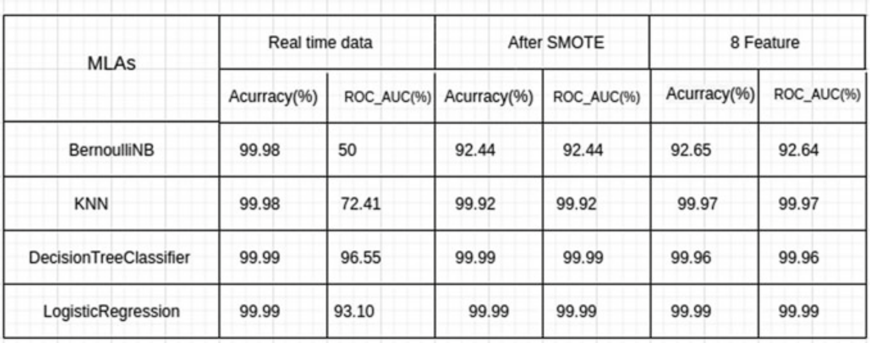
ROC curve combining SMOTE technique for LogisticRegression model (8 Feature)
true positive Rate
false positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC

Fig. 22 LogisticRegression model performance and ROC AUC graph (8 feature)

Figure 22 depicts the performance of the LogisticRegression model on a real-time Class balance data-set using the top 8 features. The accuracy and ROC AUC are lower when using the top 10 features, 99.96% and 99.96%, respectively.

7.6 Observations

Figure 23 depicts the outcomes of several MLAs on the following datasets: imbalance data-set, balance data-set, and balance data-set utilizing 8 top features. Even though we obtained excellent accuracy in our real data-set, the ROC AUC in BernouliNB and KNN was poor. This clearly shown that accuracy is ineffective when dealing with imbalanced data. Our model performance improved after using SMOTE technique. In most algorithms, except DecisionTreeClassifier, adding a different number of features did not result in a significant difference; however, when we used eight features, our accuracy improved somewhat. In our three cases, DecisionTreeClassifier and LogisticRegression were the best algorithms to utilize in botnet detection



MLAs	Real time data		After SMOTE		8 Feature	
	Acurracy(%)	ROC_AUC(%)	Acurracy(%)	ROC_AUC(%)	Acurracy(%)	ROC_AUC(%)
BernoulliNB	99.98	50	92.44	92.44	92.65	92.64
KNN	99.98	72.41	99.92	99.92	99.97	99.97
DecisionTreeClassifier	99.99	96.55	99.99	99.99	99.96	99.96
LogisticRegression	99.99	93.10	99.99	99.99	99.99	99.99

Fig. 23 Comparison of MLAs performance

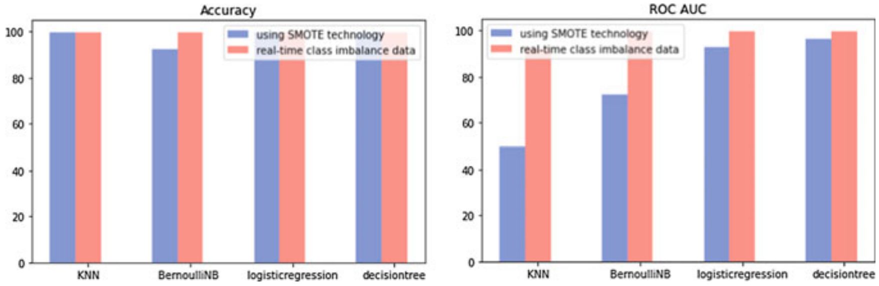


Fig. 24 Graphical presentation of machine learning performance comparison (using SMOTE technique versus without it)

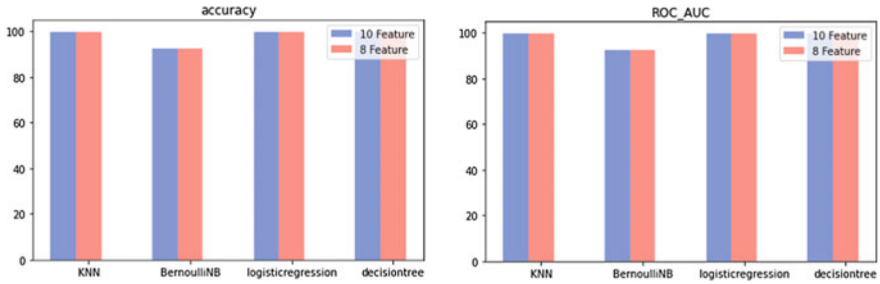


Fig. 25 Graphical presentation of machine learning performance comparison (8 feature versus 10 feature)

systems, providing high accuracy and ROC AUC. BernoulliNB was the method that performed worst.

Figure 24 demonstrates the machine learning algorithms performance comparison between the use of SMOTE and real data-set techniques.

Figure 25 compares the performance of machine learning algorithms with 8 and 10 features.

8 Conclusion

The vast number of IoT devices on the market today is often more vulnerable than traditional desktop PCs. As a result, they put information systems at risk of security breaches. One of the risks that might affect IoT devices is the IoT botnet. This paper provides a method for identifying botnets on IoT devices that employ KNN, Decision Tree, Logistic Regression, and BernoulliNB, all of which were trained using BoT-IoT data-set. We experimented on a real-time data set and a balanced data-set, and the results of the imbalance data and its impact on machine learning were given. We achieved more consistent accuracy and ROC-AUC after integrating

SMOTE technology with a similar range of precision, recall, and f1-score values. This demonstrates that by implementing SMOTE technology, we can improve the model's performance. According to the results, the Decision Tree and Logistic Regression algorithms are the most reliable in botnet identification.

References

1. Meng Y, Zhang W, Zhu H, Shen XS (2018) Securing consumer IoT in the smart home: architecture, challenges, and countermeasures. Wired [Online], Dec 2018. Available: https://www.wired.com/2017/03/cia-can-hack-phone-pc-tv-says-wikileaks/
2. Kolias C, Kambourakis G, Stavrou A, Voas J (2017) DDoS in the IoT: Mirai and other botnets [Online]. Available: https://www.computer.org/
3. Harbi Y (2021) Security in Internet of Things. Ph.D. dissertation, Department of Computer Science, Ferhat Abbas University, Setif
4. Dange S, Chatterjee M (2019) IoT botnet: the largest threat to the IoT network. In: Data communication and networks, advances in intelligent systems and computing, Singapore, p 142
5. Kansal V, Dave M (2017) DDoS attack isolation using moving target defense
6. Dibaei M, Zheng X, Jiang K, Maric S, Abbas R, Liu S, Zhang Y, Deng Y, Wen S, Zhang J, Xiang Y, Yu S (2019) An overview of attacks and defences on intelligent connected vehicles
7. Pokhrel S, Abbas R, Aryal B (2021) IoT security: botnet detection in IoT using machine learning. Macquarie University, Sydney
8. Richert W, Coelho LP. Building machine learning systems with python. Packt Publishing Ltd. ISBN 978-1-78216-140-0
9. Dey A (2016) Machine learning algorithms: a review. Int J Comput Sci Inf Technol (IJCSIT) 7(3):1174–1179
10. John Wiley, Sons Inc (2018) Machine learning for dummies [Online]. Available: https://www.wiley.com/WileyCDA/Section/id-819533.html
11. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. Elsevier [Online]. Available: https://www.elsevier.com/en-xm
12. Ayodele TO (2010) Machine learning overview. In: Zhang Y (ed) New advances in machine learning. InTech. ISBN: 978-953-307-034-6. Available from: http://www.intechopen.com/books/new-advances-in-machine-learning/machine-learning-overview
13. Gambella C, Ghaddar B, Naoum-Sawaya J (2019) Optimization models for machine learning: a survey, pp 1–40. http://arxiv.org/abs/1901.05331
14. Kotsiantis SB (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268
15. Nasteski V (2017) An overview of the supervised machine learning methods [Online]
16. Brownlee J (2017) Master machine learning algorithms. Australia
17. Omar S, Ngadi A, Jebur HH (2013) Machine learning techniques for anomaly detection: an overview. citeseerx 79(2). [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.6779&rep=rep1&type=pdf
18. Narasimha Mallikarjunan K, Bhuvaneshwaran A, Sundarakantham K, Mercy Shalinie S (2019) DDAM, detecting DDoS attacks using machine learning approach
19. Xiao L, Wan X, Lu X, Zhang Y, Wu D (2018) IoT security techniques based on machine learning
20. Brun O, Yin Y, Gelenbe E (2018) Deep learning with dense random neural network for detecting attacks against IoT-connected home environments
21. Yang K, Zhang J, Xu Y, Chao J (2020) DDoS attack detection with AutoEncoder. In: IEEE/IFIP operations and management symposium. IEEE, Piscataway Township, NJ, pp 1–9

22. Hemalatha J, Roseline SA, Geetha S, Kadry S, Damaševičius R (2021) An efficient DenseNet-based deep learning model for malware detection. Entropy 23:344
23. Cloudstor (2021) [Online]. Available: https://cloudstor.aarnet.edu.au/plus/s/umT99TnxvbpkkoE?path=%2FCSV
24. Koroniotis N, Moustafa N (2021) unsw. Available: https://research.unsw.edu.au/projects/bot-iot-dataset
25. Kotsiantis SB, Kanellopoulos D, Pintelas PE (2006) Data preprocessing for supervised learning. researchgate [Online] 1(1):1306–4428. Available: https://www.researchgate.net/
26. García S, Luengo J, Herrera F (2015) Data preprocessing in data mining [Online]. Available: https://www.springer.com/series/8578
27. Idhammad M, Afdel K, Belouch M (2018) Semi-supervised machine learning approach for DDoS detection [Online]. Available: https://www.springer.com/gp
28. Jordon J (2021) Feature selection for a machine learning model. Available: https://www.jeremyjordan.me/feature-selection/
29. Biswas S, Chakrabarty N (2020) Navo minority over-sampling technique (NMOTe): a consistent performance booster on imbalanced datasets
30. Jeatrakul P, Wong KW, Fung CC (2010) Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In: Lecture notes in computer science. Verlag, Berlin Heidelberg, p 155
31. Google (2021) Colaboratory. Available: https://research.google.com/colaboratory/faq.html#resource-limits
32. Hasan M, Islam MM, Zarif MII, Hashem MMA (2019) Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Elsevier [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660519300241
33. Hussain F, Ghazanfar S, Husnain AM, Fayyaz UU, Shahzad F, Shah GA. IoT DoS and DDoS attack detection using ResNet. researchsquare [Online]. Available: https://www.researchsquare.com/article/rs-120303/v1
34. Dasgupta D (2021) Understanding ROC (receiver operating characteristic) curve | What is ROC?[Online]. Available: https://www.mygreatlearning.com/blog/roc-curve/

AI-Based Smart Robot for Restaurant Serving Applications

Muhammad Awais Qasim, Faisal Abrar, Sarosh Ahmad, and Muhammad Usman

Abstract With magnetic trail and infrared matching techniques, limitation in paths and high processing create difficulties in restaurant serving techniques. This article provides a user-friendly serving system by adopting real-time image processing and robot guidance application. SLAM (A technology used by Japan for simultaneously localizing and mapping) can overcome the above-mentioned difficulties but is much costly and has slow processing. To overcome SLAM drawbacks, all map localization processing has been shifted on a server processor which made RSR (Restaurant Serving Robot) less costly than other techniques. RSR is fully equipped by modern innovations in AI (Artificial Intelligence). Map localization in server knows about the little movement in restaurant hall and decides appropriate path for a robot. A restaurant serving robot is a real-time path deciding robot which is designed using simulation software, camera, a database for predefined paths, an android application, a WLAN communication system, and a robot based on Arduino. Simulation software gets real time frames from the camera, declares appropriate path, and keep an eye on serving robots.

Keywords Restaurant serving robot · AI robot · Artificial intelligence

1 Introduction

Near about twenty million manufacturing jobs could be replaced by robots in the next ten years, by an idea of international economic institutes. Many applications of

M. A. Qasim (✉) · F. Abrar · S. Ahmad
Department of Electrical Engineering and Technology, Government College University Faisalabad (GCUF), Faisalabad, Pakistan

S. Ahmad
e-mail: saroshahmad@ieee.org

M. Usman
Department of Computer Science, Government College University Faisalabad (GCUF), Faisalabad, Pakistan

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_5

robots can be covered by electronics, telecommunication, and IT program. In this project, an effort to exemplify a set of rules for AI serving robots that will serve and deliver the meal to every customer. The execution is done with resources available to decrease the price of the robot [1]. This project demonstrates the idea of an automatic menu serving robot with the help of image processing, data science, and communication. This provides proper service to customers in any indoor restaurant. DOS (Digital Order System) and other applications of the digital facility will replace old service techniques i.e., waiters can take orders from customers, in upcoming years. Old service technique creates too many irregularities in serving the customer with good facilities. Restaurant serving system uses innovative techniques in a trendsetting restaurant such as camera, a robot having ATmega2560, server computer for handling data and simulation for staff operator and android application for DOS to buildup service quality and to improve customer's feast experience [2]. The proposed Restaurant Serving Robot, shown in Fig. 1, is equipped with ultrasonic sensors, RT, and speaking functions to guide the customers. DOS provides the use of online billing. Along with these facilities, the restaurant serving system includes simulation software for the ease of hotel staff. This provides an automated and great experience for customers. All communication between robot and server is done by WLAN technology [3]. The server uses an 802.11n WLAN card and the robot use Node MCU (ESP-8266) to make a DHCP (Dynamic Host Configuration Protocol) network. This network helps the server to connect to the robot every time. These all features are the backbones of RSR which makes the system move more smoothly with great performance and efficiency and made all of our time valuable [4, 5].

Restaurant Serving Robot is connected with server computer by using PS (Python Script) which is the main brain of the RSR system. Server computer reduces the cost of SLAM (A technology used by Japan for simultaneously localizing and mapping) used in multiple robots [6]. Server computers give commands for multiple robots at

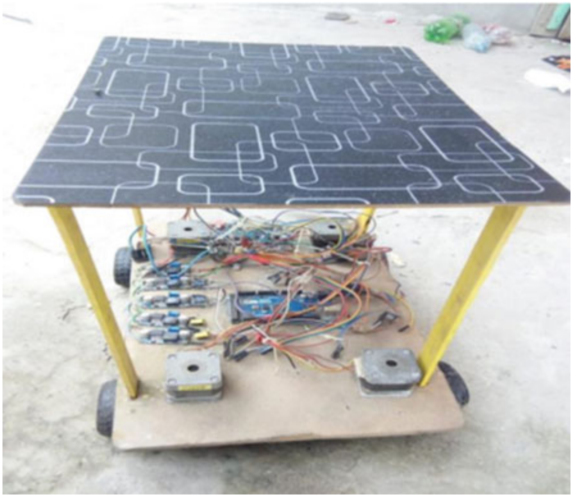


Fig. 1 Restaurant serving robot

one time and all multiple processing shifts from robot to server PC which reduces the cost for this project. A main problem of SLAM has also been covered in this project by shifting robot cameras to cameras in the hall. This algorithm knows much about the location of the person while getting images from the camera placed in the robot. These all features make the RSR work more efficiently and provide an enriching environment to the customer. RS-Robot can communicate with customers if someone is standing on its path. This robot can request and get coordinates for a new path any time when there are some complex obstacles. It is full of innovative techniques that made them more reliable and efficient than the previous technologies [7].

2 Literature Review

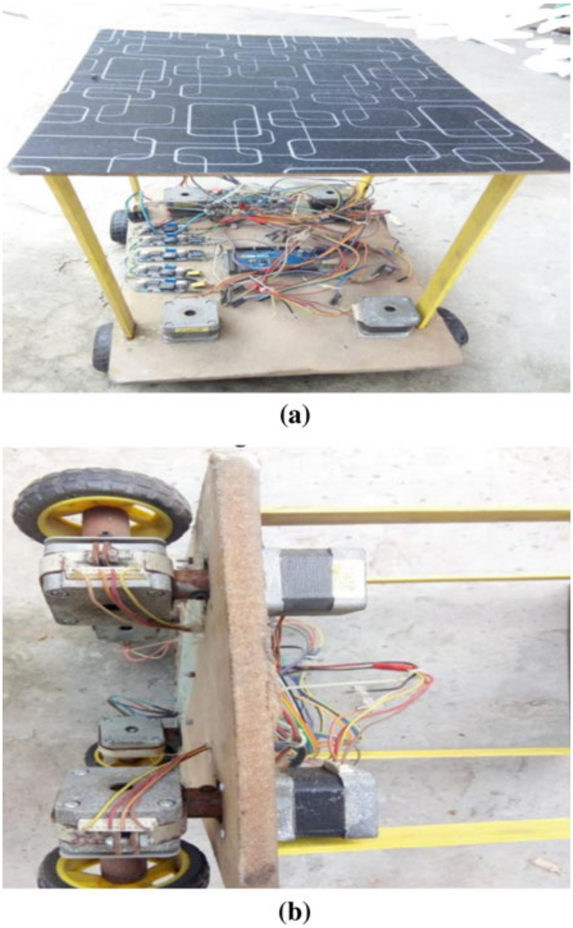
Butler Robot is used to decrease the workload of attendants in a restaurant to improve reliability and effectiveness. In [8] the author has proposed that the consumer toil on an Android application. This application is connected to the main hub and displays the fresh menu of the restaurant. The customer can check the menu and order the deal. The consumer can call the butler by a button through the application. The attendant comes to deliver the order and calculate the money for food. This whole menu can be seen in the cookhouse panel. When the given order is ready, the cook can mark the menu as cooked. After this, cooked items are displayed on the LCDs of the cashier and the waiter screen so they can serve these to the respective consumer. In [9] researcher-made research on a smart eatery for consumer center service. This system gives an online food ordering and area keeping procedure, and also a personal menu counsel service. By the use of RFID (Radio Frequency Identification) based community cards, attendants can instantly recognize. The butlers are used to take orders from the consumer and with the help of wireless LAN. Then cook cooks the order and the butler will deliver it to the consumer. When the consumer has consumed the feast, the manager will use Radio Frequency Identification-based system to identify the community-ID to count the cost of food. In [10] author put an effort into a self-service OI system based on ZigBee wireless technology. This system used Full Function Device and Reduced Function Device. FFD is a type of whole network administrator that can communicate with another device; RFD is used in the star topology network, which will work with the full function device. The author has made research, in [11], in which the eatery will be comprised of black lines followed by a robot. LED blinking and switching process are done by Arduino. The hosting robot will begin the black line when it will get bright light in the way it will turn right or left according to given commands and serve the order. After serving the food, it will again follow the dark line path and return to the default position. In [12], the author proposed a smart eatery and menu ordering system. They used an android application for a digital ordering system and the predefined path followed robot to serve customers. Pay Pal is used for billing systems and online survey systems to provide a better quality of food. In [13], the author introduced a new algorithm of DP-SLAM which includes map occupancy by using an accurate way. These developments were applied correctly in a larger and noisy environment than they tested these improvements with our earlier algorithm. They tried to test the DP-SLAM with noisier sensors and 3D maps.

The proposed RS-Robot used different sensors to give maximum output and accuracy. An ultrasonic category sensor model type. Ultrasonic Sensor, Load cell, Node MCU (*ESP8266*) are implemented. In this article, an ultrasonic sensor has been used to avert impediments. The ultrasonic sensor fulfills mainly two purposes. Firstly, if any person comes in between the pathway of the RS-Robot and customer, the sensor transmits information in the form of an indication to choose an alternative pathway or wait for a minute or two until the path is cleared. Secondly, it is extremely convenient for measuring distance and more specifically the distance between the table and admin counter. As it presents a real-time working environment, the distance obtained between the admin to the client table, by the use of an Ultrasonic sensor, will enable us to choose different paths to reach the destination. Load Cell is used for checking the availability of food on the tray or not. When the robot serves the food to the client, the load cell gives zero value then the robot knows that it has to go back to the admin table and this way serving process becomes extremely convenient.

3 Methodology

As the title describes the function of the model (Restaurant Serving Robot), it includes a robot and simulation that shows the position of the robot and gives instruction to the robot via communication. RS-Robot is used as a serving waiter in the proposed model, the main part, which is controlled by an automated script made in Python. Firstly, the model includes DOS (Digital Order System) which can send the order from a customer to cook via an android/web application. Cook receives the order and, after preparing the order, places the respective order to the Robot Tray (RT) and commands the Python Script to receive the table number and observe the suitable path, already saved in our database, for the robot via image processing and human detecting process to reach that respective table. After selecting the path, PS commands the robot to move on that selected path. PS communicates with the robot via configured wireless local area network (WLAN) module which helps to make a two-way communication between PS and the robot's microcontroller. The robot, after receives a command from PS, moves toward the destination. The front and downside views of the Proposed restaurant serving robot are shown in Fig. 2.

RS-Robot has ultrasonic sensors which continuously send distance of nearby person or objects to the PS via communication. The robot stops at a specific distance if there were some obstacles in the path. It, then picks audio from the memory and puts it on speaker, and will wait until the path clears. After the clearance of the path, the algorithm will make some calculations and the robot will start moving towards its destination. If it takes time to clear the path, an alarm signal will send to the manager or computer operator. The manager will check for path clearance itself.



(a)

(b)

Fig. 2 Front view of serving robot; **a** front view, **b** bottom View

When the robot reaches the destination, it will wait until the tray has been picked up by the customer. This process is done by load cell which placed under the RT. When an order is picked up by the customer and no weight will leave on RT then the load cell will give a signal to the microcontroller and, then the microcontroller will communicate to PS. PS then will command the microcontroller to return the robot to its default position near the cook [14].

3.1 Hardware

Designing the robot is the major part because choosing suitable electronic components is the main problem of RSR. In this section, the design of RSR and electronic circuits are discussed that are used to make it. Therefore, this portion has been divided into three sub-sections based on hardware:

Power section
Control section
Communication section.

3.1.1 Power Section

Provide accurate and continuous power to run electronic components properly is very important. As we know, most of the microcontrollers work on DC power, so providing smooth DC voltage is necessary to work it properly. If there seem some fluctuations inflow of smooth DC or miner leakage of AC voltages, then our components may not work properly or may damage the whole circuit. Therefore, provide the suitable and required power according to the given datasheet of electronic components is very important. Figure 3 shows the power setup of RS-Robot using *LM2596* modules.

Here, used 12 V, 35 amperes DC power supply (*PS-35G*) which is connected to 220 V AC power source, which is the main power supply, used to supply constant DC voltage to all circuits of RS-Robot. *PS-35G* is the best model to provide smooth DC voltage but there is some leakage of AC voltages (mV) that is enough to damage the microcontroller. To avoid this kind of damage, there implanted a DC-to-DC buck converter (*LM2596*) to supply smooth DC to every electronic component. *LM2596s* is a voltage regulator from 35 to 1.5 V. The capacitor of 100 uF is connected parallel to the output of this regulator as shown in Fig. 4.

The capacitor is used for the safety purposes of microcontrollers and motor drivers. If there appears a high voltage impulse, then the capacitor can handle it easily and pass this impulse to the ground directly.

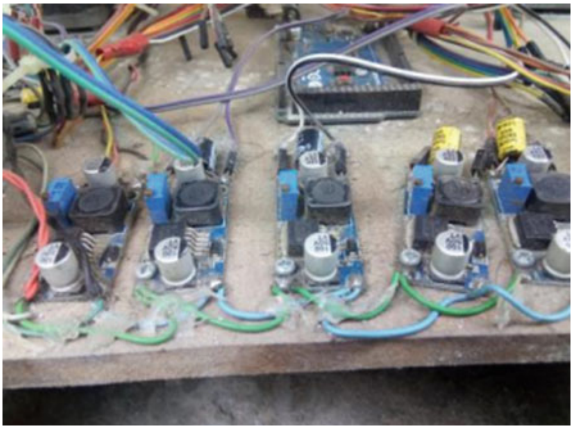


Fig. 3 Power setup using LM2596 modules

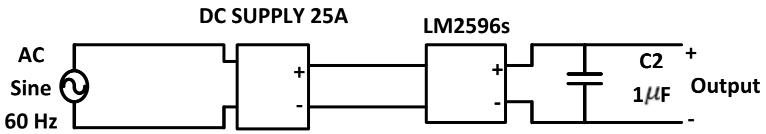


Fig. 4 Circuit diagram of power section

3.1.2 Control Section

The Control section is the heart of RS-Robot because coordination of every part of the robot is necessary to work it properly to get the required output. If there occur some abruptions in the control system, then the whole working RS-Robot will stop. Therefore, the control system should be strong enough to work properly for a long time.

1. ***Microcontroller***

The selection of suitable microcontrollers plays a vital role to get the required output. Here, the Arduino *ATmega3560* microcontroller, as shown in Fig. 5, is used to control the movement of the robot and other modules. *ATmega2560* has 54 digital pins which are enough for the project's work. One USB B-type connector is placed on it to program the controller. A power jack is already attached to connect it with the DC source. 15 out of 54 pins are PWM (Pulse Wave Modulation) pins that have little use here. Its processing speed is 16 MHz which is ok with the required speed. The huge number of pins and its processing speed is the main reason for selecting this microcontroller for RS-Robot.

Program to control motors and other sensors are uploaded in it. It will also get some commands from PS by a communication module node MCU (Micro-Controller Unit) which is connected with *ATmega2560*. When an order is ready to serve the

Fig. 5 Arduino MEGA2560

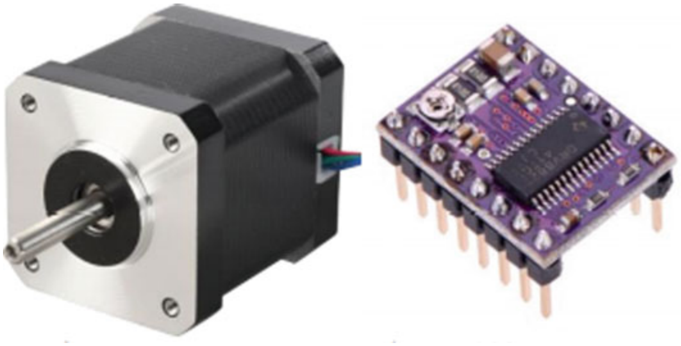


Fig. 6 Stepper motor and DRV8825

customer, Arduino receives commands to reach a specific table by using a specific path. After receiving the signal from PS, all movement of RS-Robot is controlled by *ATmega2560*. Microcontroller gives a signal to motors to move forward and backward with controlled speed. Ultrasonic sensors are also connected with the microcontroller if there are some hurdles in the way of RS-Robot. Ultrasonic sensors give a signal to the microcontroller and the motors of the robot stopped working until the hurdles get clear. Then, the microcontroller gives the command to motors to initiate again.

2. ***Motor and their Drivers***

In robotics, the selection of reliable motors has an important role to perform any physical function with machines. Here, RS-Robot used DC Stepper Motor (DC 5.6 V, 1.2 A and 1.8′/step) with motor driver *drv8825*, as shown in Fig. 6*,* for the movement of the robot.

Here one question is raised, why did we choose stepper motors for RS-Robot? The answer to this question is that it is necessary to notice the speed of motors to synchronize it with simulation in real-time. As mentioned above that this stepper motor has 1.8°/steps which make 200 steps/revolution. The revolution of the stepper motor can control by Arduino programming but a simple DC motor slows down with the passage of time and its revolution cannot be controlled by a microcontroller. On the other hand, this model of stepper motor can bear 1.5 kg weight but a simple DC motor can bear only 100–500 gm with the same voltages and is currently given to stepper motor. The circuit diagram of connections of the motor is shown in Fig. 7.

Four stepper motors are vertically attached to each corner on the base of RS-Robot and four stepper motors are connected horizontally with the shafts of vertical motors. Upper motors are used to turn RS-Robot at a specific angle and lower motors are used to move RS-Robot forward and backward very smoothly. Each stepper motor is connected with each motor driver (*DRV 8825*). *DRV8825* is a very sensitive component. It cannot bear high voltage fluctuations; therefore, the capacitor is necessary to place parallel with the motor power supply as shown in the circuit diagram in Fig. 4. Stepper motors have four wires which are connected with the motor driver at B2, B1, A1, A2 terminals. Two pins of *DRV8825* (step and dir) are connected with

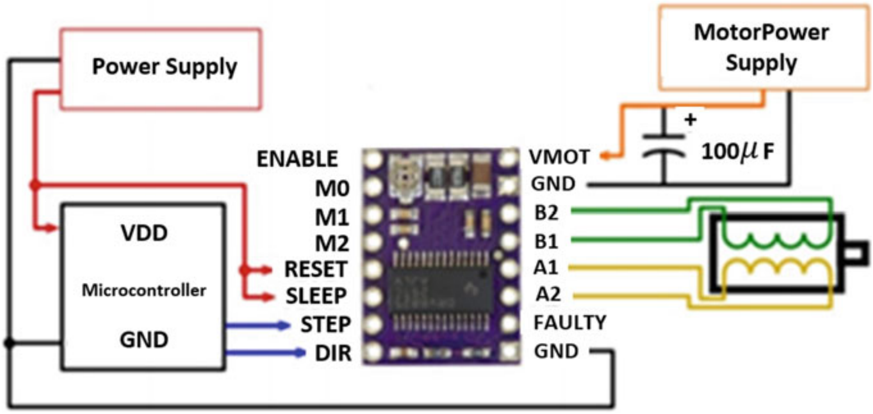


Fig. 7 Circuit diagram of motor connection

ATmega2560. These two pins are control pins that receive signals from *ATmega2560* to control the direction and steps of the motor. *DRV8825* does not have its storage. It receives signals from the controller and works according to controller commands. When an order is ready to deliver, the manager commands the robot to deliver it on a particular table and the controller sends signals to each motor driver to initiate motors with a particular speed and direction. When RS-Robot reaches the destination table, the controller sends a stop signal to the motor driver through above mentioned two pins (step, dir).

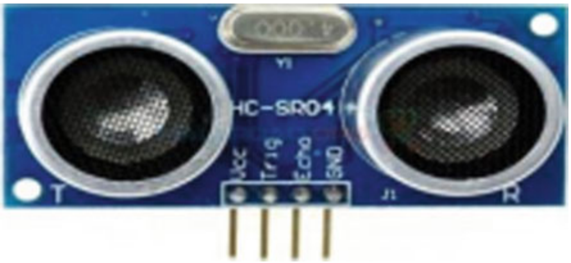
3. ***Sensors***

Different sensors are used to give the maximum output and accuracy. Sensors implemented here are Ultrasonic Sensor, Load cell, Node MCU (*ESP8266*).

(a) *Ultrasonic sensor*

An ultrasonic sensor, as shown in Fig. 8, is used to detect the obstacles. If any obstacle/person comes in the path of RS-Robot during serving the order to destination, then this sensor works and gives information to the robot. It will wait for a minute. If the path is not clear within a minute, the RS-Robot will choose another

Fig. 8 Ultrasonic sensor



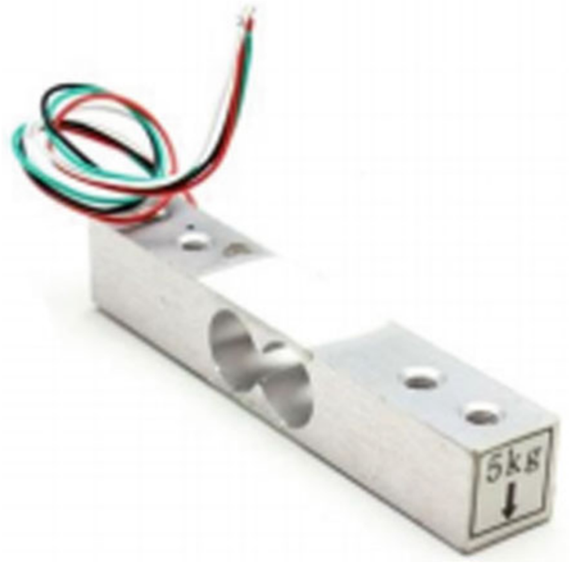


Fig. 9 Load cell module having 5 kg weight carrier capacity

path and will reach the destination on time. This sensor can also be used for another purpose such as to measure the distance i.e., the distance between the table and admin counter. If it can find the distance between the admin table to client table, it will manage different paths to reach the destination.

(b) *Load Cell*

Load Cell, shown in Fig. 9, is used to check the weightage of RT to conclude if the food is available on the tray or not. When the robot served the food, the client will pick and there will be no weight left on the RT. The load cell will give zero value of weight to the admin. The RS-Robot will then be returned to the admin table.

(c) *Speaker and Amplification*

Speaker is used here for alarm when obstacles come in RS-Robot's way. A transistor is used with a speaker for amplification. Audacity software is also used to convert the mp3 voice into the 8 kHz with 16 bits, then uploaded it into the Arduino code where it can change the voice. Some random phrases like; Please take a side etc. can also be added to its memory. The circuit diagram of the amplification circuit is shown in Fig. 10.

3.1.3 Communication Section

Node MCU (*ESP8266*) is used to communicate PC with Arduino and with the server. Initially, the connection will be established between the microprocessor with python. Then, it will connect with the server and vice versa to establish communication. Node

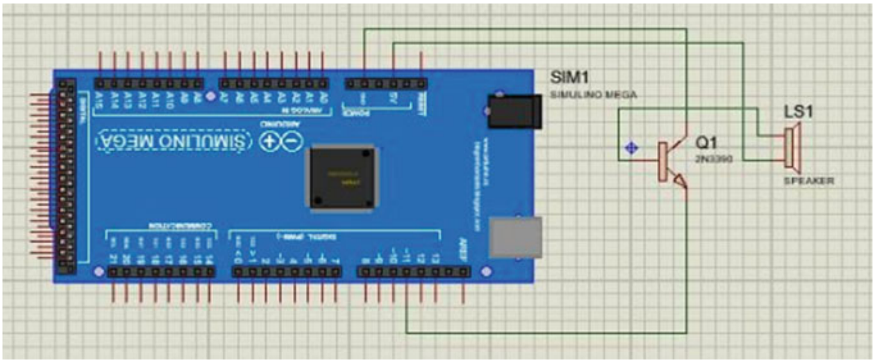


Fig. 10 Circuit diagram of amplification

MCU (*ESP8266*) is an open-source IoT platform. It includes firmware that runs on the *ESP8266* Wi-Fi SoC from Expressive systems and hardware, which is based on the *ESP-12* module. The term “Node MCU” by default refers to the firmware rather than the dev kits. The firmware uses the LUA scripting language. It is based on the e-LUA project and built on the Expressive Non-OS SDK for *ESP8266*. It uses many open-source projects, such as Luacjson and spiffs. LUA-based interactive firmware for Expressive *ESP8622* Wi-Fi SoC, as well as an open-source hardware board that contrary to the $3 ESP8266 Wi-Fi modules includes *a CP2102* TTL to USB chip for programming and debugging, is breadboard-friendly, and can simply be powered via its micro-USB port.

3.2 Software

PS is made on Python which includes the human detecting process and path selection. Initially, the script needs an input of table number and then make a process of human detecting by taking a frame from the roof of the wall using an IP camera. Then decide a path for the robot. RS-Robots get motor angle and speed in RPM (Revolutions per Minute) from the PS. After receiving the parameters, a function of RS-Robot will initiate. Microcontroller, placed on Arduino board, programmed with the help of Arduino IDE. Another mobile application made in the android studio is for DOS which gets the order from a customer to cook.

3.2.1 Arduino IDE

Arduino IDE is used to program Arduino board carry an *ATmega2560* microcontroller to design a program for RS-Robot to move on the particular parameters getting from PS. In this program, Arduino IDE is controlling eight-stepper motors. The angle

Fig. 11 Arduino IDE software Logo

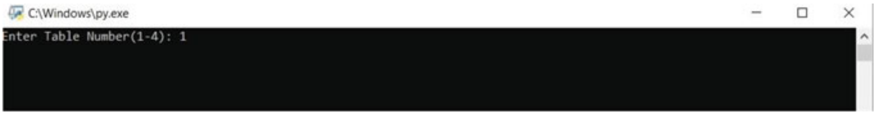
from parameters decides steps for four-stepper motors and RPM decides to remain four-step values for four-stepper motors. It continuously sends values of the ultrasonic sensor to PS by using a transmission (Tx) pin or Arduino. The logo of Arduino IDE software is shown in Fig. 11.

4 Results and Discussion

Simulation software gets real-time frames from the camera, declares appropriate paths for serving robots, and keeps an eye on RS-Robots. The serving robot has a shape like a serving table fully automated and fully equipped with different sensors like ultrasonic using high-efficiency speaker and microphone to check the obstacles, load cell under the table to check a load of tray and speaking function to communicate with customers. Simulation software has a function to show the position of RS-Robot and the crowd of people in the restaurant. It has some more features like consuming time and displays estimated time to reach the table. An android application is based on real-time menu items and adds to cart options to finalize the order.

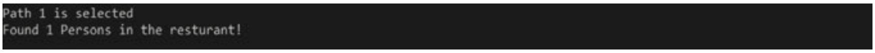
4.1 Human Detecting Process

Initially, there would be a process in the program, after receiving an input command of table number as shown in Fig. 12, is a human detecting process. This process gets a frame by using open CV2 libraries from a local IP address that is generated by an IP camera application available on the play store. The frame is filtered by using a haar-cascade filter that detects the body and stores approximate locations of persons by getting only foot values from it.



```
nter Table Number(1-4): 1
```

Fig. 12 Input table number in module



```
Path 1 is selected
Found 1 Persons in the resturant!
```

Fig. 13 Path selection

4.2 Path Selection

The path Selection process will continue its process of matching pixels between already stored path pixels and the location of persons, as shown in Fig. 13. This is achieved by making loops and by adding conditions to check different paths which have no matching. This decides the parameters for the robot program.

The first, second, and third simulation views are shown in Fig. 14.

4.3 Android Applications

4.3.1 DOS Application

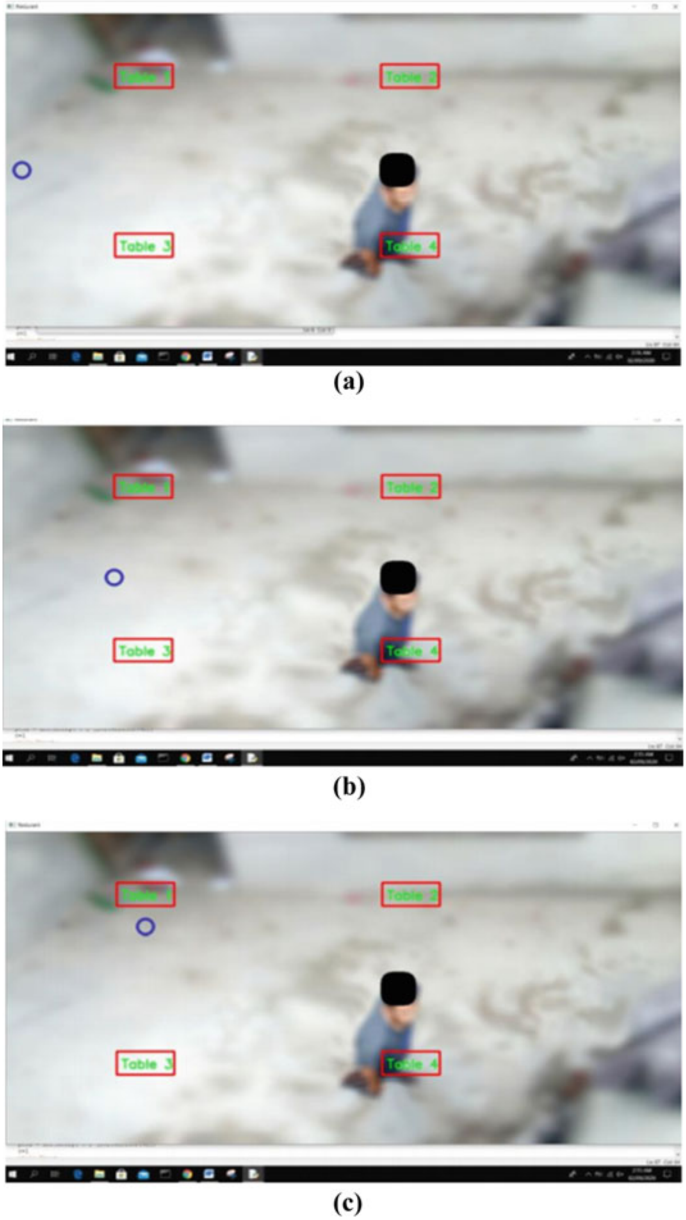
DOS is an order service application that has a menu option in which customer orders for a deal add to cart option and a helpline. DOS has another service of paying a bill via credit card. DOS is installed in android mobile which is placed on each table. It welcomes first and takes orders from the consumer and sends to the staff panel opened by chefs in the back-end system where food is going to be prepared.

4.3.2 IP Camera

IP camera continuously getting frames from the camera and generates a local IP to connect to it. This IP is then used in the script to get frames from the IP Camera application running on an android phone.

5 Conclusion

Interestingly, the idea of real-time implementation of Restaurant Serving Robot (RSR) is the need for the hour, especially during Covid-19 times. When social



(a)

(b)

(c)

Fig. 14 Simulations results of restaurant serving robot; **a** First simulation view. **b** Second simulation view. and **c** Third simulation view

distancing has been mandated by the World Health Organization (WHO) all around the world. It is scientifically proven that the places where Corona Virus spread vigorously are through public places like restaurants, pubs, and malls. It is kind of an obligation right now for society and world leaders to take immediate initiative to delegate the functions carried out by managers, waiters, and any servicemen in these public places to robots or any other medium where it almost nullifies the situation of human interaction. We are not suggesting that it can reduce contamination but it can act as a good medium to at least reduce the possibility of human interaction which is considered as one of the pivotal reasons for spreading the virus. The proposed prototype Restaurant Serving Robot works very efficiently and reliably. From Fig. 13, we can easily see the position of tables and humans. The blue circle shows the position of the RS-Robot. From Fig. 14c, it is completely clear that the RS-Robot has been reached its destination table and is delivering its order.

Apart from this valid reason, the practical demonstration of this prototype RS-Robot is not circumscribed to restaurants only, according to our vision it can be further developed to make it compatible for carrying out applications in movie theatres like in the ticket counter, in malls as a salesman. In hospitals, it can act as an artificially controlled medical service nurse which will not only increase productivity but will also ensure humanity to be better equipped for tacking any further pandemic in the future.

6 Future Work

As with any research has many scopes for enhancement by eradicating the flows, this idea of RS-Robot also falls in that category of further development in the future by which one can add a new dimension in the restaurant business all around the world. To make it possible, cost reduction can be the way, to begin with. Fortunately, this idea of RS-Robot is already be implemented at a much lower cost than the previously used SLAM module that costs around 300 dollars, a price that can be difficult for middle-income restaurant owners to manage. SLAM only knows the data of thing that is in front of its robot whereas this project uses hall cameras that know all the data of the person in the hall from above. Camera cost also reduces because the server gets data from already existed well-furnished restaurant halls.

Robot processing will become less by shifting all main processing of localizing and deciding path on the server computer. This makes the processing of microcontroller from 85 to 37% and more efficient which results in the following benefits.

- The lifetime of the robot also increases due to fewer burdens on the microcontroller.
- Many applications can be done by using this technique. RSR is just an application of this technique.

Load Lifter used in production block of industries can also be controlled by using this technique because of existing cameras for security purposes. These cameras will be used to detect objects and available paths to move lifters from one place to a destination. This feature can be further modified which will make it more feasible for commercialization in the future. End Product Package involved a serving robot, an android application, and simulation software. The availability of these items at lower cost will ensure the manufacturers all around the involved in robotics and artificial intelligence use the proposed idea. Further development of this RS-Robot at the cheapest cost without compromising the quality to ensure a fully functional restaurant serving robot that can give a new definition to the restaurant management sector in any part of the world.

References

1. Vo T (2019) Development of restaurant serving robot using line following approach. J Sci Technol Iss Inform Commun Technol 17:1. https://doi.org/10.31130/JST-UD2019-087.
2. Vo T, Dang PV, Ngo T, Le H-N, Toan D (2019) Restaurant serving robot with double line sensors following approach. 235–239. https://doi.org/10.1109/ICMA.2019.8816404
3. Shah UA, Ali F, Sohail S, Khan H (2016) Intelligent robotic waiter with menu ordering system. In: 1st International Electrical Engineering Congress (IEEC'16), May 13–14, IEP Center, Karachi, Pakistan
4. Asha S, Bhagyashree MA, Divyashree R, Hokrani R (2019) An automatic robotic restaurant. JETIR 6(5). ISSN: 2349-5162
5. Afsheen S, Fathima L, Khan Z, Elahi M (2018) A self-sufficient waiter Robo for serving in restaurants. Int J Adv Res Dev 3(5)
6. Jang H-W, Lee S-B (2020) Serving Robots: management and applications for restaurant business sustainability. Sustainability 2020, MDPI J
7. Prejitha CT, Vikram Raj N, Vibhandik H, Gayatri W, Tiple P, Kapoor K, Manduri SP, Ankush, Kulkarni MS (2020) Design of restaurant serving robot for contact less and hygienic eating experience. IRJET 7(8)
8. Pieska S, Liuska M, Jauhiainen J, Auno A (2013) Intelligent restaurant system smart-menu. In: 4th IEEE international conference on cognitive info communications, December 2–5, 2013, Budapest, Hungary
9. Tan T-H, Chang C-S, Chen Y-F (2012) Developing an intelligent e-restaurant with a menu recommender for customer-centric service. In: IEEE transactions on systems, man, and cybernetics—part c. Applications and reviews, vol 42, no 5, September 2012
10. Sun G, Song Q, Design of the restaurant self-service ordering system based on ZigBee technology. In: Communications and embedded system lab College of Information Technology and Science, Nankai University Tianjin, China
11. Noor MZH, Rahman AAA, Saaid MF, Ali MSAM, Zolkapli M (2012) The development of self-service restaurant ordering system (SROS). In: 2012 IEEE control and system graduate research colloquium (ICSGRC 2012)
12. Moravec HP, Elfes A (1985) High resolution maps from wide angle sonar. In: 1985 IEEE international conference on robotics and automation. St. Louis, Missouri. IEEE Computer Society Press, pp 116–121

13. Leonard JH, Durrant-Whyte HF (1991) Mobile robot localization by tracking geometric beacons. In: IEEE transactions on robotics and automation. IEEE, pp 376–382
14. Tan YC, Lew BF, Tan KL, Goh KV, Lee KL, Kho ZC (2010) A new automated food delivery system using autonomous track. In: Proceedings of the 2010 IEEE conference on sustainable utilization and development in engineering and technology Universiti Tunku Abdul Rahman 20 & 21 November 2010, Faculty of Engineering, Kuala Lumpur, Malaysia

AI and IoT for Smart Environment and Energy

A Novel Deep Learning Architecture Based IoT Time-Series for Energy Consumption Forecasting in Smart Households

Saloua El Motaki and Badr Hirchoua

Abstract The rapid increase in energy demand has sharply increased power utilization in today's world. Predicting the energy consumption in advance is a crucial matter to investigate to push to optimization process forward. With the ever-expanding use of electrical energy in individual households, this paper investigates the practicality of energy consumption forecasting in such environments using recent technologies such as Internet-of-Things and Artificial Intelligence. The proposed architecture uses a Long Short-term Memory neural network (LSTM) enhanced by an attention mechanism, as a recent trend of deep learning inspired by human vision to focus on specific input data selectively. After that, the attention-based LSTM is combined with the convolutional neural network, which is, on its side, adapted to the time-series data process. Several experiments are executed over four possible scenarios. The obtained results confirm that our model yields very accurate results for electric energy consumption prediction. In addition, our model hit the smallest value in all error measures, compared to the state-of-the-art methods on individual household power consumption.

Keywords Electric energy consumption · Attention mechanism · Deep neural network · Long short-term memory · Convolutional neural network

S. El Motaki (✉)
Computer Science Department, Faculty of Science Dhar El Mahraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco
e-mail: saloua.elmotaki@usmba.ac.ma

B. Hirchoua
National Higher School of Arts and Crafts (ENSAM), Industrial Engineering and Productivity Department, Moulay Ismail University (UMI), Meknes, Morocco
e-mail: hirchoua.badr@gmail.com

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_6

1 Introduction

Residential buildings are currently identified as one of the significant contributors to energy consumption and greenhouse gas (GHG) emissions worldwide. According to the Moroccan Agency for Energy Efficiency (AMEE), the building sector is among the most energy-intensive sectors in Morocco, with an energy consumption reaching 33% divided into 7% for tertiary buildings and 26% for residential buildings [1]. Similarly, households were responsible for 26.3% of the European Union's (EU) total energy consumption [2], and American households have used approximately 20.75 Quadrillion British thermal units in 2020 [3]. This consumption is subject to increase due to population growth, the creation of new cities and the constant urbanization movement.

In addition to the basic elements that can reduce building's operating costs, such as good thermal isolation, modern buildings can benefit from specific management systems based on recent technologies, including Internet of Things (IoT) and Artificial Intelligence (AI), which lead to the recently born paradigm "Smart building".

Smart buildings are becoming active to improve their energy efficiency and to fit into the new smart-grid context. Therefore, one of the major trends is to make it "intelligent" to optimize equipment management and eliminate all unnecessary consumption. This is based on the development of occupant interaction and automation systems and their integration into a global building management system based on IoT technologies and advanced learning algorithms. Moreover, recent trends suggest that Deep Learning (DL) is about to make a splash in the energy industry [4]. Although still in development, DL has demonstrated its ability to make power grids more intelligent and efficient. With these insights, many researchers have studied the potential of DL to change buildings' energy consumption and make huge savings [5].

DL has been the subject of numerous studies and has obtained significant improvements in many energy management areas [6]. Several advanced DL techniques can presently provide short-term energy consumption prediction outputs with remarkably high accuracy [7]. Which is relevant for household energy load management, energy efficiency and large-scale maintenance planning of sophisticated smart grids [8]. The most widely used DL techniques are most likely to be the Convolutional Neural Network (CNN) and LSTM. The use of such models over time-series data is back to the fact that CNN models are suitable for capturing meaningful features and are able to clean the noise in the input data, whereas LSTM models are designed to extract sequential pattern information. However, since these methods are data-driven, their reliability cannot be guaranteed by relying on household energy consumption data to forecast residential energy consumption, as the energy consumption patterns of individual households may be considerably irregular. On the other hand, traditional DL approaches, such as CNN, generally involve multidimensional inputs to ensure accurate prediction.

The present paper aims to develop a novel forecasting DL architecture that integrates attention-based LSTM neural networks with CNN to ensure an accurate household power usage prediction. The main steps to meet this purpose are as follows:

- **LSTM neural network model design**: Due to the highly volatile and univariate data related to household electricity consumption, a long-term sequence of data entries is desired to ensure reasonable accuracy. We opt for the LSTM neural network as the model's underlying mechanism, owing to its potential to handle time-series data dependency.
- **Implementation of the attention-based LSTM model**: The attention mechanism enables the LSTM framework to focus selectively on the time-series input data and to emphasize the data that contains the highest level of information. Attention mechanism constitutes a significant and recent enhancement of deep neural networks (DNNs). Attention networks have become a common element of the DL system toolbox, contributing to impressive results in machine translation applications and speech recognition [9, 10]. To our knowledge, its applicability to energy prediction, however, has so far been relatively uninvestigated.
- **Extraction and reorganization of local patterns using CNN**: Convolutional networks are shown to be highly efficient for automatic attribute processing and extraction due to the particular affinity of their operations for dealing with spatially structured data. Nevertheless, to train efficient models capable of extracting rich attributes and on non-trivial tasks (classification among thousands of time-series data), it is necessary to have large datasets and the appropriate computational resources. Despite this constraint, it is possible to take advantage of complex models already trained by using CNNs already trained by attention-based LSTM outputs to extract specific features automatically.

While the ability of DL models to perform accurately has been increased recently, the most significant performance increase has been accomplished by applying more advanced topologies to the models. There is, however, room for further improvement of the LSTM model by focusing on individual underlayers of the input time series as well as giving selective attention to the inputs. In this paper, we introduce a hybrid model that couples the attention mechanism with LSTM neural networks to achieve higher accuracy for long-term time series data forecasting and subsequently to advance the existing state-of-the-art. Additionally, we highlight the superiority of this model over other established models and suggest some other potential extensions.

The remainder of the paper is structured into the following sections: Sect. 2 introduces several relevant works addressing residential households power consumption forecasting using DL techniques. Section 3 describes briefly the baseline methods used to develop our work and outlines the proposed architecture. The results of experiments using real-world data are presented in Sect. 4 along with a detailed discussion. Finally, we conclude the paper in Sect. 5 by reviewing the advantages of our model and suggesting potential improvements.

2 Related Work

A large and growing body of literature has investigated AI methodologies for individual households electrical energy forecasting. Along with the rapid development of DL technology, prediction techniques may be usually divided into long-term, medium-term, short-term and extremely short-term forecasting predictive methods. These categories can, moreover, be based on statistical-based models [11, 12], machine learning (ML)-based models [13, 14] and DL-based models [8, 15].

To reduce the spectrum of time-series data related to household power consumption, authors in [8] have proposed a hybrid technique that joins CNN with LSTM. By extracting the space-time features of the electrical power consumed, they could predict the consumption of the household power consumption. The proposed method consists of different layers. Starting with CNN algorithm, the output of the first layer is considered as an input of LSTM layer that models the temporal information. Using the output of the LSTM unit as input to the all-in layer, they generate a predictive time-series of the electricity supply.

Similarly, the CNN-LSTM neural network combination has been adopted by [16] for residential energy consumption forecasting. They have proposed the utilization of CNN layers to extract features from multiple variables relevant to energy consumption prediction. After extracting the prominent patterns of energy consumption and filtering out the noise, they passed the output of these CNN layers as input to the LSTM layer. The output of the latter is forwarded to a fully connected layer that conveniently produces a time-series prediction of the power consumption. Besides reducing the data spectrum and extracting features, they have analyzed the variables space to decide on which attributes are the most relevant in the CNN layers.

On the other hand, instead of using time-series data as an input of the first layer of CNN, authors in [17] have considered image encoding frameworks as the first layer two-dimensional, aiming at reaching a higher accuracy. Three separate image encoding techniques, going through time-series of historical data of electrical energy load of an individual residential. Authors have concluded that the integration of image encoding techniques could improve the prediction results.

To address the DL neural networks susceptibility to overfitting (due to the number of parameters and a relatively small amount of data), Shi et al. [18] have introduced a novel Pooling-based Deep Recurrent Neural Network (PDRNN) based on a new data dimension referring to the historical load data related to the neighbouring households. The authors assume that PDRNN consists of learning from spatial information shared among interconnected agents to enable a higher amount of layer learning before any overfitting occurrence. In other words, the interactions between the neighbouring individual households and their correlations are considered within the load profile pool, which involves the generation of new features through deep layers, allowing the increase of both inputs volume and diversity.

Likewise, residential load forecasting has been addressed in [7]. To assess the difference between the system-level load and individual household electrical loads, they performed an exploratory evaluation of customer-level data. They used a density-based clustering technique to assess the inconsistency of household load profiles as well as to identify how challenging the residential load prediction process is. Moreover, they proposed a load prediction framework based on a Recurrent Neural Network (RNN) combined with LSTM, given its to learn long-term temporal connections.

In [19], a hybrid DL forecasting framework combining multiple LSTM neural networks with stationary wavelet transforms (SWT) is introduced. The authors have used SWT to stationarize the individual households data. They divide the initial signal of the power consumption into several more stable sub-signals, each of which is processed by a separate LSTM neural network. The aggregated prediction results of the LSTM neural networks using the inverse SWT constitute the final forecasting output. The work demonstrated that the stationary waveforms could improve the prediction performance of the individual LSTM neural network resulting in an enhancement of aggregate prediction accuracy.

In [20], a DL technique combined with a time-series clustering approach (K-shapes) to predict a short term energy load forecasting in individual households. The work aimed at selecting the most relevant features for improving the prediction accuracy, and it used K-shapes clustering to organize power consumers into separated clusters, following their consumption profiles. In the same way, authors in [21] have proposed an interactive clustering algorithm, Fuzzy Possibilistic C-means with Focal Point (FPCMFP), to predict the power consumption in households. The work has followed a zooming metaphor to decide on which level the data related to power consumption must be perceived to reflect the actual grouping of data.

An ensemble-based ML framework for the daily prediction of energy use by households is presented in [22]. Using a two-phase resampling scheme, the suggested ensemble learning approach produces diversity-controlled random resamples used for training single units of the Artificial Neural Networks (ANN). Afterwards, the members of the ensemble were co-trained with a linear robust combiner which includes a bias adjustment parameter. Given the finite availability of historical household energy consumption records, the findings of the analysis highlight the potential of the model to generate accurate energy consumption forecasts for a wide range of households.

Authors in [23] have proposed a model based on CNN-LSTM neural network for household power consumption forecasting. The hyperparameters, including the number of CNN kernels, the number of hidden LSTM units, and the number of units in the CNN-LSTM fully connected layer, are optimized using Practical Swarm Optimization (PSO). The experimental results, using a dataset of residential power consumption, have shown that PSO could improve the prediction accuracy of CNN-LSTM.

Also, Le et al. have established an efficient approach to forecast multi-power consumption in residential households by leveraging transfer learning and clustering-based modeling in the training of LSTM models to minimize the processing computation time. The work evolves a time-series clustering framework that employs a k-means clustering algorithm to split the training dataset into discrete groups. The resulting clusters are then used to train the LSTM model [24].

In general, predicting energy consumption in individual households using DL techniques has been, recently, an active research area. Prioritizing previously developed models or stating the best practice would be an unreasonable assumption since each model has a specific target, relevant resources and influencing parameters. However, in this work and besides proposing a new architecture, we provide a systematic comparison that demonstrates the relevance of our model.

3 Methodology

In this section, the basic concepts and properties of our model architecture are introduced. Precisely, the LSTM, the attention mechanism-based LSTM, and the CNN are detailed.

3.1 Long Short-Term Memory

RNNs are poorly effective for applications involving long time intervals, i.e., limited memory of the input data environment can occur. The main cause of this occurrence lies in the decrement or explosion of input data as it flows through the recurrent network connections. LSTM neural networks have been implemented to solve this problem, which is referred to as the "Vanishing Gradient Problem".

An LSTM network includes a memory cell, typically a layer of neurons, as well as three gates: an input gate, an output gate and a memory gate. These three gates will allow to modulate the flow of information at the input, at the output and to memorize in an analog way using a sigmoid activation function (Decide whether the LSTM unit holds, updates, or discards the information on the cell state) [25]. LSTM relies on the cell illustrated in Fig. 1 and introduces two main features. One, instead of having a single fully connected hidden layer, it has four that interact with each other within a unit. On the other hand, the LSTM has an additional memory (C_t mechanism besides the RNN's own hidden state.

The cell state represents the long-term memory of the unit where C_t undergoes only elementary operations such as sum and element-to-element product, as opposed to the hidden state at time t which represents the shorter-term memory (h_t undergoes more complex transformations, sigmoid and $tanh$ functions). The process is composed of three main phases: First, the forgetting gate F_t, given the input x_t and the hidden state h_{t-1}, allows to skip a portion of the information memorized in the cell state C_{t-1}. Second, The input gate I_t, given the input x_t and the hidden state h_{t-1}, serves to add some of the input information into the cell state C_{t-1}. Finally, the output gate O_t where a non-linear function is applied to the input x_t.

The output of the o_t network is equivalent to the hidden state at the output h_t: having the updated cell state, C_t and the result of the output gate O_t, allows to filter the input information to be transmitted in the hidden state at the output h_t . The

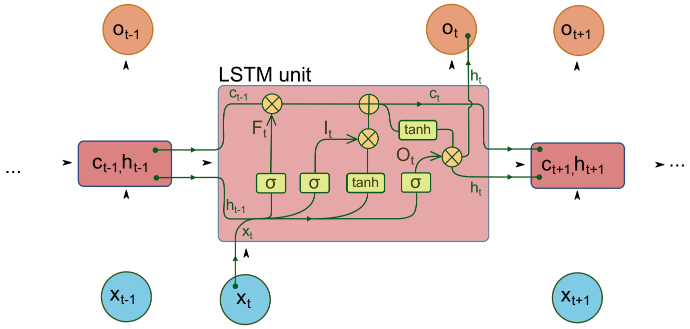


Fig. 1 The LSTM unit architecture

hidden state and the cell state are transmitted from cell to cell over time. Practically, the hidden state and the cell state are updated as follows:

$$I_t = \sigma(U^i x_t + W^i h_{t-1}) \tag{1}$$

$$F_t = \sigma(U^f x_t + W^f h_{t-1}) \tag{2}$$

$$O_t = \sigma(U^o x_t + W^o h_{t-1}) \tag{3}$$

$$C_t = \sigma(F_t \cdot C_{t-1} + I_t \cdot \tanh(U^g x_t + W^g h_{t-1})) \tag{4}$$

$$h_t = o_t = \tanh(C_t) \cdot O_t \tag{5}$$

Given σ the sigmoid function, $\cdot$ the term-to-term matrix product and the U and W correspond to weight matrices.

3.2 *Attention Model*

From the human perspective, their perception that does not directly consider all inputs from the external environment. Instead, they focus on the important elements to take only the needed information. The attention model was first applied in computer vision tasks, precisely in image recognition [26]. Recently, attention mechanisms have grown as an essential component of compelling sequence modeling in several tasks and domains. We have used the attention model to recode the input residential power consumption feature sequence. Standard network drives the encoding paradigm to an information bottleneck containing one hidden layer that encodes the entire input. The attention mechanism [27] provides a solid approach to maintain the set of internal representations which scale with the input size. The model utilizes an internal reasoning step to make a soft choice over the encoded representations. This

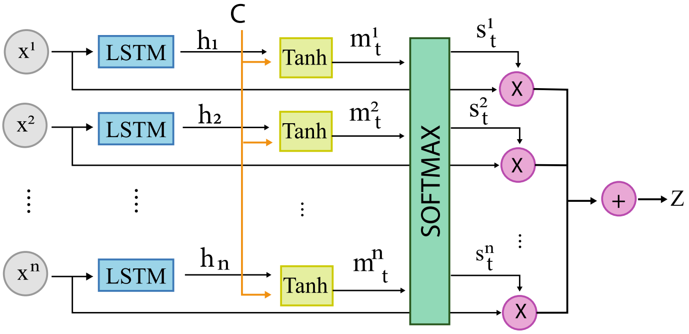


Fig. 2 The attention model

allows the model to maintain a memory balance which is crucial for scaling systems in various tasks [27] (Fig. 2).

The model begins with mapping the x_t to h_t using:

$$h_t = f(h_{t-1}, x_t) \tag{6}$$

The LSTM is adopted as f to handle the long-term dependence presented in time series prediction. The $h_t \in \mathbb{R}^s$ is the hidden state at time t with size s.

Furthermore, the attention mechanism is constructed through a deterministic attention model. Given a feature sequence $x^k = (x_1^k, x_2^k, \ldots, x_p^k) \in \mathbb{R}^p$, previous hidden state h_{t-1} and cell state C_{t-1} in the LSTM unit, we define:

$$m_t^k = v^t \tanh(W_1.[h_{t-1}, C_{t-1}] + W_2 x^k) \tag{7}$$

$$s_t^k = softmax(m_t^k) = \frac{\exp(m_t^k)}{\sum_{i=1}^{k} \exp(m_t^i)} \tag{8}$$

where the matrices W_1, W_2, and the vector v are learnable parameters. Vector m^k has length p where the ith item covers the value of the kth feature sequence at time t. The softmax function normalizes these items to deliver the attention percentages associated with each one. In other words, the attention weight s^k reflects the attention score associated with the kth feature sequences. The attention model produces at time t, i.e., the z_t weighted as the following:

$$z_t = \sum_{i=1}^{n} s_t^i x_t^i \tag{9}$$

3.3 Convolutional Neural Network

First introduced by Lecun et al. in 1998 [28], the CNN is a type of feedforward neural network that performs successfully in natural language and image processing [29]. Motivated by this success, researchers have started adopting CNN architectures to analyze time-series data [30]. Convolution can be regarded as a filter being applied and dragged over the time series. Instead of having a double dimension (width and height), filters have only one dimension (time), contrary to the case of images. In practice, if a filter of length l is convolved (multiplied) with a univariate time-series input, given filter values set to $\left[\frac{1}{l}, \ldots, \frac{1}{l}, \right]$, the convolution process leads to a moving average with a sliding gate of a length l. An implementation of a timestamp based convolution is expressed by the equation below:

$$C_t = f(w \cdot x_{t-l/2:t+l/2} + b) \quad | \quad \forall t \in [1, T] \tag{10}$$

where C represents the outcome of a convolution performed on a time-series data X of length T having an l-length filter w. b is a bias parameter, and f stands for a non-linear terminal function like the Rectified Linear Unit (ReLU). Eventually, the application of a convolutional filter on an input time-series data produces another univariate time-series C that has been subjected to a filtering operation. The application of different filters on a time-series data yields a multi-variate time-series with dimensions corresponding to the total number of filters applied. To apply multiple filters on a time-series entry is intuitively intended to learn various relevant features for either a classification or a regression task which is the purpose of our work.

3.4 The Proposed Architecture

Figure 3 highlights the general architecture of our model. We intend to leverage the strength of the attention mechanism, LSTM, and CNN in a unified model for predicting electric energy consumption.

The attention-based LSTM (AT-LSTM) model consists of the attention model and the LSTM model. The attention model inputs a sequence of energy consumption features. It adaptively selects the most critical features and assigns higher scores to the corresponding feature sequence. The output of the attention model is plugged into an LSTM model. LSTM can model the complex time series features collected from electric energy demand forecasting.

Next, CNN receives the LSTM patterns with attention weight to predict the accurate value. CNN is constructed of a convolution layer that takes LSTM outputs as inputs, an output layer that extracts features to a fully connected layer. Internally, CNN consists of a double convolution layer followed by a ReLU activation function layer and a pooling layer. Furthermore, while the backpropagation round, the gradients become smaller until they vanish, which affects the learning process. We solve

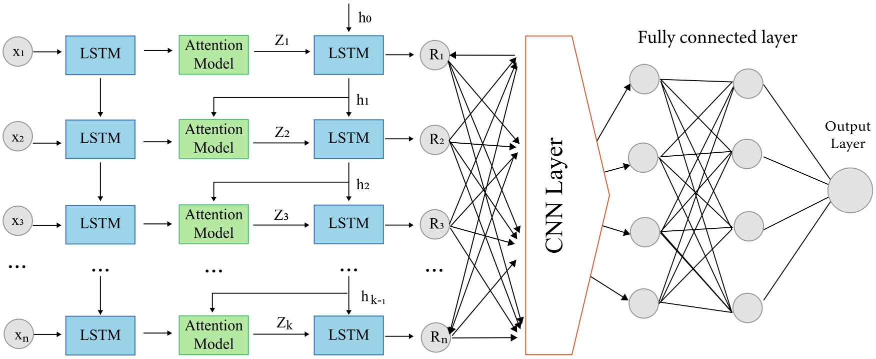


Fig. 3 The general framework of our model

this issue by using the ReLU activation function, which does not compress signals. The convolution layer employs the convolution procedure to the incoming signals and transfers the results to the next layer.

The pooling layer joins the output of a neuron cluster in the current layer into a single neuron in the following layer. The pooling layer minimizes the representation space to lower the number of parameters. We have used max-pooling that can pass the overfitting. The max-pooling layer operation is given as follows, where T is the window stride and R is the pooling size:

$$p_{ij}^{l} = \max_{r \in \mathbb{R}} y_{i \times T + rj}^{l-1} \tag{11}$$

The last layer in the proposed model is fully connected layers. It is utilized to generate power consumption for a certain period. The output of the CNN is the input to the fully connected layer.This final layer uses the following equation, where σ is a the activation function, w represents the weight of the ith node in the layer $l-1$ and the jth node for l, and bl_i^{l-1} is a bias:

$$d_i^l = \sum_j W_{ij}^{l-1}(\sigma(h_i^{l-1}) + b_i^{l-1}) \tag{12}$$

The design of the proposed model can be differently adjusted according to the nature and parameter tuning of the layers composing the model. Starting with CNN instead of LSTM affects the model performances. CNN can extract the local patterns and reorganize the order of input data. Therefore, the time series of the output from the CNN is different from the original dataset. Consequently, when this output is sent to the LSTM component, it cannot extract the temporal features, but it works as another fully connected layer. Thus, we have used the LSTM layer before the CNN layer in our proposed model. Our model is a sequence of an attention mechanism-based LSTM, LSTM layer, Convolutional layer, Pooling layer, and Dense layer.

4 Experimental Results and Discussion

In this section, we present the evaluations conducted to examine the performance of our proposed architecture.

4.1 Dataset

In this paper, we have used the individual household electric power consumption dataset.[1] To predict the validation of the introduced architecture, we utilized time-series features to predict the global active power (GAP). This dataset is presented in 1-min units with actual power consumption data obtained from a French household. The data was gathered between December 2006 and November 2010. The dataset is preprocessed to prepare the right form of input and output features. Firstly, the date and time variables are merged. Next, a total of 25,979 missing values are filled by the mean values for each column. Lastly, we have created a sub-metering reminder (SR) variable:

$$SR = (\frac{GAP \times 1000}{60}) - \sum_{i=1}^{3} sub\_metering\{i\} \tag{13}$$

The original dataset contains 7 variables (besides the date and time) as shown in Table 1. Precisely, 4 variables highlight the power consumption data, and the 3 variables obtained from energy consumption sensors. Table 2 presents a quantitative and detailed description of the dataset. The attributes collected from the energy consumption sensors are the sub-metering 1, 2, and 3. The first one corresponds to the kitchen, including the microwave oven and the dishwasher. The second one corresponds to the laundry room that includes a tumble-drier, a washing machine, and lighting. The last one represents the air conditioner and the electric water heater.

4.2 Evaluation Metrics

To investigate the performance of different models, we utilize four evaluation metrics: Mean absolute percentage error (MAPE) (Eq. 17), the mean absolute error (MAE) (Eq. 16), the mean squared error (MSE), and root-mean-square error (RMSE)

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/00235/household_power_consumption.zip.

Table 1 The individual household electric power consumption dataset features

Feature	Description
Global active power (GAP)	Household total active power (kW)
Global reactive power (GRP)	Household total reactive power (kW)
Voltage	Minute-averaged voltage (V)
Global intensity (GI)	Minute-averaged current intensity (in A)
Sub metering 1 (S1)	Active energy for the kitchen (watt-hours of active energy)
Sub metering 2 (S2)	Active energy for laundry (watt-hours of active energy)
Sub metering 3 (S3)	Active energy for climate control systems (watt-hours of active energy)

Table 2 Dataset quantitative description

Attribute	Date	GAP (kW)	GRP (kW)	Voltage	GI (A)	S1 (Wh)	S2 (Wh)	S3 (Wh)
Max	26/11/10	11.122	1.39	254.15	48.4	88	80	31
Min	16/12/06	0.076	00	223.2	0.2	00	00	00
Average	–	1.089	0.124	240.844	4.618	1.117	1.289	6.453
Std. dev	–	1.055	0.113	3.239	4.435	6.139	5.794	8.436

(Eq. 15). $\hat{Y}$ is the prediction vector from a sample of n data points, and Y is the ground truth's vector.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{15}$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{Y}_i - Y_i|}{n} \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{17}$$

4.3 Implementation Scenarios

In order to respond to the energy consumption challenge, we conduct four main experiments to cover four possible scenarios.

- **Scenario 1: Global active power (GAP) Prediction using different DNN architectures**. The set of experiments of this scenario aims to study our method's performance for predicting the GAP. To do so, we implement the existing model, run different experiments, and compare different results daily and weekly. We choose models that are capable of learning long-term dependencies and patterns from the input sequence. Precisely, for Linear regression (LR), recurrent neural network (RNN), LSTM, and Gated recurrent units (GRUs) are used for comparisons. Besides, our model is compared to Kim et al. [16] and Jin-Young et al. [31].
- **Scenario 2: Multi-step time series forecasting. Precisely, based on current power consumption, the model tries to predict the power consumption for the next week**. Technically, this scenario framing is referred to as a multi-step time series forecasting problem. We roll multi-step forecasts from all models used in scenario 1 except the LR model.
- **Scenario 3: Time resolution effect. The time resolution changes impact the model error rate**. We investigate the performance of each model while changing the time resolution. Thus, we have studied the model performances minutely, hourly, daily, and weekly. The general intuition is: the lower the resolution, the lower the error rate should be. In other words, the higher resolution ends up in a reduced dataset, while the lower one contains a considerable amount of data.
- **Scenario 4: Time resolution effect based on scenario 2 training mechanism**. We set up the experiments of different competitive benchmarks with the same time resolution. Then, the performances are summarised to point the time resolution impact on different benchmarks models.

4.4 Comparison to the State-of-the-Art Methods

Table 3 present the results obtained for the first scenario that consists of predicting energy consumption. The obtained results show that our model outperforms all state-of-the-art methods in four error metrics like MSE, RMSE, MAE, and MAPE. Experimental results prove that the proposed model achieves excellent performance for power demand forecasting. Thus, it can be considered as a competitive method for power consumption prediction.

Table 4 summarizes the results achieved for multi-step time series forecasting using a 1-week time frame. The overall obtained results reveal that our model outperforms all used benchmarks. Furthermore, all models perform better than the first scenario, which marks the usefulness of the framing and modeling as multi-step time series forecasting.

Table 3 Scenario 1 performances

Resolution	Metric	LR	RNN	LSTM	GRU	[16]	[31]	Ours
Daily	MSE	0.2526	0.0051	0.0054	0.0055	0.0049	0.0049	0.0047
	MAE	0.3915	0.052	0.055	0.055	0.053	0.054	0.0052
	MAPE	0.52	0.275	0.317	0.316	0.275	0.290	0.25
	RMSE	0.5026	0.071	0.073	0.074	0.070	0.070	0.068
Weekly	MSE	0.1480	0.010	0.015	0.015	0.0096	0.0091	0.0080
	MAE	0.3199	0.076	0.087	0.085	0.069	0.066	0.061
	MAPE	0.413	0.32	0.417	0.411	0.29	0.28	0.25
	RMSE	0.3847	0.10	0.124	0.125	0.098	0.095	0.089

Table 4 Scenario 2 performances

Metric	RNN	LSTM	GRU	[16]	[31]	Ours
MSE	0.00525	0.00547	0.00546	0.00497	0.00522	0.00494
MAE	0.053	0.0558	0.0557	0.053	0.054	0.050
MAPE	0.270	0.311	0.312	0.278	0.28	0.23
RMSE	0.0724	0.0739	0.0738	0.0704	0.0722	0.0703

To confirm the influence of changes in time resolution on the forecasting methods, we have conducted series of power consumption experiments according to time resolution. Precisely, we have considered time resolution minutely, hourly, daily, and weekly units. We have investigated the proposed model performances in terms of resolution. Table 5 exposes the performance of different models. The error rate has to be lower as the resolution decreases. Our model shows more powerful performance than all used models at all resolutions. Consequently, our model proves its value even with time resolution variations.

We demonstrate the performance of the proposed architecture with well-known competitive benchmarks. Table 6 highlights the results of competitive works. We have used the results from [16], which have set up the experiment with the same time resolution of minutely, hourly, daily, and weekly units. Marino et al. [32] examined a new energy load prediction method applying DNNs and an LSTM-based Sequence-to-Sequence (Seq2Seq). On the other hand, Mocanu et al. [33] analyzed two recently developed stochastic models for energy consumption time series prediction. Precisely, they investigated the conditional restricted Boltzmann machine (CRBM) and factored conditional restricted Boltzmann machine (FCRBM). We prove that the proposed method outperformed the competitive benchmarks.

Adjusting the model's parameters, including the number of strides and filters and the kernel size, among others, influence the model performance. To understand why our proposed model outperforms the well-known methods, we tune different model setting to end up in the highest performances. All parameters are tuned locally since the dataset is not distributed, otherwise, we have to tune and share the parameters

Table 5 Scenario 3 performances

Method	Resolution	MSE	RMSE	MAE
Linear regression	Minutely	0.404	0.63	0.41
	Hourly	0.42	0.65	0.502
	Daily	0.25	0.502	0.39
	Weekly	0.14	0.38	0.31
LSTM	Minutely	0.74	0.86	0.62
	Hourly	0.51	0.71	0.526
	Daily	0.24	0.49	0.412
	Weekly	0.10	0.32	0.243
CNN-LSTM	Minutely	0.37	0.61	0.34
	Hourly	0.35	0.59	0.33
	Daily	0.10	0.32	0.25
	Weekly	0.09	0.30	0.23
Ours	Minutely	$0.29e^{-4}$	0.0053	0.0031
	Hourly	0.005	0.072	0.054
	Daily	0.0084	0.091	0.066
	Weekly	0.011	0.106	0.081

Table 6 Scenario 4 performances

Resolution	Minutely	Hourly	Daily	Weekly
CRBMs	1.06	0.47	–	0.24
FCRBM	0.80	0.43	–	0.21
Seq2Seq	0.44	0.39	–	–
CNN-LSTM	0.37	0.35	0.103	0.095
Ours	0.043	0.055	0.074	0.098

of the network [34]. Figure 4 highlights the prediction performance of the proposed model using the 1-min data. It demonstrates that our model can predict the power consumption accurately with minimum errors. In addition, Fig. 5 shows the training and evaluation loss values for each epoch. Precisely, it presents the mean squared error of the model using the 1-min data. Furthermore, the final model has the following error values: MAE = 0.007, RMSE = 0.018, MAPE = 0.077, and MSE= 0.00036. Accordingly, the resulted model could be helpful within the household in planning expenditures. It could also be helpful on the supply side for planning electricity demand for a specific household.

Eventually, the overall obtained results of different scenarios demonstrate that our method improves the residential power consumption prediction. Moreover, it achieves far better prediction performance over various time resolutions and data framing.

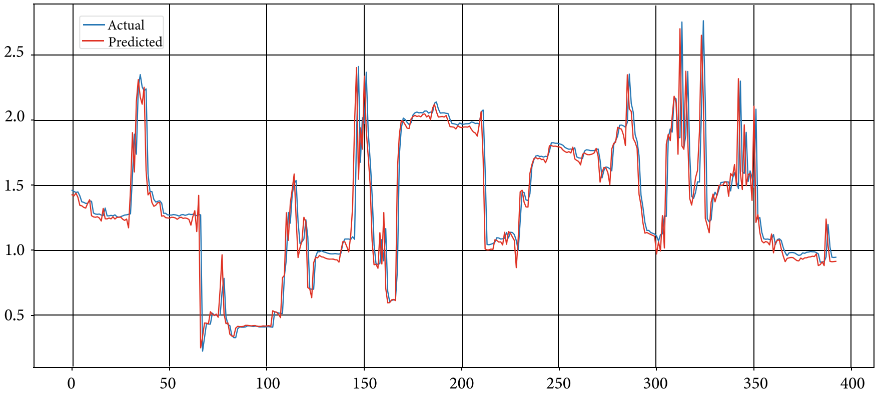


Fig. 4 The prediction performance of the proposed model using the 1-min data

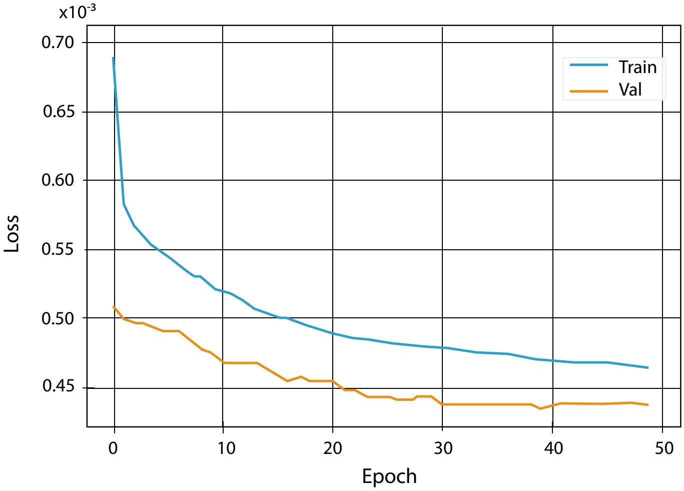


Fig. 5 The training and evaluation loss values for each epoch. We show the mean squared error of the model using the 1-min data

5 Conclusion

In this work, we have introduced an LSTM-CNN based attention network that models temporal features and the associations of multivariate features to predict residential power consumption. The predicting residential energy consumption has an imperfect trend. It is a multivariate time series, where numerous property variables influence individual prediction value. We have increased the modeling representation by combining the attention model-based LSTM with local pattern detection used by CNN to model complex features of the residential power consumption data. We have achieved the most reliable and stable forecasting performance compared to the state of the art methods. Besides, the double attention mechanism proves the efficiency of the proposed model in predicting electric energy consumption under various configurations, time resolutions, and data framing.

In future research, the proposed model will be extended by using an evolutive algorithm such as Genetic Algorithms to search for the optimal hyperparameters. In addition, we will focus on extending the input data with different scales and formats.

References

1. Efficacité énergétique dans le bâtiment. https://www.amee.ma/fr/expertise/batiment. Accessed 2021-07-05
2. Energy consumption in households. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households. Accessed 2021-07-05
3. Energy consumption of the residential sector in the United States from 1975 to 2020. https://www.statista.com/statistics/183625/us-residential-sector-energy-consumption-from-2000/. Accessed 2021-07-05
4. Shi H, Xu M, Li R (2017) Deep learning for household load forecasting a novel pooling deep RNN. IEEE Trans Smart Grid 9(5):5271–5280
5. Almalaq A, Edwards G (2017) A review of deep learning methods applied on load forecasting. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 511–516
6. Aslam S, Herodotou H, Mohsin SM, Javaid N, Ashraf N, Aslam S (2021) A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. Renew Sustain Energy Rev 144:110992
7. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y (2019) Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 10(1):841–851. https://doi.org/10.1109/TSG.2017.2753802
8. Kim TY, Cho SB (2018) Predicting the household power consumption using CNN-LSTM hybrid networks. In: Yin H, Camacho D, Novais P, Tallón-Ballesteros AJ (eds) Intelligent data engineering and automated learning—IDEAL 2018. Springer International Publishing, pp 481–490
9. Lipton ZC (2015) A critical review of recurrent neural networks for sequence learning. arXiv arXiv:1506.00019
10. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. NIPS'14. MIT Press, Cambridge, MA, pp 3104–3112
11. Amber KP, Aslam M, Hussain S (2015) Electricity consumption forecasting models for administration buildings of the UK higher education sector. Energy Build 90:127–136
12. Amjady N (2001) Short-term hourly load forecasting using time-series modeling with peak load estimation capability. IEEE Trans Power Syst 16(4):798–805. https://doi.org/10.1109/59.962429
13. Bogomolov A, Lepri B, Larcher R, Antonelli F, Pianesi F, Pentland A (2016) Energy consumption prediction using people dynamics derived from cellular network data. EPJ Data Sci 5:1–15
14. Yaslan Y, Bican B (2017) Empirical mode decomposition based denoising method with support vector regression for time series prediction: a case study for electricity load forecasting. Measurement 103:52–61
15. Alhussein M, Aurangzeb K, Haider SI (2020) Hybrid CNN-LSTM model for short-term individual household load forecasting. IEEE Access 8:180544–180557. https://doi.org/10.1109/ACCESS.2020.3028281

16. Kim T, Cho S (2019) Predicting residential energy consumption using CNN-LSTM neural networks. Energy 182:72–81
17. Estebsari A, Rajabi R (2020) Single residential load forecasting using deep learning and image encoding techniques. Electronics 9(1). https://doi.org/10.3390/electronics9010068. https://www.mdpi.com/2079-9292/9/1/68
18. Shi H, Xu M, Li R (2018) Deep learning for household load forecasting–a novel pooling deep RNN. IEEE Trans Smart Grid 9(5):5271–5280. https://doi.org/10.1109/TSG.2017.2686012
19. Yan K, Li W, Ji Z, Qi M, Du Y (2019) A hybrid LSTM neural network for energy consumption forecasting of individual households. IEEE Access 7:157633–157642. https://doi.org/10.1109/ACCESS.2019.2949065
20. Fahiman F, Erfani SM, Rajasegarar S, Palaniswami M, Leckie C (2017) Improving load forecasting based on deep learning and k-shape clustering. In: 2017 international joint conference on neural networks (IJCNN), pp 4134–4141 . https://doi.org/10.1109/IJCNN.2017.7966378
21. El Motaki S, Ali Y, Gualous H, Sabor J (2018) Possibilistic fuzzy c-means clustering under observer-biased framework. In: 2018 international conference on intelligent systems and computer vision (ISCV), pp 1–6. https://doi.org/10.1109/ISACV.2018.8354031
22. Alobaidi MH, Chebana F, Meguid M (2018) Robust ensemble learning framework for day-ahead forecasting of household based energy consumption. Appl Energy 212:997–1012
23. Kim TY, Cho SB (2019) Particle swarm optimization-based CNN-LSTM networks for forecasting energy consumption. In: 2019 IEEE congress on evolutionary computation (CEC), pp 1510–1516. https://doi.org/10.1109/CEC.2019.8789968
24. Le T, Vo MT, Kieu T, Hwang E, Rho S, Baik SW (2020) Multiple electric energy consumption forecasting using a cluster-based strategy for transfer learning in smart building. Sensors 20(9). https://doi.org/10.3390/s20092668. https://www.mdpi.com/1424-8220/20/9/2668
25. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
26. Fu J, Zheng H, Mei T (2017) Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 4476–4484. https://doi.org/10.1109/CVPR.2017.476
27. Kim Y, Denton C, Hoang L, Rush AM (2017) Structured attention networks. arXiv preprint arXiv:1702.00887
28. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
29. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
30. Gamboa J (2017) Deep learning for time-series analysis. ArXiv abs/1701.01887
31. Kim JY, Cho SB (2019) Electric energy consumption prediction by deep learning with state explainable autoencoder. Energies 12(4). https://doi.org/10.3390/en12040739. https://www.mdpi.com/1996-1073/12/4/739
32. Marino DL, Amarasinghe K, Manic M (2016) Building energy load forecasting using deep neural networks. In: IECON 2016—42nd annual conference of the IEEE industrial electronics society, pp 7046–7051. https://doi.org/10.1109/IECON.2016.7793413
33. Mocanu E, Nguyen H, Gibescu M, Kling W (2016) Deep learning for estimating building energy consumption. Sustain Energy Grids Netw 6:91–99. https://doi.org/10.1016/j.segan.2016.02.005
34. Hirchoua B, Ouhbi B, Frikh B, Khalil I (2020) A new knowledge capitalization framework in the big data context through shared parameters experiences. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 86–113. https://doi.org/10.1007/978-3-662-62199-8_4

Performances of CPV Optics in Morocco

Sarah El Himer, Mariya Ouaissa, and Mariyam Ouaissa

Abstract The optical elements are the keys to increase the geometrical and optical performances in the photovoltaic systems with concentration CPV. This article aims to develop an analysis of the CPV technologies presented along two categories: single element and two elements of concentration. Firstly, we started to test the CPV systems which is based on a single optical element for concentration, these elements are reflective (parabolic mirror), refracting (Fresnel lens) and total internal reflection (Cone, Parabolic concentrator composed, Compound parabolic concentrator with rectangular section, Pyramid, Hyperbola). We obtained that using a one stage optic for CPV systems cannot achieve our performance target for a CPV system; either the system allow having a high efficiency while the flux distribution is not homogeneous on the cell, or the opposite. This is why we have gone to simulation testing the second category of which the CPV systems are based on two optical concentration elements, (Mirror + SOEs and Fresnel lens + SOEs). As simulation results, we found that using the Fresnel lens as POE, the CPC exhibits the highest optical efficiency and the widest acceptance angle but the poorest flux distribution. The pyramid presents a homogeneous flux distribution and an optical efficiency that exceeds 80% but a slightly low acceptance angle. Concerning the systems based on parabolic mirror as POE, we found that using the parabolic mirror allows having a higher optical efficiency than the use of the Fresnel lens as POE. Regarding the flux distribution, we noticed that the pyramid and the CCPC keep the good uniformity. The compositions (mirror + hyperbola) and (mirror + dome) make it possible to have good uniformity but the acceptance angle remains low compared to the CPV system which is based on the use of Fresnel lenses as POE.

S. El Himer (✉)
Sidi Mohammed Ben Abdellah, Fez, Morocco
e-mail: sarah.elhimer@usmba.ac.ma

M. Ouaissa · M. Ouaissa
Moulay Ismail University, Meknes, Morocco
e-mail: mariya.ouaissa@edu.umi.ac.ma

M. Ouaissa
e-mail: mariyam.ouaissa@edu.umi.ac.ma

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_7

Keywords Concentrated photovoltaic · Optical efficiency · Acceptance angle · Fresnel lens · Compound parabolic concentrator · Dome · Pyramid · Cone

1 Introduction

Solar power is one of the key which help the emergent countries to spend less energy in the long term. Although, the initial cost of solar power may be higher than that of their current power generation technologies, it can finish up saving them in the long term and providing them with a more reliable power source. After the initial cost, the solar panels can produce more energy without any additional cost.

Solar energy has many advantages in emergent countries. They can help provide cleaner energy that doesn't destroy the land around them. In addition, the more energy people use from the sun and solar energy, the less CO_2 emissions are released into the atmosphere. This is beneficial for the world as a whole. Among these most promising solar energies in Morocco, we find concentrated photovoltaic.

Concentrated Photovoltaic (CPV) provides one of method's productions of electricity from solar energy. A CPV system consists of three principal elements: a concentration optics, solar cells and a tracker. Concentration optic contains generally two optical elements. The primary optic (POE) and secondary optic (SOE). Both elements are responsible for the concentration ratio, acceptance angle, irradiance uniformity and ultimately the efficiency of the module. The concentrated sunlight is usually reflected by parabolic mirrors, or refracted by lenses. Fresnel lenses can generate high solar intensity with less weight, and lower cost compared to mirrors.

The literature shows various theoretical studies about the performance of systems based on Fresnel lens and SOEs. Ning et al. [1] discussed a photovoltaic concentrators composed of Fresnel lens as primary optical element and dielectric totally internally reflecting concentrator (DTIR) as secondary optical element. Results showed that the concentrator system studied a larger acceptance angle, higher concentration ratio and a uniform flux distribution than the use of Fresnel lens alone. Buljan et al. [2] designed a CPV system by using a dielectric solid concentrator consisting of a combination of refractive, Total Internal Reflection (TIR), reflective surfaces (RXI-type SOE) and a flat Fresnel lens. They claimed good irradiance uniformity with an acceptance angle of 1.2× and an optical efficiency of 82.86%.

Victoria et al. [3] proposed a comparison study of the optical performances of concentrators using the same circular Plano-convex aspheric lens and different SOEs. They showed that the compound parabolic concentrator (CPC) presents the high optical efficiency and the largest acceptance angle than other SOEs, but the worst irradiance uniformity over the solar cell. Chen et al. [4] presents a study concerns three solar concentrators composed of Fresnel lens as POE and three SOEs: kaleidoscope (KOD), refractive truncated pyramid (RTP) and open truncated pyramid (SP). The target of this study is the determination of the optimum parameters of the SOEs. They found that the optimized KOD gives the highest acceptance angle (1.7°) as high as optical efficiency which arrives to 87%. Ferrer-Rodríguez et al. [5] present a complete optical modeling procedure using detailed optical simulation methodology including wavelength-dependent properties for four different refractive SOEs (Dome, SILO-Pyramid, refractive truncated pyramid, and trumpet) fabricated from PMMA associated with Fresnel lens. As a result, they found that the RTP and the SILO-Pyramid give the largest acceptance angle 1.11° and 1.13°respectively. Ferrer-Rodríguez et al. [6], El Himer et al. [7, 8] and Yahyaoui et al. [9] present an optical modeling of four different HCPV units composed of Fresnel lens and four refractive SOEs made from PMMA. They analyzed their performances while considering the sub-cells current density generation. They also experimentally compare the optical and electrical performances of the different Fresnel-based HCPV systems in another work [10]. Sellami et al. [11] designed a solar photovoltaic concentrator: a Square Elliptical Hyperboloid Concentrator (SEHC). For a SEHC with a concentration ratio of 4×, they showed that this optical element presents a large acceptance angle of 120° with stable optical efficiency of 40%, within which, this configuration has an advantage which is allowing collection of both diffuse and direct radiation for all days. For different dimensions of SHE concentrator allow to obtain only 50° as acceptance angle and a higher optical efficiency of 70%.

This paper analyses two categories of CPV systems based on three characteristics: optical efficiency, acceptance angle and flow distribution. The first category is that the CPV system consists of a single optical element which can be reflective (parabolic mirror), refracting (Fresnel lens) and five elements which is based on total internal reflection (CPC, CCPC, Cone, Hyperbole and pyramid).

The second category is a CPV system composed of two optical elements. For this test, we have chosen six secondaries among the most used, associated with a flat circular Fresnel lens on the one hand and on the other hand, the same SOEs associated with a parabolic mirror. Besides the optimal placement of each SOE relative to the FL and parabolic mirror is determined to find the best placement.

The rest of the paper is organized as follows: the second section introduces high concentrated photovoltaic; Sect. 3 is dedicated to the one stage concentrator and systems using two stages concentrators are discussed in the fourth section. Section 5 highlights the major results in a conclusion. The used characters in equations and in the text are all defined in Table 1.

2 High Concentrated Photovoltaic

The CPV system is based on the use of optical elements that can be refractive (Fresnel lenses), reflective (Mirror), or both and each CPV system is characterized by its concentration ratio which given by:

$$\mathrm{C} = \eta \times \mathrm{C_g} \tag{1}$$

where $\mathrm{C_g}$ is the geometrical concentration and η is the optical efficiency which are respectively given by:

Table 1 Nomenclature

a	Exit radius of the SOEs	MCPV	Micro concentrated photovoltaic
aN	Half of the difference between the	n1	Index of the input medium
A, B	Input radius of the hyperbolic	n2	Index of the exit medium
A_{in}	Input aperture area	n	Index of the secondary optical element
A_{out}	Exit or receiver area	nr	Number of reflection inside the cone
C	Whole concentration ratio	P_{in}	Power at the input of the system
C_g	Geometric concentration ratio	P_{out}	Power at the output of the system
C_{opt}	Optical concentration ratio	PMMA	Poly Methyl MethAcrylate
CR	Concentration ratio of the hyperbolic	POE	Primary optical element
CPC	Compound parabolic concentrator	r'	The angle between the normal of the prism surface and the incident ray
CCPC	Crossed compound parabolic concentrator	R_{con}	Entrance radius of the Cone
CPV	Concentrated photovoltaic	R_{cpc}	CPC entrance radius
d	Diameter of Fresnel lens	SOE	Secondary optical element
D	Parabolic mirror diameter	θ_i	Entrance angle of the SOEs
dm	Parabolic mirror depth	θ_{in}	Solar angle
f	Focal length of Fresnel lens and parabolic mirror		
f_{cpc}	Focal of the CPC	θ_f	Exit angle of the pyramid
FL	Fresnel Lens	β	Refractive angle
F/#	F_number of Fresnel lens	θ_{out}	Exit angle
H	Length of the SOEs	α	Angle of the pyramid

$$C_g = \frac{A_{in}}{A_{out}} = \frac{n_2 \sin\theta_{out}}{n_1 \sin r'} \tag{2}$$

$$\eta = \frac{P_{out}}{P_{in}} \tag{3}$$

where A_{in} and A_{out} are the input and the output aperture of CPV system as illustrated in Fig. 1

CPV systems can be classified depending on their geometry configuration: One stage concentration (use of one optical element) and the two stages of concentration: use of two optical elements to concentrate the solar radiation on solar cells. Table 2 summarizes the different basic components used in the optical systems of the CPV. As described in the table, the use of a plane reflector, homogenizer, prism and a light concentrator alone (without the addition of another optical element), results in a low concentration rate. On the other hand, Fresnel lenses and parabolic mirrors can be used alone for medium and high concentration.

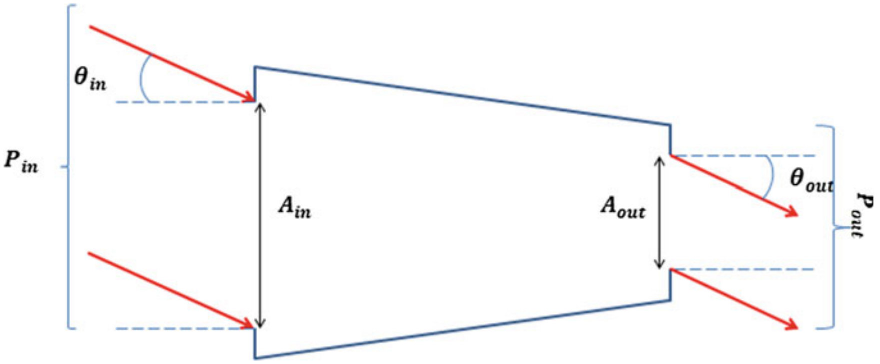


Fig. 1 Non imaging optic

Table 2 Characteristics of concentrators

Types	Characteristics	Concentration
Plan reflector	Reflection	Low and medium
Homogenizer	Refraction with TIR	Low
Parabolic mirror	Reflection	Medium and high
Fresnel lens	Refraction	Medium and high
CPC	Refraction with TIR	Medium
Prism	Refraction with TIR	Low
waveguide	Refraction with TIR	Low

3 One Stage CPV Concentrators

In this type of concentrator, the solar rays can be concentrated by using single optical element which can be reflective (parabolic mirror), refractive (Lens), luminescent and total internal reflection (CPC, cone etc.…) or any elements capable of concentrating solar radiation on an area equivalent to that of a solar receiver. Here we will list the optical element that used alone in concentrated photovoltaic system:

3.1 Fresnel Lens and Mirror Parabolic

The Fresnel lens is characterized by its focal distance f, its number of facets per unit of length and its diameter d.

The Fresnel lens is defined by its f number, F/#, where:

$$\mathrm{F}_{\#} = \frac{\mathrm{f}}{\mathrm{d}} \tag{4}$$

And its opening angle 2θi given by:

$$\tan\theta_i = \frac{d}{2f} \tag{5}$$

The Fresnel lens consists of a series of prisms [11] organized as shown in Fig. 2.

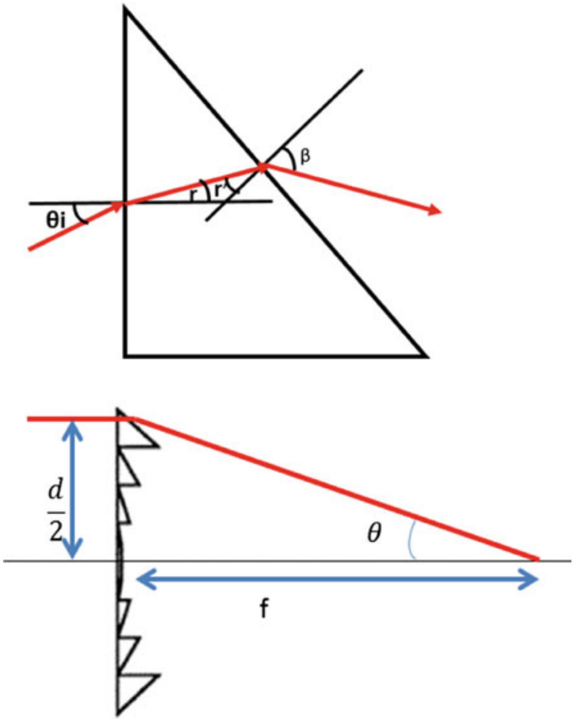
Figure 2 illustrates one prism which shows the path of a light ray. β is the angle between the normal direction of the prism surface and the refracted ray, r′ is the angle between the normal of the prism surface and the incident ray, θ_{in} is the angle of collection between the optical axis and the central radius of the prism and f is the focal length of the Fresnel lens. Each prism of the lens is designated by the following equations [12]:

$$\beta = \theta_i + r' \tag{6}$$

$$n \sin r' = \sin \beta \tag{7}$$

$$\tan r' = \frac{1}{2(n\sqrt{1 + \left(\tan^{-1}\theta_i\right)^2} - \tan^{-1}\theta_i)} \tag{8}$$

Fig. 2 Fresnel lens



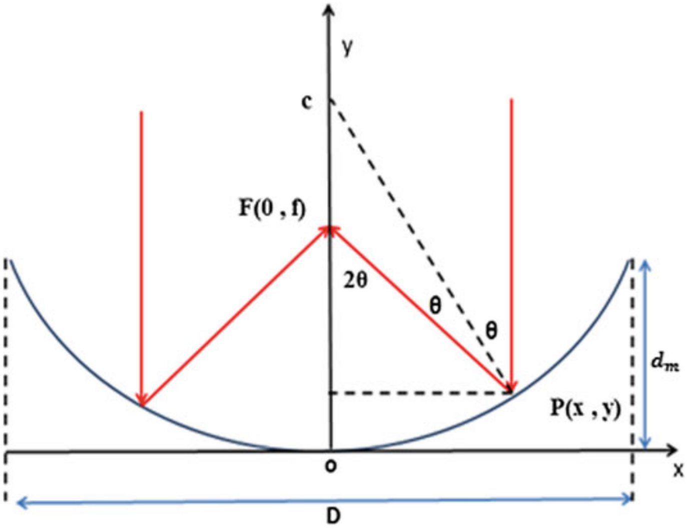


Fig. 4 Schematic view of a parabolic mirror concentrator

3.2 Mirror Parabolic

The Parabolic mirrors or dish is used to concentrate solar rays either to produce beams of parallel light at a desired point. This dish is defined by its diameter D, focal distance f and depth d_m, as illustrated in Fig. 4. These parameters are related by following equations [13]:

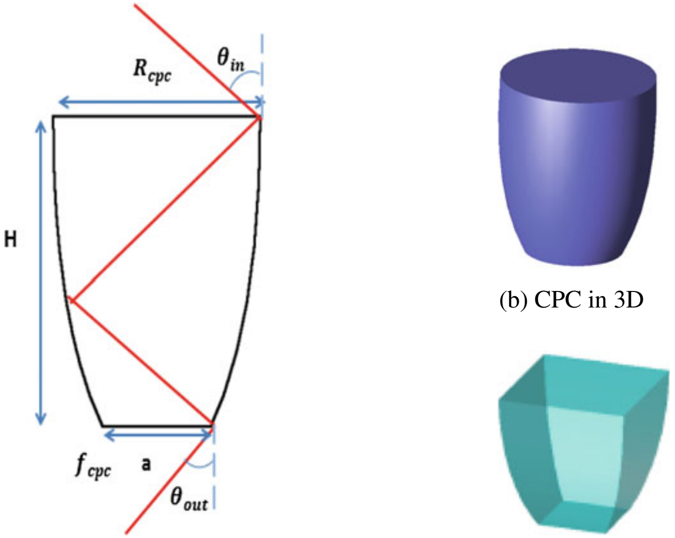
$$d_m = \frac{D^2}{16f} \tag{9}$$

$$\tan \theta_i = \frac{D}{2f} \tag{10}$$

3.3 Compound Parabolic Concentrator

The Compound Parabolic Concentrator (CPC) (Fig. 5) is demarcated by its geometrical parameters related by the next equation, simply obtained by using the Snell–Descartes laws for refraction and the outside angles of this optical element:

$$R_{cpc} = \frac{a}{\sin\theta_i} \tag{11}$$



(a) 2D view of the CPC/CCPC

(b) CPC in 3D

(c) CCPC in 3D

Fig. 5 Schematic view of CPC/CCPC

$$H = \left(R_{cpc} + a\right)\cot\theta_i \tag{12}$$

$$f_{cpc} = r(\sin\theta_{out} + \sin\theta_i) \tag{13}$$

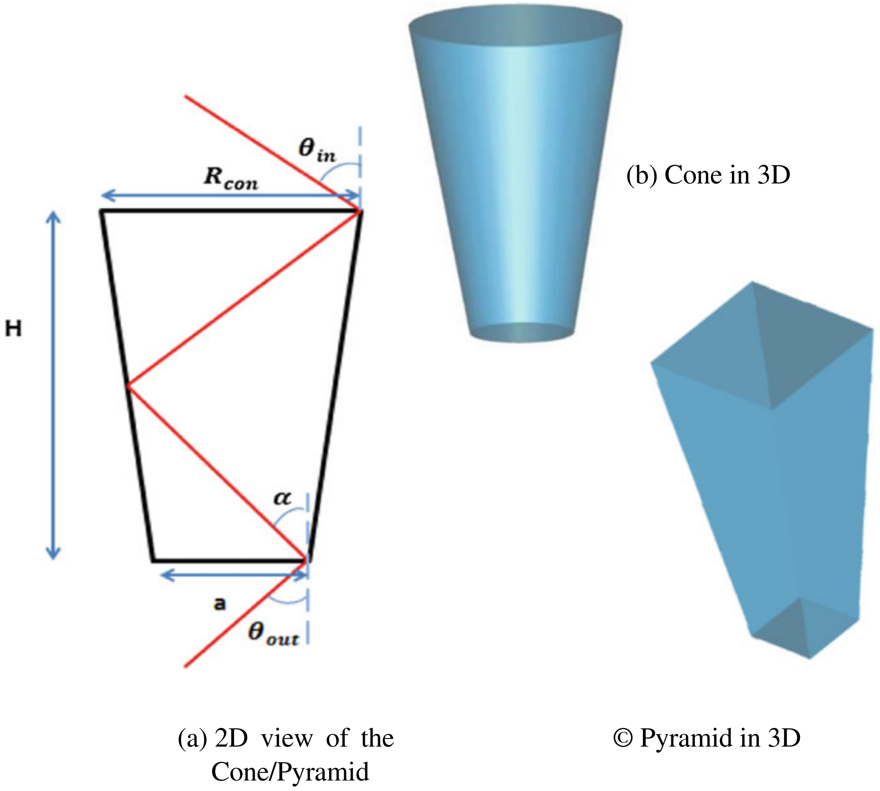
3.4 Conic Shape

The pyramid and the cone (see Fig. 6) are secondary optics. They are characterized by the output and input angle θ_f and θ_i, the output and input radius a, and R_{con}, the angle of cone or pyramid α and there length H. The latter is dependent on the inside number of ray reflections, n_r:

$$\theta_{out} = \theta_i + 2n_r\alpha \tag{14}$$

$$R_{con} = a + \sum_{n=1}^{n_{max}} a_n \tag{15}$$

$$H = \frac{R_{con} - a}{\tan\alpha} \tag{16}$$



(b) Cone in 3D

(a) 2D view of the Cone/Pyramid

© Pyramid in 3D

Fig. 6 Schematic view of cone/pyramid

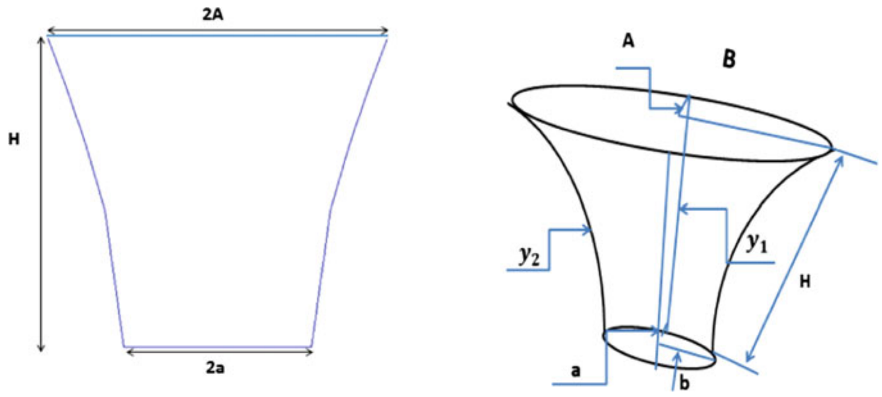
3.5 Hyperboloid Shape

The hyperbolic concentrator is an element which is habitually used for thermal application more than photovoltaic. Its design (see Fig. 7) was created on the following Eqs. (17)–(20). The coordinates of the profiles connecting the ends of the major axis, A and a, and the minor axis, B and b, are given by:

$$\mathrm{Y}_1 = \sqrt{\left(\left(\frac{\mathrm{x}}{\mathrm{a}}\right)^2 - 1\right)\mathrm{H}^2(\mathrm{CR} - 1)^{-1}} \tag{17}$$

$$\mathrm{CR} = \left(\frac{\mathrm{A}}{\mathrm{a}}\right)^2 \tag{18}$$

$$\mathrm{Y}_2 = \sqrt{\left(\left(\frac{\mathrm{x}}{\mathrm{b}}\right)^2 - 1\right)\mathrm{H}^2(\mathrm{CR} - 1)^{-1}} \tag{19}$$



(a) 2D view of the hyperbole (b) Hyperbole in 3D

Fig. 7 Schematic view of the hyperbole

$$\mathrm{CR} = \left(\frac{\mathrm{B}}{\mathrm{b}}\right)^2 \tag{20}$$

3.6 Dome

The concept of the Dome (see Fig. 8) is based on designing its input surface in a way that every ray coming from the extreme points of the primary optic $\mathrm{A}(x_1, y_1)$ reaches the opposite extreme point of the cell $\mathrm{B}(x_2, y_2)$. Moreover, imposing the conservation of the optical path length stated in Fermat's principle to all rays that pass through the point A and B nevertheless of their incidence angle [14].

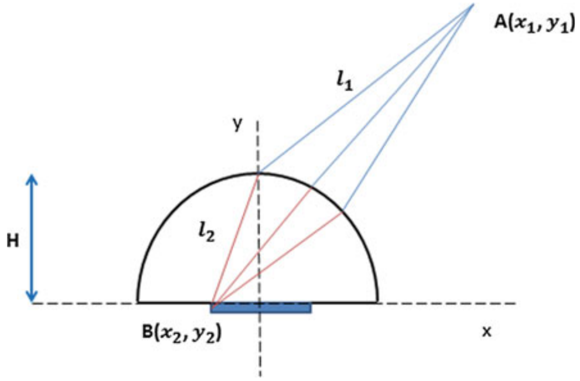


Fig. 8 Schematic view of the dome

$$l_1 = \sqrt{(x - x_1)^2 + (y - y_1)^2} \tag{21}$$

$$l_2 = \sqrt{(x - x_2)^2 + (y - y_2)^2} \tag{22}$$

The height of the dome H, is the initial point to construct the profile. It is fixed by imposing the tangent of the bidimensional profile to be horizontal at the origin.

$$(x_0, y_0) = (0, H) = \frac{x_0}{\tan\left[\arcsin\left(\frac{\sin\left(\mathrm{arctg}\left(\frac{1}{2F/\neq}\right)\right)}{n_{SOE}}\right)\right]} \tag{23}$$

3.7 *Simulation Results*

3.7.1 Fresnel Lens and Mirror Parabolic

In order to judge the performance relevance of one stage concentrators, we use a flat circular Fresnel lens and parabolic mirror with 350 mm as diameter and focal length of 270 mm associated to a square solar cell of 10 × 10 mm^2. The solar cell has been placed to catch the extreme rays, which come from the edge of the lens or the mirror and fall at the edges of solar cell in the worst case. The *z*-axis is set as the optical axes and the origin z = 0 at position of the Fresnel lens and the parabolic mirror.

The maximum position of the cell is dictated by the extreme rays which give a maximum distance z_{max} and are given by Eq. (24). However, it is also possible to put the cell in a z-position smaller than z_{max} without generating any losses if the cell diameter is large enough to catch the whole focalized beam. This condition gives the minimum position z_{min} (Eq. 25). Thus, the position z of the optical element has to be included between z_{min} and z_{max}. The irradiance used to measure the optical efficiency is 1000 W/m^2.

$$z_{max} = f + \frac{R}{\tan\theta} \tag{24}$$

$$z_{min} = f - \frac{R}{\tan\theta} \tag{25}$$

Table 3 resumes the optical performances of Fresnel lens and parabolic mirror, we notice that the optical efficiency of mirror parabolic is higher by 10% comparing to Fresnel lens, this is due to the fact that in the prisms of the Fresnel lens, the solar rays are reflected by the TIR and this decreases the intensity of the solar rays and that some rays don't reach the solar cell. We observe also that the two optical elements present the high intensity when the solar cell placed in focal plan. Concerning the

Table 3 Performances of optical elements used alone

	Fresnel lens		Parabolic mirror	
	Optical efficiency (%)	Acceptance angle (°)	Optical efficiency (%)	Acceptance angle (°)
$Z = Z_{min}$	81.51	0.6	91.32	0.6
$Z = f$	85	0.6	94.86	0.8
$Z = Z_{max}$	80.25	0.6	90.64	0.6

acceptance angle, the parabolic mirror presents the larger (0.8°) in case of the solar cell placed on focal plan, however this angle decreased to 0.6° when we set the solar cell in Z_{max} or Z_{min} which is the same angle presented by Fresnel lens whatever the position of the solar cell.

Figures 9 and 10 show the flux distribution on solar cell for three positions, In case of parabolic mirror, the flux distribution is uniform when the cell is placed on maximum or minimum distance, but in focal length, we observe a pic with high intensity on cell center. For the Fresnel lens, the flux distribution is non-uniform in the three positions.

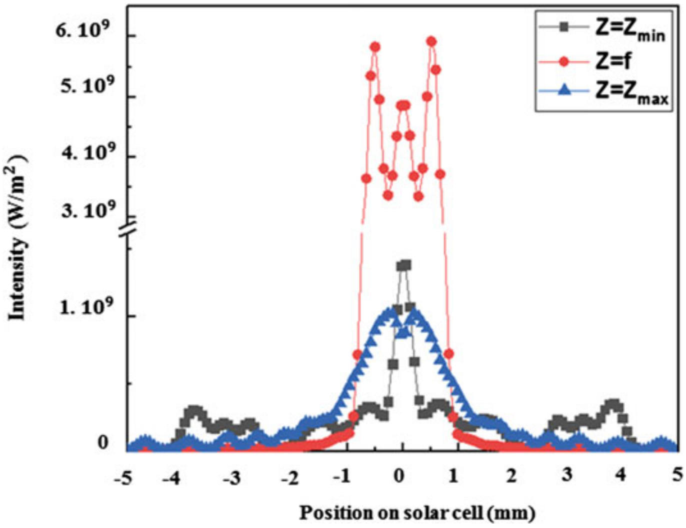


Fig. 9 Flux distribution using Fresnel Lens

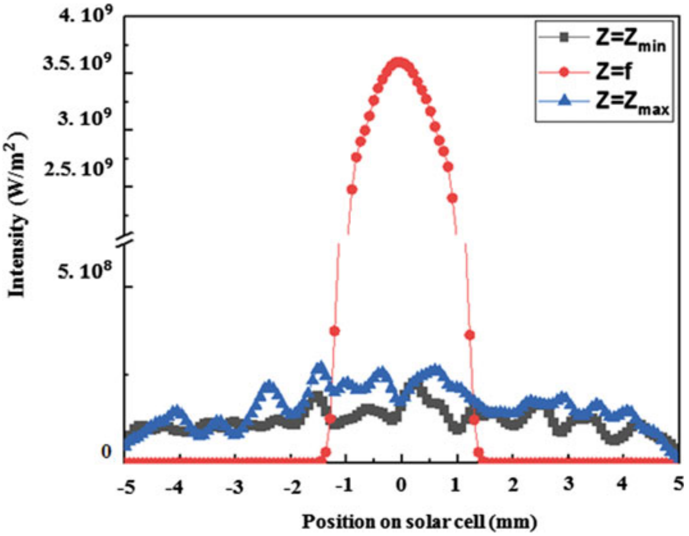


Fig. 10 Flux distribution using parabolic mirror

3.7.2 CPC, Conic and Hyperboloid Shapes

We propose to study here the performance of these elements (CPC, CCPC, the cone, the pyramid and the hyperbole) made from B270 as a function of the entry angle. The entry angle is varied in the range 10–70°.

We remember that the cone and the pyramid are characterized by the number of reflections which will be varied to optimize the two concentrators. To compare their performances, the five elements considered must have the same length, which is why we varied the number of reflections from 2 to 8 in the case of the pyramid and the Cone.

Figure 11 illustrates the evolution of the flux distribution at the exit of each optical element for different input angles. The size of each element is summarized in Table 4, in which, we find the input diameter, the length, the exit diameter of each element is fixed to 10 mm, as well as the number of reflection chosen in the case of the cone and the pyramid, this choice is made so that the five elements have the same length.

In the case of the CPC, the cone and the hyperbole, for an input angle of 10°, the flux is concentrated in a point on the center of the output, but from 30°, the flux is distributed over a circular surface with a diameter equal to the receiver side. In the case of square shapes, the flux is distributed over the entire surface of the receiver but beyond 50°, the flux begins to be concentrated in the center. These results can be explained by the reflection of the rays in the optical elements. In the case of entry angles of 50° and 70°, the areas of the entry and exit openings are almost equal, and then the rays don’t reflect, on the other hand, in the case of small entry angles, we notice that the rays are reflected because of the difference between the input and output opening.

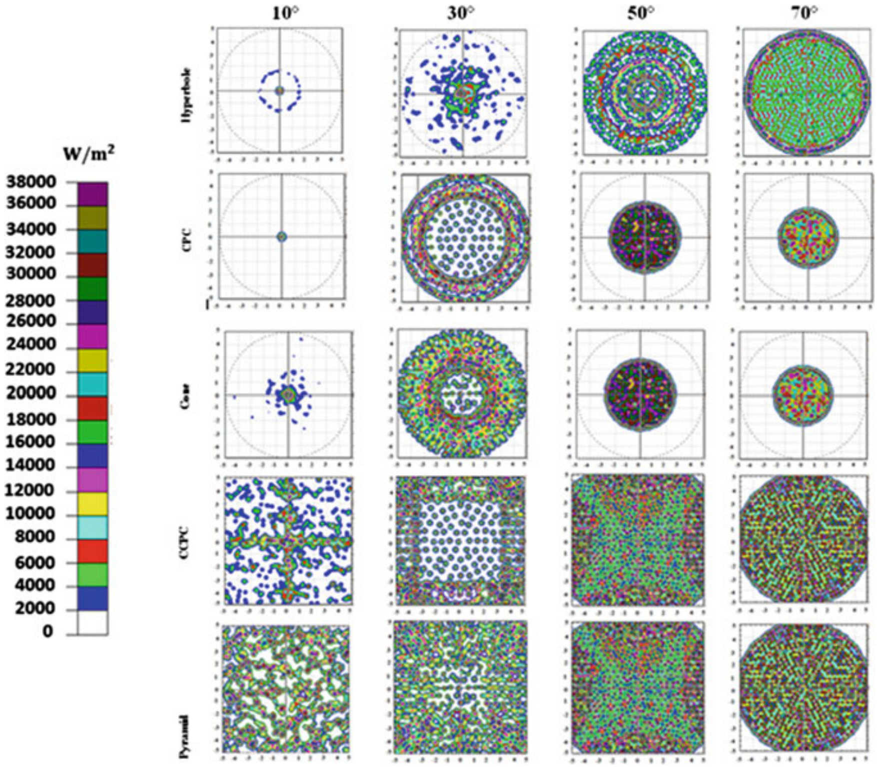


Fig. 11 Flux distribution at output of each optical element versus input angle

Table 4 Size of the optical elements with output diameter of 10 mm

Acceptance angle (°)		10	30	50	70
Length		191.65	25.98	9.67	3.75
Input diameter	CPC/CCPC	57.5877	20	6	5.32
	Cone/Pyramid	40	16.44	13.04	10.42
	Hyperboloid	57.5862	19.9991	13.0485	10.6061
Reflection number of Cone/Pyramid		7.9	3.2	2.3	2

Table 5 summarizes the optical efficiency values of the five elements for different entry angles. We can notice that the optical efficiency is high for all five elements and reach their maximum efficiency at 50°.

To summarize, we presented in this section several CPV concentrators that using one optic. We note that a one stage optic for CPV systems couldn't achieve the overall performance criteria of CPV system; either the system can allow a high efficiency while the flux distribution is not homogeneous on the cell, or the opposite.

Table 5 Optical efficiency of the five optical elements

	10	30	50	70
CPC	95.61	95.98	95.98	95.67
CCPC	90.20	94.83	95.67	95.27
Cone	96.3	95.28	95.5	95.23
Hyperbole	93.13	95.93	95.93	95.62
Pyramid	93.88	94.04	95.54	94.17

4 Two Stage CPV Concentrators

Multistage concentrators are systems that use two or more optical elements. The first one is named the Primary Optic (POE), and the second is the Secondary Optic (SOE) [15, 16].

4.1 Concentrators Based on Fresnel Lenses as a Primary Element

In this part, we will study six two-stage concentrators, dedicated to high concentration photovoltaic systems. These six concentrators are composed of the same Fresnel lens as primary optic and six of the most used as secondary ones: CPC, CCPC, Pyramid, cone, hyperbola and Dome. We use a flat circular Fresnel lens with 350 mm as diameter and focal length of 270 mm associated to a square solar cell of 10 × 10 mm^2. The irradiance used to measure the optical efficiency is 1000 W/m^2. Including wavelength (0.4–1.2 μm).

Table 6 summarizes the size of the six SOEs used, the pyramid and the cone are designed with nr = 2.3 reflections.

Table 7 summarizes the optical performance of the six optical elements as secondary optical elements, from this table it can be seen that:

Table 6 SOEs sizes

	Input diameter (mm)	Output diameter (mm)	Length (mm)
CPC	30.39	10	49.51
CCPC	30.39	10	49.51
Cone	27.94	10	49.51
Pyramid	27.94	10	49.51
Hyperbole	30.39	10	49.51
Dome	26.08	10	13.04

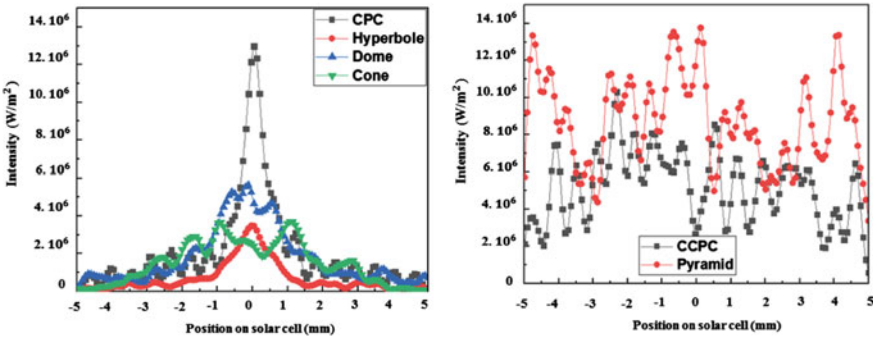
Table 7 Comparison performances of the six secondary optical elements

		Acceptance angle (°)	Optical efficiency (%)
CPC	$Z = Z_{MIN}$	1.4	81.4
	$Z = f$	1.4	83.16
	$Z = Z_{Max}$	1.4	81.32
CCPC	$Z = Z_{MIN}$	0.6	51.51
	$Z = f$	0.8	78.57
	$Z = Z_{Max}$	0.6	50.65
Cone	$Z = Z_{MIN}$	0.6	79.5
	$Z = f$	0.8	82.76
	$Z = Z_{Max}$	0.6	80.3
Pyramid	$Z = Z_{MIN}$	0.6	78.95
	$Z = f$	0.8	81.14
	$Z = Z_{Max}$	0.6	79.47
Hyperbole	$Z = Z_{MIN}$	1.2	67.11
	$Z = f$	1.2	70.37
	$Z = Z_{Max}$	1.2	68.32
Dome	$Z = Z_{MIN}$	0.6	77.16
	$Z = f$	0.75	80.59
	$Z = Z_{Max}$	0.6	76.88

- The CPC has the highest optical efficiency (83%) and the widest acceptance angle (1.4°) regardless of its position relative to the Fresnel lens.
- The cone, the pyramid and dome presents an optical efficiency which exceeds 80% but the acceptance angle is low compared to the CPC and this angle decreases if the optical elements are placed at the maximum and minimum distances.
- Hyperbola and CCPC have the lowest optical efficiency (around 78%) this is due to the number of reflections, in fact, these two forms increase the number of reflections of the solar rays inside these two elements and because of this increment, ray intensity decreases. Regarding the acceptance angle, we notice that the hyperbola has a wide acceptance angle (1.2°), thanks to its large entry opening.

Figure 12 show the repartition of the light intensity when scanning the centerline (vertical and horizontal) on the receiver obtained for the six elements. Under normal light incidence, for circular exit, we observe a high central intensity in the receiver center, which drastically decreases while moving away to the left and right borders. Regarding the square exit (Fig. 11b), the flux density is centred around 9×10^6 W/m^2 for the pyramid and 3×10^6 W/m^2 for the CCPC, with some small variations.

From this comparative study of the six optical elements associated with a Fresnel lens as POE, we found that:



(a) Circular output

(b) Square output

Fig. 12 Flux distribution at SOEs output

- CPC exhibits the highest optical efficiency and the widest acceptance angle but the poorest flux distribution.
- The pyramid presents a homogeneous flux distribution and an optical efficiency that exceeds 80% but a slightly low acceptance angle

4.2 Concentrators Based on Parabolic Mirror as a Primary Element

In this part, we will replace the Fresnel lens by a mirror parabolic as POE which associated with the same six SOEs. We use a parabolic mirror with 350 mm as diameter and focal length of 270 mm associated to a square solar cell of 10 × 10 mm^2 and we keep the same simulation characteristics as the previous section.

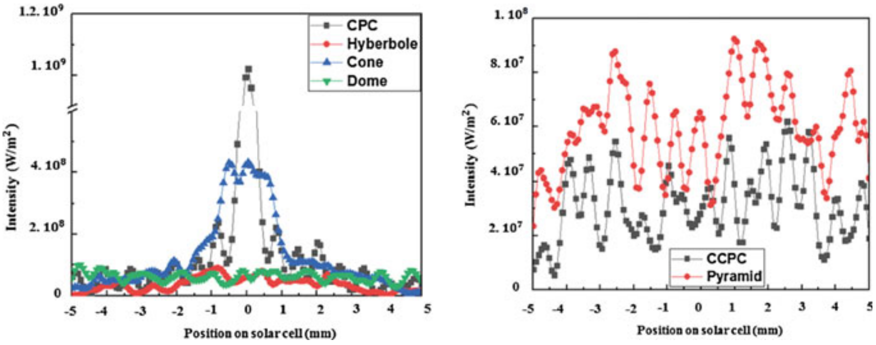
Table 8 shows the optical performance of the six optical elements as secondary optical elements, and then we can observe that using the parabolic mirror as POE, we found that.

- The cone has the highest optical efficiency (90%) but the acceptance angle is low compared to the CPC.
- The CPC, the pyramid and dome presents an optical efficiency which exceeds 80%. Concerning the acceptance angle, the CPC and the pyramid present the largest one (1.4°) in focal plan and it decreases in maximal and minimal distances. For the dome, the acceptance angle reaches only 0.8° in focal plan.
- Hyperbola and CCPC have the lowest optical efficiency (around 77%) this is due to the number of reflections, in fact, these two forms increase the number of reflections of the solar rays inside these two elements and because of this increment, ray intensity decreases. Regarding the acceptance angle, we notice that the hyperbola has a wide acceptance angle (1°), thanks to its large entry opening.

Table 8 Comparison performances of the six secondary optical elements

		Acceptance angle (°)	Optical efficiency (%)
CPC	$Z = Z_{MIN}$	1.2	81.4
	$Z = f$	1.4	85.35
	$Z = Z_{Max}$	1.2	81.32
CCPC	$Z = Z_{MIN}$	0.6	78.88
	$Z = f$	0.8	47.11
	$Z = Z_{Max}$	0.6	79.23
Cone	$Z = Z_{MIN}$	0.6	81.44
	$Z = f$	0.8	90.56
	$Z = Z_{Max}$	0.6	80.85
Pyramid	$Z = Z_{MIN}$	0.8	79.23
	$Z = f$	1.4	88.87
	$Z = Z_{Max}$	0.8	80.04
Hyperbole	$Z = Z_{MIN}$	0.8	75.68
	$Z = f$	1	76.39
	$Z = Z_{Max}$	0.8	74.22
Dome	$Z = Z_{MIN}$	0.6	82.68
	$Z = f$	0.8	83.62
	$Z = Z_{Max}$	0.6	83.2

Figure 13 show the repartition of the light intensity when scanning the centerline (vertical and horizontal) on the receiver obtained for the six elements. Under normal light incidence, for CPC and cone, we observe a high central intensity in the receiver center, which drastically decreases while moving away to the left and right borders. For the hyperbole, dome and the square exit, the flux density is centred around 6



(a) Circular output (b) Square output

Fig. 13 Flux distribution at SOEs output

× 10^7 W/m^2 for the pyramid, 3 × 10^7 W/m^2 for the CCPC, 6 × 10^7 W/m^2 for the hyperbole and 4 × 10^7 W/m^2 for the dome with some small variations.

By comparing these results with that of the previous section, we can say that the use of the parabolic mirror allows having a higher optical efficiency than the use of the Fresnel lens as POE. Regarding the flux distribution, we noticed that the pyramid and the CCPC keep the good uniformity.

The compositions (mirror + hyperbola) and (mirror + dome) make it possible to have good uniformity but the acceptance angle remains low compared to the CPV system which is based on the use of Fresnel lenses as POE.

5 Conclusion

We have analyzed two categories of CPV systems according to the 3 characteristics: acceptance angle, distribution of flux and optical efficiency, and. The first category is that the CPV system consists of a single optical element which can be reflective (parabolic mirror), refracting (Fresnel lens) and five elements based on total internal reflection (CPC, CCPC, cone, hyperbola and pyramid). The category is a CPV system consisting of two optical elements. For this test, we chose six secondaries among the most used, associated with a flat circular Fresnel lens on the one hand and, on the other hand, with the same SOEs associated with a parabolic mirror. Besides the optimal position of each secondary relative to the FL and the parabolic mirror is determined to find the best placement. As simulation results, we noticed that the optical efficiency of mirror parabolic is higher by 10% comparing to Fresnel lens, this is due to the fact that in the prisms of the Fresnel lens, the solar rays are reflected by the TIR, we found also that the CPC exhibits the highest optical efficiency and the widest acceptance angle but the poorest flux distribution. The pyramid presents a homogeneous flux distribution and an optical efficiency that exceeds 80% but a slightly low acceptance angle. Then, we can say that a single stage for CPV systems don't get the three performance keys of a CPV system; either the system allows having a high optical efficiency while the distribution of flux is not uniform on the cell, or reverse. Concerning the two stages, and by comparing simulation results, we found that using the parabolic mirror allows having a higher optical efficiency than using Fresnel lens as primary. Apropos, the distribution of the flux, we noticed that the pyramid and the CCPC keep a good uniformity. The combination (mirror + hyperbola) and (mirror + dome) make it possible to have good uniformity but the acceptance angle remains low compared to the CPV system which is based on the use of Fresnel lenses as POE.

References

1. Ning X, O'Gallagher J, Winston R (1987) Optics of two-stage photovoltaic concentrators with dielectric second stages. App Opt 26:1207–1212
2. Buljan M, Miñano JC, Benítez P, Mohedano R, Chaves J (2014) Improving performances of Fresnel CPV systems: Fresnel-RXI Köhler concentrator. Opt Express 22:A205–A210
3. Victoria M, Domínguez C, Antón I, Sala G (2009) Comparative analysis of different secondary optical elements for aspheric primary lenses. Opt Express 17(8):315, 6487–6492
4. Chen YC, Chian HW (2005) Design of the secondary optical elements for concentrated photovoltaic units with Fresnel lenses. Appl Sci 5(4):770–786
5. Ferrer-Rodríguez JP, Baig H, Fernández EF, Almonacid F, Mallick T, Pérez-Higueras P (2017) Optical modeling of four Fresnel-based high-CPV units. Sol Energy 155:805–815
6. Ferrer-Rodríguez JP, Fernández EF, Baig H, Almonacid F, Mallick T, Pérez-Higueras P (2018) Development, indoor characterisation and comparison to optical modelling of four Fresnel-based high-CPV units equipped with refractive secondary optics. Sol Energy Mater Sol Cells 186:273–283
7. El Himer S, Ahaitouf A, El-Yahyaoui S, Mechaqrane A, Ouagazzaden A (2018) A comparative of four secondary optical elements for CPV systems. In: 14th international conference on concentrator photovoltaic systems (CPV-14), AIP conference proceedings 2012 030003-1-030003-7
8. El Himer S, El-Yahyaoui S, Benmohammadi Z, Mechaqrane A, Ahaitouf A (2016) Performance analysis and comparison of the CCPC and pyramid shaped solar concentrators for CPV, Technologies and materials for renewable energy, environment and sustainability AIP conference proceedings 1758, 020004-1-020004-8. https://doi.org/10.1063/1.4959380
9. El-Yahyaoui S, Ahaitouf A, El Himer S, Mechaqrane A (2019) Indoor characterization of pyramid- and cone-type secondary optics. AIP Conf Proc 2149:070003. https://doi.org/10.1063/1.5124202
10. Chen YC, Chiang HW (2015) Design of the secondary optical elements for concentrated photovoltaic units with9 Fresnel lenses. App Sci 5:770–786
11. Sellami N, Mallick TK, McNeil DA (2012) Optical characterisation of 3-dstatic solar concentrator. Energy Convers Manage 64:579–586
12. James LW (1989) Use of refractive secondaries in photovoltaic concentrators. Tech. Rep. SAND, 89-7029, Sandia National Laboratories, Alburquerque, NM, USA
13. Leutz R, Suzuki A, Akisawa A, Kashiwagi T (1999) Design of a non-imaging Fresnel lens for solar concentrators. Sol Energy 65:379–387
14. Ritou A (2016) Développement, fabrication et characterisation de modules photovoltaïques à concentration à ultra haut rendement à base de micro-concentrateurs. PhD thesis Grenoble Alpes University
15. El Himer S, Ahaitouf A, El-yahyaoui S, Mechaqrane A (2018) Parametric optimization and performances analysis of four secondary optical elements for concentrator photovoltaic systems. Int J Renew Energy Res IJRER 8:2289–2298
16. Ahaitouf A, Chevallier C, Salvestrini JP, Ougazzaden A (2014) Contribution to solar concentrators design for photovoltaic application. In: Proceedings of the 2014 international renewable and sustainable energy conference (IRSEC), Ouarzazate, Morocco, 17–19 October, pp 78–81

Artificial Intelligence Based on Particle Swarm Optimization for Optimal Wind Turbine Power Control Using Doubly Fed Induction Generator

Elmostafa Chetouani, Youssef Errami, Abdellatif Obbadi, and Smail Sahnoun

Abstract The system under investigation is a wind turbine of 5 MW connected via a gearbox to a doubly-fed induction generator (DFIG). The stator of this DFIG is connected directly to the grid, while the rotor uses back-to-back converters to connect to the grid. This chapter focuses on controlling the active and reactivepower generated by a variable wind power plant and the power transferred between the electrical grid and the system. A maximum power point tracking (MPPT) technique is also utilized to get the maximum power of the fluctuating wind speed. The rotor side converter (RSC) and grid side converter (GSC) decoupled vector control is principally established by a traditional Proportional-Integral (PI) and with an intelligent PI whose parameters are modified using the particle swarm optimization technique (PSO). Through Matlab/Simulink, the performances and results obtained by classical PI are studied and compared to those obtained by PSO tuned PI controller.

Keywords AI · DFIG · GSC · MPPT · PSO · PI-controller · RSC · WECS

1 Introduction

Around the world, societies are on the point of experiencing a significant and desperately needed revolution in how they create and utilize energy. This transition is shifting the world away from fossil fuel consumption (which contributes to climate change and other environmental and social issues) and toward cleaner,

E. Chetouani (✉) · Y. Errami · A. Obbadi · S. Sahnoun
Laboratory: Electronics, Instrumentation and Energy – Team: Exploitation and Processing of Renewable Energy – Department of Physics, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco

Y. Errami
e-mail: errami.y@ucd.ac.ma

A. Obbadi
e-mail: obbadi.a@ucd.ac.ma

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_8

renewable energy sources. Many reasons have supported the rapid adoption of renewable energy, including driving economic development, improving energy security, expanding energy access, and reducing climate change. Most countries, including Morocco, already use Renewable Energy (RE) to generate electricity for their daily life (Internet, Electric Cars, heat and cool buildings, etc.). This transition is also going with the technology development, which has been known great importance, thanks to the immense progress made in scientific research, especially in the Artificial Intelligence (AI) domain. Wind energy is a renewable energy source that is created by converting the kinetic energy of the wind. Intensive research and investigations into wind systems have recently resulted in a variety of wind energy configurations [1]. The most common variable speed Wind Energy Conversion System (WECS) architecture is one with a Doubly-Fed Inductor Generator (DFIG) whose stator is connected directly to the power grid. However, the back-to-back converters connect the rotor to the grid, allowing for bidirectional power transfer, as seen in Fig. 1. The benefit of this configuration is that the utilized converters are designed to allow only a percentage of the system's total power to flow through, reducing losses in power electronic components [1]. The rotor side converter (RSC) and grid side converter (GSC) are the two components of these converters, both of which are made up of Insulated-Gate Bipolar Transistors (IGBTs) switches and operate using the Sinusoidal Pulse Width Modulation (SPWM) methodology [2]. The design of an acceptable control scheme is a difficult and challenging issue due to the nonlinear characteristics of the Doubly Fed-Induction Generator. For controlling the powers of the DFIG, the vector control based on the Proportional-Integral (PI) controller is a widely used technique. However, this technique has limits because the PI controller gains are very dependent on the generator characteristics. As a result, any change in the generator's setting has an impact on the control's performance [3]. Various artificial intelligence-based control systems for controlling the power of DFIG have recently been developed

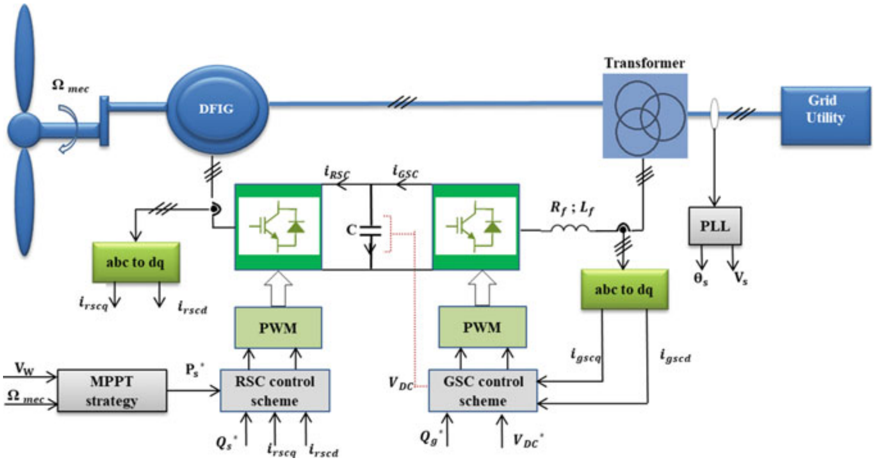


Fig. 1 Configuration of a wind energy converter system (WECS) using a DFIG

in the literature (Artificial Neural Networks (ANN), Fuzzy logic Controller (FLC), Particle swarm optimization algorithms, etc.) to overcome vector control disadvantages. In [4], a neural network controller has been established to control the vector control based on PI controller for controlling the dynamic response of the DFIG based variable speed wind turbine. The authors have been proposed a comparative analysis of the suggested control with the PI controller and RST polynomial controller to demonstrate the effectiveness of the NN controller. Besides, the authors of [5] used AI based on adaptive neuron fuzzy inference system (ANFIS) to control and improve the DFIG's performance in the face of grid defects. This method was compared to the Fuzzy Logic Controller (FLC) and the Back Propagation Network (BPN) by the authors. In addition, an artificial intelligence fractional-order robust control has been proposed and compared with Sliding Mode in [6].

The PSO has been broadly utilized to pilot the powers of the DFIG. This control method minimizes the calculating time, keeps excellent precision, is robust against the parameters of system variation, and can be implemented in a low-cost microcontroller [7]. Also, it has few parameters to adapt, can run the parallel calculation, it can be powerful, it has superior probability and efficiency in finding the global optima, it can meet the solution quickly, it does not overlap and mutate, it has brief computational time, and it can be powerful in solving problems showing difficulty to get correct mathematical models. In [8], an optimal PID regulator for controlling the active and reactive power of DFIG by using the PSO algorithm has been suggested.

Several control systems have been developed to accomplish the maximum power point (MPP) approach, which can be classified into two groups [9]. To develop the optimal speed reference, the first one required knowledge of the characteristic aerodynamic curve of wind turbine speed, whereas the second one does not require any information about wind speed. The MPP with Optimal Tip Speed Ratio (OPTSR), MPP with Optimal Torque Control (OTC), MPP with Power Signal Feedback (PSF), and MPP with Perturbation and Observation (P&O) are examples of these techniques [10]. The purpose of this chapter is to use particle swarm optimization to regulate active and reactive power. Furthermore, a comparison of the PSO algorithm and the Open Loop Indirect Field Oriented approach for managing the wind energy system based on the DFIG is provided. The following is how the chapter is organized: The modeling of the wind energy system is done in Sect. 2. The MPP with a speed control approach is proposed in Sect. 3. Indirect Field-oriented control and the Particle Swarm optimization approach are used to control the powers in Sects. 4 and 5, respectively. Section 6 compares and contrasts the simulation findings. Section 7 concludes with a conclusion and some perspectives.

2 Wind Power Plant System Components Modeling

A Wind Turbine (WT), a generator, power electronics converters, DC Link Capacitor, filter circuits, and a control system are typical components of a Wind Energy

Conversion System based on Doubly Fed Induction Generator. The configuration of the DFIG-based WECS is depicted in Fig. 1.

2.1 Wind Turbine Modeling

WT is the most significant component of WECS, as it generates electrical energy from the kinetic energy of wind speed. The wind's aerodynamic power is given as follows [11]:

$$P_{Tu} = \frac{1}{2}.Cp(\lambda, \beta).\rho.\pi.R^2.V_w^3 \tag{1}$$

The turbine efficiency performance coefficient of power is denoted by Cp. The empirical formula (2) can be used to calculate this coefficient. It is determined by the pitch angle (β) and the ratio of the blades' linear speed to the wind speed (λ), which is stated in Eq. (3) [11]:

$$\begin{aligned} C_P(\beta, \lambda) = [0.5 - 0.0167.(\beta - 2)].\sin\left(\frac{\pi(\lambda + 0.1)}{18.5 - 0.3.(\beta - 2)}\right) \\ - 0.00184.(\lambda - 3).(\beta - 2) \end{aligned} \tag{2}$$

$$\lambda = \frac{R.\Omega_{Tu}}{V_w} \tag{3}$$

Supposing that the overall mechanical dynamics of the system are brought back to the turbine shaft, the following equations describe the model [12]:

$$J_{tot}.\frac{d\Omega_{mec}}{dt} + f.\Omega_{mec} = T_g - T_{em} \tag{4}$$

$$T_g = \frac{T_{Tu}}{G_B} \quad \text{and} \quad G_B = \frac{\Omega_{mec}}{\Omega_{Tu}} \tag{5}$$

The wind turbine mechanical torque output T_{Tu} is written as below:

$$T_{Tu} = \frac{P_{Tu}}{\Omega_{mec}} \tag{6}$$

where the overall inertia of WECS is J_{tot}, T_{Tu} is the turbine torque, T_g is the gearbox torque, Tem is the electromagnetic torque of the DFIG, and f is the overall viscosity coefficient of friction.

2.2 Doubly Fed Induction Generator Modeling

The doubly-fed induction generator is known for its reliability and is used in a variety of industrial applications, but it has a complicated equation system to understand and manage. Using some simplifying assumptions and the Park transformation, a simpler model of the DFIG may be obtained, allowing this complexity to be removed.

Assuming d-q axis rotating with the synchronous speed (ω_s), the generator voltage and flux equations are given by Chetouani et al. [11].

The stator and rotor voltages equations can be formulated as follows:

$$\begin{cases} V_{sd} = R_s.i_{sd} + \frac{d\varphi_{sd}}{dt} - \omega_s\varphi_{sq} \\ V_{sq} = R_s.i_{sq} + \frac{d\varphi_{sq}}{dt} + \omega_s\varphi_{sd} \\ V_{rd} = R_r.i_{rd} + \frac{d\varphi_{rd}}{dt} - \omega_r\varphi_{rq} \\ V_{sq} = R_r.i_{rq} + \frac{d\varphi_{rq}}{dt} + \omega_r\varphi_{rd} \end{cases} \tag{7}$$

The stator and rotor field magnetic flux equations can be written as follows [11]:

$$\begin{cases} \varphi_{sd} = L_s.i_{sd} + M.i_{rd} \\ \varphi_{sq} = L_s.i_{sq} + M.i_{rq} \\ \varphi_{rd} = L_r.i_{rd} + M.i_{sd} \\ \varphi_{rq} = L_r.i_{rq} + M.i_{sq} \end{cases} \tag{8}$$

The electromagnetic torque can be given as follows [11]:

$$T_{em} = -p\frac{M}{L_s}(i_{rq}\varphi_{sd} - i_{rd}\varphi_{sq}) \tag{9}$$

2.3 Modeling of the RSC

The RSC is a three-level converter, with each arm consisting of two complimentary regulated switches. The RSC converter's input voltages between phases can be characterized as a function of the voltage V_{DC} and the interrupt states Sj [13]:

$$\begin{cases} U_{ab} = (S_a - S_b).V_{DC} \\ U_{bc} = (S_b - S_c).V_{DC} \\ U_{ca} = (S_c - S_a).V_{DC} \end{cases} \tag{10}$$

The simple input voltages (V_a, V_b, V_c) equations can be written as follows [13]:

$$\begin{cases} V_a = \frac{2.S_a - S_b - S_c}{3}.V_{DC} \\ V_b = \frac{2.S_b - S_a - S_c}{3}.V_{DC} \\ V_c = \frac{2.S_c - S_a - S_b}{3}.V_{DC} \end{cases} \tag{11}$$

where (S_a, S_b, S_c) present the control signals of the switches. From Eq. (11), the model the RSC can be given under the matrix presented below:

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \frac{V_{DC}}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} . \begin{bmatrix} s_a \\ s_b \\ s_c \end{bmatrix} \tag{12}$$

The system equation presented in Eq. (12) is utilized to model the RSC under Matlab/Simulink.

2.4 Modeling of the GSC and the DC Link

Through a Resistance-Inductance filter, the GSC is connected to the power grid. However, describing the grid side converter and DC-link components is required for a better understanding of GSC control. The converter's receptor convention is taken into account, as shown in Fig. 2. The mathematical model of the filter (R_f, L_f) can be written as follows [14]:

$$\begin{cases} V_{gsca} = -R_f.i_{gsca} - L_f.\frac{di_{gsca}}{dt} + V_{ga} \\ V_{gscb} = -R_f.i_{gscb} - L_f.\frac{di_{gscb}}{dt} + V_{gb} \\ V_{gscc} = -R_f.i_{gscc} - L_f.\frac{di_{gscc}}{dt} + V_{gc} \end{cases} \tag{13}$$

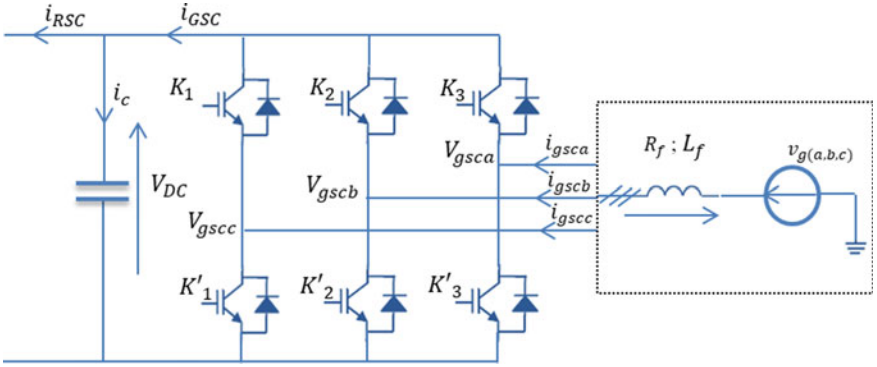


Fig. 2 Grid Side Converter associated with (R_f, L_f) filter

The simple input voltages (V_{gsca}, V_{gscb}, V_{gscc}) equations can be represented as follows [13]:

$$\begin{cases} V_{gsca} = \frac{2.S_a - S_b - S_c}{3}.V_{DC} \\ V_{gscb} = \frac{2.S_b - S_a - S_c}{3}.V_{DC} \\ V_{gscc} = \frac{2.S_c - S_a - S_b}{3}.V_{DC} \end{cases} \tag{14}$$

The system Eq. (13), which provide the mathematical formulation of the GSC in the d-q synchronous frame, can be written by using the park and Laplace transformations as follows [13]:

$$V_{gscd} = -[R_f + L_f.s].i_{gscd} + \omega_g.L_f.i_{gscq} + V_{gd} \tag{15}$$

$$V_{gscq} = -[R_f + L_f.s].i_{gscq} - \omega_g.L_f.i_{gscd} + V_{gq} \tag{16}$$

The components of Park of the converter input voltages are V_{gscd} and V_{gscq}, and the grid voltages in the d-q referee are V_{gd} and V_{gq}. Equations (15) and (16) can be written by using the park transformation as follows:

$$V_{DC}.\ S_d = -[R_f + L_f.s].i_{gscd} + \omega_g.L_f.i_{gscq} + V_{gd} \tag{17}$$

$$V_{DC}.S_q = -[R_f + L_f.s].i_{gscq} - \omega_g.L_f.i_{gscd} + V_{gq} \tag{18}$$

where S_d and S_q are the GSC converter's switching functions in the d-q referee, which are represented as follows using the Park transformation:

$$S_d = \frac{1}{\sqrt{6}}(2.s_a - s_b - s_c).\cos(\omega t) + \frac{1}{\sqrt{6}}(s_b - s_c).\sin(\omega t) \tag{19}$$

$$S_q = \frac{1}{\sqrt{6}}(s_b - s_c).\cos(\omega t) - \frac{1}{\sqrt{6}}(2.s_a - s_b - s_c).\sin(\omega t) \tag{20}$$

The electric power transferred between the grid and GSC can be calculated as follows [14]:

$$P_g = V_{gd}.i_{gscd} + V_{gq}.i_{gscq} \tag{21}$$

$$Q_g = V_{gq}.i_{gscd} - V_{gd}.i_{gscq} \tag{22}$$

The following equations were used to simulate the bus continuous:

$$i_C = i_{GSC} - i_{RSC} \tag{23}$$

$$V_{DC} = \frac{1}{C.S}.i_C \tag{24}$$

3 Maximum Power Point Tracking Methodology

In this chapter, the Maximum Power Point Tracking (MPPT) technique is used to maximize the extracted power under normal conditions. This approach regulates the rotation speed by controlling the DFIG's electromagnetic torque. Control without speed regulation and control with speed regulation are the two forms of MPPT control. The first relies entirely on the generator's rotation speed to determine the best turbine speed for generating maximum generator torque. The wind speed has been calculated. The second is to use the PI controller to keep the DFIG rotation speed at a constant reference speed. An anemometer is used to measure the wind speed in this mode. If the coefficient C_p is optimized, the reference speed is maximized. The relative speed λ is the most significant parameter to optimize in this regard.

The second technique is investigated in this chapter. The electromagnetic torque Tem created by the DFIG is equal to its optimal (reference) value $T_{em\text{-}opt}$ imposed by the command [15]:

$$T_{em} = T_{em\text{-}opt} \tag{25}$$

As illustrated in Fig. 3, the optimum electromagnetic torque Tem-opt for obtaining a rotation speed equal to the optimal speed is:

$$T_{em\text{-}opt} = \left[K_p + K_i.\frac{1}{S}\right].\left[\Omega_{mec\text{-}opt} - \Omega_{mec}\right] \tag{26}$$

The PI controller parameters K_P and K_i are used here. The following equations produce the optimal speed ($\Omega_{mec\text{-}opt}$):

$$\Omega_{mec\text{-}opt} = G_B.\Omega_{Tu\text{-}opt} \tag{27}$$

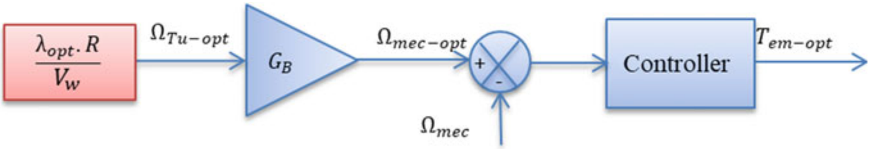


Fig. 3 MPPT scheme with speed regulation

$$\Omega_{\text{Tu-opt}} = \frac{V_{\text{wind}}.\lambda_{\text{opt}}}{R} \tag{28}$$

where G_B is the gearbox coefficient used to match the slow speed of the turbine shaft to the high speed of DFIG.

3.1 Classical PI Controller for MPPT

Figure 4 depicts the rotational speed regulation loop. The turbine torque (T_{Tu}) is seen as a disturbance that the controller must compensate for. The latter is used to achieve a zero static error and lower response time while keeping in mind the system's stability, which must be ensured.

The pole compensation method is used to determine the PI controller parameters. The system's time constant is written as follows:

$$T_{\text{sys}} = \frac{J_{\text{tot}}}{f} \tag{29}$$

Because of the J_{tot} and f values, the system's dynamic is quite slow. As a result, the system constant time is divided by 10^3 to provide a quick response.

$$K_{\text{imppt}} = \frac{1}{\tau.f}; \quad \text{with } \tau = \frac{T_{\text{sys}}}{10^3} \tag{30}$$

$$K_{\text{pmppt}} = \frac{-K_{\text{imppt}}.J_{\text{tot}}}{f} \tag{31}$$

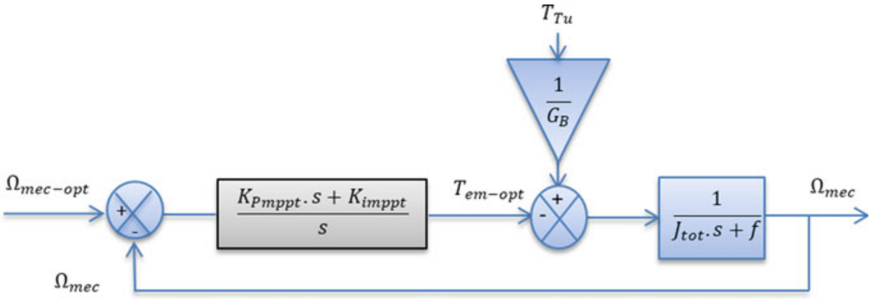


Fig. 4 Rotation speed regulation loop

4 Vector Control of the DFIG-Based Wind Energy Converter

4.1 Control of the RSC

The orientation of the stator flux vector, which is oriented along the axis d, is used to operate the Rotor Side Converter (RSC). The voltage in the electrical network is assumed to be constant, and the resistance of the stator windings is neglected [11].

$$\varphi_{sd} = \varphi_s,\ \varphi_{sq} = 0 \tag{32}$$

$$V_{sd} = 0, \quad V_{sq} = V_s = \omega_s.\varphi_s \tag{33}$$

The rotor voltages can be expressed as follows [11, 16]:

$$V_{rd} = \left[R_r + \left(L_r - \frac{M^2}{L_s}\right).s\right]i_{rd} - g.\omega_s\left(L_r - \frac{M^2}{L_s}\right)i_{rq} \tag{34}$$

$$V_{rq} = \left[R_r + \left(L_r - \frac{M^2}{L_s}\right).s\right]i_{rq} + g.\omega_s\left(L_r - \frac{M^2}{L_s}\right)i_{rd} + g\frac{V_sM}{L_s} \tag{35}$$

The rotor and stator current expressions are derived from Eqs. 7 and 8:

$$i_{rd} = \left[V_{rd} + g.\omega_s\left(L_r - \frac{M^2}{L_s}\right)i_{rq}\right] \Big/ \left[R_r + \left(L_r - \frac{M^2}{L_s}\right).s\right] \tag{36}$$

$$i_{rq} = \left[V_{rq} - g.\omega_s\left(L_r - \frac{M^2}{L_s}\right)i_{rd} - g.\frac{V_sM}{L_s}\right] \Big/ \left[R_r + \left(L_r - \frac{M^2}{L_s}\right).s\right] \tag{37}$$

$$i_{sd} = -\frac{M}{L_s}i_{rd} + \frac{\varphi_s}{L_s} \tag{38}$$

$$i_{sq} = -\frac{M}{L_s}i_{rq} \tag{39}$$

The goal is to control the active and reactive power pumped into the grid from the stator independently. The powers of the stator can be stated as follows [16]:

$$P_s = -V_{sq}\frac{M}{L_s}i_{rq} = -V_s\frac{M}{L_s}i_{rq} \tag{40}$$

$$Q_s = \frac{V_{sq}^2}{\omega_s L_s} - V_{sq}\frac{M}{L_s}i_{rd} = \frac{V_s^2}{\omega_s L_s} - V_s\frac{M}{L_s}i_{rd} \tag{41}$$

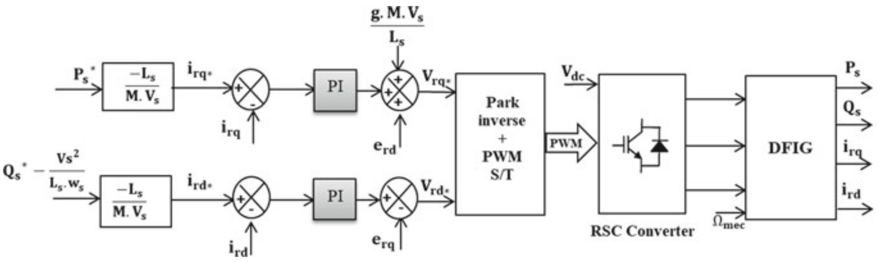


Fig. 5 Scheme of RSC control

An open loop controls the active and reactive power, as shown in Fig. 5. The rotor currents (i_{rq}, i_{rd}) are, nevertheless, controlled by a closed loop. Indirect vector control is the name given to this method, which is based on a two-series PI controller. The current $i_{rq}{}^{*}$ is generated from the input power reference to control active power. Regulating $i_{rd}{}^{*}$, which is computed from the $Q_s{}^{*}$, controls the reactive power. Equations (40) and (41) can be used to derive the following current references [17]:

$$i_{rq}^{*} = -\frac{L_s}{M.V_{sq}}.P_s^{*} \tag{42}$$

$$i_{rd}^{*} = -\frac{L_s}{M.V_{sq}}.\left(Q_s^{*} - \frac{V_{sq}^{2}}{w_s L_s}\right) \tag{43}$$

The maximum extracted power from the turbine is $P_s{}^{*}$, which is calculated using the MPPT algorithm ($P_s{}^{*} = P_{Tu}{}^{*}$). For a null reactive power transfer, the reactive power reference ($Q_S{}^{*}$) is set to zero. The voltage references can be simply defined using Eqs. (34) and (35) in the following order:

$$V_{rq}^{*} = \left[i_{rq}^{*} - i_{rq}\right].\left[K_{p\text{-}rsc1} + K_{i\text{-}rsc1}.\frac{1}{S}\right] + e_{rd} + Vs' \tag{44}$$

$$V_{rd}^{*} = \left[i_{rd}^{*} - i_{rd}\right].\left[K_{p\text{-}rsc2} + K_{i\text{-}rsc2}.\frac{1}{S}\right] + e_{rq} \tag{45}$$

where

$$e_{rd} = g.\omega_s.\left(L_r - \frac{M^2}{L_s}\right).i_{rd}, \quad e_{rq} = g.\omega_s.\left(L_r - \frac{M^2}{L_s}\right).i_{rq}, \text{ and}$$
$$Vs' = g.\frac{V_s.M}{L_s} \tag{46}$$

The electromotive forces e_{rd} and e_{rq} represent a decoupling term that must be compensated.

4.1.1 Determination of the PI Controller’s Gains

The settings of the PI controllers in Fig. 5 are identical. The following is the definition of the system time constant:

$$\mathrm{T_s} = \left(\mathrm{L_r} - \frac{\mathrm{M^2}}{\mathrm{L_s}}\right)/\mathrm{R_r} \tag{47}$$

The PI parameters can be found as follows using $\mathrm{T_{rsc}} = \frac{\mathrm{T_s}}{100}$ for a rapid reaction of the system and the pole compensation technique:

$$\mathrm{K_{prsc}} = \frac{1}{\mathrm{T_{rsc}}}.\left(\mathrm{L_r} - \frac{\mathrm{M^2}}{\mathrm{L_s}}\right) \tag{48}$$

$$\mathrm{K_{irsc}} = \frac{\mathrm{K_{prsc}}.\mathrm{R_r}}{\left(\mathrm{L_r} - \frac{\mathrm{M^2}}{\mathrm{L_s}}\right)} \tag{49}$$

4.2 Control of the GSC

Controlling DC-link voltage and reactive power exchanged with the grid is the job of the grid side converter. The grid voltage is assumed to be oriented to the q-axis axis in order to elaborate the control scheme of the GSC converter shown in Fig. 6. As a result, the grid voltages can be written as:

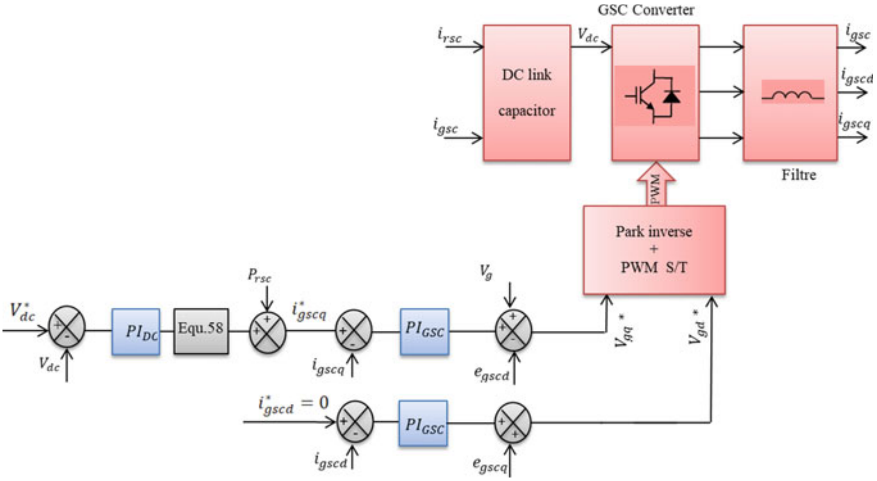


Fig. 6 Scheme of GSC control

$$V_{gd} = 0 \quad \text{and} \quad V_{gq} = V_g \tag{50}$$

The current reference frame is regulated to gain control. As a result, Eqs. (15), (16), (21), and (22) for voltages and powers can be simplified as follow:

$$V_{gscd} = -[R_f + L_f.s].i_{gscd} + \omega_g.L_f.i_{gscq} \tag{51}$$

$$V_{gscq} = -[R_f + L_f.s].i_{gscq} - \omega_g.L_f.i_{gscd} + V_g \tag{52}$$

$$P_g = V_g.i_{gscq} \tag{53}$$

$$Q_g = V_g.i_{gscd} \tag{54}$$

The relationship between the powers of converters can be represented as shown in Eq. (55) by multiplying Eq. (23) by the DC-link voltage (V_{DC}) [14]:

$$V_{DC}.i_C = P_{GSC} - P_{RSC} \tag{55}$$

The following is how the grid side converter power can be computed:

$$P_g = P_{GSC} = V_{DC}.i_C + \ P_{RSC} \tag{56}$$

where P_{RSC} denotes the power of the rotor side converter, which is defined as:

$$P_{RSC} = V_{DC}.i_{RSC} \tag{57}$$

As a result, the DC-link power ($P_{dc}{}^{*}$) can be written as:

$$P^{*}_{dc} = V_{DC}.i^{*}_{c} \tag{58}$$

The grid current references can be calculated using Eqs. (50), (51), (53), and (54) as follows [14]:

$$i^{*}_{gscq} = \frac{1}{V_g}.\left(V^{*}_{DC}.i^{*}_{c} + \ P_{RSC}\right) \tag{59}$$

$$i^{*}_{gscd} = \frac{Q^{*}_{g}}{V_g} \tag{60}$$

The voltage references are written like this:

$$V^{*}_{gd} = \left[i^{*}_{gscd} - i_{gscd}\right].\left[K_{p\text{-}gsc2} + K_{i\text{-}gsc2}.\frac{1}{S}\right] + e_{gscq} \tag{61}$$

$$V^{*}_{gq} = \left[i^{*}_{gscq} - i_{gscq}\right].\left[K_{p\text{-}gsc1} + K_{i\text{-}gsc1}.\frac{1}{S}\right] - e_{gscd} + V_{g} \tag{62}$$

where:

$$e_{gscq} = \omega_{g}.L_{f}.i_{gscq} \quad \text{and} \quad e_{gscd} = \omega_{g}.L_{f}.i_{gscd} \tag{63}$$

The grid current expressions can be obtained from Eqs. (17) and (18) as shown below:

$$i_{gscq} = \frac{1}{[R_{f} + L_{f}.s]}.\left(V^{*}_{gq} - \omega_{g}.L_{f}.i_{gscd} - V_{DC}.S_{q}\right) \tag{64}$$

$$i_{gscd} = \frac{1}{[R_{f} + L_{f}.s]}.\left(V^{*}_{gd} + \omega_{g}.L_{f}.i_{gscdq} - V_{DC}.S_{d}\right) \tag{65}$$

4.2.1 Determination of the PI_{DC} Controller's Gains

The PI_{DC} controller is used to keep the DC-link voltage at its reference, as shown in Fig. 6. As a result, the following are the PI controller (K_{p-DC}, K_{i-DC}):

$$K_{p-DC} = 2.\xi.\omega.c, \tag{66}$$

where ξ is the damping coefficient

$$K_{i\text{ - }DC} = \omega^{2}.c \tag{67}$$

4.2.2 Determination of the PI_{GSC} Controller's Gains

The PI_{GSC} controller in inner loop regulates the current i_{gscq} and i_{gscd} passing through the Resistance-Inductance (RL) filter. As a result, the controlled system's temporal constant is stated as:

$$T_{s} = \frac{L_{f}}{R_{f}} \tag{68}$$

The PI controller parameters (K_{pgsc}, K_{igsc}) are as follows when $T = \frac{T_s}{10}$ is used for a quick system dynamic response:

$$K_{pgsc} = \frac{L_{f}}{T} \tag{69}$$

$$K_{igsc} = \frac{R_f}{T} \tag{70}$$

5 Optimal PI Tuning Gains Using PSO Algorithm

5.1 Overview of PSO

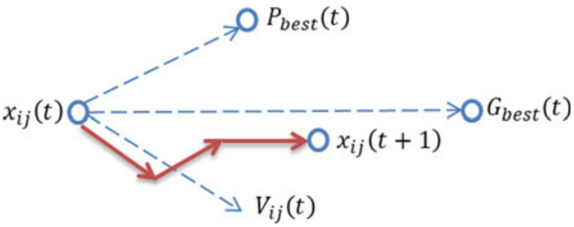
Particle swarm optimization (PSO) is an evolutionary meta-heuristic computation technique inspired by examining the behavior of bird flocking or fish schooling when they are hunting for food or keeping away from enemies [17, 18]. Kennedy and Eberhart discovered it for the first time in 1995. The PSO is an artificial intelligence technique based on the study of decentralized and self-organized collective behavior. The population candidate solutions are used in the PSO technique to produce an optimal solution for the problem [18]. A fitness function is used to determine the optimality level. The PSO algorithm works on the principle that each particle is searching for the objective in D-dimensional space. At high speed, the other particles will adopt the best and most ideal position identified. The swarm is initially spread randomly in the search space, with each particle moving at a random speed. At each step, each particle tries to remember its best position and asks its neighbors for the swarm's global best position, which it adopts with a suitable velocity [18].

5.2 PSO Mathematical Model

Each swarm particle x_{ij} is represented as a point in the Cartesian coordinate system, randomly specified by beginning velocity and position, as shown in Fig. 7. To reach their goal, the particle searches the research space through iterative test placements. The mathematical expressions are given in [18, 19] as:

$$\begin{aligned} V_{ij}(t+1) = W.V_{ij}(t) + C_1.r_1.\big(\mathrm{Pbest}_{ij} - x_{ij}(t)\big) \\ + C_2.r_2.\big(\mathrm{Gbest}_j - x_{ij}(t)\big) \end{aligned} \tag{71}$$

Fig. 7 Update diagram of each particle in the swarm



$$x_{ij}(t+1) = V_{ij}(t+1) + x_{ij}(t) \tag{72}$$

where $V_{ij}(t)$ denotes the particle's velocity with a j dimension at iteration t, $x_{ij}(t)$ denotes the particle's location with a j dimension at iteration t, P_{best} denotes the particle's best previous position, and G_{best} denotes the best particle in the population. W is the inertia weight factor that controls the exploration and exploitation of the search because it dynamically adjusts velocity. (C_1 and C_2) are the cognitive and social parameters' acceleration constants (positive constants), (r1 and r2) are random values in the range [0–1], d is the problem's dimension, and n is the swarm's size.

5.3 Tuning PI Parameters Using PSO

The particle swarm optimization approach determines the optimal PI parameters [K_p, K_i]. To ensure the system's stability in a closed loop, each swarm particle aims towards a minimum of the fitness function. Integral Absolute Error (IAE), Integral Square Error (ISE), Integral Time Square Error (ITSE), and Integral Time Absolute Error (ITAE) are some of the performance indices that can be utilized to determine the parameters [20]. The Integral Absolute Error criterion (IAE) is employed as the fitness function for the algorithm and is defined as follows:

$$Fitness = IAE = \int_0^{\infty} |e(t)|.dt \tag{73}$$

The signal error e(t) is calculated as a difference between the reference input and the actual output value, as shown in Fig. 8, which represents the block diagram of the approach for optimal controller design. For the provided errors in Eq. (74) the IAE criterion is used to define five objective functions.

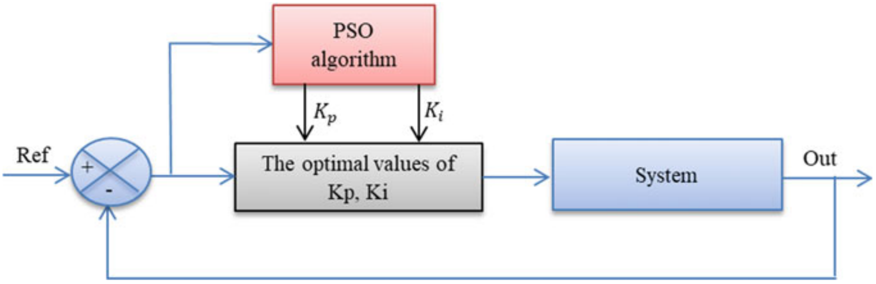


Fig. 8 Block diagram of the each PI controller with PSO

$$\begin{cases} \mathrm{e_{RSC1}} = \mathrm{i_{rq\text{-}ref}} - \mathrm{i_{rq}} \\ \mathrm{e_{RSC2}} = \mathrm{i_{rd\text{-}ref}} - \mathrm{i_{rd}} \\ \mathrm{e_{GSC1}} = \mathrm{i_{rq\text{-}ref}} - \mathrm{i_{rq\text{-}grid}} \\ \mathrm{e_{GSC2}} = \mathrm{i_{rd\text{-}ref}} - \mathrm{i_{rd\text{-}grid}} \\ \mathrm{e_{dc}} = \mathrm{V_{dc\text{-}ref}} - \mathrm{V_{dc}} \end{cases} \tag{74}$$

A PI controller for DC link voltage, two controllers for the rotor side converter (RSC), and two controllers for the grid side converter control the wind energy conversion system (GSC). There are two parameters to optimize for each controller (K_p, K_i). The original idea for establishing the PSO code consists of creating five swarms of particles, each with its objective function to reach.

5.4 Design of the Algorithm

The steps of the PSO technique's searching operation are as follows [21]:

- **Step 1**: Create an initial population at random.
- **Step 2**: Once a certain number of iterations have been completed, the algorithm is terminated.
- **Step 3**: Evaluate each particle's objective function and register each particle's best previous position (Pi) as well as the global best position (Pg).
- **Step 4**: For each particle, update the improved velocity of formula (71) and the position of formula (72).
  - Use the following formulas to check the velocity constraint conditions:

$$\mathrm{V_i(t+1)} = \begin{cases} \mathrm{V_{max}} & \text{if } \mathrm{V_i(t+1)} > \mathrm{V_{max}} \\ \mathrm{V_i(t+1)} & \text{if } \mathrm{V_{min}} < \mathrm{V_i(t+1)} < \mathrm{V_{max}} \\ \mathrm{V_{min}} & \text{if } \mathrm{V_i(t+1)} < \mathrm{V_{min}} \end{cases} \tag{75}$$

  - Check the particle position constraint conditions as follows:

$$\mathrm{x_i(t+1)} = \begin{cases} \mathrm{x_{max}} & \text{if } \mathrm{x_i(t+1)} > \mathrm{x_{max}} \\ \mathrm{x_i(t+1)} & \text{if } \mathrm{x_{min}} < \mathrm{V_i(t+1)} < \mathrm{x_{max}} \\ \mathrm{x_{min}} & \text{if } \mathrm{x_i(t+1)} < \mathrm{x_{min}} \end{cases} \tag{76}$$

- **Step 5:** Go back to step 2.

The PSO-PI flowchart control employed in this system is shown in Fig. 9 [21].

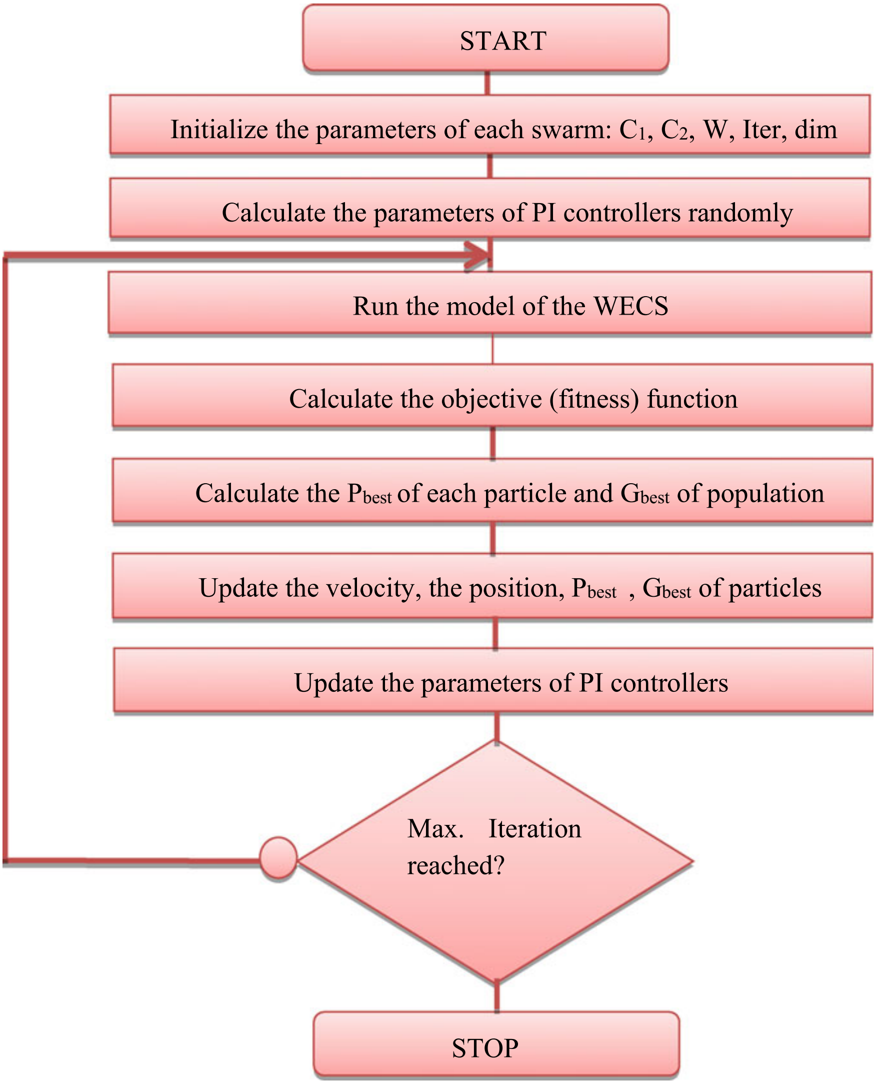

START
Initialize the parameters of each swarm: C_1, C_2, W, Iter, dim
Calculate the parameters of PI controllers randomly
Run the model of the WECS
Calculate the objective (fitness) function
Calculate the P_{best} of each particle and G_{best} of population
Update the velocity, the position, P_{best} , G_{best} of particles
Update the parameters of PI controllers
Max. Iteration reached?
STOP

Fig. 9 Flowchart of PSO algorithm

6 Simulation and Results

Matlab/Simulink is used to model and run the entire system (WECS). The wind profile shown in Fig. 10 is used. The DC bus voltage reference is 1200 V (Vdc^* = 1200 V), which is chosen as a constant. The rated mechanical power reference produced from the MPPT method is shown in Fig. 11. The rated output power is 5 MW when the wind profile is equivalent to 12.5 m/s, as can be observed easily.

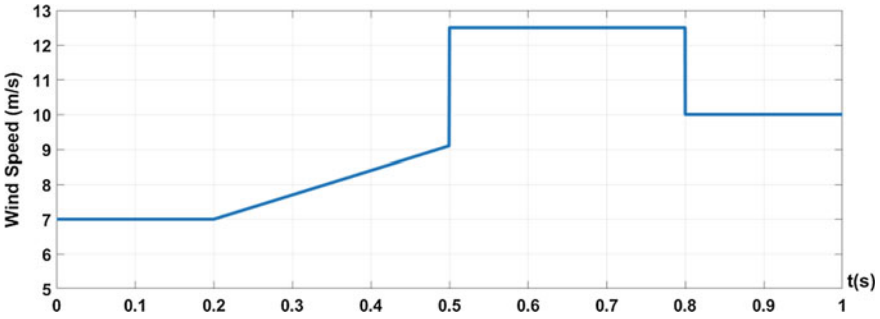


Fig. 10 Wind speed profile (m/s)

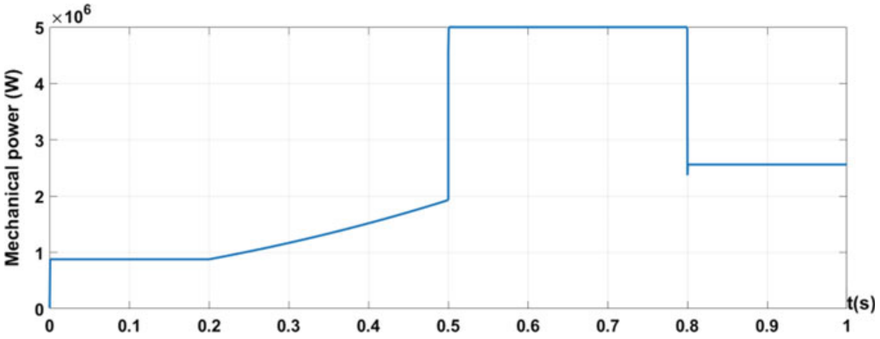


Fig. 11 Mechanical power (W)

The speed rotation of the multiplier output drives the doubly-fed inductor generator, which has the mechanical speed depicted in Fig. 12 as the input speed produced by the MPPT algorithm. As a result, zooming in on Fig. 12 reveals that the mechanical speed tracks its reference with a response time of about 0.3 (ms) and zero static error. To ensure a unity power factor, the reactive power references Qg* and Qs* are set to 0

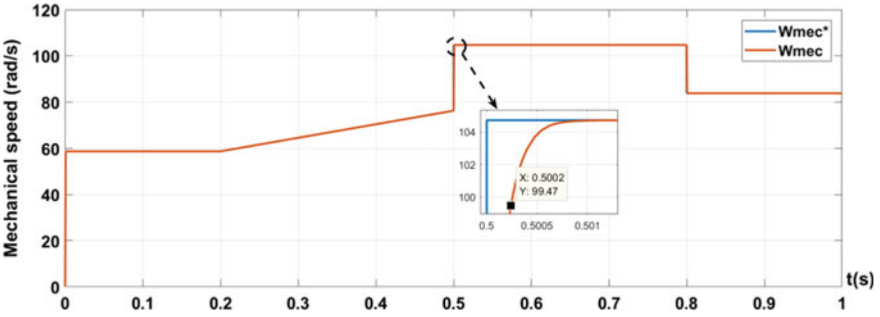


Fig. 12 Mechanical speed computed by the MPPT algorithm

VAR. The pole compensation calculates the parameters of the several PI controllers used to control the wind energy conversion system. Then, using the particle swarm optimization (PSO) algorithm, the parameters listed in Tables 1 and 2 are fine-tuned. To create the PSO program, the setups shown in Table 3 were chosen.

The results of the traditional PI and the tuning PSO controller PI are obtained and compared. The active and reactive powers produced by the DFIG are examined in this chapter. When comparing the results in Fig. 13, it is clear that the active power computed by Smart PI (Ps-PSO) closely follows the reference (Ps*) than the active power calculated by conventional PI controllers (Ps-PI). The response time has been reduced to 5.1 m s from 13.3 m s, and the static error has been reduced to 0.16% from 2.04%. Figure 14 shows how the PSO technique assures good reactive power tracking and decreases time response significantly, especially when the wind profile suddenly changes at t = 0.5 s. The DC-link voltage computed by the classical PI and the smart PI is shown in Fig. 15. The overshot of the DC link voltage is decreased to 233 V from 459 V, and the PSO improves the response time.

The results of the two proposed controls are compared in this section. Table 4 shows the characteristics of the wind energy conversion system components employed in the simulation.

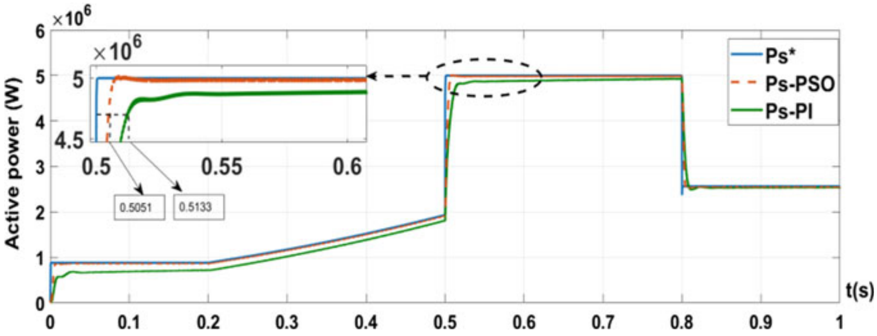


Fig. 13 Stator active power (W)

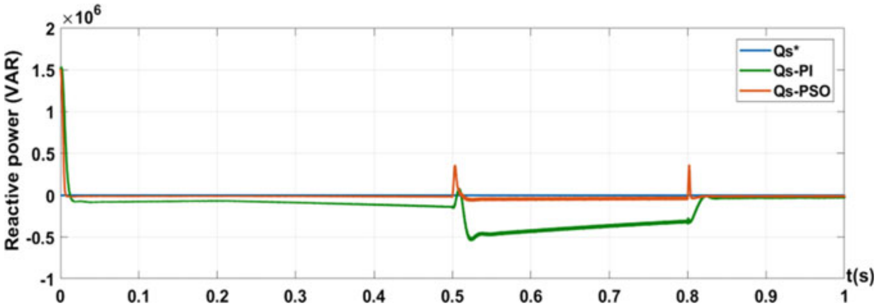


Fig. 14 Stator reactive power (VAR)

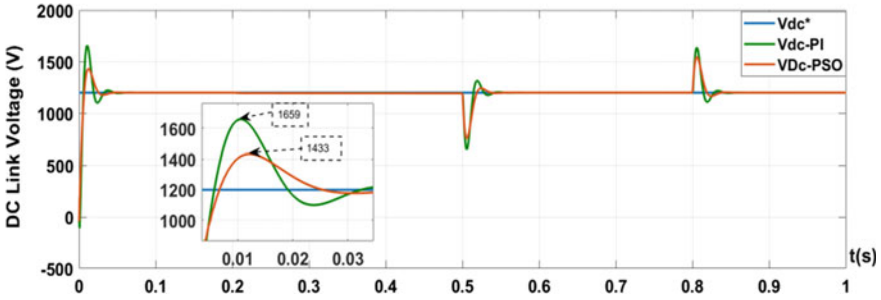


Fig. 15 DC-link voltage

7 Conclusion

The Wind Energy System Conversion (WECS) is modeled and simulated using variable wind speed in this chapter to demonstrate the superiority of the proposed method. The grid is connected directly to the stator of the doubly-fed inductor generator, and the rotor is attached via back-to-back converters. Traditional PI controllers are used to establish stator flux-oriented control. On the other hand, intelligent PI controllers are being developed to optimize PI controller gains using Artificial Intelligence (AI) based on the Particle Swarm Optimization (PSO) method. For the wind chain energy based on a DFIG of 5 MW, the outcomes of the two methodologies are compared and evaluated. The simulation results show that the PSO tuning technique not only generates satisfying and fascinating results, notably when the wind speed varies abruptly or when the system characteristics change, but it also optimizes the PI controller parameters for the system under consideration. The PSO technique outperforms the traditional PI controller, and it is a suitable technique for tracking the variation of a stochastic and changeable energy wind profile.

Appendix

See Tables 1, 2, 3 and 4.

Table 1 PI controller gains computed by pole compensation and by PSO methods

PIs gains	Kpdc	Kidc	Kprsc_1	Kirsc_1	Kprsc_2
Without PSO	1.848	396	0.1446	0.2376	0.1446
With PSO	2	355.46	1	0.418	1

Table 2 PI controller gains computed by pole compensation and by PSO methods

PIs gains	$Kirsc_2$	$Kpgsc_1$	$Kigsc_1$	$Kpgsc_2$	$Kigsc_2$
Without PSO	0.2376	200	5e + 4	200	5e + 4
With PSO	0.9354	134,74	3.6205.e + 7	127.03	6.7585.e + 6

Table 3 Set of parameters used for establishing the PSO algorithm

Parameters	Value
Population size	15
Number of parameters	10
Number of iterations	20
W	0.9
$C_1 = C_2$	2

Table 4 Set of parameters of the WECS utilized for the simulation

Components of the WECS	Parameters	Symbol	Value
Turbine	Radius of blade	R	51.583 m
	Coefficient of multiplier	G_B	47.23
	Total moment of inertia	J_{tot}	1000 kg m^2
DFIG	DFIG rated power	Ps	5 MW
	Stator leakage inductance	Ls	1.2721 mH
	Rotor resistance	Rr	1.446 mΩ
	Rotor leakage inductance	Lr	1.1194 mH
	Mutual inductance	M	0.55187 mH
	Stator line to line voltage	Vs	950 V
Capacity	DC-link capacitance	C	4400 μF
	DC link Voltage reference	V_{dc}^*	1200 V
Filter RL	Resistor of the filter	R_f	20 Ω
	Inductance of the filter	L_f	0.08 H

References

1. Alhato MM, Bouallègue S (2019) Direct power control optimization for doubly fed induction generator based wind turbine systems. Math Comput Appl 24(3):77
2. Bouazza H, Bendaas ML, Allaoui T, Denai M (2020) Application of artificial intelligence to wind power generation: modelling, control and fault detection. Int J Intell Syst Technol Appl 19(3):280–305
3. Oliveira CMR, Aguiar ML, Monteiro JRBA, Pereira WCA, Paula GT, Almeida TEP (2016) Vector control of induction motor using an integral sliding mode controller with anti-windup. J Control Autom Electr Syst 27(2):169–178
4. Arama FZ, Bousserhane IK, Laribi S, Sahli Y, Mazari B (2018) Artificial intelligence control applied in wind energy conversion. System 9(2):571–578
5. Muthusamy M, Parvathy AK (2020) Artificial intelligence techniques-based low voltage ride through enhancement of DFIG. J Mech Continua Math Sci 15(3):125–139
6. Nassimuallah, Irfan S, Chowdhury MDS, Techato K, Alkhammash H (2017) Artificial intelligence integrated fractional order control for Doubly Fed induction Generator based wind energy system. IEEE Access 04
7. Ben Belghith O, Sbita L, Bettaher F (2016) MPPT design using PSO technique for photovoltaic system control comparing to fuzzy logic and P&O controllers. Energy Power Eng 08:349–366
8. Bharti OP, Saket RK, Nagar SK (2017) Controller design for doubly fed induction generator using particle swarm optimization technique. Renew Energy 114:1394–1406
9. Baba AO, Liu G, Chen X (2020) Classification and evaluation review of maximum power point tracking methods. Sustain Futures 2(February)
10. El Filali A, Zazi M (2021) Arduino implementation of MPPT with P and O algorithm in photovoltaic systems. Int J Eng Appl Phys (IJEAP) 1(1):9–17
11. Chetouani E, Errami Y, Obbadi A, Sahnoun S (2021) Backstepping and indirect vector control for rotor side converter of Doubly Fed-induction generator with maximum power point tracking. In: Motahhir S, Bossoufi B (eds) Digital technologies and applications. ICDTA 2021, Lecture Notes in Networks and Systems, vol 211, pp 1711–1723
12. Errami Y, Obbadi A, Sahnoun S (2020) Control of PMSG wind electrical system in network context and during the MPP tracking process. Int J Syst Control Commun 11(2):200–225
13. Ihedrane Y, El Bekkali C, Bossoufi B, Bouderbala M (2019) Control of power of a DFIG generator with MPPT technique for wind turbines variable speed. In: Derbel N, Zhu Q (eds) Modeling, identification and control methods in renewable energy systems, green energy and technology. Springer, Singapore, pp 105–129
14. Elazzaoui M (2015) Mode modeling and control of a wind system based Doubly Fed induction generator: optimization of the power produced. J Electr Electron Syst 4(1):1–8
15. Dahbi A, Nait-Said N, Nait-Said M (2016) A novel combined MPPT-pitch angle control for wide range variable speed wind turbine based on neural network. Int J Hydrogen Energy 41(22):9427–9442
16. Bouderbala M, Bossoufi B, Lagrioui A, Taoussi M, Alami H, Ihedrane Y (2018) Direct and indirect vector control of a doubly fed induction generator based in a wind energy conversion system. Int J Electr Comput Eng (IJECE) 9(3):1531–1540
17. Laina R, Lamzouri F, Boufounas E, El Amrani A, Boumhidi I (2018) Intelligent control of a DFIG wind turbine using a PSO evolutionary algorithm. Procedia Comput Sci 127(2018):471–480
18. Chetouani E, Errami Y, Obbadi A, Sahnoun S (2021) Optimal tuning of PI controllers using adaptive particle swarm optimization for Doubly-Fed induction generator connected to the grid during a voltage dip. Bull Electr Eng Inform 10(5):2367–2376
19. Labdai S, Hemici B, Nezli L, Bounar N, Boulkroune A, Chrifi-Alaoui L (2019) Control of a DFIG Based WECS with optimized PI controllers via a duplicate PSO algorithm. In: 2019 international conference on control, automation and diagnosis, ICCAD 2019—Proceedings

20. Bekakra Y, Attous DB (2014) Optimal tuning of PI controller using PSO optimization for indirect power control for DFIG based wind turbine with MPPT. Int J Syst Assur Eng Manag 5(3):219–229
21. Dai H, Chen D, Zheng Z (2018) Effects of random values for particle swarm optimization algorithm. Algorithms 11(23):1–20

A Comparative Study Between NARX and LSTM Models in Predicting Ozone Concentrations: Case of Agadir City (Morocco)

Anas Adnane, Amine Ajdour, Radouane Leghrib, Jamal Chaoufi, and Ahmed Chirmata

Abstract It is well known that air pollution has become a significant global environmental problem causing numerous chronic diseases to humans. Many countries worldwide, including Morocco, are making a lot of effort in order to minimize air pollutants emissions and many other actions to decrease the damages caused by these noxious gases. Predicting the concentrations of an air pollutant in general, and ozone in particular, is of a high interest due to the fact that it will provide legislators an idea about the future situation so that they can take all the necessary measures. This paper proposes a comparative study between two machine learning algorithms in predicting ambient ozone concentrations over Agadir City: Non-linear autoregressive network with exogenous inputs (NARX) and long short-term memory (LSTM). The results show that both models perform very efficiently with a small upper hand for LSTM model.

Keywords Air pollution · Ozone · Nonlinear autoregressive exogenous model (NARX) · Long short-term memory (LSTM) · Long-term prediction

1 Introduction

Surface ozone in the troposphere is a toxic and dangerous air pollutant. It is a secondary pollutant formed and produced under strong solar radiation driven chemical reactions involving carbon monoxide (CO), nitrogen oxides (NOx), and volatile

A. Adnane (✉) · A. Ajdour · R. Leghrib · J. Chaoufi
LETSMP, Department of Physics, Faculty of Science, Ibn Zohr University, Agadir, Morocco
e-mail: anas.adnane@edu.uiz.ac.ma

A. Ajdour
e-mail: amine.ajdour@edu.uiz.ac.ma

R. Leghrib
e-mail: r.leghrib@uiz.ac.ma

A. Chirmata
Department of Energy and Environment, Wilaya of Agadir, Agadir, Morocco

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_9

organic compounds (VOCs) [1, 2]. It is well established that high ozone concentrations can cause serious damage to human health, such as respiratory and cardiovascular diseases and infections, and the natural ecosystem [3]. Predicting such air pollutants is a big step towards taking all the necessary measures in order to reduce their concentrations and the resulting damages [4].

Atmospheric conditions have a significant role in the formation of ozone as well as its transport, dispersion, and accumulation at the surface [1]. For example, it is found that low ozone concentrations are strongly correlated with high wind speed [5]. Another example is that of solar radiation and temperature: high values of these two meteorological parameters are very linked to elevated ozone concentrations [6].

In recent years, machine learning algorithms have been increasingly adopted in order to model and predict the complex and nonlinear relationships underlying air pollution in general, and ozone in particular. Among these ML algorithms we can cite the non-linear autoregressive network with exogenous inputs (NARX) which was applied by Adnane et al. in order to predict the ambient ozone concentrations in Agadir city (Morocco) [7], the Long Short Term Memory architecture which was used by Ribeiro in order to predict the maximum ozone concentrations in the East Austrian region [8] and the Random Forest [9].

In this paper, we conduct a comparative study between NARX model and LSTM model in terms of modeling and predicting surface ozone concentrations over the city of Agadir. This work can be summarized as follows: First, we give and idea about the study area and the data collection and analysis. Second, we describe the two models used in this study. Third, we run the two models and analyze the results given and compare between them.

2 Materials and Methods

2.1 Study Area and Data Collection

The Moroccan kingdom is located in the southern part of the Mediterranean basin, and it is considered as one of the most vulnerable countries to trans-boundary air pollution [10]. Our study area is Agadir city, which is located in the southwest of Morocco with a population of over than 900.000 residents [11]. We illustrated the geographical location of our study area in Fig. 1. We choose Anza site (−9.658485; 30.448091) due to the fact that it represents the most polluted zone in Agadir city. We picked the period (01/05/2016–30/06/2016) since it contains fewer missing data.

We reconstructed the missing data in our database by using the nearest neighbor method, which consists of the substitution of the lacking values by the average of data from days $d - 1$ and $d + 1$. Additionally, we converted the wind speed and direction into the zonal and meridional wind, with the aim of taking into account the cyclical characteristics of the wind.

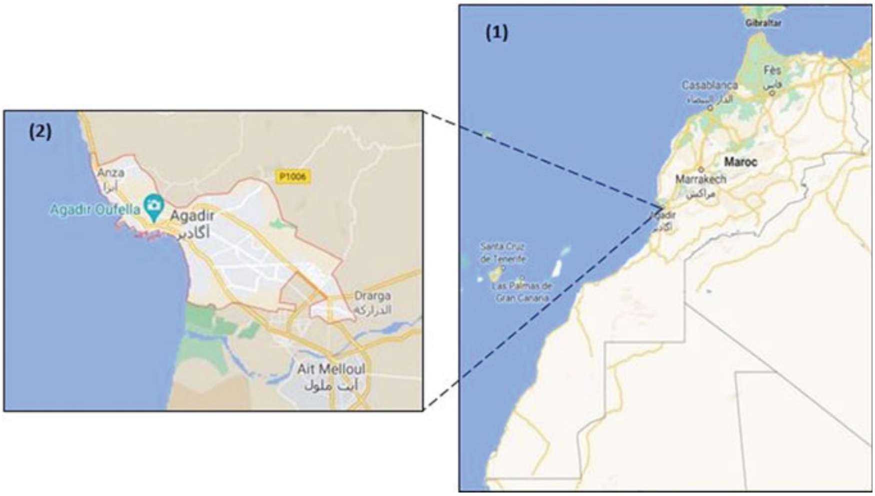


Fig. 1 The geographical location of the study area. (1) Morocco, (2) Agadir

2.2 LSTM

Recurrent Neural Networks (RNNs) are well known for their ability to exhibit temporal dynamic behavior. The RNN structure can be thought of as numerous copies of the same network, each passing the information to its successor. The main advantage of RNNs is their capability to remember each information through time. Nonetheless, RNNs can perform unsuccessfully when it comes to long-term dependencies [12]. Long-Short Term Memory (LSTM) cells proposed by S. Hochreiter and J. Schmidhuber in 1997 belong to recurrent neural networks architecture targeting the vanishing gradient issue [13]. The added feature of LSTM is the presence of a memory cell that can preserve information in memory for long duration. In LSTM, a set of gets is used to control the addition or deletion of the information in the cell [14]. A common LSTM unit has three gates: an input gate, a forget gate, and an output gate. The forget gate determines what information we are going to throw away from the cell state. The input gate has the role of deciding what new information will be stored in the cell state, and the output gate chooses what value we desire to output. Equations related to the LSTM gate are presented below:

Forget gate:

$$f_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right), \tag{1}$$

Input gate:

$$i_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right), \tag{2}$$

Output gate:

$$o_t = \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right), \tag{3}$$

where f_t, i_t, o_t denote the forget gate, input gate and output gate, respectively. w represents the relevant weight in every gate associated with each LSTM block, σ is the sigmoid function, x_t is the current input vector at time-step t, h_{t-1} denotes the previous output at t − 1, and b is the bias of each gate. The equations for the final output, the cell state and the candidate cell state are the following:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{4}$$

$$\tilde{C}_t = \tanh\left(w_c.\left[h_{t-1}, x_t\right] + b_c\right), \tag{5}$$

$$h_t = o_t * \tanh(C_t), \tag{6}$$

where $\tilde{C}_t$ represents candidate for cell state at t, C_t and C_{t-1} refer to the new and precedent cell states (memory) at t and t − 1, and * is the element wise multiplication of the vectors (Fig. 2).

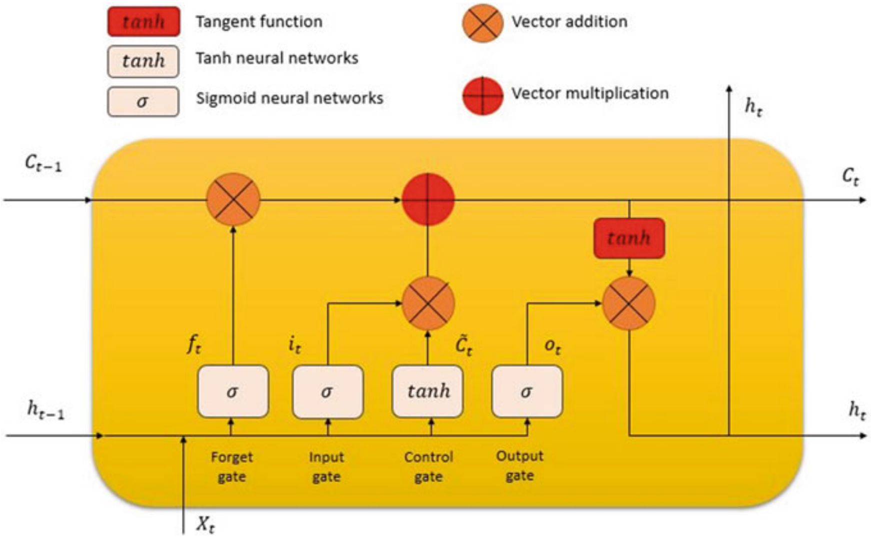


Fig. 2 The general scheme of LSTM model

2.3 NARX

NARX is none other than a recurrent neural network architecture. It is largely employed in modelling and system identification applications with complicated nonlinearities [15]. NARX models are composed of a two-layered feed-forward network. The activation functions used in the hidden layer and the output layer are the sigmoid and the linear transfer functions respectively. NARX models can predict the actual value of the output variable using a nonlinear function of the model's antecedent inputs and outputs. Among the most NARX strong points is their rapid convergence and their considerable capacity to produce great long-term predictions. We used MATLAB 2015 in order to design the NARX models. A NARX model is calculated employing the following equation:

$$\mathrm{y}'(\mathrm{t}) = \mathrm{F}[\mathrm{y}(\mathrm{t}-1), \ldots, \mathrm{y}(\mathrm{t}-\mathrm{n_y}), \mathrm{u}(\mathrm{t}-1), \ldots, \mathrm{u}(\mathrm{t}-\mathrm{n_u}), \tag{7}$$

where y(t) is the actual output variable and y′(t) its one-step forecast. $\mathrm{n_u}$ and $\mathrm{n_y}$ represent the maximum lags of the exogenous and endogenous variables u and y, and F is approximated by the FFNN.

We used, in this work, the Levenberg-Marquardt backpropagation procedure. This training algorithm is well known for its stability and its fast convergence [16, 17]. The LMBP algorithm was designed to approximate the second-order derivative with no need to compute the Hessian matrix [18]. This approximation is given with the following formula:

$$\Delta w = \left[J^T(w)J(w) + \lambda I\right]^{-1} J^T(w)e(w) \tag{8}$$

where, J represents the Jacobian matrix and J^T its transpose, w is the weight vector, e is the error vector and I denotes the identity matrix, λ is the learning constant, adjusted iteratively to find the least error.

Figure 3 illustrates the flowchart of finding the most optimal NARX model. As of Fig. 4, it shows the general scheme of a NARX model**.**

2.4 Performance Indices

In order to assess the performance and the prediction capabilities of the models used in this research work, different performance indices have been selected. The Root Square Mean Error (RMSE) quantifies how close the forecasted values equal the actual ones. The Coefficient of Correlation (CC) illustrates the strength of the link between the estimations and the real measurements. The Mean Absolute Error (MAE) represents the mean of the absolute values of the individual prediction errors. The more the MAE and RMSE are low and the CC is close to 1, the better the model

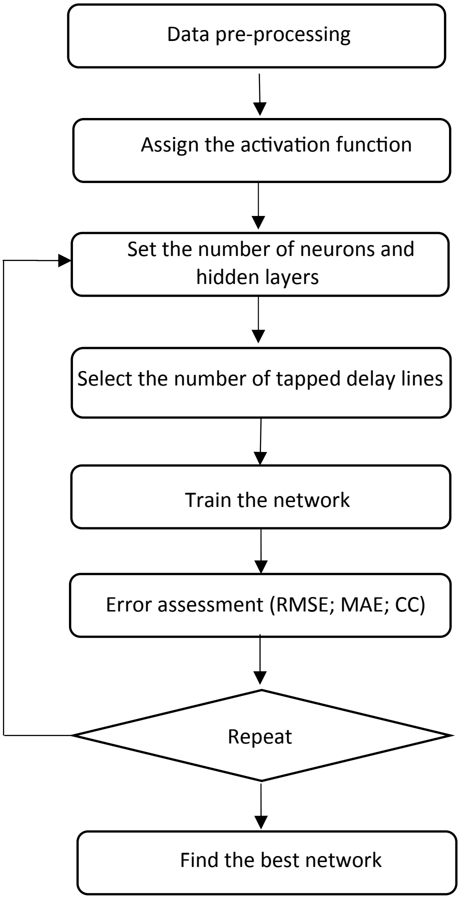

Data pre-processing
Assign the activation function
Set the number of neurons and hidden layers
Select the number of tapped delay lines
Train the network
Error assessment (RMSE; MAE; CC)
Repeat
Find the best network

Fig. 3 The flowchart of finding the mist optimal NARX model

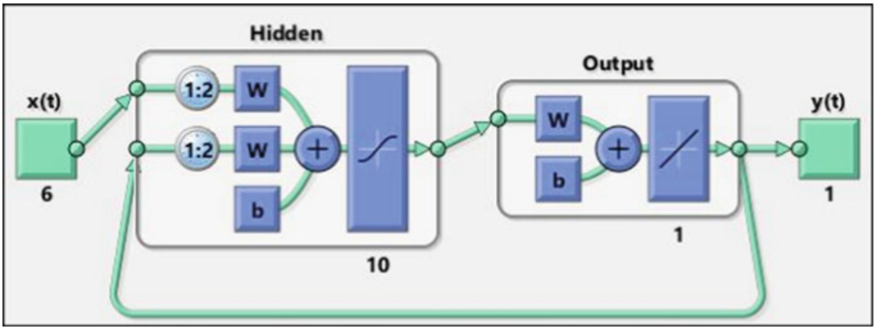


Fig. 4 The general scheme of NARX model

is. These metrics are calculated using the following formulas:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{9}$$

$$CC = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)(x_i - \overline{x}_i)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2 (x_i - \overline{x}_i)^2}} \tag{10}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{11}$$

We denote by x_i and y_i the values of the observed and forecasted values respectively, and $\overline{x}_i$ and $\overline{y}_i$ the mean values of the observed and forecasted values respectively.

3 Results and Discussion

The variables selected to be in the input layer of our network were the meteorological variables such as temperature, humidity, zonal wind and meridional wind, alongside with some air pollutants such as CO and SO_2. Ozone was the only variable in the output layer.

Our database was divided into two major sections: 85% of the data were allocated to the training of the network, and the remaining 15% were attributed to the test and the validation of the performance of the network.

We used the logistic and linear sigmoid activation functions in the hidden layers and the output layer respectively. The most challenging task in using artificial neural networks is to find the most suitable number of neurons as well as the number of the hidden layers. We tested several NARX and LSTM architectures, and we represented the most optimal ones in Table 1.

LSTM mainly solves sequence data processing. The first parameter of the LSTM layer is the number of neurons or nodes we want in the layer. In our model, we use two hidden layers with 12 neurons for each hidden layer. The second parameter is return_sequences, which is set to true since we will add more layers to the model. Finally, we need to compile our LSTM before we can train it on the training data. We call the compile method on the Sequential model object, which is "model" in our

Table 1 Performance of the types of artificial neural networks: NARX and LSTM

Model	Structure	CC (%)	RMSE	MAE
NARX	6-10-1	83,37	4,77	3,71
LSTM	6-12-12-1	91,47	3,93	3,30

case. We used the mean absolute error as the loss function, and to reduce the loss or optimize the algorithm, we used the Adam optimizer. We programmed our LSTM algorithms using Python.

Table 1 represents the performance criteria of the most optimal NARX and LSTM models. We notice that both models perform very well, but with a slight advantage for the LSTM model with a CC of 91.47%, an RMSE of 3.93 and an MAE of 3.30. This advantage can be explained by the fact that NARX has a higher forecasting error due to predicted output is fed back to the feedback network rather than the actual output which accumulates the error further to the following samples prediction.

We illustrated the observations of ambient ozone concentrations for the first 36 h of the test period alongside with the predictions given by the NARX and LSTM models in Fig. 5. We notice that both models give very good predictions. We notice also that LSTM model fail sometimes to predict well the extreme values. We observe also a tiny upper hand for LSTM model when it comes to long-term predictions. This can be explained by the fact that deep learning artificial neural networks, such as LSTM, can link input to output more accurately by learning the long-term dependencies present in the data. However, this does not prevent us from mentioning that NARX maintains the trend compared to LSTM. Generally, it can be noted that LSTM prediction is often backward, while NARX prediction is forward in comparison to the measurements, especially for the maximum concentrations.

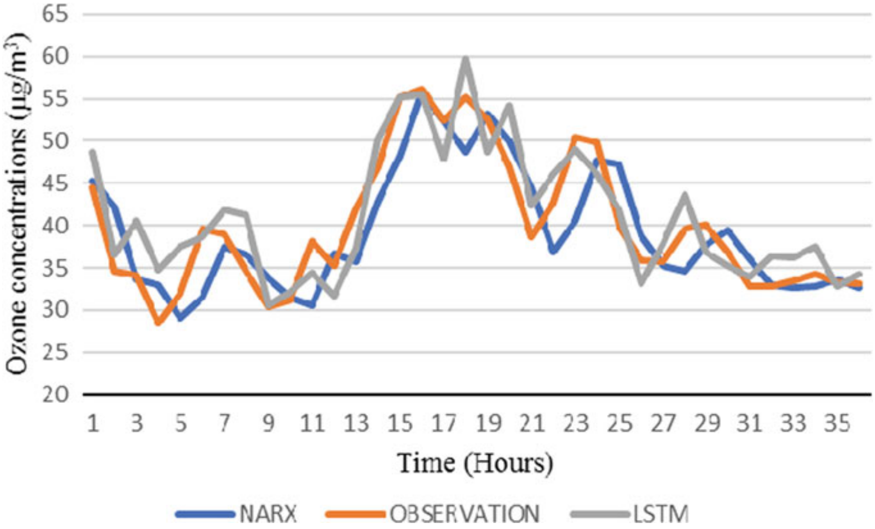


Fig. 5 Representation of observed values of ozone concentrations alongside with those predicted by NARX and LSTM models

4 Conclusion

To conclude, we tried in this paper to compare between two of the most powerful machine learning models (NARX and LSTM) in their capability of predicting ambient ozone concentrations over the city of Agadir. We used as inputs a combination of meteorological parameters and air pollutants. In order to establish meaningful interpretations and conclusions concerning the accuracy of the results, we used several performance metrics.

Both models gave very accurate predictions despite the limited amount of data. A small advantage for the LSTM model was highlighted in predicting in long-term the ozone concentrations. As of NARX, it maintains efficiently the general tendency compared to LSTM. We conclude that the obtained results are encouraging regarding the use of recurrent neural network in modeling air pollutants concentrations.

Acknowledgements We acknowledge the Wilaya and the Region of Souss Massa for their support for this research work, under the leadership of the Wali and the President of the Souss Massa region, to whom we present our special thanks.

References

1. Sari D, Incecik S, Ozkurt N (2020) Analysis of surface ozone episodes using WRF-HYSPLIT model at Biga Peninsula in the Marmara region of Turkey. Atmos Pollut Res 11(12):2361–2378. https://doi.org/10.1016/j.apr.2020.09.018
2. Kleinman LI (1994) Low and high NOx tropospheric photochemistry. J Geophys Res 99(D8):831–838. https://doi.org/10.1029/94jd01028
3. Zhang W, Qian CN, Zeng YX (2014) Air pollution: A smoking gun for cancer. Chin J Cancer 33(4):173–175. https://doi.org/10.5732/cjc.014.10034
4. Ajdour A, Leghrib R, Chaoufi J, Chirmata A, Menut L, Mailler S (2019) Towards air quality modeling in Agadir City (Morocco). Mater Today Proc 24:17–23. https://doi.org/10.1016/j.matpr.2019.07.438
5. Mao J et al (2020) Meteorological mechanism for a large-scale persistent severe ozone pollution event over eastern China in 2017. J Environ Sci (China) 92(February):187–199. https://doi.org/10.1016/j.jes.2020.02.019
6. Zanis P et al (2014) Summertime free-tropospheric ozone pool over the eastern Mediterranean/middle east. Atmos Chem Phys 14(1):115–132. https://doi.org/10.5194/acp-14-115-2014
7. Adnane A, Leghrib R, Chaoufi J (2020) The Use of a Recurrent Neural Network for Forecasting Ozone Concentrations in the City of Agadir (Morocco). J At Mol Condens Matter Nano Phys 7(3):197–206. https://doi.org/10.26713/jamcnp.v7i3.1545
8. Ribeiro S, Alquézar R (2002) Local maximum ozone concentration prediction using neural networks. OGAI J (Oesterreichische Gesellschaft fuer Artif Intell 21(2):3–6
9. Hsieh MH, Hsieh MJ, Chen CM, Hsieh CC, Chao CM, Lai CC (2018) Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. Sci Rep 8(1):1–7. https://doi.org/10.1038/s41598-018-35582-2
10. Oufdou H, Bellanger L, Bergam A, Khomsi K (2021) Forecasting daily of surface ozone concentration in the grand Casablanca region using parametric and nonparametric statistical models. Atmosphere (Basel) 12(666). https://doi.org/10.3390/atmos12060666

11. World Population Review (2021) https://worldpopulationreview.com/world-cities/agadir-population
12. Luo J, Zhang Z, Fu Y, Rao F (2021) Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results Phys 27:104462. https://doi.org/10.1016/j.rinp.2021.104462
13. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
14. Cocianu CL, Avramescu M-S (2020) The use of LSTM neural networks to implement the NARX model. A case study of EUR-USD exchange rates. Inform Econ 24(1/2020):5–14. https://doi.org/10.24818/issn14531305/24.1.2020.01
15. Wang H, Song G (2014) Innovative NARX recurrent neural network model for ultra-thin shape memory alloy wire. Neurocomputing 134:289–295. https://doi.org/10.1016/j.neucom.2013.09.050
16. Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. Q Appl Math 2
17. Alsumaiei AA (2020) A nonlinear autoregressive modeling approach for forecasting groundwater level fluctuation in urban aquifers. Water 12(3):1–16. https://doi.org/10.3390/w12030820
18. Roghanchi P, Kocsis KC (2019) Quantifying the thermal damping effect in underground vertical shafts using the nonlinear autoregressive with external input (NARX) algorithm. Int J Min Sci Technol 29(2):255–262. https://doi.org/10.1016/j.ijmst.2018.06.002

Spatiotemporal Prediction of $PM_{2.5}$ Concentrations Based on IoT Sensors

Abdellatif Bekkar, Badr Hssina, Samira Douzi, and Khadija Douzi

Abstract Not only in the West but elsewhere, Particulate Matter (PM) pollution as one of the staid health problems in East Asia has increasingly received due attention by different kinds of scholars over the past decade. Its severity has seriously questioned an array of researchers in academia. Including our efforts, this proposed account focuses on Taiwan, as an emerging country, which itself has not yet escaped the poor air quality. To this end, this study uses Facebook's Prophet Model for Forecasting $PM_{2.5}$ concentration in Taiwan more specifically in Meinong city. The large-scale database remains the key source for the evaluation of the suggested model in this account. Database consists in the dataset of the Environmental Protection Administration, and Air box IoT dataset (1677 sensors). Based on the past records of $PM_{2.5}$, meteorological data, $PM_{2.5}$ concentration of neighboring sensors, the forecasting model have effectively proven their adequacy for short time series. Even the experimental analysis of the forecasting for $PM_{2.5}$ levels has efficaciously amount to the effectiveness of proposed model; its manifestations hence occur not only in the significant decrease not only in MAPE (from 19.220 to 1.543), but also RMSE (from 14.892 to 4.732) for the proposed method.

Keywords Air pollution · Airbox · $PM_{2.5}$ · Internet of Things · Forecasting · Facebook prophet

A. Bekkar (✉) · B. Hssina · K. Douzi
Laboratory LIM, IT Department FST Mohammedia, Hassan II University,
Casablanca, Morocco
e-mail: badr.hssina@fstm.ac.ma

S. Douzi
FMPR, Mohammed V University in Rabat, Rabat, Morocco
e-mail: s.douzi@um5r.ac.ma

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_10

1 Introduction

More than ever before, worldwide countries have interminably become under the threat of homicidal effects of air pollution. Asia as a case in point has witnessed unprecedented jeopardies due to such issue [1]. According to WHO survey carried out in 2016, air pollution caused around 29% were due to heart disease, 27% stroke, 22% chronic obstructive pulmonary disease, 14% lung cancer, and 8% pneumonia among 2.2 million air pollution-related deaths in this Region [2]. Like other Asian emergent countries, Taiwan itself as an industrial country has experienced the issue of air pollution. In the last decade, the country has witnessed the highest air pollution. Taken Taipei as an example, $PM_{2.5}$ air pollution brought about 1100 deaths since January 1, 2021 [3].

Already defined as follows, air pollution by and large is said to the presence of pollutants (gaseous or particulate matther) in the atmosphere. Its severe effects extend to the environment and health. Air pollution sources originate in either natural or anthropogenic; that is the latter is linked to human activity by the direct release of gaseous compounds (representing more than 95% of the global masses pollutants released into the air) come from industrial, domestic, agricultural, transport, various combustions, etc. [4]. Accurately, $PM_{2.5}$, the so-called dust elsewhere, has a wide assortment of sources. Of many different chemical compounds $PM_{2.5}$ composes. The combustion of wood, fossil fuels in vehicles, thermal power stations, or industries are all grounds for $PM_{2.5}$ production. (PM stands for Particulate Matter while the number stands for diameter). Due to their smaller size (2.5 μm) they can penetrate into the pulmonary alveoli and even the blood. It results in stroke, ischemic heart disease, lung cancer, respiratory infection, and chronic obstructive pulmonary disease [5–7]. $PM_{2.5}$ riskiness extends to its ability, as an important vector, to accelerate the spread of the COVID19 [8].

The Internet of Things (IoT) has speedily been developed thanks to the development of 4G communication, Wi-Fi, and wireless sensor networks (WSN), and the popularization of portable devices adding to other numerous environmental monitoring devices [9, 10]. Even the most major industrialized countries have identified IoT as a key technological development area. Environments, healthcare, transportation, homes, and cities are all smart imperative aspects of the future development of the Internet of Things [11]. Air pollution monitoring benefits from these devices. By virtue of connecting a group of sensors, the monitoring network used to collect and send air quality information to the station within a stitch of time [11–13].

Monitoring and meanwhile reducing the concentration of $PM_{2.5}$ were of the uppermost priority for Taiwan government. Taiwan Environmental Protection Administration has installed 77 air quality monitoring stations. Many monitoring sensors are fixed in these stations which require a greater open area. The building of monitoring stations and the higher costs of their maintenance makes it a tough job to construct in many places. Therefore, in terms of spatial distribution in various counties and cities is also uneven. Due to the above conditions, the space-time resolution of these traditional stations is relatively low.

Taiwan has deployed a large number of the Airbox [14] Project sensors which is a micro-air quality sensor. Combining with the IoT and cloud technologies makes it possible to identify the concentration of $PM_{2.5}$ and monitor it every five minutes. Two issues with the data from Airbox sensors occur. One of them is that some of the sensors become unable to properly function or defective on the one hand. The sensors on the other hand may be located near the pollution resources. Compared to Traditional EPA, Airbox sensors are relatively poor in terms of accuracy. Before big data analysis takes place, heterogeneous environmental monitoring data from many scales and sources, data fusion or assimilation is required. It is possible hence to improve future analysis and forecasting precision by including environmental big data.

The forecasting of time-related data is a difficult problem due to the complexity of the molecule. In this research thus, $PM_{2.5}$ observation data from both Taiwan EPA and Air Box sensor are used aiming at predicting the concentration of $PM_{2.5}$. The conventional time-series analysis approaches are insufficient to analyze complex changes of $PM_{2.5}$ which pushes us to use the Facebook Prophet to this end thanks to its fastness in fitting procedure and robust to the larger outliers. Not only so, Facebook Prophet Model has can also be adaptable to multivariate datasets, missing values, dramatic changes, and holidays. These features make it of a great significance to be isolated study for the analysis of $PM_{2.5}$.

The remainder in this research seeks to by and large outline its key parts. The work is organized as follows. Section 2 addresses the description of the dataset, preprocessing, and features selection technique. The attention is shifted in Sect. 3 to the outline of the proposed work of using Facebook Prophet to build an effective $PM_{2.5}$ multivariate forecasting method. In Sect. 4, we describe our experimental setup and results and plotting the performance of the proposed system. Finally, the conclusions drawn from our study are given in Sect. 5.

2 The Area of Study and Data

2.1 The Area of Study

Geographically speaking, Taiwan, which is an emerging country located in East Asia. It consists of 36,197 km^2 and spans the Tropic of Cancer. Its population reaches approximately 23 million people (with a population density: 656/km^2). Related to the key sources of air pollution data used in this study, they are as follows:

- The Taiwanese Environmental Protection Agency (TWEPA). The database Composes of more than 260,000 samples recorded in 2017. 77 air monitoring stations as shown in Fig. 1. Six criteria pollutants including sulfur dioxide (SO_2), nitrogen oxides (NO_x), ozone (O_3), carbon monoxide (CO), PM_{10} (particles 10 µm in diameter), $PM_{2.5}$, and Wind Speed, and Wind Direction are recorded.
- The cooperation between the Academia Sinica and businesses, e.g., Edimax, the low-cost $PM_{2.5}$ sensor launched the Airbox project in 2015. Its establishment

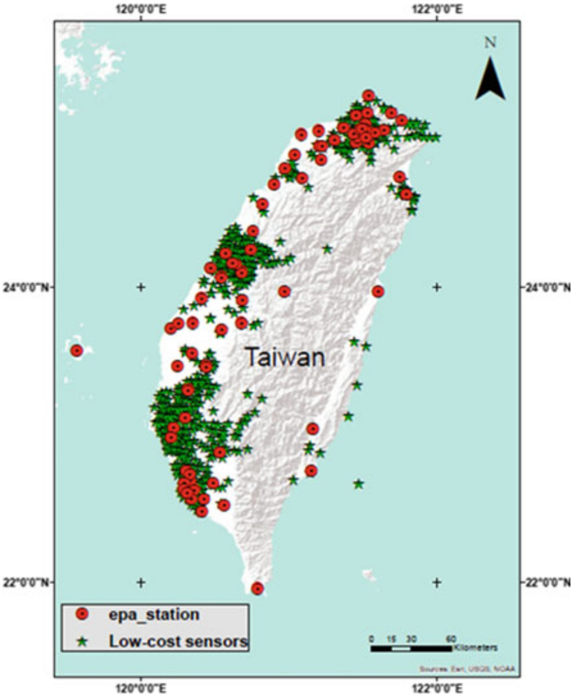


Fig. 1 Research area and monitoring station distribution map

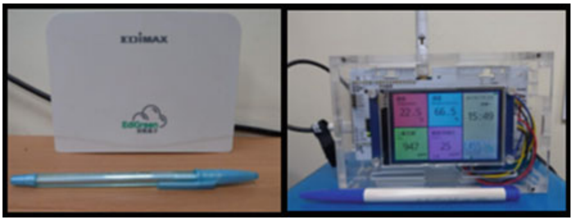


Fig. 2 Low-cost sensors of Airbox project

has led to the installation of thousands of boxes in Taiwan. The Realtek Ameba development board with the PMS5003 optical particulate matter sensor as shown (Fig. 2) used by a great number of the air quality detection sensors. 2963 low-cost sensors were set up (Fig. 1) the latter records the $PM_{2.5}$ concentration, temperature (°C), and relative humidity (%) every 5 min.

2.2 Data Preprocessing EPA Dataset

2.2.1 Missing Value

The object of this study is the station of Meinong District. Its dataset is made up of 9082 records with multi-features such as date, the concentration of SO_2, CO, O_3,

PM_{10}, $PM_{2.5}$, NO_2, NO_x, NO, Wind Speed, and Wind Direction. The allocated time ranges from January 1st, 2017, to September 09th, 2017 [15]. However, this dataset incorporates many missing values due to some uncontrollable reasons. To resolve this occurring issue, the following tasks take place.

Firstly, we reassembled the hourly dataset from an hourly basis to a daily basis by calculating the average of 24 h. Secondly; the linear interpolation method is used. The equation of the linear interpolation function is:

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_1 - x_0) \tag{1}$$

where x is the independent variable, x_1 and x_0 are known values of the independent variable, and $f(x)$ is the value of the dependent variable.

2.2.2 Feature Selection

Different methods can be applied to select the targeted features. One of the most used methods in precedent works is applied mathematical correlation. The latter is utilized to calculate the relations between input and output variables [16, 17].

Reducing the complexity and meantime ameliorating the performance remains the key aim which is feature selection. It, however, requires finding the correlation between output value and input features. The Pearson correlation is the most popular method used. The following equation can calculate its coefficient r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where$x$ and y represent variables, and $\bar{x}$ and $\bar{y}$ represent the mean of the variables (Fig. 3).

Different historical air quality data parameters increased the atmosphere's concentration of $PM_{2.5}$. We calculate Correlations between the criteria pollutants, and we find A high correlation between $PM_{2.5}$, PM_{10}, CO, NO_2, and NO_x.

Air quality is affected also by Weather parameters (wind speed, wind direction). For example, high wind speed will reduce the concentration of $PM_{2.5}$ [17]. Therefore, the Weather parameters are a primordial significance of the $PM_{2.5}$ forecasting task.

2.3 Airbox Dataset

2.3.1 Missing Value

This dataset includes 1677 low-cost sensors. Each sensor has an average of 9394 records with multi-features. The period is from January 1st, 2017, to March 31st, 2017 [18]. Including date, the concentration of $PM_{2.5}$, PM_{10}, PM_1, Temperature,

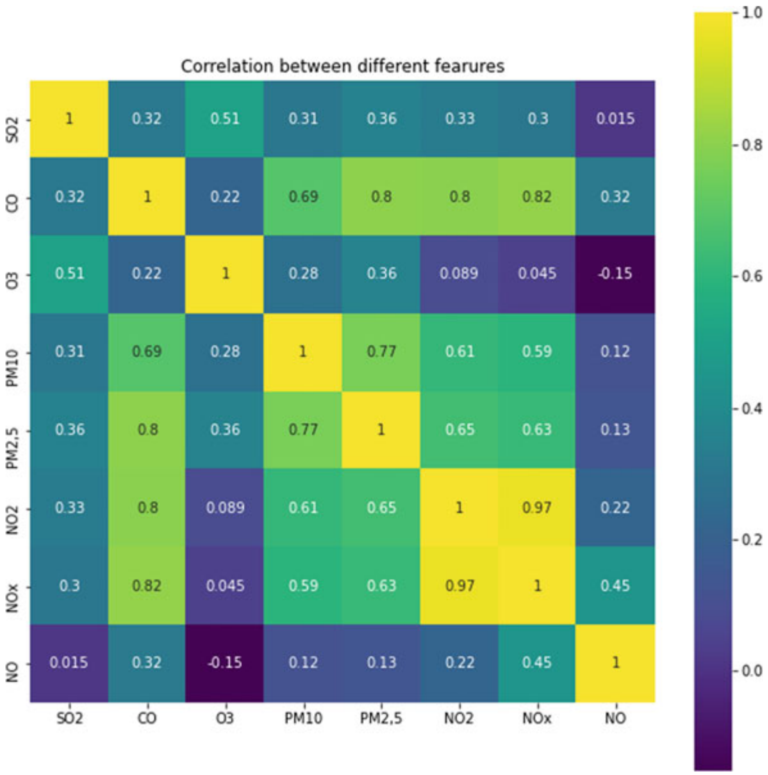


Fig. 3 The correlation matrix of the air quality features

Humidity, deviceId, latitude, and longitude. However, this dataset incorporates many missing values due to low-cost sensors.

Firstly, we resampled the dataset on a daily basis and secondly, we fill the gaps with the linear interpolation method.

2.3.2 Calculation of the Distance Between Sensors and EPA Station

The sensor nodes are deployed over various geographical locations. The distance between sensors and EPA station is tightly linked related to the diffusion of $PM_{2.5}$. See Fig. 4.

According to the latitude and longitude relations between each sensor and EPA station the shortest distance between each collection station could be calculated using the haversine formula [19].

The haversine formula is used to calculate the distance between two points on the Earth's surface specified in longitude and latitude:

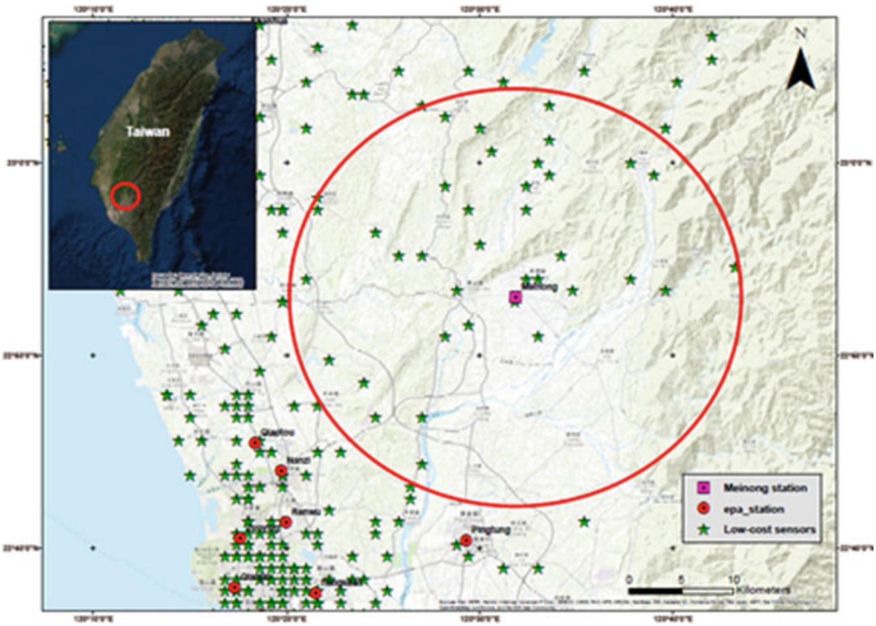


Fig. 4 Spatial distribution of AirBox devices (green start) in Meinong metropolitan area

$$Distance = 2R\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_i - \phi_j}{2}\right) + \cos(\phi_i)\cos(\phi_j)\sin^2\left(\frac{\varphi_i - \varphi_j}{2}\right)}\right) \tag{3}$$

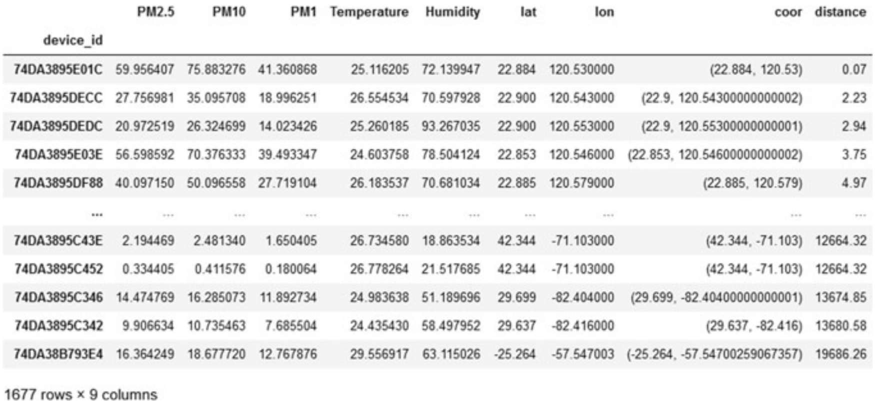
where (R) represent earth radius, ϕ and φ are correspondingly the latitudes and the longitudes of points (i, j).

Using the above equations, the distance between each sensor and EPA station is calculated and sorted sensors by distance Fig. 5 we fix 20 km threshold as area. 14 sensors are nearest to the station.

2.3.3 PCA Selection

The statistical approach Principal Components Analysis [20] is used to find the principal features of a distributed dataset based on the total variance [21]. More precisely, this method has a significant role in reducing the dimensionality of a data set on the premise of retaining the main variance. In multivariate time series analysis, the PCA can be used to generate detailed information about the major components of a set of variables.

In this study PCA was applied to firstly carry out spatial delineation, secondly to simplify the dimensions of variables, and thirdly to figure out the possible sources of $PM_{2.5}$ in a specific region with the massive monitoring data (Fig. 6).



	PM2.5	PM10	PM1	Temperature	Humidity	lat	lon	coor	distance
device_id									
74DA3895E01C	59.956407	75.883276	41.360868	25.116205	72.139947	22.884	120.530000	(22.884, 120.53)	0.07
74DA3895DECC	27.756981	35.095708	18.996251	26.554534	70.597928	22.900	120.543000	(22.9, 120.54300000000002)	2.23
74DA3895DEDC	20.972519	26.324699	14.023426	25.260185	93.267035	22.900	120.553000	(22.9, 120.55300000000001)	2.94
74DA3895E03E	56.598592	70.376333	39.493347	24.603758	78.504124	22.853	120.546000	(22.853, 120.54600000000002)	3.75
74DA3895DF88	40.097150	50.096558	27.719104	26.183537	70.681034	22.885	120.579000	(22.885, 120.579)	4.97
...	...	...	...	...	...	...	...	...	...
74DA3895C43E	2.194469	2.481340	1.650405	26.734580	18.863534	42.344	-71.103000	(42.344, -71.103)	12664.32
74DA3895C452	0.334405	0.411576	0.180064	26.778264	21.517685	42.344	-71.103000	(42.344, -71.103)	12664.32
74DA3895C346	14.474769	16.285073	11.892734	24.983638	51.189696	29.699	-82.404000	(29.699, -82.40400000000001)	13674.85
74DA3895C342	9.906634	10.735463	7.685504	24.435430	58.497952	29.637	-82.416000	(29.637, -82.416)	13680.58
74DA38B793E4	16.364249	18.677720	12.767876	29.556917	63.115026	-25.264	-57.547003	(-25.264, -57.54700259067357)	19686.26

1677 rows × 9 columns

Fig. 5 The distance between EPA Airbox sensors

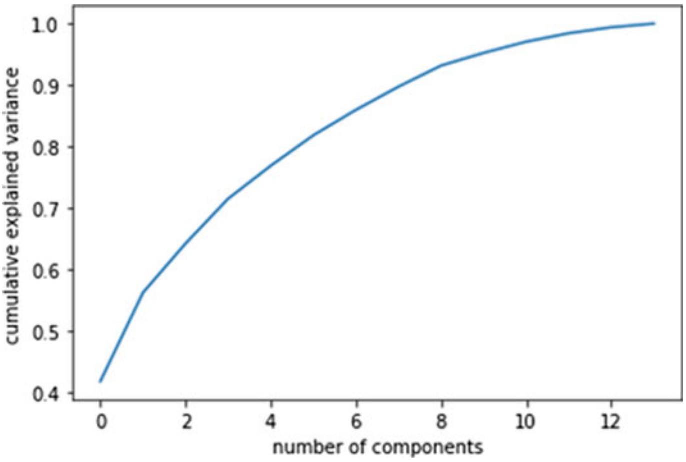


Fig. 6 The cumulative sum of PCA components' variance quantity

More than 40% of the total variance contained in the first component, while eight components are needed to reach more than 90% of the information given by 14 sensors. The analysis of this section lets us expect good results from the forecasting phase.

3 Methodology

These are the key steps involved the prediction of $PM_{2.5}$ concentration in a short time series. First we preprocess the dataset (adjustment, Cleansing, Resampling). Next we detect the nearest neighbor low-cost sensor to the regulatory station. Subsequently,

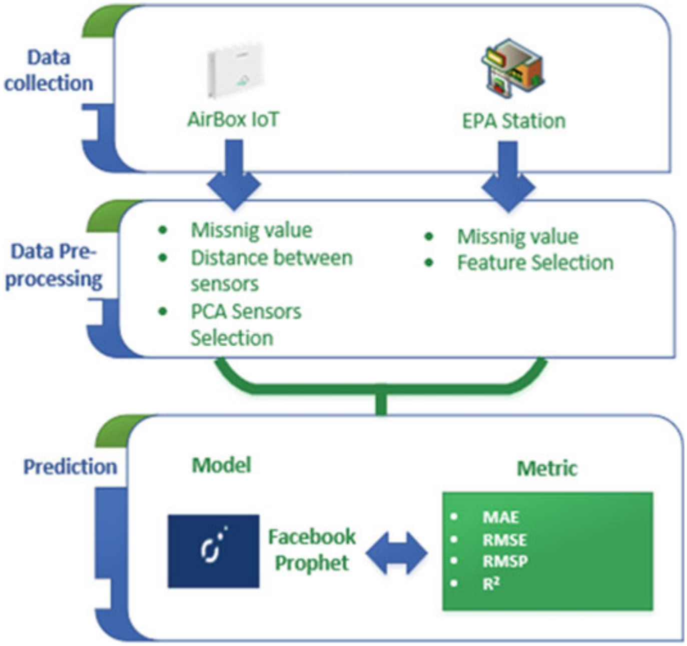


Fig. 7 Flowchart for the hybrid model

the Forecasting with Facebook prophet is performed. After the parameterization of the model and the validation with the statistical metrics, and finally the $PM_{2.5}$ concentrations are estimated. The procedure is shown in the flowchart (Fig. 7).

Facebook prophet model bears a data frame with only two columns. They are ds for Date Time and y for values which necessities being numeric. Based on the two columns, the model learns through training on the historical data.

The core part of the model builds a new data frame for saving the new predicted and predicting values. After that, the validation of the forecast comes into practice by the actual data.

3.1 Facebook Prophet

Facebook Prophet is a model that provides functionality from both generalized linear models (GLM) and additive models (AM). It was specified by Taylor and Letham [22].

Prophet is based on two concepts. The first is developed over many iterations of forecasting a variety of data at Facebook. The second, it helps analysts to make incremental improvements by checking the model manually with a system for measuring and tracking forecast accuracy and flagging forecasts. Moreover, Prophet implements the decomposition of the time series model with three main model components: trend, seasonality, and holidays. These components can be expressed as in equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \tag{4}$$

where $g(t)$ represents the non-periodic changes in the value of the time series, the seasonality of the model's (t) (which can be daily, weekly, monthly, annual or other), $h(t)$ represents the effects of holidays and $\varepsilon(t)$ is the error term.

3.2 Statistical Metrics

Four indicators are used in this paper with the objective of evaluating the effectiveness of the model. They are stated below: mean absolute error (MAE), Mean absolute percent error (MAEP), root mean square error (RMSE), and coefficient of determination (R^2).

3.2.1 MAE

Mean absolute error MAE: is a measure of the average errors between paired observations (the true value) and the model prediction value. The calculation formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

3.2.2 MAEP

The mean absolute percentage error (MAPE) is one of the most popular measures of the forecast accuracy. It is recommended in most scholarly articles [23, 24], MAPE is the average of absolute percentage errors (APE). It usually expresses the accuracy as a ratio defined by the formula:

$$MAE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{6}$$

3.2.3 RMSE

Root mean square error stands for the square root of the mean of the square of all of the errors. As being characterized, it is an excellent error metric for numerical predictions. It is of a sound reflection of the precision of the prediction error. The calculation formula is shown below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

3.2.4 R-Squared

R-Squared (R^2 or the coefficient of determination) is defined as a statistical measure in a regression model. R-squared functions as a representation of the proportion of the variance for a dependent variable which is explained by an independent variable. In other words, R-squared shows how effectively the data fit the regression model. See the calculation formula below:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{8}$$

In all these equations, n is the sample size, (yi) and $\hat{y}_i$ represent the real value and predicted value at time, respectively; $\bar{y}_i$ denotes the mean of all real values.

4 Results and Discussions

In what follows, the attention is shifted towards the comparison of the Facebook prophet model performance. In the first stage, none of the features were included, yet in the second phase the concentration of pollutants and metrological data were inclusively of a considerable significance. In the eventual stage data $PM_{2.5}$ low-costs sensors were added.

The hyper-parameter of Facebook prophet including a linear model and weekly seasonality and the change point scale, which was experimentally determined to be 0.7, were encompassed.

4.1 $PM_{2.5}$ Prediction Univariable

In next 5 days prediction, the MAE was 9.741 and the RMSE 12.995, and the MAPE 17.673 (Table 1 and Figs. 8 and 9).

Table 1 Comparison of prediction performance without including any features

Time	MAE	MAEP	RMSE	R-square
2 days	14.774	19.220	14.892	Nan
5 days	9.741	17.673	12.995	Nan
10 days	10.903	30.689	12.277	Nan

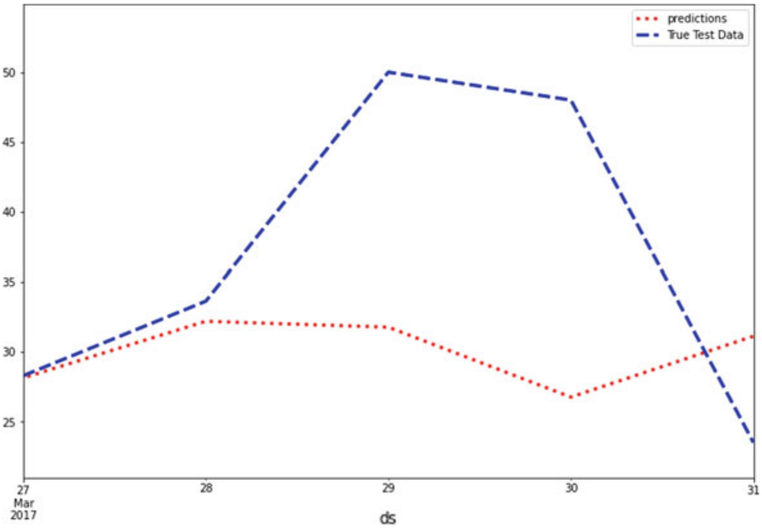


Fig. 8 $PM_{2.5}$ concentration forecasting results for 5 days without including any features

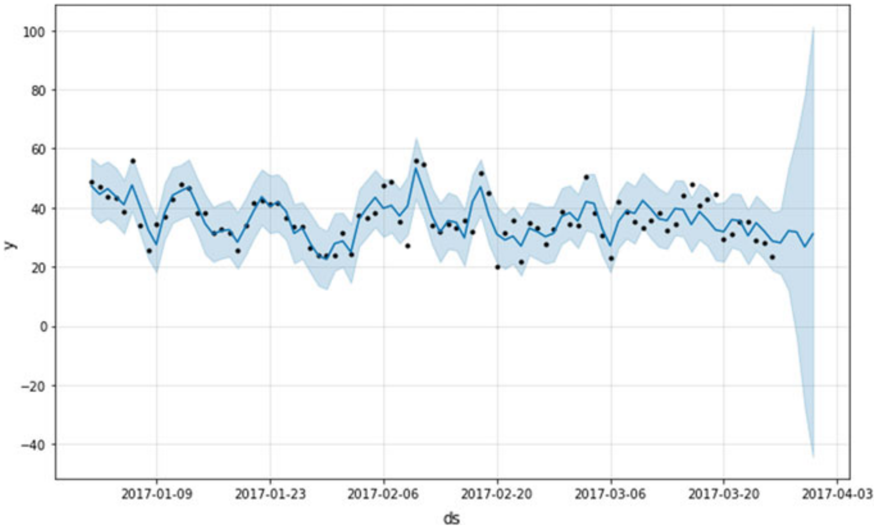


Fig. 9 $PM_{2.5}$ concentration forecasting results without including any features

4.2 $PM_{2.5}$ Prediction Concentration with Other Pollutants

In the next 5 days' prediction, the MAE was 3.649 and the RMSE 12.995, and the MAPE 5.277, and the R-square is 0.867 (Table 2).

Since prediction time increased from 2 to 10 days, MAE and RMSE increased to 2.168 and 9582, respectively (Table 3). The R-square also decreased to 22.1% as the prediction interval increased from 2 to 10 days (Table 2). Both increased error and decreased R-Square are the outcome of unforeseeable conditions that may occur as the time progresses (Figs. 10 and 11).

4.3 $PM_{2.5}$ Prediction Concentration with Other Pollutants and Spatiotemporal Features (IoT)

In the next 5 days' prediction, the MAE was 2.420 and the RMSE 3.365, and the MAPE 5.355, and the R-square is 0.948 (Table 3).

While prediction time increased from 2 to 10 days, MAE and RMSE increased to 4315 and 1692, respectively (Table 3). R-square also decreased to 30.9% as the prediction interval increased from 2 to 10 days (Table 3 and Figs. 12 and 13).

The three above stages showcase that Facebook Prophet Model suit shorter prediction. Once the prediction interval increases, the error increases as well, the reverse is valid. While the low-costs sensors $PM_{2.5}$ concentration were added, the errors decreased and meanwhile the precision increased. Such demonstrates that Facebook prophet model is extremely accurate for short-term predictions and relatively accurate for long-term predictions with spatiotemporal features.

Table 2 Comparison of prediction performance with the concentration of other pollutants

Time	MAE	MAEP	RMSE	R-square
2 days	2.880	5.978	3.686	0.954
5 days	3.649	5.277	12.995	0.867
10 days	5.048	5.978	13.268	0.733

Table 3 Comparison of prediction performance with the concentration of other pollutants and spatiotemporal features

Time	MAE	MAEP	RMSE	R-square
2 days	1.542	1.543	4.732	0.992
5 days	2.420	5.355	3.365	0.948
10 days	5.857	14.598	6.424	0.683

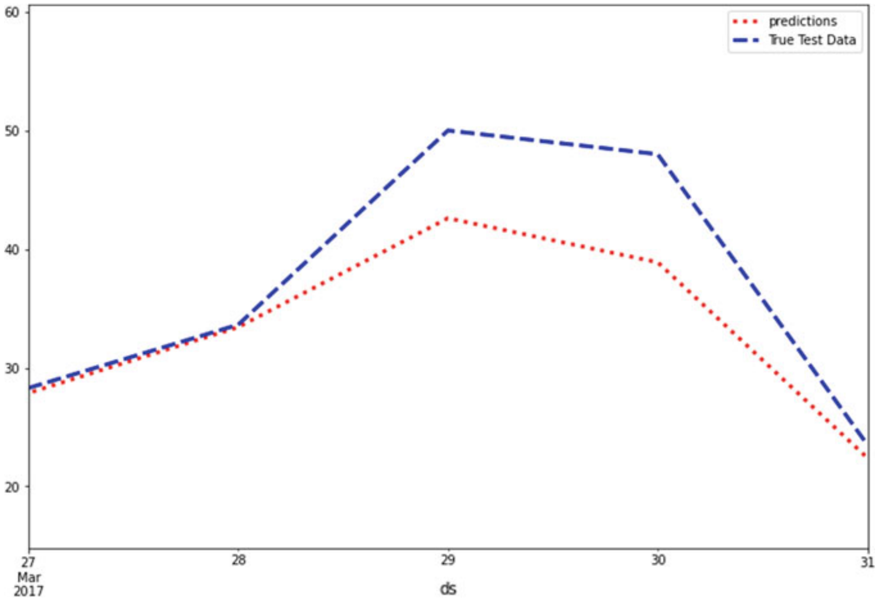


Fig. 10 $PM_{2.5}$ concentration forecasting results for 5 days with the concentration of other pollutants

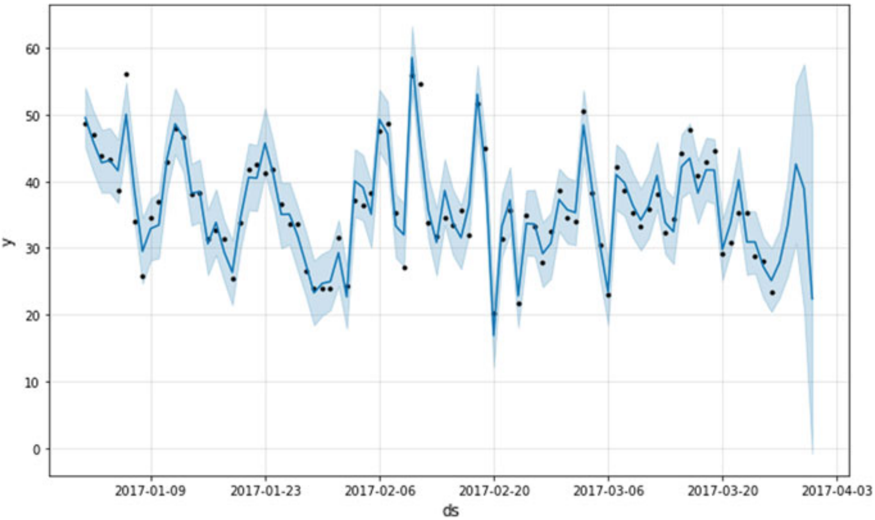


Fig. 11 $PM_{2.5}$ concentration forecasting results with the concentration of other pollutants

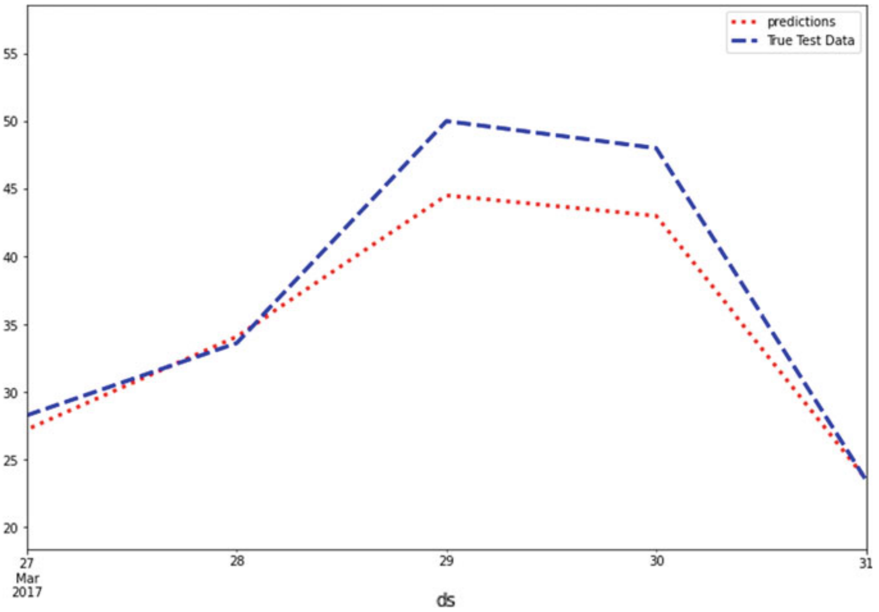


Fig. 12 $PM_{2.5}$ concentration forecasting results for 5 days with the concentration of other pollutants and spatiotemporal features

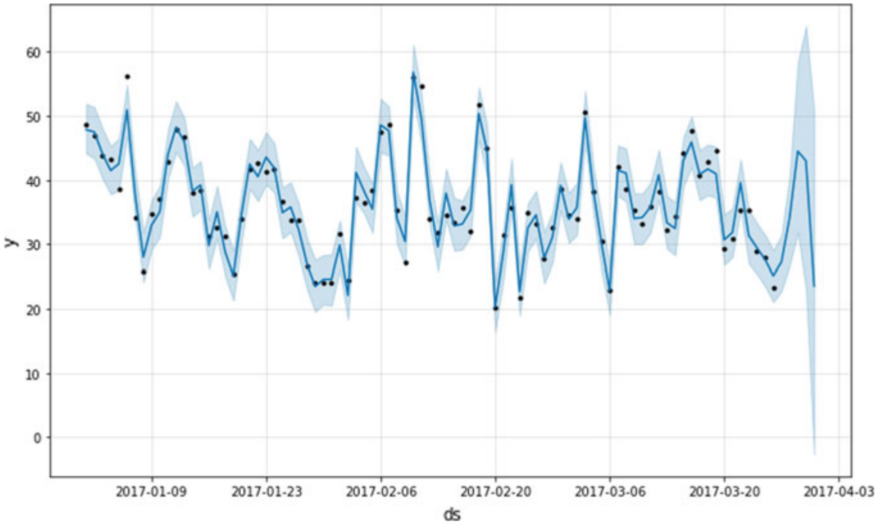


Fig. 13 $PM_{2.5}$ concentration forecasting results with the concentration of other pollutants and spatiotemporal features

5 Conclusion

Be it in the long-term or the short-term, the accurate prediction of $PM_{2.5}$ remains an integral step in minimizing the impairment brought about by air pollution. It is thus a push factor for the government organizations and businesses to readily prepare for any economic or health disturbances that might be arisen due to air pollution-related phenomena. As already figured out by our paper, $PM_{2.5}$ concentration in both the long and the short term precisely estimated in regard to seasonality with the inclusion of low-costs sensors in Facebook prophet model. Despite the fact that Facebook prophet model has the ability to accurately forecast the $PM_{2.5}$ concentration both in the short-term and the long-term, several limitations exist. The Airbox data was published by Taiwan three Months of data. This prevents us from being able to accurately predict in a precise way into 30 days due to the absence of available data. A point of a prominence to mention in this research is that Facebook prophet model ability to predict air pollution in Taiwan cannot be restricted to the targeted area of study, say Tawain. It nonetheless can be applied to other regions in which air pollution poses a hazardous threat.

References

1. WHO (2013) Policy implications for countries in eastern Europe, Caucasus and central Asia. World Health Organization Regional Office for Europe, Copenhagen. Health effects of particulate matter. Available from: http://www.euro.who.int/__data/assets/pdf_file/0006/189051/Health-effects-of-particulate-matter-final-Eng.pdf. Accessed 13 Aug 2021
2. Nery MCD. One third of global air pollution deaths in Asia Pacific. Available from: https://www.who.int/westernpacific/news/detail/02-05-2018-one-third-of-global-air-pollution-deaths-in-asia-pacific. Accessed 13 Aug 2021
3. Tracking the cost of air pollution. Available from: https://www.greenpeace.org/international/campaign/tracking-cost-air-pollution. Accessed 13 Aug 2021
4. Lu HY, Wu YL, Mutuku JK, Chang KH (2019) Various sources of PM2.5 and their impact on the air quality in Tainan City, Taiwan. Aerosol Air Qual Res 19:601–619. https://doi.org/10.4209/aaqr.2019.01.0024
5. Xing YF, Xu YH, Shi MH, Lian YX (2016) The impact of PM25 on the human respiratory system. J Thorac 8(1):69–74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19
6. Apte JS, Marshall JD, Cohen AJ, Brauer M (2015) Addressing global mortality from ambient PM2.5. Environ Sci Technol 49:8057–8066
7. Conibear L, Butt EW, Knote C, Arnold SR, Spracklen DV (2018) Residential energy use emissions dominate health impacts from exposure to ambient particulate matter in India. Nat Commun 9:617
8. Nor NSM, Yip CW, Ibrahim N et al (2021) Particulate matter (PM2.5) as a potential SARS-CoV-2 carrier. Sci Rep 11:2508. https://doi.org/10.1038/s41598-021-81935-9
9. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54(15):2787–2805
10. Alam N, Vats P, Kashyap N (2017) Internet of Things: a literature review. In: 2017 recent developments in control, automation & power engineering (RDCAPE). IEEE
11. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities. IEEE Internet Things J 1(1):22–32

12. Al-Ali AR, Zualkernan I, Aloul F (2010) A mobile GPRS-sensors array for air pollution monitoring. IEEE Sens J 10(10):1666–1671
13. AoT is now an anchor partner in a new NSF-funded project called SAGE. Available from: https://arrayofthings.github.io/index.html
14. Lim M. AirBox: collaborative success in monitoring air quality. Available from: https://nrev.jp/2020/08/06/airbox-collaborative-success-in-monitoring-air-quality/. Accessed 13 Aug 2021
15. Civil IoT Taiwan. Available from: https://ci.taiwan.gov.tw/dsp/en/environmental_en. Accessed 13 Aug 2021
16. Tao Q, Liu F, Li Y, Sidorov D (2019) Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. IEEE Access 7:7669076698
17. Freeman BS, Taylor G, Gharabaghi B, The J (2018) Forecasting air quality time series using deep learning. J Air Waste Manag Assoc 68(8):886. https://doi.org/10.1080/10962247.2018.1459956. PMID: 29652217
18. Chen L-J, Ho Y-H, Lee H-C, Wu H-C, Liu H-M, Hsieh H-H, Huang Y-T, Candice Lung S-C (2017) An open framework for participatory PM2.5 monitoring in smart cities. IEEE Access J 5:14441–14454
19. Mathpal MG (2018) A landmark based shortest path detection by using A* and Haversine formula. Int J Recent Innov Trends Comput Commun 6(7):98–101. https://doi.org/10.17762/ijritcc.v6i7.1690
20. Jolliffe IT (1986) Principal component analysis. Springer, New York
21. Bao B-K, Liu G, Xu C, Yan S (2012) Inductive robust principal component analysis. IEEE Trans Image Process 21(8):3794–3800
22. Taylor SJ, Letham B (2017) Forecasting at scale. PeerJ Preprints 5:e3190v2
23. Bowerman BL, O'Connell RT, Koehler AB (2004) Forecasting, time series and regression: an applied approach. Thomson Brooks/Cole, Belmont, CA
24. Hanke JE, Reitsch AG (1995) Business forecasting, 5th edn. Prentice-Hall, Englewood Cliffs, NJ

Comparative Study Between Different Recommendation Systems in Smart Agriculture

Mohamed Bouni, Badr Hssina, Khadija Douzi, and Samira Douzi

Abstract With the advancement of information and communication technologies, the agriculture sector is changing. Efforts are being made to improve productivity and cut losses by implementing cutting-edge technology and equipment. Because the majority of farmers are unaware of current technology and techniques, several expert systems have been built around the world to assist farmers. These expert systems, on the other hand, rely on the knowledge base that has been stored. In the world today, the Internet of Things (IoT) plays a significant role in human existence by enabling devices to be sensed or controlled remotely over existing networks. The goal of the study was to analyze recommendation system in smart agriculture. Farmers should change the kind of pesticides and water supplies they use on a regular basis. The method can aid farmers in the production of appropriate crops. This allows them to enhance their lifestyles and contribute more to society.

Keywords Recommendation system · Smart agriculture · IoT · Crops recommendation

M. Bouni (✉) · B. Hssina · K. Douzi
Laboratory LIM, IT Department FST Mohammedia, Hassan II University, Casablanca, Morocco
e-mail: mohamed.bouni1-etu@etu.univh2c.ma

B. Hssina
e-mail: badr.hssina@fstm.ac.ma

S. Douzi
FMPR, Mohammed V University in Rabat, Rabat, Morocco
e-mail: s.douzi@um5r.ac.ma

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_11

1 Introduction

In Agriculture, RSs have a significant impact on managing and using the resources efficiently, such as fertilizers, agrochemicals, irrigation. In a fertilizer RS was developed to enrich the soil and increase its productivity. In recent years, agriculture and the food system have seen fundamental changes, reflected by price increases and driven by income and population expansion, migration and urbanization, as well as speculation. Without a doubt, the world needs to invest in agriculture [1].

As a result of technical developments such as sensors, gadgets, machinery and information technology, modern farms and agricultural operations are substantially different than those of several decades before. Agriculture today frequently employs sophisticated technologies such as robotics, temperature and moisture sensors, aerial photography, GPS technology, and a plethora of complex IoT devices. Agriculture's modern technology allow businesses and farmers to be more profitable, efficient, safe, and environmentally friendly [2].

The growth of digital agriculture and its associated technologies has created a plethora of new learning opportunities. Remote sensors, cameras, and other linked equipment will collect data 24 h a day, seven days a week throughout an entire farm or area. These will keep track of plant health, soil conditions, temperature, and humidity, among other things. The amount of data generated by these sensors will be enormous. This allows farmers to have a better knowledge of the state of affairs on the ground by using innovative technology to provide them with more precise and timely information about their condition [3].

Environmental data collected by remote sensors is processed by algorithms and statistical data that can be comprehended and used by farmers to make decisions and keep track of their farms. The algorithmic rule's ability to anticipate outcomes improves as more inputs and statistical data are collected. Farmers are expected to employ these technologies to achieve their goal of improved yield by making better field selections [4].

2 AI and IoT for Smart Agriculture

Agriculture research is being improved in a variety of ways in order to enhance the quantity and quality of agricultural performance. Researchers have worked on a variety of initiatives involving soil characteristics, weather conditions, and crop scouting. Some experiments took place in actual farm fields, while others took place in polyhouses [5].

Carrnige Mellon University researchers used Wireless Sensor Technology to work on a plant nursery. It is described a wireless sensor network-based polyhouse monitoring system that makes use of environment temperature, humidity, CO_2 level, and sufficient light detection modules. This control technique in polyhouse allows automatic polyhouse adjustment [6].

The authors have created and implemented a strategy in developing a real-time crop monitoring system to boost rice plant production in some projects, such as an effective method for crop monitoring using wireless sensor networks. To monitor the dampness of the leaves, this system used motes with sensors. IoT provides a platform for researchers to keep real-time data and transmit notifications to farmers promptly. The Internet of Things allows for quick access to data from sensor nodes. The Internet of Things is also used in the product supply chain business process [7].

Cloud architecture helps IoT preserve big data in agriculture, such as historical information, soil qualities, and fertilizer distribution, image cultivation via camera and information received through sensors, and recording information, among other things. For the design of a standard work model, the authors analyzed collected data to identify a correlation between environment, work, and yield. Detection of faults and monitoring for negative signs [8].

In the field of agriculture, spatial data mining is used. Temperature and rainfall data are collected as initial geographical data and analyzed to reduce crop losses and boost agricultural yield. For spatial association analysis, they employed an optimization method to show ongoing refinement [9].

This study proposes a method in which all of the important aspects are considered at the same time and a solution is found so that the system is not overly confusing for the user. Unlike other models proposed by earlier researchers, this system evaluates all of the important components that are essential for plant growth and processes them all at the same time using various algorithms, whereas other models consider only parameters at a time while keeping the other factors constant [10].

Some studies, for example, are carried out to determine the rate of evaporation and how plant growth is affected when there is insufficient water. As a result, a derived equation is presented.

$$ETo = Kpan \times \text{Epan}$$

ETo: reference crop evapotranspiration.

Kpan: pan coefficient.

Epan: pan evaporation.

Despite the fact that an equation is proposed, it has several drawbacks. This might mostly be done for a land with a smaller area. This is not suited for commercialization since the profit will be minimal if the cultivation area is reduced. It would be an issue if there was adequate water for plants but no temperature, because the key environmental elements have a mutual link in plant growth. The system's task is finished once the farmer or user has cultivated the projected crop kind. However, the approach suggested in this study includes a feedback system. The system can track plant growth and provide feedback if the farm is malnourished even after recommending the optimal crop kind. In order for the user to take the required safeguards ahead of time [11].

3 Related Work

Raja et al. [12] suggested a recommendation system for crops based on rainfall, temperature, and other region factors. Farmers should aim for high yields and earnings based on crop output and current market pricing. Mokkaram and Arefin employed an algorithm called Upazila Selection for harvest suggestion, which incorporates the present seasonal detection. It uses agro-climatic data and agro-cultural context to recommend adequate crops at the right season.

Pudumalar et al. [13] analyzed a data set containing soil properties and other test findings and applied precision farming approaches to boost production and profitability. Back then, they based their suggestions on the meteorological data and climate circumstances. The weather was represented by The Hidden Markov model and predict advices. Their weather study reveals a vastly enhanced suggestion mechanism.

Okayasu et al. [14] employed a scalable graph-based collaborative filtering recommendation method, by opposition to a central recommender system that suggest the correct representation utilizing a graph and another image format. This improves accuracy and coverage area. Takashi Okayasu et al. discussed how they used an algorithm and devices to create a low-cost field environmental monitoring and plant growth measurement system for smart agriculture. Plant height, leaf color, and other parameters are measured using the sensors.

Yan and Shou-hua [15] used IoT technology in conjunction with a remote monitoring system with the goal of collecting and alerting the agricultural production environment in real time via SMS. This publication offered farmers with a productive grasp of how to sow and care for their crops. Lai and Zeng improved crop yield by using technology such as Big Data, IoT, and cloud computing. This model examines the crop sequence, the next crop to be planted, fertilizer needs, and so on. This approach makes it easier to estimate total production and fertilizer requirements per crop, and control agricultural product costs.

Mekala and Viswanathan [16] introduced a study that uses a humidity and humidity control-based on recommendation system to collect the data sets required from the geographical area to be cultivated and the kind of fertilizer needed. Increase the crop's productivity. Mekala and Palanisamy created an algorithm to forecast crop growth rates based on available resources, indicating the needs to be given throughout growth based on the crop type.

4 Proposed Method

New ways for maximizing yield and nutrition while conserving water resources can be developed by utilizing existing infrastructure and applying cutting-edge methodologies and technology, such as improved phenotyping and genetics. Farmer/academia collaboration, test beds, and platforms for implementing and testing new productive agriculture technologies for monitoring, planting, and harvesting will insure technology uptake [17].

These test beds are also valuable in evaluating solutions for field sensing, data science, and visual analytics. Furthermore, a new era for a sustainable and healthy community can be developed by targeted researches to control contamination of agricultural goods at all phases of production. Pilot studies are also required for pest-resistance advances in grain, fruit, and vegetable storage and transportation [18].

Environmental researchers must develop and deploy new technologies to improve detection and control of PFAS (Fire Protection Accreditation Scheme) with less inputs in order to satisfy the demand for realistic approaches to manage the possible environment repercussions of PFAS. Control, treatment, destruction, and removal of PFAS in soils requires improved ideas that are more practical and efficient. To accomplish this complex and difficult task, multidisciplinary efforts combining several environmental science fields are required to develop such instruments and execute them in the field [19].

For better comprehension the carbon and nitrogen cycles, new technologies are required. To accomplish this goal, technology researchers and soil scientists must work together to develop improved in-situ systems capable of assessing physical, chemical, and biological aspects of soil in large scale fields comprising various types of soil. In this context, current advances in health and energy can be applied to precision agriculture [20].

Innovation and automation in subterranean sensing and secure communications, data gathering, analysis, and visualization will be critical in developing technology-aware, advanced precision agriculture methods. The development of sensors for soil and water quality across networked landscapes is required. Developments in smart agriculture data analytics, in-situ and remote sensing, indigenous and local knowledge, should be integrated into operational systems [21].

5 Methodology

5.1 Dataset Collection

Because the essential environmental variables have a mutual link in plant growth, it would be a problem if there was enough water for plants but no warmth. Once the farmer or user has cultivated the expected crop kind, the system's job is done. The approach proposed in this paper, however, includes a feedback system. Even after selecting the best crop kind, the system can track plant growth and provide feedback if the farm is undernourished. To allow the user to take the necessary precautions ahead of time [22].

5.2 Collecting Environment Factors

The environmental elements needed to be obtained in order to compare and predict the initial data set. Arduino microcontrollers are used to collect environmental data. Because the temperature and humidity sensors are both made up of a single microcontroller, data is collected using four sensors. Sunlight intensity sensor, soil moisture sensor, soil pH sensor, and humidity and temperature sensor are the sensors. The sensors are linked to an Arduino Wi-Fi module, and the data collected is relayed to a database. The data is cleaned and processed using clustering and other techniques before being passed on to the next component of crop recommendation and saved in the database [23].

5.3 Crop Prediction

Because environmental circumstances vary by region, a machine learning model is employed to forecast the optimal crop variety for the chosen land. Machine learning algorithms are used to train the crop suggesting model with the data obtained from the Arduino sensors in order to identify the best crop to cultivate with the highest likelihood of growing. The optimal crop type is chosen using the Nave Bayes and Support vector machine techniques. It was chosen what type of crops the farmer should cultivate based on this model. This is accomplished through the examination of parameters like as humidity, temperature, soil moisture, pH level, and sunlight [24].

Using two machine learning algorithms, the system primarily recommends four crop types based on the above-mentioned characteristics. Naive Bayes is a classifier model construction technique that assigns class labels to problem cases represented as vectors of feature values, with the class labels selected from a finite set. The goal of the support vector machine is to find a hyper plane in N dimensional space (N = the number of features) that distinctly classifies the input points [25].

5.4 Monitoring and Feedback

According to the environmental variables of the chosen piece of land, this proposed system product would primarily identify four types of crops. The soil condition or any other modifications in the specified land would be the explanation for obtaining a probability of greater than 90%. However, the farmer's feedback mechanism is integrated in the system to avoid these influences affecting crop prediction [26].

Following the recommendation of a crop type, the farmer is asked for more information and input on a regular basis via the mobile application, which is used to guide the farmer with essential safeguards. The feedback mechanism is employed in the mobile application by selecting the crop type to deliver the appropriate feedback. As a result, the product's overall accuracy and reliability improve with time [27].

6 Comparison of Diverse Recommender System for Smart Agriculture

The strategies employed by several smart agriculture researchers with IoT are depicted in Table 1. It also includes a list of recommendations that we believe should be adopted in the system in the future.

7 Conclusion

In a modern environment with limited space and agricultural knowledge, all elements are examined from the perspective of the farmer and the plant, and the farmer is appropriately guided till harvesting. It is critical to have information and understanding of the factors that affect plant culture, as well as how to maintain or regulate them, before choosing a plant to grow. These characteristics are automatically processed by this system, which determines the crop kind to be planted. Once the plant has been cultivated, the farmer is asked for comments on a monthly basis. The system is self-trained by this feedback, and the accuracy with time and data collected is improved. This technology eliminates the need for a specialist's direction, and it requires minimal maintenance. As a result, deploying this system will have no additional financial impact on the user. Village farmers may have cultivated the "same" crop for centuries, yet weather patterns, soil conditions, and pest and disease epidemics varied with time. Using the proposed approach, farmers will be able to cope with and even benefit from these changes by receiving updated information.

Table 1 Comparison of diverse recommender system for agriculture using

No.	Methods	Concept used	Results/outcome	Advantage	Inconvinients	Recommendation
1	Recommendation based on the climate yield and the farmer for rainfall and temperature	IoT	By implementing this strategy, farmers will be able to increase the yield of their crops, which will benefit them	It increases the crop productivity by 3%	More sensors may be used to get the full picture of the crop's health, product, and soil conditions. For every node, the sensors require batteries, resulting in a higher battery consumption	Solid ph. (soil ability to resist change in acidity) can be implemented here
2	The Pearson correlation similarity technique is used to calculate several agro-ecological and agro-climatic variables in this user's area	IoT	crop recommendations are generated at the upazil level based on the algorithm and data sets acquired	It increases the crop productivity by 7%	the best way to increase the crop productivity is to use an expert system and IoT	Advanced technologies, such as physiographic mechanisms, can be applied in this case
3	Precision agriculture is the approach utilized here, which is a modern farming technology	K nearest Neighbors and naive random tree technique	This method was more productive and profitable, and it assisted many people in planting the proper crop at the right time	Expert system and IoT is an excellent way to improve the crop rate	The expert system does provide recommendations for herbicides and pesticides based on the disease. There is also a possibility that the battery usage will be very significant, as well	We can optimize crop rate by using technology such as camera, pH rate, and sunshine

(continued)

Table 1 (continued)

No.	Methods	Concept used	Results/outcome	Advantage	Inconvinients	Recommendation
4	Recommendation based on local weather data and people's interest in similar conditions in the past	IoT, hidden Markov model	The weather context greatly enhances the recommended system	the accuracy of the weather data is increased	Due to the Li-Fi sophistication, it takes time to install in underdeveloped countries, but once it is, it can be an upgrade to the sector. Furthermore, cattle sensors are dangerous because the animals may fidget with them, in the meantime, it remains unproductive	We can use li-fi technology to increase the accuracy of the results
5	Recommendation based on graphic analysis and a trust enhanced factor	IoT	The addition of trust to the RS improves accuracy as well as coverage	It increases IOT network accuracy	The biggest downside is that playhouses cannot be utilized on large fields since they raise the expense. When we consider a broad region, the drip irrigation setup cost may also rise	IOT networking should be increased
6	Recommendation based on trust and regulation model	IoT	It aims to provide protection against IOT attackers and tackle security challenges	It improves security	Security challenges can be faced	Security should be increased

(continued)

Table 1 (continued)

No.	Methods	Concept used	Results/outcome	Advantage	Inconvinients	Recommendation
7	Proposed technique for personalizing agricultural information based on user clustering	IoT, data mining, data warehousing	The framework significantly improves the system for personalizing Agriculture information	It improves the website quality service	The transmission distance is shorter, which can present issues if connectivity is poor	It has been enhanced by the addition of efficient and optimized code
8	The similarity baring capacity is calculated on the energy state	dynamic tag, IoT, precision	The following theory was validated by user similarity and service resource value	It improves the precession	The ubimote sensor used in the system can fail, which may cause the entire system to fail because it gathers and communicates the data to the farmers	We can use improved services
9	Design of a knowledge base and applications for fertilization recommendations	IoT	The data gathered by the sensors was easily accessible via the Android application, and the project was deemed a success	It increases the fertilization efficiency	N/A	A higher-quality camera can be used, as can the master node concept have been described in paper 2. Battery efficiency can be boosted by programming the microcontroller to stand by when the system is not in use, allowing the batteries to be charged via sunlight at that time

(continued)

Table 1 (continued)

No.	Methods	Concept used	Results/outcome	Advantage	Inconvinients	Recommendation
10	Cognitive approach for recommended system	IoT, the cognitive system sensors	The framework is an engine observing the things like sensors by considering a used item	It increases stability in the system	It has security issues	N/A
11	Recommendation system for brown plant hopper control	IoT, multiskilling,	It decreases the bph (Buffer pH) rate in the break down	It slows down the bph (Buffer pH) growth rate	Battery consumption will rise as more sensors are deployed to aid farmers in their work	In order to decrease the bph (Buffer pH) rate, we can increase the number of measures
12	Recommendation based on the application and improvement of intelligence recommendation of agriculture knowledge	IoT, the major algorithm	The sensors and Raspberry PI were linked and wireless communication was created using IoT	It increased the improvement in information for agriculture	The power supply can be a drawback if the entire system fails due to a lack of power supply. Excess sunlight well impairs photographs during the day, while images cannot be taking accurately at night	N/A

References

1. Shahzadi R, Ferzund J, Tausif M, Asif M (2016) Internet of Things based expert system for smart agriculture. Int J Adv Comput Sci Appl 7(9). https://doi.org/10.14569/IJACSA.2016.070947
2. Li C, Niu B (2020) Design of smart agriculture based on big data and Internet of things. Int J Distrib Sens Netw 16(5):155014772091706. https://doi.org/10.1177/1550147720917065
3. Abbas K, Afaq M, Ahmed Khan T, Song W-C (2020) A blockchain and machine learning-based drug supply chain management and recommendation system for smart pharmaceutical industry. Electronics 9(5):852. https://doi.org/10.3390/electronics9050852
4. Suresh G, Senthil Kumar A, Lekashri S, Manikandan R (2021) Efficient crop yield recommendation system using machine learning for digital farming. Int J Mod Agric 10(1):906–914
5. Bu F, Wang X (2019) A smart agriculture IoT system based on deep reinforcement learning. Future Gener Comput Syst 99:500–507. https://doi.org/10.1016/j.future.2019.04.041
6. Pratyush Reddy KS, Roopa YM, Rajeev KLN, Nandan NS (2020) IoT based smart agriculture using machine learning. In: 2020 second international conference on inventive research in computing applications (ICIRCA), Coimbatore, India, juill. 2020, pp 130–134. https://doi.org/10.1109/ICIRCA48905.2020.9183373
7. Nithiya S, Annapurani K (2021) Optimized fertilizer suggestion in smart agriculture system based on fuzzy inference rule. Acta Agric Scand Sect B-Soil Plant Sci-Ence 71(3):191–201
8. Krishna K, Silver O, Malende WF, Anuradha K (2017) Internet of Things application for implementation of smart agriculture system. In: 2017 international conference on SMAC IoT Soc. Mob. Anal. Cloud-SMAC, p. 54-59, 2017.
9. Mashal I, Alsaryrah O, Chung T-Y, Yuan F-C (2020) A multi-criteria analysis for an internet of things application recommendation system. Technol Soc 60:101216. https://doi.org/10.1016/j.techsoc.2019.101216
10. Sinha BB, Dhanalakshmi R (2022) Recent advancements and challenges of Internet of Things in smart agriculture: a survey. Future Gener Comput Syst 126:169–184. https://doi.org/10.1016/j.future.2021.08.006
11. Lopez-Ridaura S, Frelat R, van Wijk MT, Valbuena D, Krupnik TJ, Jat ML (2018) Climate smart agriculture, farm household typologies and food security. Agric Syst 159:57–68. https://doi.org/10.1016/j.agsy.2017.09.007
12. Raja SKS, Rishi R, Sundaresan E, Srijit V (2017) Demand based crop recommender system for farmers. In: 2017 IEEE technological innovations in ICT for agriculture and rural development (TIAR), Chennai, pp 194–199. https://doi.org/10.1109/TIAR.2017.8273714
13. Pudumalar S, Ramanujam E, Rajashree RH (2016) Crop recommendation system for precision agriculture, p 5
14. Okayasu T, Alsaryrah O, Chung TY, Yuan FC, Affordable field environmental monitoring and plant growth measurement system for smart agriculture. Proc Int Conf Sens Technol 1–4
15. Yan Z, Shou-hua B (2012) Research on optimizing recommend system for agriculture information personalization based on user clustering. In: Proceedings of 2012 international conference on industrial control and electronics engineering, pp 1477–1480
16. Mekala M, Viswanathan P (2017) A novel technology for smart agriculture based on IoT with cloud computing. In: 2017 international conference on SMAC IoT Soc Mob Anal Cloud -SMAC, pp 75–82
17. Chandra A, McNamara KE, Dargusch P (2018) Climate-smart agriculture: perspectives and framings. Clim Policy 18(4):526–541. https://doi.org/10.1080/14693062.2017.1316968
18. Cicioğlu M, Çalhan A (2021) Smart agriculture with internet of things in cornfields. Comput Electr Eng 90:106982. https://doi.org/10.1016/j.compeleceng.2021.106982
19. Barasa PM, Botai CM, Botai JO, Mabhaudhi T (2021) A review of climate-smart agriculture research and applications in Africa. Agronomy 11(6):1255. https://doi.org/10.3390/agronomy11061255

20. Kumar A, Sarkar S, Pradhan C (2019) Recommendation system for crop identification and pest control technique in agriculture. In: 2019 international conference on communication signal-Cessing ICCSP, pp 0185–0189
21. Lavanya G, Rani C, Ganeshkumar P (2020) An automated low cost IoT based Fertilizer intimation system for smart agriculture. Sustain Comput Inform Syst 28:100300. https://doi.org/10.1016/j.suscom.2019.01.002
22. Aniley AA, Kumar A (2019) Advanced sensor materials based real-time soil moisture content and temperature monitoring using IoT technology in smart agriculture. Indian J Environ Protect 39:639–644
23. Sinha A, Shrivastava G, Kumar P (2019) Architecting user-centric internet of things for smart agriculture. Sustain Comput Inform Syst 23:88–102. https://doi.org/10.1016/j.suscom.2019.07.001
24. Dinesh D, Frid-Nielsen S, Norman J, Mutamba M, Loboguerrero Rodriguez AM, Campbell BM, Is climate-smart agriculture effective? A review of selected cases. CCAFS Work Pap, no 129
25. Marcu I, Suciu G, Balaceanu C, Banaru A (2019) IoT based system for smart agriculture. In: 2019 11th international conference on electron comput artif intel ECAI, pp 1–4
26. Pratama B, Fenrianto S, Fajar AN, Amyus A, Nurbadi R (2018) A smart agriculture systems based on service oriented architecture. In: 2018 3rd international conference on information technology & electronic engineering, ICITISEE, pp 281–286
27. Thakare B, Rojatkar DV (2021) A review on smart agriculture using IoT. In: 2021 6th international conference on communication and electronics systems, ICCES, pp 500–502

Industry 4.0 and Intelligent Transportation

Configuration Security for Sustainable Digital Twins of Industrial Automation and Control Systems in Emerging Countries

Zhihan Lv, Jingyi Wu, Dongliang Chen, and Anna Jia Gander

Abstract Digital twins (DTs) under industry 4.0 provide the manufacturing industry with the mapping relationship between products in physical space and virtual space, as well as the process of recording, simulating, and predicting the operation trajectory of the all life cycle of objects in the physical world and digital virtual space. This paper is to analyze and study the configuration security of industrial automation and control systems to contribute to the security of the world's industrial control networks. Also, it is hoped to the world's industries as soon as possible to get rid of the risk of invasion of industrial control systems. This paper uses an improved artificial bee colony (ABC) algorithm combined with support vector machine technology for research, and expects to achieve good attack detection results. In the case of small-scale data, the performance of this method is more general. However, in the case of large-scale data, the detection accuracy, false alarm rate, and detection time of this method are all excellent. Compared with other attack detection methods, the method proposed has certain advantages in various aspects. Security situation awareness can be used to detect, analyze, and visualize the security situation of industrial control network platform and data flow process, and analyze the security threat intelligence from the time and space dimensions through DTs technology. The attack detection method of industrial control system based on ABC algorithm can effectively detect the attack state, and provides an important theoretical basis for the research of attack detection methods.

Keywords Industrial automation · Control system · Attack detection · Artificial bee colony algorithm

Z. Lv (✉) · J. Wu · D. Chen
College of Computer Science and Technology, Qingdao University, Qingdao, China

J. Wu
e-mail: wjy_515151@163.com

A. J. Gander
Department of Applied IT, The University of Gothenburg, Gothenburg, Sweden

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_12

1 Introduction

As the information industry is expanded and the information technology develops, the Internet has become essential in people's lives. People's daily life is more closely connected with the Internet, but in the development of the Internet, the requirements for network equipment have become higher [1–3]. In the last century, the Internet has not developed fast, and people's understanding of Internet technology is not mature enough. Therefore, the application of the Internet in industrial automation and control systems was rare at that time [4, 5]. The Internet at that time could not meet the requirements of industrial automation and control systems. Its instantaneity of information was poor. Also, each research institution built behind closed doors and did not form a benign development system. No agreement was reached on various protocols, software and hardware, and development paths. Thus, the research on industrial automatic control systems lags behind and the impact on society is minimal [6–8]. As one of the key enabling technologies to realize intelligent manufacturing, digital twins (DTs) are very suitable for industrial network environment. DTs are usually defined as building a virtual model equivalent to physical entities, adding or expanding new capabilities for physical entities through virtual real interaction feedback between physical entities and virtual models, data fusion analysis, decision iterative optimization, etc.

Applying DTs system promotes the development of industrialization. While improving production efficiency and product quality, it also brings many industrial safety problems. The frequent occurrence of production accidents makes industrial safety paid more attention to. In the process of the gradual connection and integration of Internet technology as well as traditional industrial automation and control systems, major enterprises have also applied the Internet to improve the efficiency of their own enterprises. Since the internet technology progresses fast in the twenty-first century, industrial control systems have been applied to infrastructure fields such as electricity, natural gas, and aviation. Also, the industrial control system is also undergoing various updates and developments with the Internet's renewal [9]. The current industrial control system has not only been used as a connection network within the company, but its role has been extended to various aspects. Therefore, its connection has changed from a small area to the entire Internet. One of the important tasks is to communicate and exchange information with the Internet within the company [10]. At present, more than 80% of the infrastructure systems in the world need to rely on industrial control systems for various automation tasks [11].

Therefore, the deeper and deeper connection between industrial control systems and the Internet brings more challenges. The open network environment makes industrial control systems affected by a large amount of network information. The data exchange process also loses stronger closedness and security [12]. Thus, the network configuration security of industrial control systems has received unprecedented attention. The industrial control network is the most important part of the industrial control system. It is also constantly improved with the equipment update of the industrial control system and continues to assume more important responsibilities [3]. In recent years, attacks on industrial control systems have been countless and triggered a series of security incidents. In 2010, the Iranian nuclear power plant was attacked by a US hacker, and its control system was successfully infected with the network virus used by the hacker. It caused a large number of centrifuges in the control system of the nuclear power plant to be damaged, which brought serious losses to the economy and security of the nuclear power plant. Such industrial control system attack events emerge in endlessly.

Therefore, this paper analyzes and studies the configuration security of industrial automation and control systems to contribute to the security of the world's industrial control networks. Also, it is hoped to the world's industries as soon as possible to get rid of the risk of invasion of industrial control systems.

2 Related Works

The deeper and deeper connection between industrial control systems and the Internet brings more challenges. The open network environment makes industrial control systems affected by a large amount of network information. The data exchange process also loses stronger closedness and security [13].

As the problem of the industrial control network becomes more and more severe, the research on the network problem of the industrial system is also more and more. Galloway and Hancke [14] described the industry in detail, particularly emphasizing the differences between enterprise and industrial networks. They briefly introduced the history of industrial networks, further explained some operations of industrial networks, outlined popular protocols currently in use, and described current hotspots [14]. McLaughlin et al. [15] explored the prospects of network security for integrated circuits. These include a brief history of industrial control system network attacks, as well as current trends in industrial control system attacks and defenses [15].

In addition to the description and research of industrial control systems and networks, many scholars have also conducted research on the security inspection of industrial control systems. Ponomarev and Atkison [16] described the security measures of industrial control systems and proposed a method for detecting network intrusions of additional industrial control systems. By measuring and verifying the data transmitted on the network, on the same network, the accuracy of the intrusion detection system developed to distinguish attackers from engineer machines reached 94.3%. The accuracy of distinguishing attackers and engineer machines on the Internet was 99.5% [16]. Zhang et al. [17] developed a network attack detection system based on the defense-in-depth concept. This attack detection system provides multiple layers of defense to gain valuable time for defenders before unrecoverable consequences occur in the physical system. Intrusion detection results show that k-nearest neighbors, guided aggregation, and radiofrequency have a low false alarm and false alarm rates for man-in-the-middle and denial-of-service attacks. It can provide detection of these network attacks [17]. Shen et al. [18] proposed a hybrid enhanced device fingerprint recognition method to enhance the traditional intrusion detection mechanism in industrial control networks. Utilizing the simplicity of program processing and the stability of hardware configuration, the intrusion classification and detection methods based on device fingerprints were evaluated. Forged attacks and intrusions were performed on the method to verify the robustness and effectiveness of the method [18].

The research on the risk of industrial control systems is also the top priority of the research on industrial control networks. Zhang et al. [19] researched the risk propagation model of industrial control systems. They proposed a fuzzy probabilistic Bayesian network method for dynamic risk assessment and an approximate dynamic reasoning algorithm suitable for dynamic assessment of integrated circuit network security risks. The effectiveness of this method was verified after experiments [19]. Volkova et al. [20] conducted a comprehensive investigation of the security of the most important control system communication protocols. To achieve comparability, a universal test method based on attacks that exploit known control system protocol vulnerabilities was created for all protocols [20].

Therefore, the research here will be based on previous research. However, due to space and time, it will only conduct intrusion detection research on industrial control networks. This paper uses an improved artificial bee colony (ABC) algorithm combined with support vector machine technology for research and expects to achieve good attack detection results.

3 Methods

3.1 Industrial Control Network in DTs Environment

The development of industrial control networks is a step-by-step process. From the traditional control network to the more advanced Fieldbus, and later with the advancement of scientific and technological civilization, it has developed into the current research hotspot of industrial Ethernet and wireless network control [21]. The development of future industrial networks requires efforts from the real-time, security, and reliability of communications. It is not easy to reach this level. Realizing multi-total line integration, the real-time heterogeneous network is also an important direction for future development [22]. Compared with the traditional industrial Internet platform, the industrial DTs system is a complex virtual real iterative system with the integration of man, machine, and environment. The importance of man-machine remote control security, man-machine cooperation security, and environmental security in the system is more than that of the traditional industrial Internet platform. There are many human-computer interaction (HCI) modes and technologies in industrial DTs system, including human-computer cooperation, personalized biometric, omni-directional perception, etc. The application of these new modes and technologies makes the HCI-oriented DTs system show the characteristics of people-centered, uncertain system, and urgent need for security control. Figure 1 presents the remote interactive information security mechanism of DTs system. The whole security architecture adopts the intelligent module unit architecture. Each intelligent module unit has independent security control capability. When a problem occurs in an intelligent module unit in the system, it will not spread to other intelligent module units.

Nowadays, with the emergence and development of the Internet and related technologies, the way companies do business has changed a lot. It enables information and communication to surround the entire social life and penetrate it to a large extent. Where there is civilization in the world, there is an industrial control network. In some fields, such as the field of office automation, Internet technology has appeared in office equipment.

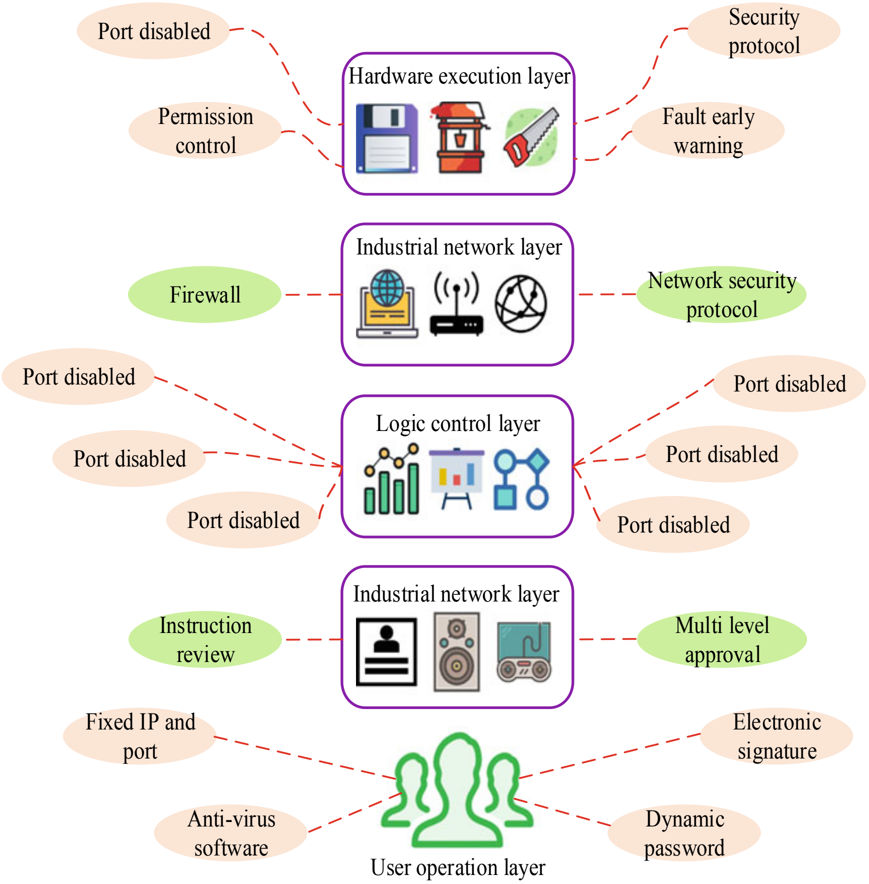


Fig. 1 The security mechanism for remote interactive information of DTs system

The industrial control network consists of three parts: enterprise network, monitoring network, and detection network. Figure 2 presents the specific distribution.

The control layer network is directly connected to the automation control system. In the early days of the development of industrial control networks, their similarity to the Internet was low. The industrial control system was completely independent.

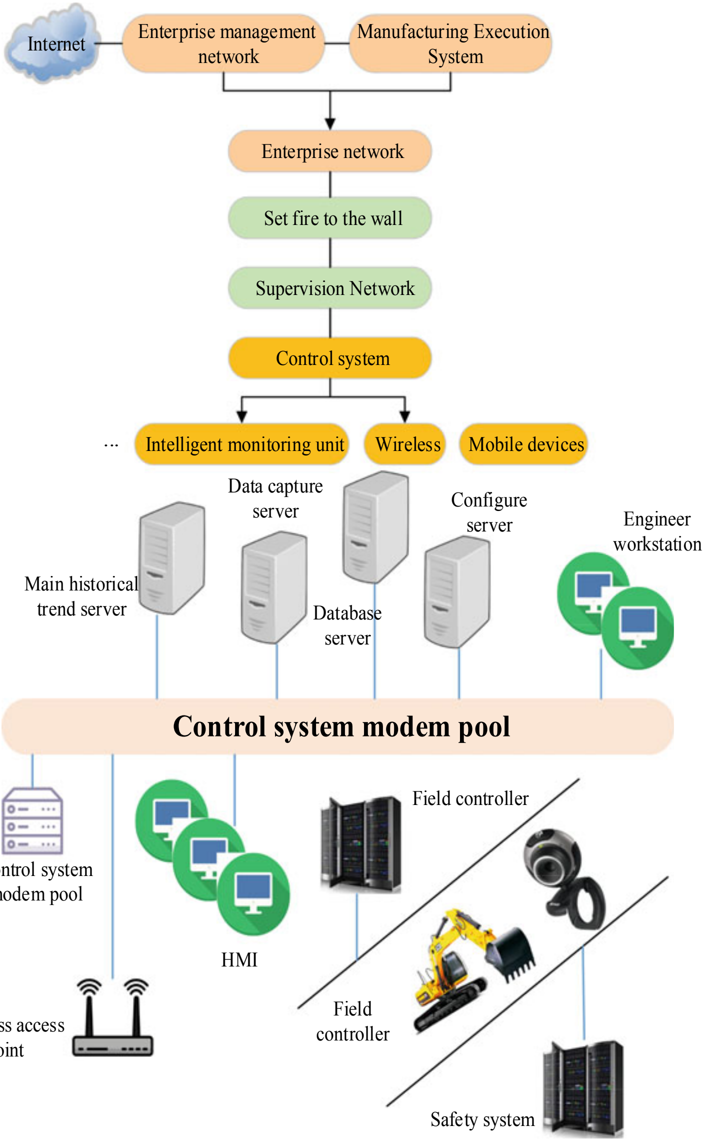


Fig. 2 Industrial control network distribution

Its communication protocol and hardware were not shared with the Internet at all. With the development of technology and the integration between various industries, the two gradually become consistent and can be well compatible with each other [23].

In addition, in the manufacturing and processing industry, based on the Internet, open and transparent business operations are the development direction of new technologies. The open network environment makes industrial control systems affected by a large amount of network information. The data exchange process also loses stronger closedness and security [24]. In recent years, attacks on industrial control systems have been countless and triggered a series of security incidents.

The consequences of the invasion of industrial control networks are also serious, ranging from affecting the company's effectiveness to causing environmental disasters or even casualties. Industries that have been invaded by industrial control networks are found in petrochemical, electric, nuclear, and hydraulic sectors. Also, these directions will be related to a country's economic, political and national security issues. Therefore, the problem of industrial control network invasion in these aspects should be paid more attention.

3.2 Industrial Network Security

From the distribution of the industrial control network, it can be seen that the industrial control network consists of an enterprise network, a monitoring network, and a detection network. Therefore, the security threats in the industrial control network may come from the following: internal and external threats to the upper computer system, communication security threats, and lower computer system threats. The detailed source is shown in Fig. 3.

Among them, in the host computer system, the confidentiality of the transmitted data cannot be guaranteed between the central control system and the station control system. Also, the configuration security in the control system cannot be guaranteed. In particular, the security awareness of operators is more difficult in the security of the host computer system. Security in a communication network depends primarily on the reliability of its communication protocols. The main security vulnerability in the lower computer system lies in the use of various devices.

3.3 Industrial Control System Attack Detection

The main ways of abnormal attack operation of industrial control network include Modbus master station, field equipment, and network communication path (Fig. 4). This paper designs the corresponding intrusion detection algorithm for the intrusion situation that the external system of the host computer will encounter. Because the connection between the external system of the host computer and the Internet is closer,

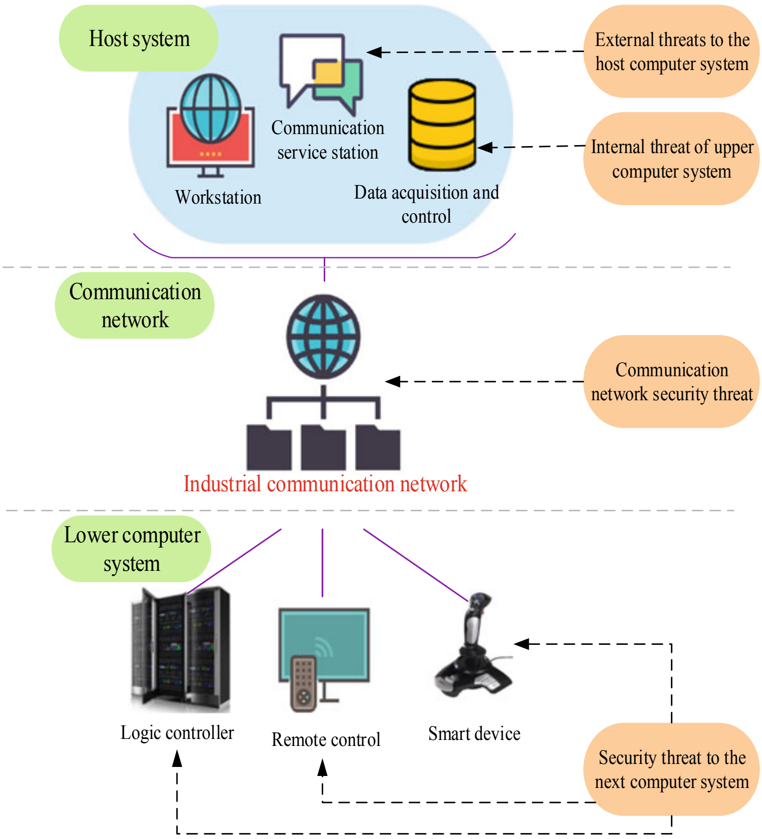

Host system
Communication service station
Workstation
Data acquisition and control
External threats to the host computer system
Internal threat of upper computer system
Communication network
Communication network security threat
Industrial communication network
Lower computer system
Logic controller
Remote control
Smart device
Security threat to the next computer system

Fig. 3 Threat sources of the industrial control network

the probability of being attacked is also higher. This paper analyzes the contents of the data packet to perform corresponding attack detection and builds a detection model suitable for this paper. Also, it detects the designed model to ensure its rationality and effectiveness.

In industrial control systems, the data they access is multi-source, complex and high-dimensional. Therefore, in the process of data processing, the real-time requirements for processing are high. In the process of attack detection, high-dimensional, large, and complicated data will seriously slow down attack detection time, resulting in a serious impact on the detection speed and detection accuracy [25]. Thus, before attack detection, the data must be reduced. This paper will use the method of feature selection for data dimensionality reduction.

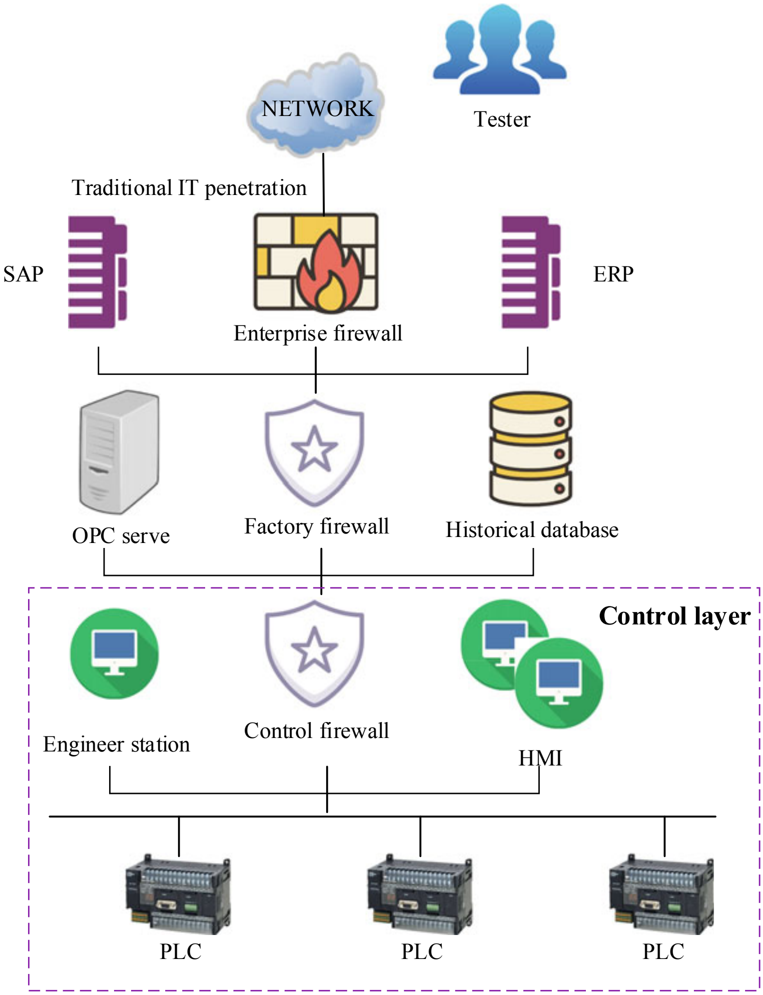


Fig. 4 Schematic diagram of abnormal attack of industrial control network system

3.4 Feature Selection Based on Improved ABC Algorithm

Intuitively, an ABC originates from the honey behavior of bee colonies. Bees carry out different activities according to their respective divisions of labor. Also, they share and exchange information on bee colonies to find the optimal solution to the problem [12]. The ABC algorithm is a kind of swarm intelligence algorithm.

The standard ABC algorithm classifies artificial bee colonies into three categories by simulating the actual honey-picking mechanism of bees: honey-bees, observation-bees, and scout-bees [26]. The entire bee colony aims to find the honey source with the largest amount of nectar [27]. In the standard ABC algorithm, the honey-bees use the previous honey source information to find new honey sources and share the honey source information with the observation-bees. The observation-bees wait in the hive and look for new sources of honey based on the information shared by the honey-bees. The task of scout-bees is to find a new valuable honey source, and they randomly look for honey sources near the hive [13].

It is assumed that the solution space of the problem is D-dimensional. The number of honey-bees and observation-bees is SN, and equal to the number of honey sources. Then, the standard ABC algorithm treats the solving process of the optimization problem as a search in the D-dimensional search space. The position of each honey source represents a possible solution to the problem, and the amount of nectar from the honey source corresponds to the fitness of the corresponding solution. To execute the ABC algorithm, first, initialize and randomly generate SN initial solutions, that is, the number of bees and food sources. Each solution is a D-dimensional vector. D represents the number of parameters to be optimized. Equation (1) is used to uniformly generate the initial food source:

$$x_{ij} = x_{j\,\min} + rand(0, 1)(x_{j\,\max} - x_{j\,\min}) \tag{1}$$

After initialization, it is necessary to start to optimize the food source, including three stages: bee collection, observation, and investigation. A honey-bee corresponds to a honey source. The honey-bee corresponding to the i-th honey source searches for a new honey source according to Eq. (2).

$$x'_{id} = x_{id} + \phi_{id}(x_{id} - x_{kd}) \tag{2}$$

i = 1,2, …, SN, d = 1, 2, …, D. ϕ_{id} is a random number on the interval $[-1, 1]$, and k $\neq$ i. The standard ABC algorithm compares the newly generated possible solution $X'_i = \{x'_{i1}, x'_{i2}, \ldots, x'_{iD}\}$ with the original solution $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}$ and uses a greedy selection strategy to retain a better solution. Each observation-bee selects a honey source based on the probability. The probability is calculated as Eq. (3).

$$p_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \tag{3}$$

fit_i is the fitness value of the possible solution X_i. For the selected honey source, the observation-bee searches for a new possible solution according to the above probability equation. When all the honey-bees and observation-bees search the entire search space, if the fitness value of a honey source is not increased within a given step (defined as the control parameter "limit"), the honey source is discarded. The honey-bee corresponding to the honey source becomes a scout-bee. The scout-bee searches for new possible solutions through Eq. (4).

$$x_{id} = x_d^{\min} + r(x_d^{\max} - x_d^{\min}) \tag{4}$$

r is a random number on the interval [0, 1]. $x_d^{\min}$ and $x_d^{\max}$ are the lower and upper bounds of the d-dimension.

Because the traditional ABC algorithm is easy to fall into a local optimal situation, this paper improves its probability equation accordingly, as Eq. (5).

$$p(x_i) = \frac{fit(x_i) - \overline{fit(x_i)}^2}{\sum_{j=1}^{SN} fit(x_i) - \overline{fit(x_i)}^2} \tag{5}$$

Using this method can prevent the colony from becoming too premature. This paper uses the packaging mode for feature selection.

3.5 Attack Detection Model

After performing the data dimension reduction of the data feature selection method of the packaging mode, the optimal feature subset is obtained. In the selection of the optimal features, a support vector machine (SVM) method is used. The SVM attack detection model is trained based on the new data set generated. It is the attack detection model designed, referred to as TABC-SVM.

The model designed is mainly divided into two parts: training and testing. The training part mainly completes the processing of the original data, the feature selection, and the screening of the optimal data. The testing part mainly judges whether the input data is normal or attacked, thereby judging whether an alarm is issued. Figure 5 displays the flowchart.

3.6 Simulation

The programming language used is JAVA. The software environment used is the Windows operating system. The hardware environment is Intel (R) Core (TM) i7-6700HQ CPU, NVIDIA GeForce GTX1060 graphics card and 8 GB memory.

For the standard SVM algorithm, given the training dataset $T = \{(x_i, y_i), (x_2, y_2) \ldots (x_N, y_N)\}$, through the nonlinear mapping of data samples x_i, the classification hyperplane with maximum soft interval is solved to correctly divide the data categories y_i. The initial solution of the optimal partition hyperplane is transformed into the optimization problem of convex quadratic programming, as Eqs. (6) and (7).

$$\min_{\omega,b,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N} \xi_i \tag{6}$$

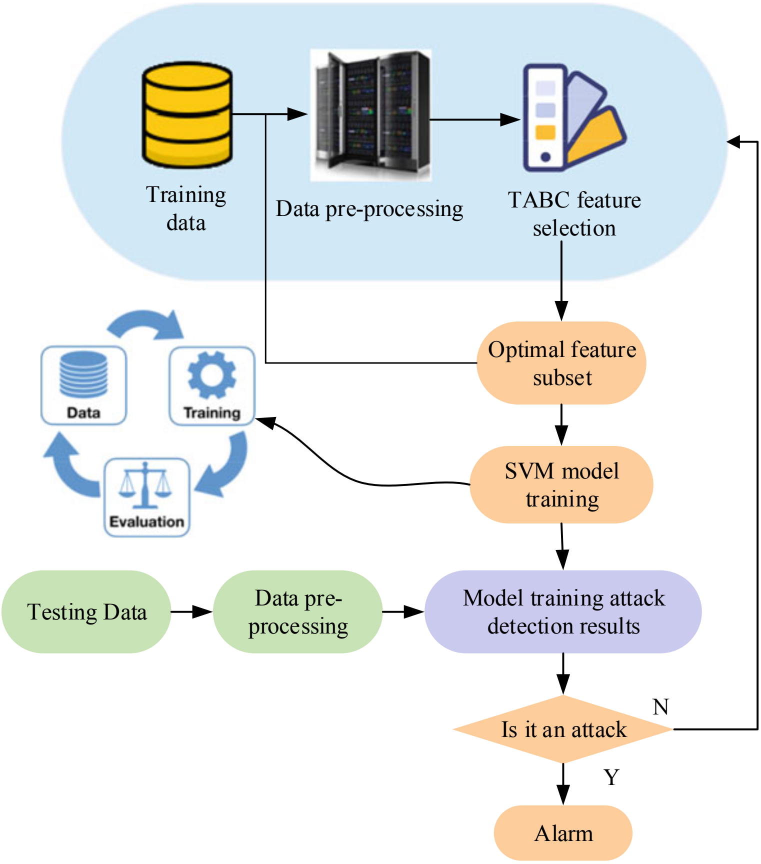


Fig. 5 Attack detection flowchart

$$s.t.\ y_i(\omega \cdot \phi(x_i) + b) \geq 1 - \xi_1 \tag{7}$$

$\phi(x_i)$ is the data mapping using kernel function. According to Lagrange's duality principle, the minimization of the original problem is transformed into the maximization of the dual problem. Lagrange function is introduced to solve the problem, as Eq. (8).

$$L(\omega, b, \xi, \alpha, \beta) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\omega \cdot \phi(x_i) + b) - 1 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i \tag{8}$$

First, find the partial differential of ω, b, ξ for Lagrange function, as Eqs. (9)–(11).

$$\nabla_{\omega} L(\omega, b, \xi, \alpha, \beta) = \omega - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \tag{9}$$

$$\nabla_{b} L(\omega, b, \xi, \alpha, \beta) = -\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{10}$$

$$\nabla_{\xi} L(\omega, b, \xi, \alpha, \beta) = C - \alpha_i - \beta_i = 0 \tag{11}$$

Equation (12) can be obtained.

$$\begin{cases} \omega = \sum_{i=1}^{N} \alpha_i x_i y_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha_i = C - \beta_i \end{cases} \tag{12}$$

By substituting (12) into the Lagrange function and solving the maximization of the dual problem, a new optimization problem can be obtained as Eqs. (13)–(15).

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j x_i y_i k(x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \tag{13}$$

$$s.t \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{14}$$

$$k(x_i \cdot x_j) = \phi(x_i) \cdot \phi(x_j) \tag{15}$$

Obtain the optimal solution α, select a Lagrange factor of positive component $0 < \alpha_j < C$, and calculate as Eq. (16).

$$b = y_j - \sum_{j=1}^{N} \alpha_i y_i k(x_i \cdot x_j) \tag{16}$$

The final decision function is expressed as Eq. (17).

$$f(x) = sign\left(\sum_{j=1}^{N} \alpha_i y_i k(x \cdot x_i) + b \right) \tag{17}$$

The default SVM parameters are used during the simulation experiments. The bee colony size is set to 20 and the maximum number of iterations is set to 60. The detection time, relevance ratio and false alarm rate of the model designed for small data volume and large data volume are analyzed and compared with other commonly used methods.

The attack types tested are Probe attack, User-to-Root (U2R) attack, Denial of Service (DoS) attack, and Remote-to-Login (R2L) attack.

For the performance analysis of TABC-SVM algorithm, the four standard test functions are as follows:

Sphere unimodal function:

$$f(x) = \sum_{i=1}^{D} x_i^2 - 100 \leq x_i \leq 100 \tag{18}$$

Rosenbrock unimodal function:

$$f(x) = \sum_{i=1}^{D-1} (100(x_{x+1} - x_i^2)^2 + (x_i - 1)^2) - 50 \leq x_i \leq 50 \tag{19}$$

Rastrigin multimodal function:

$$f(x) = \sum_{i=1}^{D} (x_i^2 - 10\cos(2\pi x_i) + 10) - 5.12 \leq x_i \leq 5.12 \tag{20}$$

Griewank multimodal function:

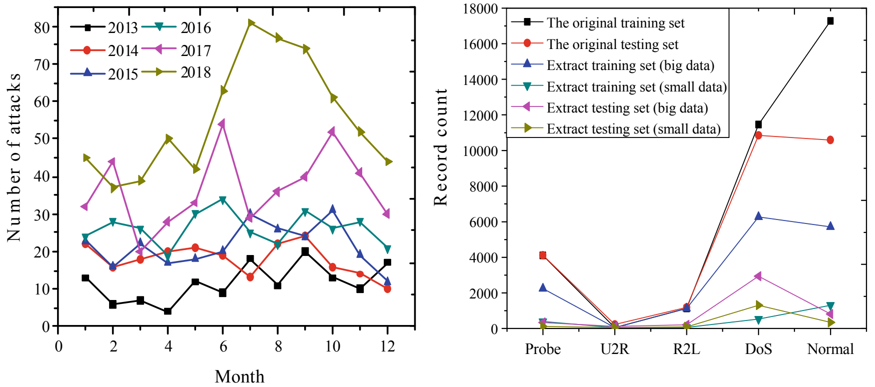
$$f(x) = \frac{1}{4000} \sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} \cos\left[\frac{x_i}{\sqrt{i}}\right] + 1 - 600 \leq x_i \leq 600 \tag{21}$$

4 Results and Discussion

4.1 Basic Data Acquisition

According to statistics, from 2013 to 2018, the incidence of industrial control network intrusions in the world rises (Fig. 6a). Figure 6a suggests that there are many industrial control network intrusion events in 2018. According to statistics, the frequency of industrial control network intrusion incidents increased rapidly from 2013 to 2018 and reached 665 times in 2018. Therefore, it is urgent to solve the security problems of industrial control networks.

To have enough data for subsequent simulation experiments, this paper first conducts simulation experiments for four different attack situations and normal situations. After the simulation, multiple pieces of feature data are generated (Fig. 6b).



(a) Industrial control network intrusion events between 2013 and 2018

(b) Number of training and testing records

Fig. 6 Basic data acquisition results

The feature data in the normal case in Fig. 6b come from different network busy levels. The four types of attack feature data also include different attack strengths, different numbers and positions of attackers, and different victim positions.

4.2 Comparison Results of the Small-Data Case

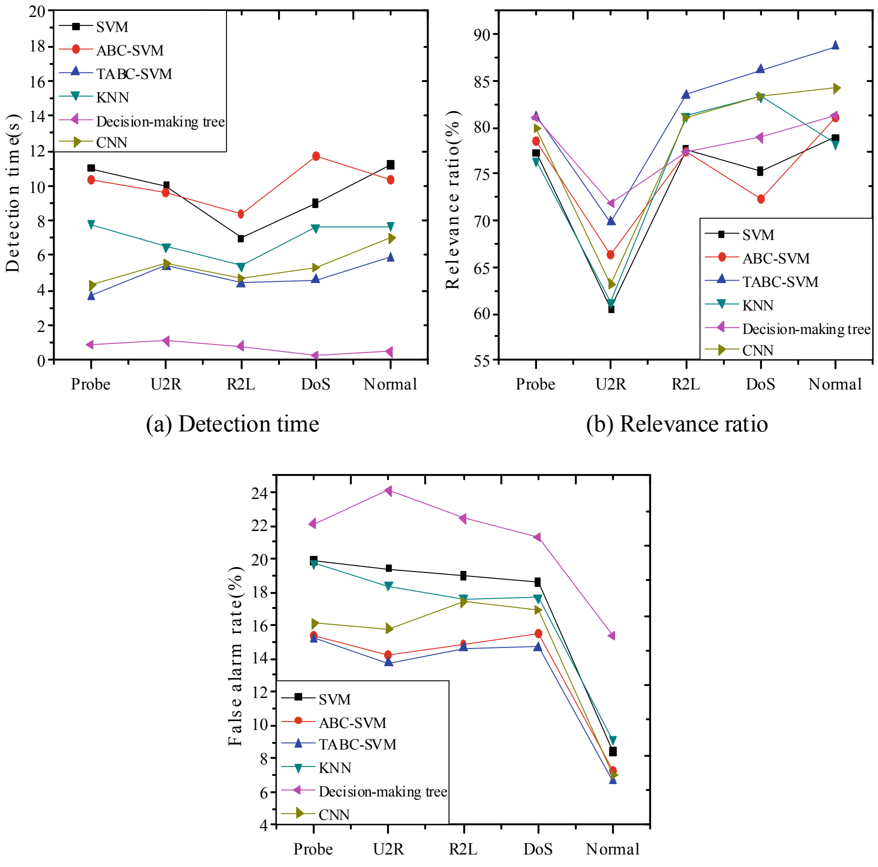
In research, traditional SVM, ABC-SVM, K-Nearest Neighbor (KNN), decision tree, and convolutional neural network (CNN) are used as controlled experiments to study the detection time, relevance ratio, and false alarm rate of various methods in the case of small data (Fig. 7).

Figure 7 suggests that the algorithm proposed is superior in detection time. Due to the nature of the decision tree algorithm, its detection time is extremely short, but this also brings high false alarm rates. However, the algorithm performs well in all aspects in the case of a small data scale.

4.3 Comparison Results of the Big-Data Case

The detection time, relevance rate and false alarm rate of various methods in the case of big data are studied and compared (Fig. 8).

Figure 8 indicates that the algorithm proposed has only a higher detection time than the decision tree algorithm and has a better relevance ratio. Also, the false alarm rate is better. Although the decision tree algorithm has a short detection time, its relevance ratio and false alarm rate perform poorly.



(a) Detection time

(b) Relevance ratio

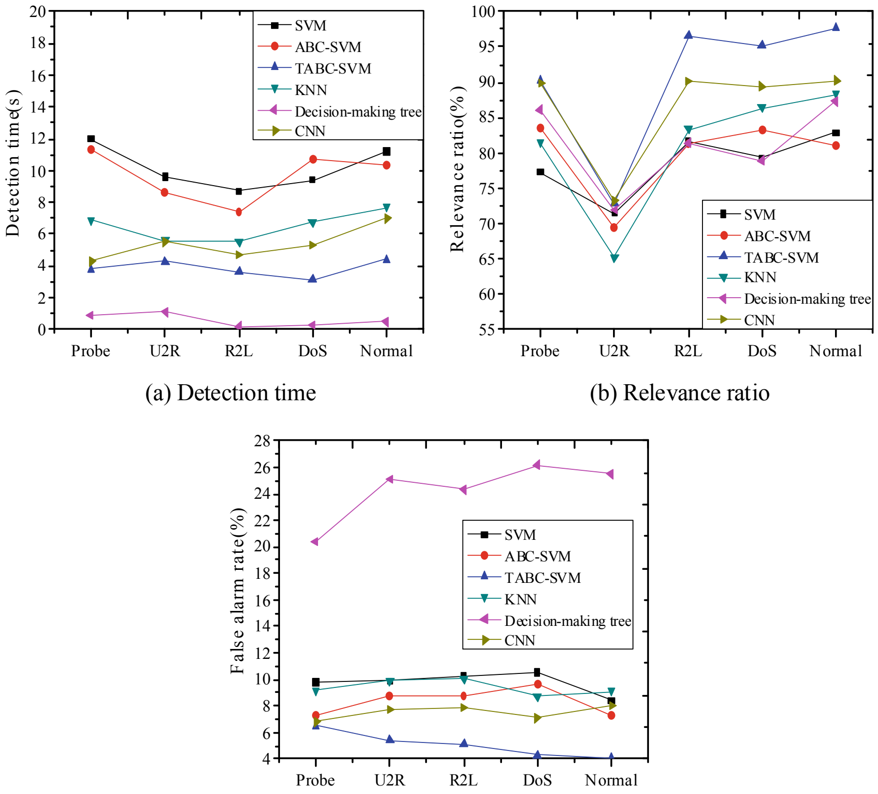
(c) False alarm rate

Fig. 7 Comparison results of the small-data case

4.4 Comparison Results of Unknown Conditions

For unknown conditions, that is, the detection comparison of no specific attack type is performed. An attack type is randomly selected for attack detection, thereby detecting the comprehensive ability of the algorithm proposed (Fig. 9).

Figure 9 reflects that the detection capabilities of various algorithms have declined in unknown conditions, especially for the detection of U2R type attacks. Due to the small number of training and testing of this type, the detection capability of this type of attack is low. But in other aspects, the algorithm proposed is better than other algorithms.



(a) Detection time

(b) Relevance ratio

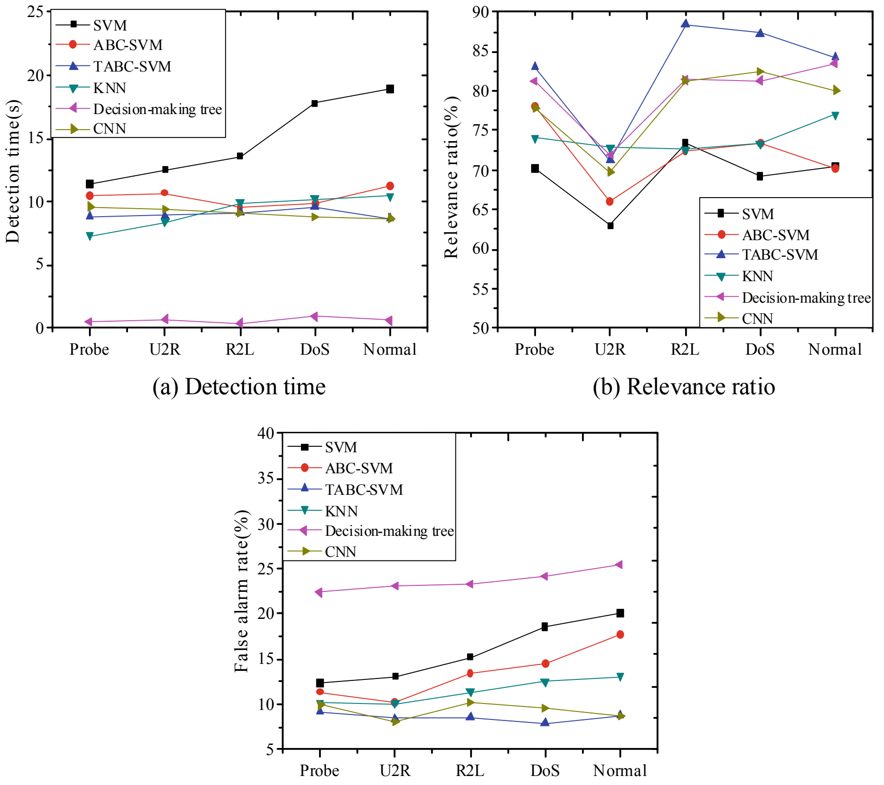
(c) False alarm rate

Fig. 8 Comparison results of the big-data case

4.5 *Performance Evaluation of Attack Detection Algorithm of Industrial Control Network*

The simulation results are compared and analyzed regarding the classification accuracy and false positive rate to verify the attack detection performance of the TABC-SVM algorithm proposed. The compared algorithms include KNN algorithm, SVM algorithm, GA-SVM algorithm, and decision tree algorithm. Figure 10a, b show the detection rates of different algorithms for different types of attacks. Figure 10c shows the comparison results of the overall classification accuracy of different classification methods.

TABC-SVM algorithm proposed is higher than other classification methods in the overall classification accuracy, with an accuracy of 96.94%. The reason is that the improved algorithm selects sub-classifiers from two aspects: the base classifier and the complementarity between classifiers, to improve the classification accuracy of the algorithm as a whole. In terms of attack detection, the classification accuracy



(a) Detection time

(b) Relevance ratio

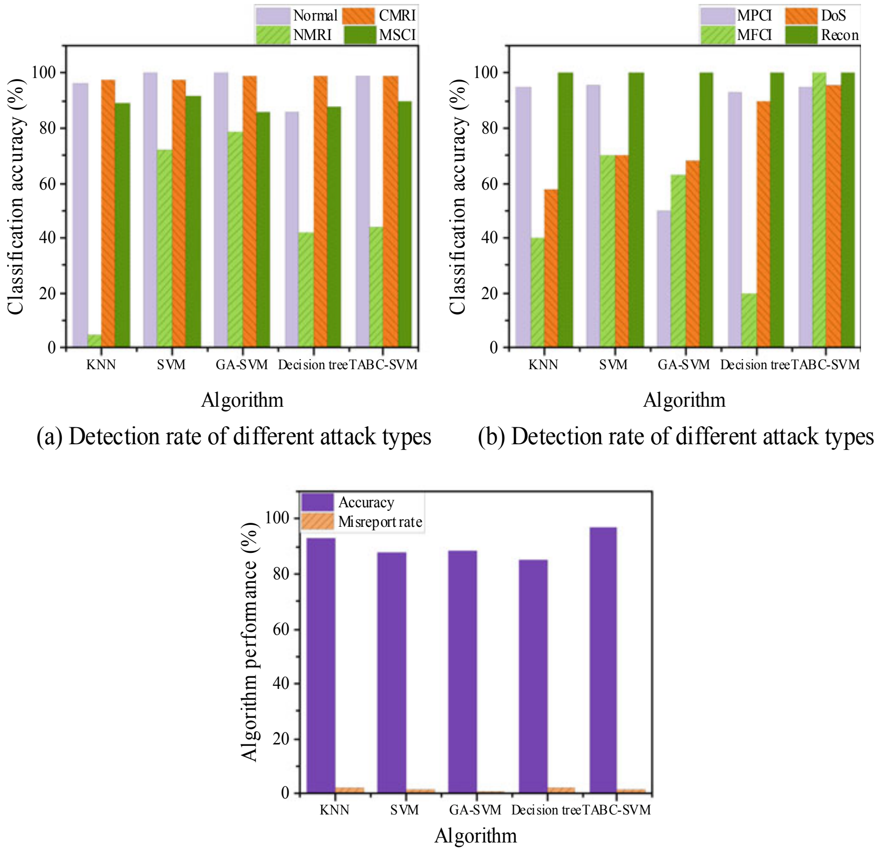
(c) False alarm rate

Fig. 9 Comparison results of unknown conditions

of the algorithm proposed is low only in NMRI, the detection effect of other types of attacks is significantly higher than other algorithms, and the classification accuracy is relatively stable.

5 Conclusion

This paper analyzes and studies the configuration security of industrial automation and control systems to contribute to the security of the world's industrial control networks. Also, the world's industries expect to get rid of the risk of invasion of industrial control systems. This paper uses an improved ABC algorithm combined with SVM for research and expects to achieve excellent attack detection results. In the case of small-scale data, the performance of this method is more general. However, in the case of large-scale data, the detection accuracy, false alarm rate, and



(a) Detection rate of different attack types

(b) Detection rate of different attack types

(c) Comparison of overall classification accuracy

Fig. 10 Attack classification accuracy and attack detection performance of TABC-SVM algorithm

detection time of this method are all excellent. Compared with other attack detection methods, the method proposed has certain advantages in various aspects. The attack detection method of the industrial control system based on the ABC algorithm can effectively detect the attack state and provides an important theoretical basis for the research of attack detection methods. Although some achievements have been made, there are still some shortcomings. Although the method proposed performs well in all aspects, due to the time relationship, this paper fails to identify various attacks. Therefore, next, various attacks will be identified and detected to consolidate the research results.

References

1. Pan F, Pang Z, Luvisotto M, Xiao M, Wen H (2018) Physical-layer security for industrial wireless control systems: basics and future directions. IEEE Ind Electron Mag 12(4):18–27
2. Kleinmann A, Wool A (2017) Automatic construction of statechart-based anomaly detection models for multi-threaded industrial control systems. ACM Trans Intell Syst Technol (TIST) 8(4):1–21
3. Luvisotto M, Pang Z, Dzung D (2019) High-performance wireless networks for industrial control applications: New targets and feasibility. Proc IEEE 107(6):1074–1093
4. Samaddar A, Easwaran A, Tan R (2020) SlotSwapper: a schedule randomization protocol for real-time WirelessHART networks. ACM SIGBED Rev 16(4):32–37
5. Combita LF, Cardenas AA, Quijano N (2019) Mitigating sensor attacks against industrial control systems. IEEE Access 7:92444–92455
6. Li H, Savkin AV (2018) Wireless sensor network based navigation of micro flying robots in the industrial internet of things. IEEE Trans Industr Inf 14(8):3524–3533
7. Coutinho RW, Boukerche A, Vieira LF, Loureiro AA (2018) Underwater wireless sensor networks: a new challenge for topology control-based systems. ACM Comput Surv (CSUR) 51(1):1–36
8. Giraldo J, Urbina D, Cardenas A, Valente J, Faisal M, Ruths J, Candell R (2018) A survey of physics-based attack detection in cyber-physical systems. ACM Comput Surv (CSUR) 51(4):1–36
9. Jiang X, Pang Z, Zhan M, Dzung D, Luvisotto M, Fischione C (2019) Packet detection by a single OFDM symbol in URLLC for critical industrial control: a realistic study. IEEE J Sel Areas Commun 37(4):933–946
10. Seetanadi GN, Oliveira L, Almeida L, Arzen KE, Maggio M (2018) Game-theoretic network bandwidth distribution for self-adaptive cameras. ACM SIGBED Review 15(3):31–36
11. Angle MG, Madnick S, Kirtley JL, Khan S (2019) Identifying and anticipating cyberattacks that could cause physical damage to industrial control systems. IEEE Power Energy Technol Syst J 6(4):172–182
12. Yue D, Han QL (2019) Guest editorial special issue on new trends in energy internet: artificial intelligence-based control, network security, and management. IEEE Trans Syst Man Cybern: Syst 49(8):1551–1553
13. Zolanvari M, Teixeira MA, Gupta L, Khan KM, Jain R (2019) Machine learning-based network vulnerability analysis of industrial Internet of Things. IEEE Internet Things J 6(4):6822–6834
14. Galloway B, Hancke GP (2012) Introduction to industrial control networks. IEEE Communications surveys & tutorials 15(2):860–880
15. McLaughlin S, Konstantinou C, Wang X, Davi L, Sadeghi AR, Maniatakos M, Karri R (2016) The cybersecurity landscape in industrial control systems. Proc IEEE 104(5):1039–1057
16. Ponomarev S, Atkison T (2015) Industrial control system network intrusion detection by telemetry analysis. IEEE Trans Dependable Secur Comput 13(2):252–260
17. Zhang F, Kodituwakku HADE, Hines JW, Coble J (2019) Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data. IEEE Trans Industr Inf 15(7):4362–4369
18. Shen C, Liu C, Tan H, Wang Z, Xu D, Su X (2018) Hybrid-augmented device fingerprinting for intrusion detection in industrial control system networks. IEEE Wirel Commun 25(6):26–31
19. Zhang Q, Zhou C, Tian YC, Xiong N, Qin Y, Hu B (2017) A fuzzy probability Bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems. IEEE Trans Industr Inf 14(6):2497–2506
20. Volkova A, Niedermeier M, Basmadjian R, de Meer H (2018) Security challenges in control network protocols: A survey. IEEE Commun Surv Tutor 21(1):619–639
21. Li X, Zhou C, Tian YC, Qin Y (2018) A dynamic decision-making approach for intrusion response in industrial control systems. IEEE Trans Industr Inf 15(5):2544–2554
22. Dezfouli B, Radi M, Chipara O (2017) REWIMO: a real-time and reliable low-power wireless mobile network. ACM Trans Sensor Netw (TOSN) 13(3):1–42

23. Taleb T, Afolabi I, Bagaa M (2019) Orchestrating 5G network slices to support industrial internet and to shape next-generation smart factories. IEEE Network 33(4):146–154
24. Beckett R, Gupta A, Mahajan R, Walker D (2019) Abstract interpretation of distributed network control planes. Proc ACM Programm Lang 4(POPL):1–27
25. Wang B, Li X, de Aguiar LP, Menasche DS, Shafiq Z (2017) Characterizing and modeling patching practices of industrial control systems. Proc ACM Meas Anal Comput Syst 1(1):1–23
26. Jin X, Guan N, Xia C, Wang J, Zeng P (2018) Packet aggregation real-time scheduling for large-scale WIA-PA industrial wireless sensor networks. ACM Trans Embedded Comput Syst (TECS) 17(5):1–19
27. Ma Y, Gunatilaka D, Li B, Gonzalez H, Lu C (2018) Holistic cyber-physical management for dependable wireless control systems. ACM Trans Cyber-Phys Syst 3(1):1–25

An Empirical Investigation on Lean Method Usage: Issues and Challenges in Afghanistan

Salim Ahmadzai and Muhammad Bakhsh

Abstract Lean model is an effective software development method for enhancing the productivity in Software's area and can be used in many software development companies to take away unnecessary waste, empower the team and increase efficiency. The main purpose of this study was to find and assess the issues and challenges such as (lack of education, communication, training etc.) of lean software development approach while implementing in various Software development projects in Afghanistan. The study finds that the lack of communication is seen as a major issue in order to identify the requirements of product. The issue of language seems to exist only when both Non-English talks on. Long-term meetings help to better understand the demands, specifications, experience and previous domains of product.

Keywords Agile development · Lean method · Software companies · Issues and challenge lean implementation

1 Introduction

Lean is a framework for software development with an end-to-end focus on creating customer satisfaction, improving efficiency, deliver fast, amplify learnings, optimizing value sources and empower the team [1].

Lean approach has many issues and challenges companies have been tried to introduce many viable solutions to address these issues, challenges and barriers such as inadequate communication, lack of confidence, employee attitude, poor morale as well as many other distributing problems.

With Agile popularity, more people are interested in using its benefits to address some of these problems. Similarly, lean was also introduced in the creation of distributed system development to resolve problems [2]. While the results produced in global software development were modified by each approach, several problems remained and the expected growth effects were not provided [3].

S. Ahmadzai (✉) · M. Bakhsh
Department of Computer Science, University of Abasyn, Peshawar, Pakistan
e-mail: muhammad.bakhsh@abasyn.edu.pk

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_13

This research addresses to review the existing methods of the integration of lean presenting in the company and how it can be applied in the global software development. The research focuses on problems and challenges faced by implementing lean and effective sustainable implementation of lean in a global product development environment [4].

Many application development models have been developed in the software company [5]. Each type has various features and each type is different with other one; in general, these approaches are categorized as either a Compact and strong approach.

The strong techniques in product development, also called traditional techniques, are based on the idea that the boundary between system and environment is defined and immutable. However, this idea, no longer works in today's unpredictable open-world environments, which require techniques that allow product to respond to changes by nature-organizing its layout and nature-adapting its behavior [6]. Because of their simple, methodical and organized design the conventional approaches are still commonly used in the software company they have proven that the can provide predictability, consistency and reliable (see Fig. 1).

Nevertheless, they have a range of main drawbacks, including weak adaptation to rapidly evolving market needs, and a propensity to be over-budget and even behind plan, with less functionality than the specifications defined. Despite ambiguous product expectations, it is also a major challenge for conventional approaches to require a complete range of specifications before implementation [7].

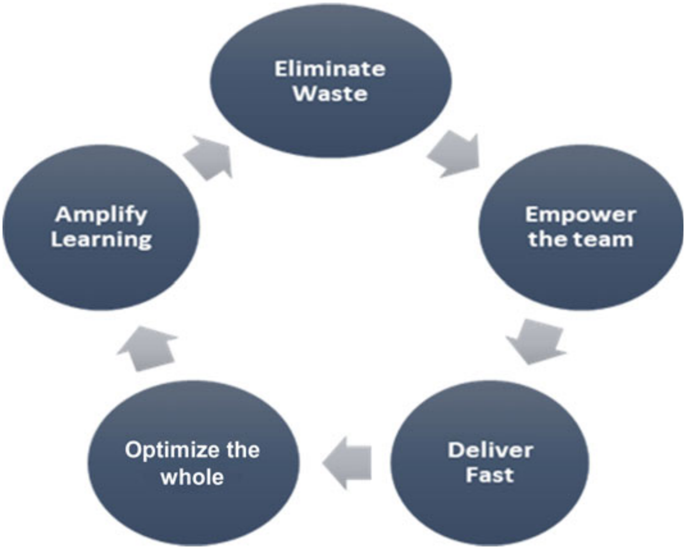


Fig. 1 Lean principles

The agile product development method, including Crystal, Kanban, Scrum, Crystal and Adaptive Software Development (ASD), have been built and developed by researcher since the 1990s as a solution for the limitations of conventional approaches designed to accept, rather than reject, a high rate of change. Such new approaches are primarily oriented towards integrative and incremental development [8].

2 Literature Review

The aim of this literature review is to analyze existing literature to assess if there are sufficient proof on issues and challenges in lean software development approaches [9].

A comprehensive review of literature with the key stages of preparing the study and announcing the findings of the review will be the first phase of the research. Review of a literature is a structured and iterative approach for gathering and analyzing information [10]. In addition, this study will review all solutions in a collaborative environment that will facilitate the efficient application of lean practices and identify gaps in existing research [11]. Experience from other practitioners is vital for using a new approach which explains what's relevant, what's possible and what's not possible about Lean in Afghanistan, where we did a qualitative investigation to figure out what problems early adopters had and how to solve them [12]. This study would include both outcomes and results from an empirical way of evaluating data in order to recognize, review and analyse all existing study in the field of agile lean approach for handling distributed development [13].

Harbour is known to have significant difficulties introducing lean manufacturing like every other efficiency development initiative. For instance, [14] highlighted the complexity and controversy of applying one of the many lean development strategies called as just-in-time. A lack of formal analytical process and evaluation of value-adding capabilities within organizations, such as the lean concept can further compound this problem.

Technology plays an important part in people's lives and in extreme industry rivalry. That's why software engineering (SE) is a field that is not vigilant when releasing new methodology, and is quickly identifying with great achievement on complicated project.

The problem of business process creation and optimization has been debated over the last few decades are how it should function to produce quicker, safer, and cheaper solution.

Large numbers of various methods have been developed for software development, of which only few have endured for use today. So-called agile methodologies have become more and more popular in recent years. These approaches are designed to support experimentation rather than conventional phases of development approach and rely on ongoing engagement between developers and clients [15].

3 Framework Development

In order to present key issues and challenges in lean Software Process Technology, we will use a conceptual and terminological framework. The framework adopted in the research has been depicted in Fig. 2.

The selection of events sites was accompanied by an online search of selected instances to obtain primary relevant data about the events. By the mean of Conduction sites, Project sites were defined, with the naming of the relevant authority who can really respond to questions, usually identified as project manager.

Before visit starting the aim of the visit explained and also explained the relevant issues and challenges faced by project members for not implementing lean framework in software companies. After that, a comprehensive meeting to the software companies and other areas such as governmental organizations, private organizations were conducted for further observations.

The visits were mostly 4–6 h long. Multiple visits were carried out to each plant to collect some further details as well as to explain doubts. Often the visits lasted 5–6 h. Multiple visits were carried out for each plant to acquire any further information and to clear any doubts. The collection of gathered information and synthesis of information was accompanied by the administration of case studies. The outcomes and result were eventually planned.

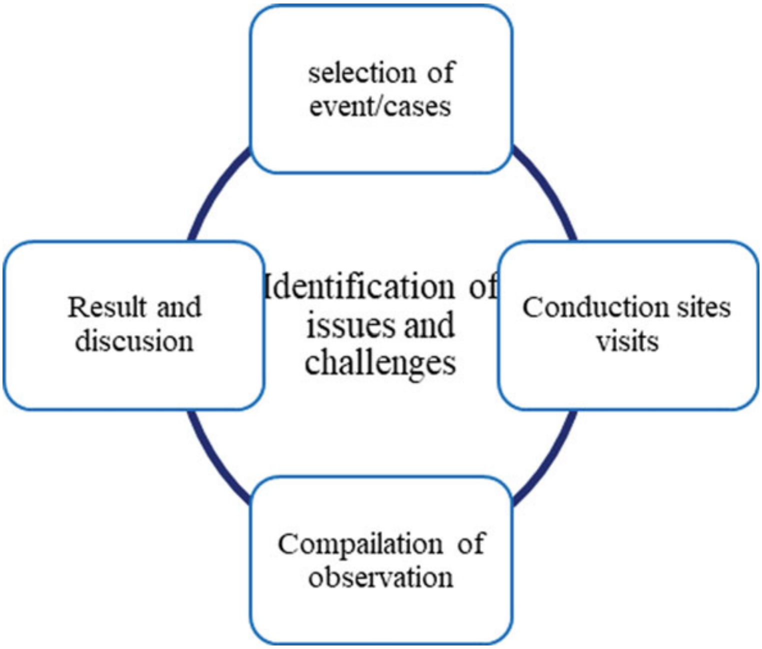


Fig. 2 Development framework

4 Research Methodology

The main method of this study was a qualitative analysis and survey of empirical data collected via questionnaires, interview and google form sent to companies' project managers.

In this study the method has been proposed to find and determine the issues in lean software development implementation in Afghanistan companies and to propose the paradigm for lean software development implementation with the concentration of determining the issues and challenges of lean software development.

A comprehensive "questionnaire" was arranged for finding issues and challenges of lean software development in Afghanistan companies to test the effectiveness of the Questioners, the questionnaire was pre-tested on a corralled sample.

The methodology adopted in the research has been depicted in Fig. 3.

The study approach is based on a survey method is often seen as a simple method to study. The survey method will use a number of approaches to address the research question. A questionnaire or a list of question is developed on lean software development method issues and challenges within Afghanistan companies. Data collected through mails, post, and face to face interview. Popular survey methods include face-to-face interviews, mail questionnaires and interviews by telephone.

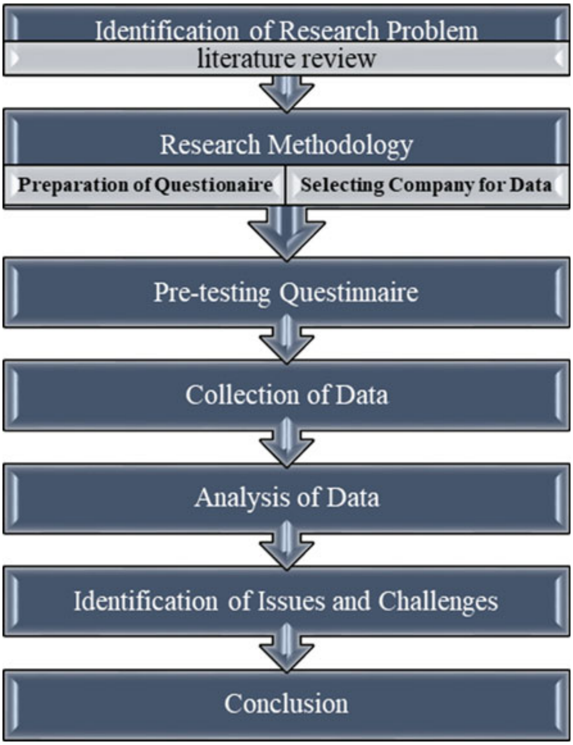

Identification of Research Problem
literature review
Research Methodology
Preparation of Questionaire
Selecting Company for Data
Pre-testing Questinnaire
Collection of Data
Analysis of Data
Identification of Issues and Challenges
Conclusion

Fig. 3 Research methodology

The survey method will use a number of approaches to address the research question. Popular survey methods include face-to-face interviews, mail questionnaires and interviews by telephone. However, as with any other method and approach of study, it's possible to perform a low quality survey, rather than a high quality and real value survey.

5 Results and Discussion

After study literature of review and by knowing the lean method with its benefits, weakness and issues a list of questions (questionnaire) was prepared. Questionnaire contained total of 65 questions. All the questions were formed in MCQs, yes/no and subjective type format so it could be answered simply and wouldn't take long time.

The questionnaires were sent to 35 software companies for data collection and taking interviews with stakeholders including: project managers, engineers, customer, employees, contractors etc. all of them were guided to answer the questions on their own point of view. There were quick responses from the project managers, engineers, customer and some of them neglected. In order After that the answers to questions were gained and compiled the collected data. the collected data was categorized and analyzed on the base of features and probability of their occurrence and achieved the clear result about the most common major issues and challenges as depicted in the following list and chart for implementing lean software development method in Afghanistan companies (Fig. 4).

1. Lack of knowledge about lean method.
2. Manager pressure issues.
3. Lack of training for employees.
4. Lack of consultants.
5. Long implementation time needed.
6. Lack of communication.
7. Financial issues.

Survey result indicate how many sessions of training are given on lean in the companies. In Afghanistan many organizations organize lean training sessions for their workers, most of those are 6–7 long hours; detailed depicted in Fig. 5.

Research revel the efficiency of teams using lean in their companies. Many companies find lean useful, but a large number of companies assume that there is no massive Change has taken place in the success of their team; detailed depicted in Fig. 6.

The research survey illustrates the impact of lean on project growth in companies. Approximately 29% of companies have discovered that lean has no positive or negative effects on the growth of their companies' projects; detailed depicted in Fig. 7.

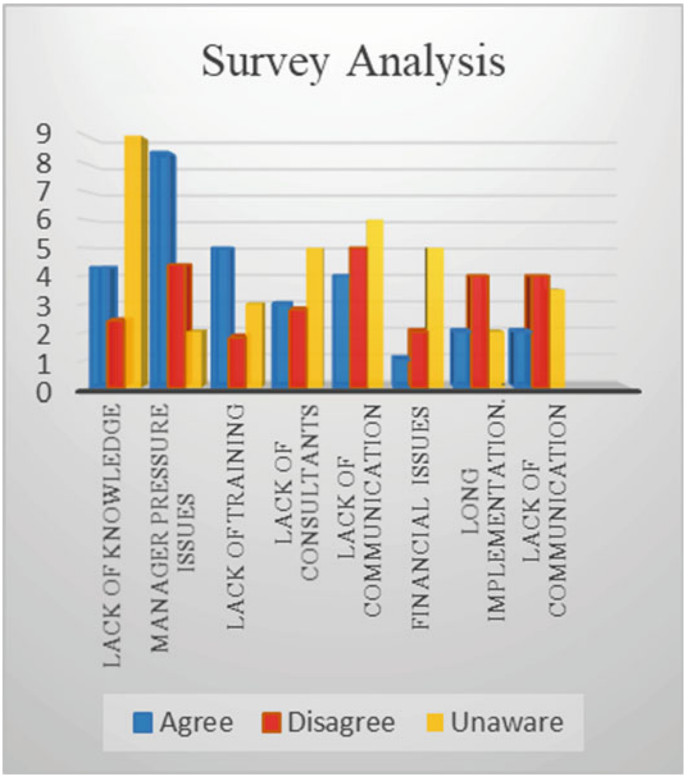


Fig. 4 Issues and challenges

6 Conclusions

After our survey and study we have found that there are many issues and challenges (Lack of lean understanding, Employee attitude issues, Fiscal issues etc.) while implementing the lean framework in software companies in Afghanistan. About 90% of the respondents are not understand about the lean framework and methods, and almost 84% of the companies are facing business stress which make stay in software project time and duration.

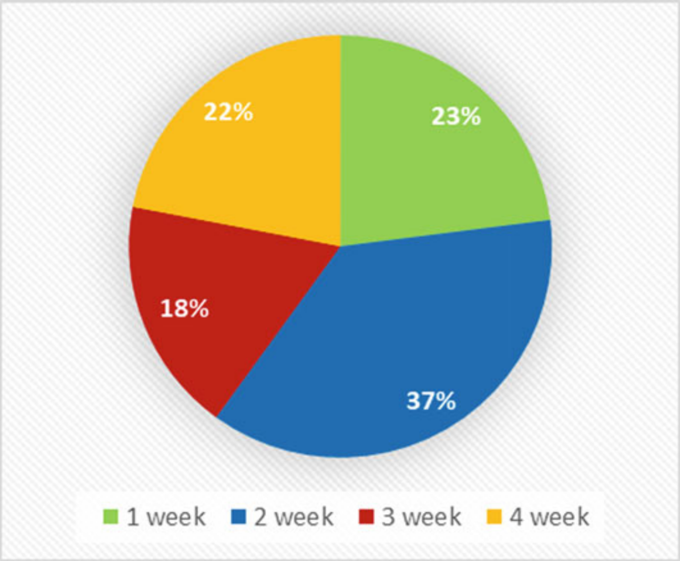


Fig. 5 Training provided on scrum

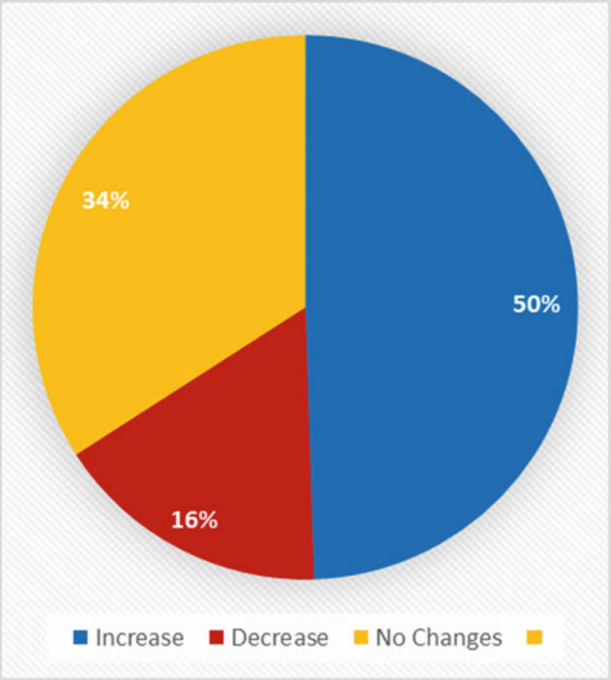


Fig. 6 Team performance

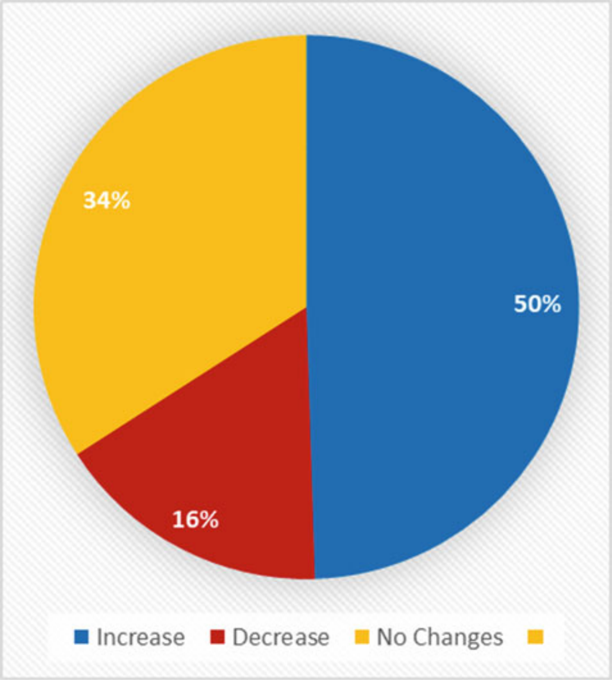


Fig. 7 Effect on project development

Acknowledgements Foremost, I cannot express sufficient gratitude to my supervisor and co-supervisor for their continued support of consultant and motivation: Dr. Muhammad Bakhsh and Dr. Mohsin Masood. I give my heartfelt thanks for the chances committee provided for learning. Finally, to my caring, loving and supportive father Hazrat Wali; my heartfelt appreciation, your motivation When the occasions are roughly valued and blandly acknowledged It was a good feeling and satisfaction to know that while I accomplished my studies, you were able to provide care of our home activates, My sincere gratitude.

References

1. Hajjdiab H, Taleb AS (2011) Adopting Agile software development: issues and challenges. Int J Manag Value Supply Chains (IJMVSC) 02(03):10
2. Asnawi AL, Gravell AM, Wills GB (2011) Empirical investigation on Agile methods usage: issues identified from early adopters in Malaysia. In: International conference on agile software development, England
3. Issa UH (2013) Implementation of lean construction techniques for minimizing the risks effect on project construction time. Alexandria Eng J 704
4. Fullerton RR, Wempe WF (2014) Lean manufacturing, non-financial performance measures, and financial. Int J Oper Prod Manag
5. Jiméneza LRMDMMEM (2015) Methodology implementation in the laboratories of an industrial. Safety Sci 78

6. Nowotarski JPP (2015) Barriers in running construction SME—case study on introduction of Agile methodology to electrical. Proc Eng 122(8):2
7. Achanga P, Shehab E, Roy R, Nelder G (2015) Critical success factors for lean implementation within SMEs. J Manuf Technol Manag 17(4):21
8. Brioso X (2015) Integrating ISO 21500 guidance on project management, lean construction and PMBOK. Proc Eng 123
9. Panwar A, Jain R, Rathore APS (2016) Obstacles in lean implementation in developing countries—some cases from the process of sector of india 02(01)
10. Avinash Panwar RJ (2016) Obstacles in lean implementation in developing countries—some cases from the process sector in India. Int J Lean Enterp Res 2(1):26
11. Sarhan JG (2017) Lean construction implementation in the Saudi Arabian construction industry, vol 17. Queensland University of Technology
12. Gupta P (2017) Applying Agile lean to global software development 10
13. Kleszcz D (2018) Barriers and opportunities in implementation of lean manufacturing tools in the ceramic industry. Prod Eng Arch 05
14. Maqbool B, Rehman FU, Abbas M (2018) Implementation of scrum in Pakistan's IT industry. no. 978-1-4503-5431-8/18/01…$15.00, p. 08
15. Nifla RPK (2019) Barriers for implementing lean concepts in Indian construction industry. Int J Eng Res Technol (IJERT) 08(05):03

Optimization of the Effects Oscillation Welding: Sinusoidal and Triangular Beam During Laser Beam Welding of 5052-H32 Aluminum Alloy

Radouane El Kinani, Herinandrianina Ramiarison, Noureddine Barka, and Abderrazak El Ouafi

Abstract Aluminum alloys are increasingly replacing steel in marine, automotive and aerospace industries to reduce the weight of components, which leads to gains in terms of energy savings. The 5xxx series is an alloy characterized by its low density and high corrosion resistance. In this work oscillating laser welding was used to join a 2 mm thick AA5052-H32 aluminum alloy in butt configuration. Two beam oscillation patterns, the sinusoidal pattern and triangular pattern, have been studied and compared. The macro- and microstructure of the two welding patterns as well as the micro-hardness have been studied. The weld hardness profiles of both patterns showed a W-shaped appearance, and the minimum hardness was measured in the heat affected zone (HAZ). Statistical analysis of the weld quality of AA 5052-H32 in a process window (laser power: 1800–2000–2200 W; oscillation frequency: 200–300–400 Hz; oscillation amplitude: 1–1.5–2 mm) as well as the development of a regression model analyzing and predicting the tensile strength (TS) of the weld were developed for each oscillation pattern. Single effects and interactions of parameters on (TS) were analyzed. The (TS) is more sensitive to values of laser power between 1800 and 2000 W at oscillation frequency between 300 and 400 Hz with a large amplitude value more than 1 mm. The maximum tensile strength is (~183 MPa) for the sinusoidal pattern and (~178 MPa) for the triangular pattern. These two maximum values of the (TS) were found for the same parametric combination applying a laser power of 2000 W, an oscillation frequency of 400 Hz and an amplitude of 2 mm.

R. El Kinani (✉) · H. Ramiarison · N. Barka · A. El Ouafi
Departement of Mathematics, Computer Science and Engineering, Université du Québec à Rimouski, 300, allée des Ursulines, Rimouski, QC G5L 3A1, Canada
e-mail: Radouane.Elkinani@uqar.ca

H. Ramiarison
e-mail: herinandrianina.ramiarison@uqar.ca

N. Barka
e-mail: noureddine_barka@uqar.ca

A. El Ouafi
e-mail: abderrazak_elouafi@uqar.ca

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_14

Keywords Laser beam welding · Aluminum alloy 5052-H32 · Beam oscillation · Tensile strength · Interaction effects · Regression model

1 Introduction

Welding is the most polyvalent and realistic joining method applicable to the manufacturing of products in all industrial fields. Therefore, “laser welding” is recognized as an advanced process for joining materials using a high-power, high-energy density laser beam.

Laser welding applications have successfully moved from specialized research laboratories to production sites. This modern welding process has now reached industrial maturity, as shown by the numerous applications in various sectors of activity: automotive, aeronautics, boiler making, etc. Such a development was only possible thanks to the development of increasingly powerful, reliable, and competitive laser sources, and above all, thanks to a better theoretical knowledge of the laser welding process.

Laser welding has established itself as an advanced technique for joining metals such as steels [1], titanium alloys [2], and magnesium alloys [3, 4]. This is due to the advantages it offers over conventional welding techniques, including shielded metal arc welding and tungsten electrode welding.

LBW is the abbreviation for laser beam welding, which is used to fuse metal or plastic plates with a focused laser beam of high-power density, which acts as a heat source to melt the dimensions of the plates. This welding process generates a molten pool that cools down very quickly when the laser is moved over the surface of the metal parts. This welding technique is characterized by its exceptional advantages, noting that this welding process has low distortion and low porosity, which can be zero in some cases [5], also this technique is marked by a smaller heat-affected zone (HAZ) and a strong metallurgical bond.

The laser welding process was used to assemble the different materials, including aluminum and aluminum alloys, various steel grades, magnesium alloys, titanium alloys and nickel-based superalloys. Also, the optimization of process parameters to improve weld quality such as tensile strength, the hardness the penetration depth of the weld, the aspect ratio of the weld seam and the corrosion resistance have been carried out thoroughly. The key parameters influencing the welding qualities were identified comprising the laser power, the oscillation frequency, and the amplitude. However, it is difficult to obtain a good quality weld of an aluminum alloy using the laser beam welding technique because of the high reflectivity of aluminum alloys. This can be explained by the microstructure of the aluminum alloy which often causes fractional absorption of the incident radiation [6]. Plus, the high thermal conductivity causing rapid heat transfer into the workpiece and limits the energy concentration in the weld pool, and the low viscosity limiting the growth of the weld pool before solidification of the thin aluminum sheet restrict the laser weldability and weld quality of aluminum alloys [7]. Among the solutions proposed in the past, the use of lasers with shorter wavelengths around 1000 nm was proposed to improve the absorptivity of the laser in the welding of aluminum alloys [8]. Accordingly, Nd:YAG laser (λ = ~1070 nm) is more preferable to CO_2 laser (λ = ~1006 nm). Recently, fiber lasers (λ = 1020–1070 nm) are increasingly used because of the other advantages they offer namely the ability to focus the beam to a very small round spot in the range of 0.3–1.53 nm [9], a better optical quality and easy transport of the laser in the fibers, which together result in higher energy efficiency and greater penetration into the weld [10]. Laser welding is often performed by deep penetration or by conduction. The latter is often preferred for welding thin aluminum alloy sheets due to its ability to minimize distortion, prevent loss of alloying elements, and produce fine-grained structures through rapid solidification [11, 12]. In addition, establishing the optimal laser welding parameters for different series of aluminum alloys (AA) has been shown to be effective in improving the quality of welds [13]. The AA 5000 series (non-heat treatable AA) is characterized by high strength due to the inclusion of a large amount of Mg in the aluminum solid solution. The AA 5000 series (e.g. AA 5052, AA 5083) is also known for its high marine corrosion resistance, high thermal conductivity and good formability [14]. For those reasons, they are progressively replacing steel in the marine and automotive industries for the manufacture of structural parts. There is extensive literature on the laser weldability of the AA 5000 series and the effects of different welding parameters on the quality of the weld to improve the mechanical properties of the welded parts. Li and his colleagues [15] investigated the effect of the laser head oscillation parameters (the circular shape) on the porosity, morphology, and microstructure as well as the mechanical properties of the AA 5083 weld, 4 mm thick and 100 × 150 mm in size, using a CW fiber laser with a maximum power of 6 kW. In their studies, they showed that when the oscillation frequency was higher than 200 Hz and the oscillation diameter was higher than 2 mm, the porosity of the weld can be suppressed by the laser beam oscillation. They also stated that the microstructure analysis performed by backscattered electron diffraction and electron probe microanalyzer revealed that the laser beam oscillation could refine the grains and promote the uniform distribution of β(Mg2Al3) element in the weld melting zone, thus leading to an increase of the microhardness in the weld zone. The results of the tensile test showed that the tensile strength and elongation of the laser oscillation weld joints were significantly higher than those without oscillation, the authors claim that the increase in tensile strength is due to the suppression of porosity. Three types of oscillating laser beam welding of 5A06 aluminum alloy were investigated by Wang et al. [16], the results showed that oscillating laser beam welding using linear, circular, and infinity (∞) paths reduce the porosity of the welded joint compared to the joint realized without oscillation. They also showed that among the three oscillation modes, the infinity mode was the best in terms of decreasing the porosity in the weld and enhancing the tensile strength, they deduced that when the oscillation frequency and width increase from 100 to 300 Hz and from 1 to 3 mm, respectively, the weld depth to width ratio decreases meaning that the stirring effect caused by the laser oscillation has a great influence on the casting of the liquid metal, especially on the width of the weld. The effect of oscillation patterns (transverse, longitudinal and circular) of the laser beam on the improvement of the surface morphologies of the 4 mm thick aluminum alloy AA6061-T6 weld in butt configuration was studied by Wang et al. [17]. They showed that the beam oscillation improved the weld morphologies and promoted the formation of equiaxed grains in the fusion zone due to the stirring effect. The main conclusion is that beam oscillation has almost no effect on the tensile strength of the weld. The tensile strengths of all welds were in the range of 220 to 231 MPa, or about 70%, of the base metal. However, it had an obvious effect on the deformation of the weld. The deformation of the circular oscillation weld reached 8%, which is 38%, higher than that of the weld without beam oscillation. The increase in strain was attributed to the increase in the fraction of equiaxed grains in the weld.

Extensive studies have been conducted to reduce porosity formation and improve the mechanical properties of welds. In order to achieve those goals, the researchers in this field of laser welding have used different approaches, such as the optimization of protective gases [18], the application of the two-beam technology [19], the application of a modulated laser power system [20], the use of an oscillating beam [21], the utilization of devices to generate Lorentz forces in the weld pool [22]. Among these methods, laser beam oscillation welding has attracted significant academic and industrial interests in recent years. Busuttil et al., concluded that beam oscillation could reduce both the melt temperature gradient and the weld sensitivity to hot cracking during 6xxx laser welding [23]. Wang et al., also developed circular oscillation welding and obtained sound welds, they concluded that beam oscillation was responsible for increasing the weld ductility [21]. Rubben et al., obtained healthy tailor blanks by using mechanical oscillation of the laser beam to expand the melt [24]. All these studies mentioned before having shown that oscillating laser welding is useful to stabilize the keyhole and improve the quality of the weld. Most of this work focus on the study of the individual effects of the main laser parameters (laser power, speed, focal position, etc.) on the quality of the weld using Nd:YAG lasers and the interaction effects of these parameters. Only a few works have performed a systematic analysis of the effects of the oscillation parameters (frequency, amplitude, power) of the laser beam on the characterization of the weld, including the mechanical properties. Until now, the effects of laser power, oscillation frequency and oscillation amplitude and their interactions on the weld quality of aluminum alloys (5000 series) have not been thoroughly studied. Currently, AA 5052-H32 thin sheet is used to manufacture the TV screen in the rear seat of the aircraft cabin.

In this paper, two types of laser beam oscillations (sinusoidal path, triangular path) were done to weld 5052-H32 aluminum alloy sheets to observe the welding performance. The effects of laser power, oscillation frequency and amplitude, and the interaction effects of these factors on the tensile strength and micro-hardness of the welds were studied. In addition, the comparison between two welding pattern (Fig. 1) was discussed.

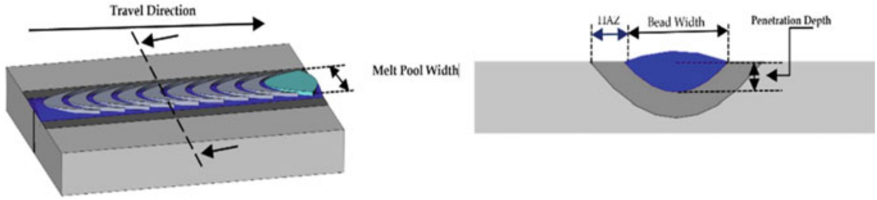


Fig. 1 Geometry of the weld bead (left) with its corresponding cross section (right)

Table 1 Chemical composition (% by weight) determined by XRF analysis of sheet AA 5052-H32

Elements	Al	Mg	Si	Ti	Cr	Mn	Fe	Ni	Cu
AA 5052-H32	69.69	2.62	0.29	0.02	0.18	0.06	0.15	0.01	0.02

Table 2 Mechanical properties of 5052-H32 aluminum alloy sheet

Base material	Ultimate tensile strength (MPa)	Yield strength (MPa)	% Elongation
AA 5052-H32	228 232	171 180	15.3 15

2 Experimental

2.1 *Materials*

In this investigation 5052-H32 aluminum alloy sheets of 2 mm thickness and size 100 mm × 25 mm, were used as the base metal for the experiments. The chemical composition determined by XRF analysis and mechanical properties of 5052-H32 Aluminum alloy sheet are represented in Tables 1 and 2, respectively. The two base metal sheets were welded without gap in a butt joint configuration using the Nd-YAG laser ($\lambda = 1070$ nm) in continuous mode, then after the welding process, the material was machined and cut to obtain specimens that will be used for the characterization, as shown in Figs. 2 and 3.

2.2 *Laser Welding Processing*

The entire welding process has been carried out by adopting a Nd-YAG laser source, the HIGHYAG BIMO laser head powered by an IPG YLS-3000-ST2 fiber laser and mounted on a FANUC M-710IC robot, The robotic arm that controls the movement of the laser beam can move on 6 axes. The laser power is transferred through an optical fiber with a diameter of 200 μm. The ILVDC-Scanner system mounted on

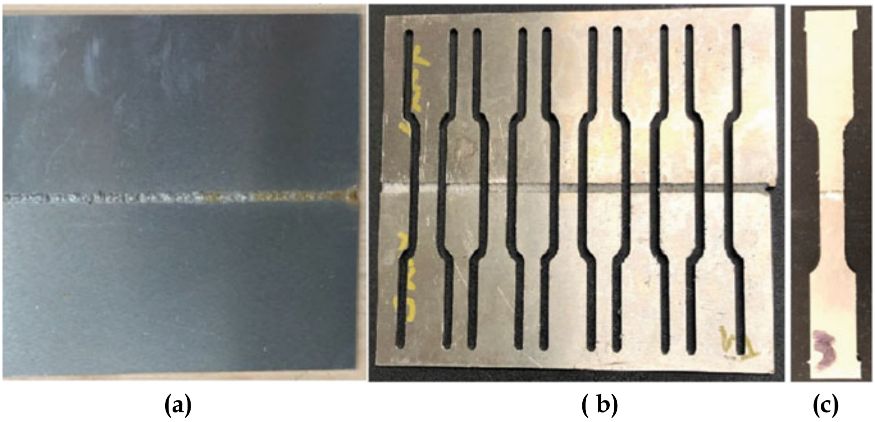


Fig. 2 **a** The two 5052-H32 aluminum alloy sheets after welding, **b** the two sheets after machining and **c** the type of specimens used for our experiments

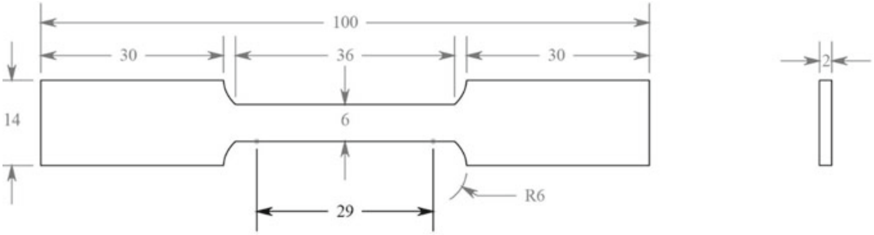


Fig. 3 Tensile test specimen

the laser head is designed to perform the desired scanning patterns using a direct drive motor that rotates the mirror scanner through a controlled back and forth angle to scan the surface being processed. The ILVDC-Scanner can work with frequency up to 1000 Hz. The laser source used has a maximum power of 3 kW, a wavelength of 1070 nm and a focal length of 310 mm and a variable welding speed, the speed chosen in our work is 0.04 m/s note (V). The diameter of the focal spot used in this study is 0.45 mm, and it can be modified by the collimator attached to the laser head. Figure 4 shows the two patterns of beam spot motion considered in this study. The main components of the laser cell are shown in Fig. 5.

The main parameters selected in this study are shown in Table 3. It includes the laser power, the amplitude of oscillation and the frequency. The three levels of each factor are the maximum, middle and minimum values as shown in Table 3, the range of parameters used for the welding process was chosen after some preliminary experimental tests. The parameter combinations shown in Table 4 were obtained using the L9 Taguchi orthogonal method. Three welding experiments were performed for each condition to improve the reliability of the experimental results

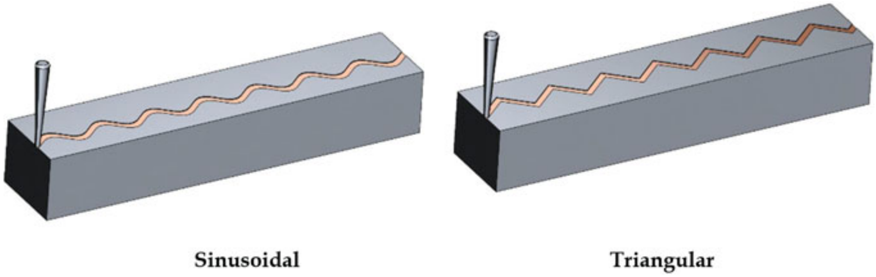


Fig. 4 Schematic representation of two patterns used for laser welding

Fig. 5 LASER cell (Nd-YAG 3 kW) mounted on a FANUC robot with 6 axes

Table 3 Symbols and values of the treatment parameters to be adopted for the welding process

Processing parameter	Symbol	Minimum value level 1	Centre value level 2	Maximum value level 3
Laser power (W)	P	1800	2000	2200
Oscillation frequency (Hz)	F	200	300	400
Oscillation amplitude (mm)	A	1	1.5	2

Table 4 The nine treatment conditions used in the Taguchi method for the laser welding process

Exp. no	Power (W)	Frequency (Hz)	Amplitude (mm)
1	1800	200	1.0
2	2000	200	1.5
3	2200	200	2.0
4	2000	300	1.0
5	2200	300	1.5
6	1800	300	2.0
7	2200	400	1.0
8	1800	400	1.5
9	2000	400	2.0

of the tensile tests. The specimens used for the examination of the micro-hardness were cut, through the weld, using a cutting machine.

2.3 Tensile Test

The tensile test of specimens from different welding patterns (sinusoidal and triangular) were performed according to ASTM E8, using an MTS 810 testing machine at a speed rate of 0.025 mm/min. Three specimens, as represented in Fig. 3 were machined, and tested for each experimental series with the objective of improving the reliability of the result, by calculating the average of the three measurements.

2.4 Regression Model Analysis

In this work, the tensile test results of the two welding modes (sinusoidal and triangular) were used to generate mathematical models using ANOVA and regression analysis. Analyses were performed using Minitab 18 software. The 95% confidence interval was used to determine the significance of factors and their interactions on the responses. Non-significant factors and interactions were removed to develop a valid model that fit the data. To fit the model we use the hierarchy method, this strategy consists in eliminating the factors and the non-significant interactions. Similarly, we adopted a regression analysis to generate mathematical equations that link significant factors to responses. A development of two-dimensional contour plots was done to show the relationship between two continuous factors and the fitted responses. The main effects graphs were studied to interpret the impact of each parameter level on a given average response.

3 Results and Discussion

3.1 Micro-hardness

Hardness measurements were performed using the Vickers micro-hardness technique. The polished specimens were placed in a Vickers hardness tester (ST-2000) to measure the micro-hardness in the region along the median plane of the weld (perpendicular cross-section via the tool direction of the welded aluminum alloy plates), which were 5 mm to the right and left of the middle of the weld. The load applied was 100 kgf for a dwelling time of 10 s. Figure 6 shows the Vickers hardness profile of the welds of sinusoidal and triangular patterns from different welding parameters. The Vickers hardness distribution was found to be symmetrical about the center of the keyhole, showing a W-shaped appearance.

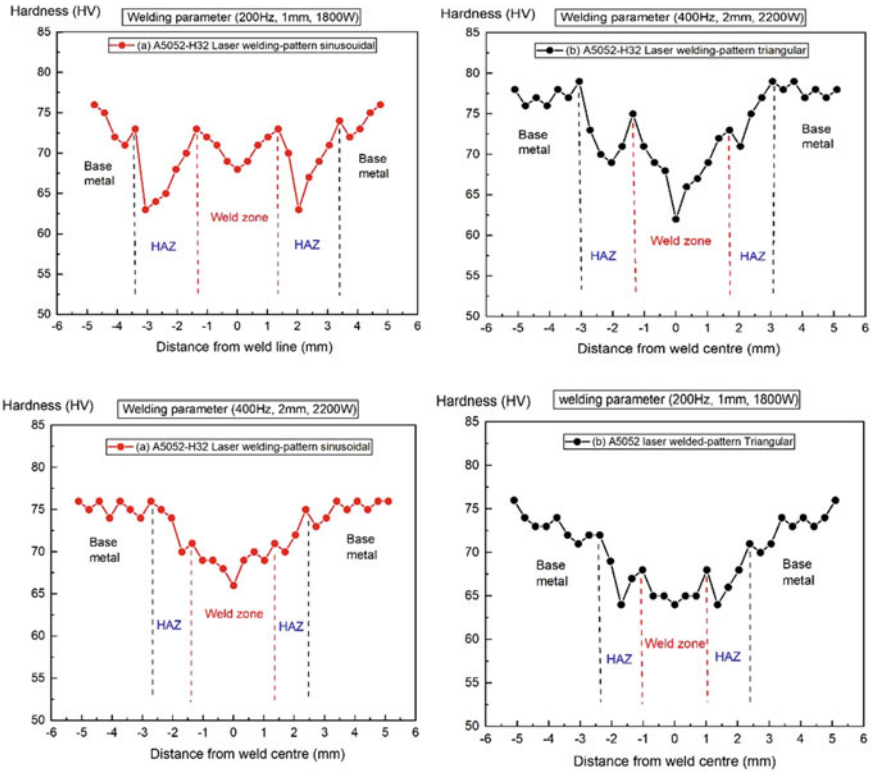


Fig. 6 Vickers hardness profile of the welds of both sinusoidal and triangular welding patterns for different welding parameters

Note that the micro-hardness is lowered at the melting edge and then gradually increases from the heat affected zone to the base metal. The low value of microhardness in the center of the melting zone can be interpreted by the fact that alloying elements are vaporized due to the high temperatures in the melting zone, and the existence of resulting porosity. The average micro-hardness of the weld fusion zone is about 72 HV for the sinusoidal oscillation and 69 HV for the triangular oscillation with the welding parameters involving 2200 W of power, 2 mm of amplitude and 400 Hz of frequency. And for the parametric combination (1800 W, 1 mm, 200 Hz) we found that the micro-hardness is (70.88 HV) for the sinusoidal pattern and (65.71 HV) for the triangular pattern. Another remark to add is that when passing from the first combination to the second one there is an increase of the oscillation amplitude that explains the width of the fusion zone in both welding patterns. The second explanation is that the oscillation of the beam enlarges the interaction area of the laser beam and increases the heat input.

3.2 Macro- and Micro-structures

In this study, the Weck reagent is used as etchant to analyze the macro- and microstructure of our specimens. Figure 8 shows the optical images of the cross sections perpendicular to the direction of movement of the laser head, (a) triangular pattern and (b) sinusoidal pattern. In Fig. 7, we can clearly see the difference between the profiles of the weld seam. This difference is related to the type of pattern used. In addition, we notice that the top surfaces of the welds are concave and rough for both triangular and sinusoidal welding patterns, and the root of the welds are rough for both patterns. These defects can be improved by the circular pattern of the laser beam, [17].

The characteristic microstructures of the weld for the two patterns (sinusoidal and triangular) are presented in Fig. 8. The microstructure of the two scan modes revealed three distinct regions, namely the fusion zone (FZ), the heat-affected zone (HAZ) and the base metal (BM). We observe that the microstructure of the two welding modes significantly differs between the three zones. The nugget zone and the heat affected zone are characterized by the evolution of columnar dendrites. The β(Mg3Al2) dendrites shown in Fig. 8d were formed and exhibited a non-uniform distribution due to the high heat input and melting degree during the laser welding process. Figure 8c shows that the structure of the (HAZ) zone includes larger grains compared to the one found in the weld zone, which can be explained by the grain coarsening generated by the heat [17]. The presence of pores in the fusion zone can be explained by surface contamination and the inclusion of hydrogen. The existence of porosity can but also by the non-use of shielding gas in our welding process.

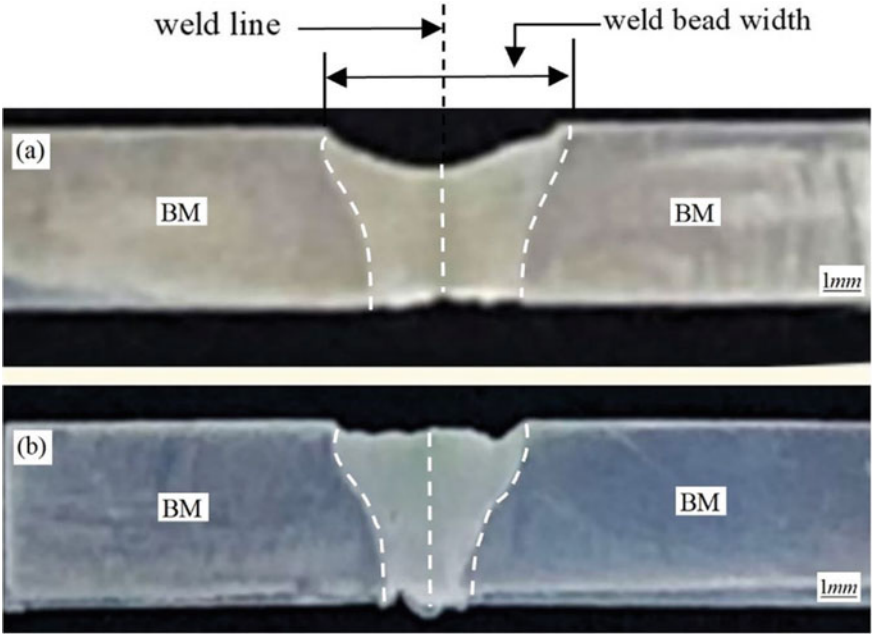


Fig. 7 Optical macrographs of cross sections perpendicular to the direction of movement of the laser head of 5052-H32 aluminum plates welded in butt configuration with parameters (2000 W–400 Hz–2 mm), **a** triangular pattern, **b** sinusoidal pattern

3.3 Measured Experimental Values of Tensile Strength

Figures 2 and 3 show the AA 5052-H32 samples and their dimensions that were used in our experiments. The tensile strength results are obtained from three welding experiments that were performed for each condition and for each welding mode. The experimental tests of the two patterns are respectively presented in Tables 5 and 6. The tensile strength of the welds of both sinusoidal and triangular welding patterns is between (108.3–182.7) and (104.4–177.9) MPa respectively corresponding to a thermal energy by unit length of weld of (45–50) and (45–50) kJ/m respectively. The linear energy by unit length of weld (H) was calculated using the following equation:

$$H = P/V \tag{1}$$

where P is the laser power in Watts and V is the welding speed in m/s. For reference as shown in Table 2, the tensile strength of the base metal is 228 MPa.

The parametric combination of the experimental cycle number 6, in both welding patterns (see Tables 5 and 6), gives a low tensile strength value of 108.3 MPa for the sinusoidal pattern and 104.4 MPa for the triangular pattern. On the other hand, the parametric combination of the experimental cycle number 9, in the two welding

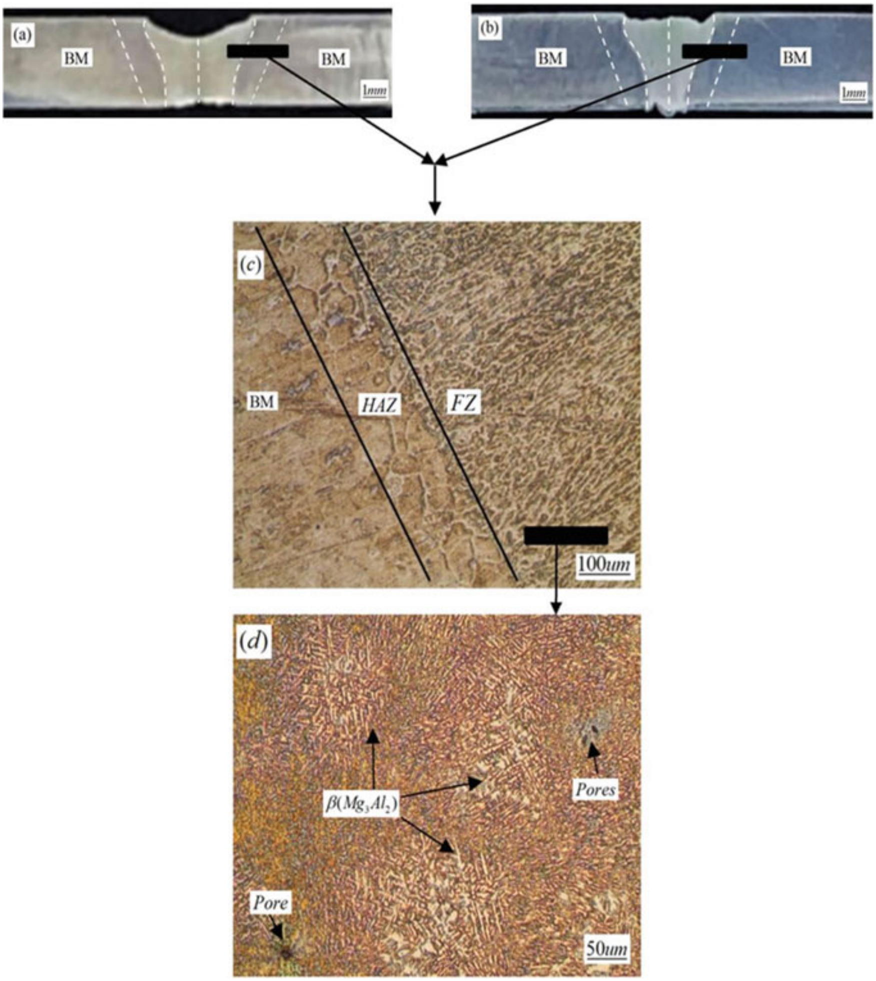


Fig. 8 **a** Microstructure of a weld cross section made under welding, **b** conditions (2000 W–400 Hz–2 mm), **c** Low magnification image of the joint cross section with three distinct regions, (BM), (HAZ), (FZ), **d** Magnification image of the fusion zone

patterns (see Tables 5 and 6), gives a value of the maximum tensile strength which is 182.7 MPa for the sinusoidal pattern and 177.9 MPa for the triangular pattern. In these two parametric combinations 6 and 9, we notice that there is an increase in the tensile strength even though the value of the amplitude factor which dictates the oscillation diameter remains unchanged (fixed). It is concluded that the effect of amplitude factor on tensile strength is very low. On the contrary, the variation of the two other factors power and frequency has a very important role in the variation of the value of the tensile strength. Figure 9 shows a direct comparison between two welding patterns (sinusoidal and triangular). We can clearly see that the sinusoidal

Table 5 Tensile tests performed and experimental results

Scanning pattern sinusoidal				
Exp no	Power (W)	Frequency (Hz)	Amplitude (mm)	Tensile strength value (MPa)
1	1800	200	1.0	127.2
2	2000	200	1.5	177.9
3	2200	200	2.0	147.7
4	2000	300	1.0	152.1
5	2200	300	1.5	140.0
6	1800	300	2.0	108.3
7	2200	400	1.0	150.7
8	1800	400	1.5	109.1
9	2000	400	2.0	182.7

Table 6 Tensile tests performed and experimental results

Scanning pattern triangular				
Exp no	Power (W)	Frequency (Hz)	Amplitude (mm)	Tensile strength value (MPa)
1	1800	200	1.0	110.7
2	2000	200	1.5	175.0
3	2200	200	2.0	152.4
4	2000	300	1.0	168.6
5	2200	300	1.5	143.9
6	1800	300	2.0	104.4
7	2200	400	1.0	147.8
8	1800	400	1.5	137.0
9	2000	400	2.0	177.9

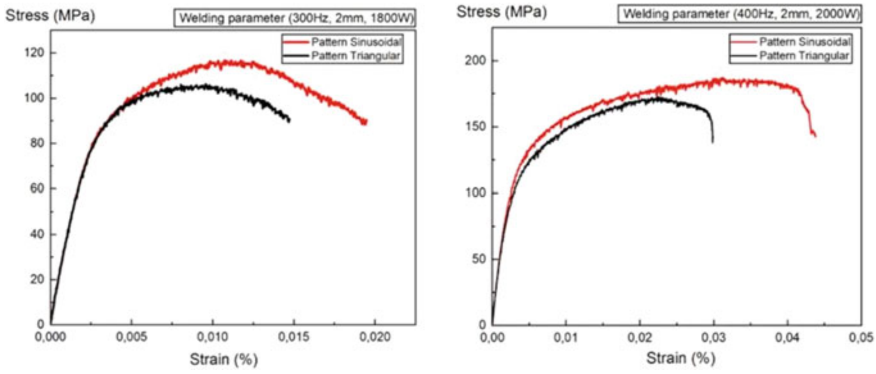


Fig. 9 Comparison of the tensile strength between the two oscillation patterns has two different parametric combinations

oscillation overcomes the triangular oscillation because of its higher value of the tensile strength.

3.4 *Development of the Regression Model*

We applied a statistical study based on two-way analysis (ANOVA) to establish a linear relationship (mathematical model) between the welding parameters and the tensile strength of AA 5052-H32, on the values of the tensile strength measured (experimental) for the two welding patterns (sinusoidal and triangular) as presented in Tables 5 and 6 respectively. In this work, we used the step-by-step method allowing to exclude the non-significant terms after each iteration with requirement of a hierarchical model, a level of statistical significance of 95% was used.

3.4.1 Sinusoidal and Triangular Scanning Pattern

Tables 7 and 8 respectively show the results of the bidirectional analysis models (two-factor ANOVA) of the two welding patterns (sinusoidal and triangular), after removing the three-factor model. Now the models have values R2 = 99.18% (for the boss sinusoidal) and R2 = 99.53% (for the triangular pattern). However, the hierarchy rule was used to eliminate non-significant interactions in order to refine the model.

From Table 7, and after an analysis of main factors first [Laser power (P) − Oscillation frequency (F) − Oscillation amplitudes (A)], we let us see that the factor (P) comes in first place with the greatest contribution which is worth 26.37%, and with a value of P (0.007) < 0.05 for the tensile strength (RT) measured, the second

Table 7 Two-way interaction model for sinusoidal pattern and ANOVA result

Variable	DDL	Square sum	Contribution (%)	Medium square	F-value	P-value
Power (W)	1	1466.41	26.37	3086.85	135.42	0.007
Frequency (Hz)	1	17.68	0.32	674.11	29.57	0.032
Amplitude (mm)	1	12.61	0.23	160.39	7.04	0.118
Two-factor interaction: P * P	1	3264.32	58.69	3264.32	143.20	0.007
Two-factor interaction: F * F	1	496.12	8.92	496.12	21.76	0.043
Two-factor interaction: P * F	1	259.08	4.66	259.08	11.37	0.078
Error	2	45.59	0.82			
Total	8	5561.82	100			
R-sq = 99.18%; R-sq (adj) = 96.72%						

Table 8 Two-way interaction model for pattern triangular and ANOVA result

Variable	DDL	Square sum	Contribution (%)	Medium square	F-value	P-value
Power (W)	1	1410.67	25.77	3718.98	287.18	0.003
Frequency (Hz)	1	100.86	1.84	35.59	2.75	0.239
Amplitude (mm)	1	9.63	0.18	64.69	5.00	0.155
Two-factor interaction: P * P	1	3383.90	61.82	3383.90	261.30	0.004
Two-factor interaction F * F	1	249.39	4.56	249.39	19.26	0.048
Two-factor interaction: P * F	1	293.76	5.37	293.76	22.68	0.041
Error	2	90.59	1.65			
Total	8	5474.11	100			
R-sq = 99.53%; R-sq (adj) = 98.11%						

place is reserved for the factor (F), with a contribution of 0.32% which is low but which remains significant because its P (0.032) < 0.05. On third place we find the factor (A) with a very small contribution equal to 0.23% and which is not significant because its P (0.118) > 0.05, like us explained before, according to the analysis of the results of Tables 5 and 6 the variation of factor (A) has no influence on the (RT). Secondly, we analysed the effects of two-way interactions, we remark that the interaction (P * P), arrives in primary rank with a big value contribution of 58.69% and a value of P (0.007) < 0.05, this clearly reflects the importance from factor (P) and interaction (P * P). The second rank is reserved for interaction (F * F) with a contribution of 8.92%, which is significant because P (0.043) < 0.05. The last rank for the interaction (P * F) because its contribution does not exceed 4.66% which is not significant with the value of P (0.078) > 0.05.

The summary of the coefficients generated and used to develop the sinusoidal sweep linear regression model is shown in Table 9. The linear regression mathematical equation developed to estimate laser welding tensile strength (RT) for AA5052-H32 is given in Eq. (2). Terms such as welding power (P), oscillation frequency (F), interaction (P * P), interaction (F * F) have a noticeable effect on tensile strength,

Table 9 Coefficient of two-way interaction model for pattern sinusoidal

Term	Cofficient	P value
Constant	−3603	0.009
P	3.965	0.007
F	−1.980	0.032
A	13.08	0.118
P * P	−0.001010	0.007
F * F	0.001575	0.043
P * F	0.000509	0.078

however the term amplitude (A) and the interaction between (P * F) they do not have a synergistic effect on tensile strength.

$$RT = -3603 + 3.965P - 1.980F + 13.08A \\ - 0.0010P * P + 0.0015F * F + 0.000509P * F \quad (2)$$

The model fit was verified by analyzing the plots of the residuals, as shown in Fig. 10. It is evident from the normal probability plot (Fig. 10 (left)) that the errors are normally distributed because the data (i.e., the residuals) are distributed along the straight line. Based on results from Zhao et al. [25], the assumption that the residuals are normally distributed was satisfied. The verification of the regression model was further clarified based on the analysis of the fits against (Fig. 10 (right)). The residuals appear to be randomly distributed between the high and low fitted values. This indicates that the regression model assumption was satisfied over the entire range of fitted values.

According to the analysis of the results presented in Table 8 we noted firstly that the factor (P) comes first with the contribution which is worth 25.77%, and with a value of P (0.003) < 0.05 for tensile strength (RT) measured, the second place is reserved for the factor (F), with a contribution of 1.84% which is low and which is not significant because its P (0.718) > 0.05, the last main factor (A) also is not significant because its P (0.239) > 0.05.

For the analysis of the effects of two-factor interactions we notice that the interaction (P * P), has a large contribution value of 61.82% and a value of P (0.004) < 0.05. This clearly reflects the importance of factor (P) and interaction (P * P). The second place is reserved for the interaction (P * F) with a contribution of 5.37%, which is significant because P (0.048) < 0.05. The interaction (F * F) is in third place because its contribution does not exceed 4.56% which is also significant with the value of P (0.041) < 0.05.

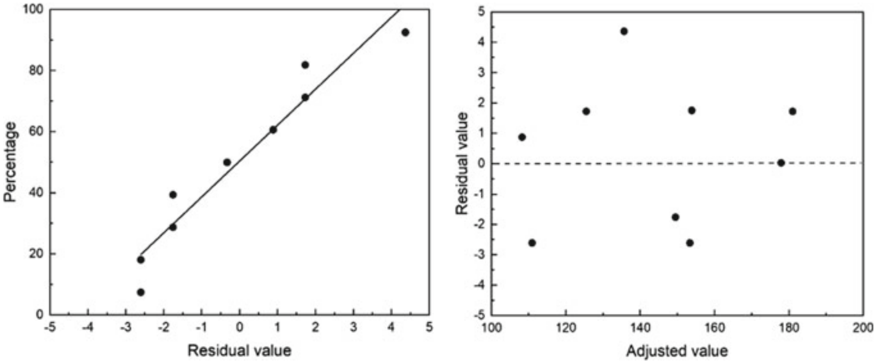


Fig. 10 Residual of normal probability (left), residual of normal probability as a function of adjusted value (right)

Table 10 Coefficient of two-way interaction model for pattern Triangular

Term	Coefficient	P value
Constant	−4325	0.004
P	4.353	0.003
F	0.455	0.239
A	−8.31	0.155
P * P	−0.001028	0.004
F * F	0.001117	0.048
P * F	−0.000542	0.041

The summary of the coefficients generated and used to develop the linear regression model for the triangular scan is shown in Table 10. The linear regression mathematical equation developed to estimate the tensile strength (TS) of laser welding for AA5052-H32 is given in Eq. (3). The only terms that have a noticeable effect on tensile strength are the laser power factor and the interaction (P * P), the interaction (F * F), the interaction between (P * F), however the frequency (F) and amplitude (A) of oscillation do not have a synergistic effect on tensile strength.

$$RT = -4325 + 4.353P + 0.455F - 8.31A - 0.001028P * P \\ + 0.001117F * F - 0.000542P * F \quad (3)$$

Likewise, the triangular mode welding model was verified by analyzing the residual plots, as shown in Fig. 11. In the normal probability plot (Fig. 11 (left)) by noticing that the errors are normally distributed, because the data (i.e., residuals) are distributed along the straight line, then according to the results of Zhao et al. [25], the assumption that the residuals are normally distributed was satisfied. The verification of the regression model was further elucidated on the basis of the fit analysis with respect to (Fig. 11 (right)). The residuals appear to be randomly distributed between

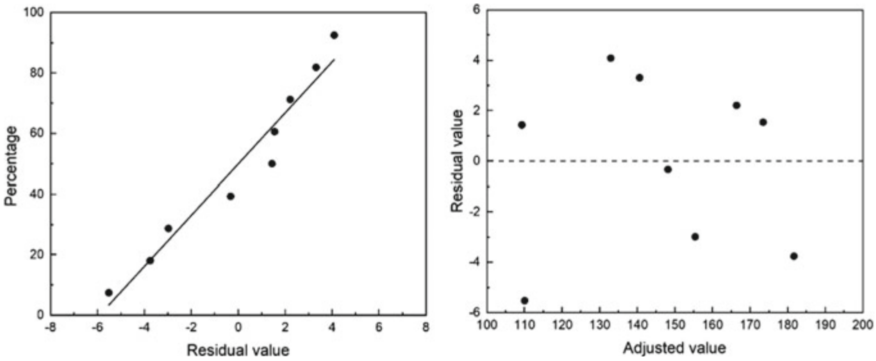


Fig. 11 Residual of normal probability (left), residual of normal probability as a function of adjusted value (right)

the high and low fitted values. This indicates that the regression model assumption was satisfied over the entire range of fitted values.

Before finishing this sub-section of the development of the regression model and starting the sub-section which deals with the graphic analysis of the main effects of each model, which means of each welding pattern. We can say that the sinusoidal and triangular modes are good candidates for predicting tensile strength. This conclusion is based on the summaries of the coefficients generated and used to develop each model. As in this work we are interested in the study of the welding performance more particularly the mechanical properties we can say that the sinusoidal sweep model is more advantageous, because it's this which gives greater values of the tensile strength.

3.4.2 Graphical Analysis of the Main Effects

The main effects graph displays the mean of each level of factors and graphically interprets the relative importance of the contribution of the different factors on the total variation of the response. The straight line connecting each two levels reflects the influence of each factor level increase on the measured responses. The steeper the slope between the two levels, the more the level of the factor in question affects the response and vice versa. Figure 12 shows the response variation curves (tensile strength RT) of the two sinusoidal and triangular welding modes as a function of the main effects related to the factors experienced, laser power (P), oscillation frequency (F) and the oscillation amplitude (A). These graphs allow us to perform a simple analysis by observing the effects of the parameters studied on the evolution of the response. The significance of a factor is determined by the change in the response, following the change in the level of the factor, the greater the change in the response, the more the change from a low level to a higher level of the factor occurs fast and is manifested by a steep slope in the main effects graph [26, 27].

Figure 12 shows the effects of factors P, F and A on tensile strength. In this case, the main effects graph indicates that the three factors for each welding mode follow different trends, we also observe that the lines connecting the three levels of each factor are not horizontal, so we can conclude that the three Factors studied affect the tensile strength but with different degrees. According to Fig. 12, it's obvious that the increase in laser power (P) and more precisely the passage from level 1 to level 2 for the two welding patterns positively affects the tensile strength, while the increase in the oscillation frequency (F), when going from level 1 to level 2 for both welding patterns negatively affects tensile strength (RT). For the amplitude of oscillation (A) of triangular pattern we notice that its increase when it goes from level 1 to level 2 affects positively tensile strength. On the contrary, the change from level 1 to level 2 of the amplitude (A) has a weakly negative effect at the (RT) for the sinusoidal pattern, this is explained by the weak slope of the line. Switching from level 2 to level 3 of (P) for both sweep modes seem to have an opposite effect to that of the first passage from level 1 to level 2, as does increasing (F), when it goes from level 2 at level 3, it seems to have an opposite effect than that of the first passage from level

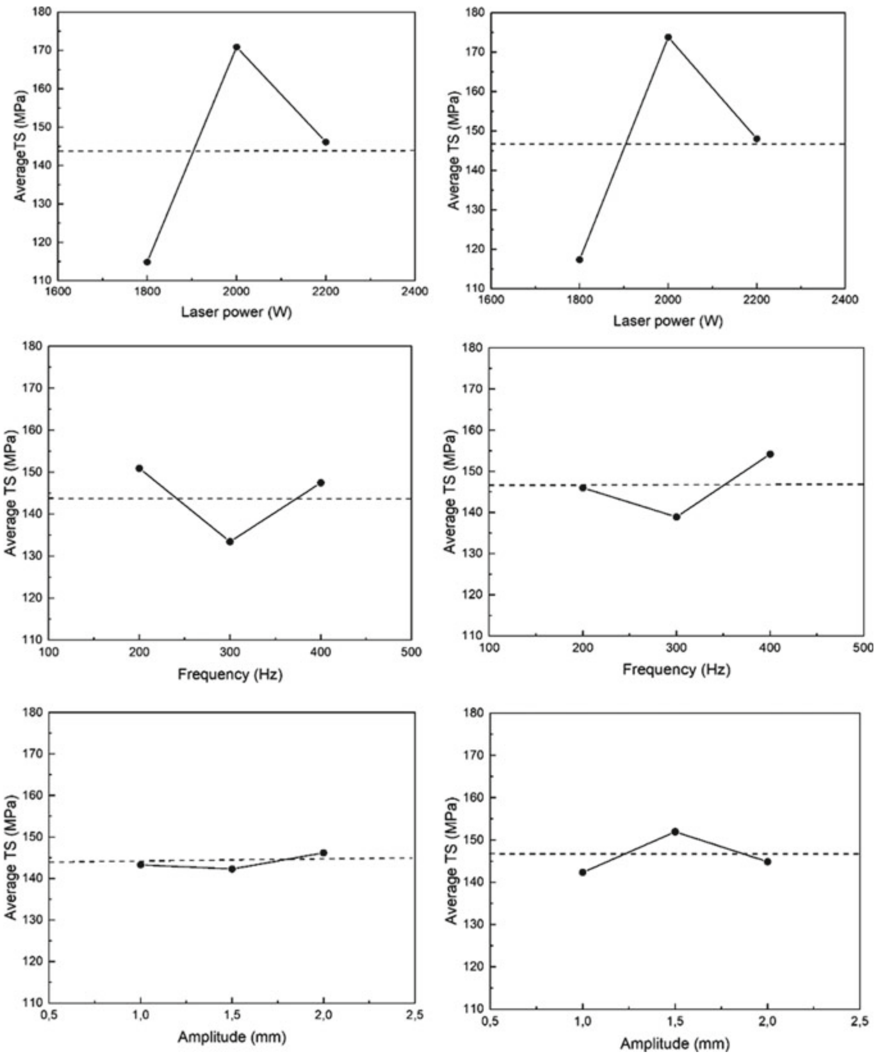


Fig. 12 Main effects graph for tensile strength (left), sinusoidal pattern, (right) triangular pattern

1 to level 2. As well as the passage from level 2 to level 3 for the amplitude factor (A) for the two welding modes, it has the opposite effect to that of the first passage from level 1 to level 2.

3.5 Variation of Tensile Strength with Sinusoidal and Triangular Welding Process Parameters

Prior to the analysis of the welding process parameters of the two oscillation patterns, the predicted values for all 9 experimental series were calculated and compared to the measured experimental values. As shown in Tables 11 and 12, the deviations between the predicted values of the tensile strength and the experimental values for all 9 parametric combinations are less than 4%. This indicates that both regression models can predict the tensile strength of AA 5052-H32 with good accuracy. Figure 13 shows the scatter plot of the measured and predicted tensile strength (TS) for the two welding

Table 11 Experimental and predicted values of sinusoidal pattern tensile strength given by the regression model

Exp no	Processing parameters			Tensile strength (MPa)		Error (%)
	Power (W)	Frequency (Hz)	Amplitude (mm)	Experimental	Model predicted	
1	1800	200	1.0	127.2	124.9	1.83
2	2000	200	1.5	177.9	177.2	0.38
3	2200	200	2.0	147.7	148.7	0.69
4	2000	300	1.0	152.1	153.2	0.74
5	2200	300	1.5	14	134.9	3.77
6	1800	300	2.0	108.3	110.4	1.88
7	2200	400	1.0	150.7	152.6	1.25
8	1800	400	1.5	109.1	107.7	1.30
9	2000	400	2.0	182.7	180.4	1.30

Table 12 Experimental and predicted values of triangular pattern tensile strength given by the regression model

Exp no	Processing parameters			Tensile strength (MPa)		Error (%)
	Power (W)	Frequency (Hz)	Amplitude (mm)	Experimental	Model predicted	
1	1800	200	1.0	110.7	111.9	1.10
2	2000	200	1.5	175	175.4	0.24
3	2200	200	2.0	152.4	156.7	2.72
4	2000	300	1.0	168.6	172.5	2.27
5	2200	300	1.5	143.9	142.9	0.68
6	1800	300	2.0	104.4	107.4	2.80
7	2200	400	1.0	147.8	151.5	2.46
8	1800	400	1.5	137	137.7	0.50
9	2000	400	2.0	177.9	179.5	0.89

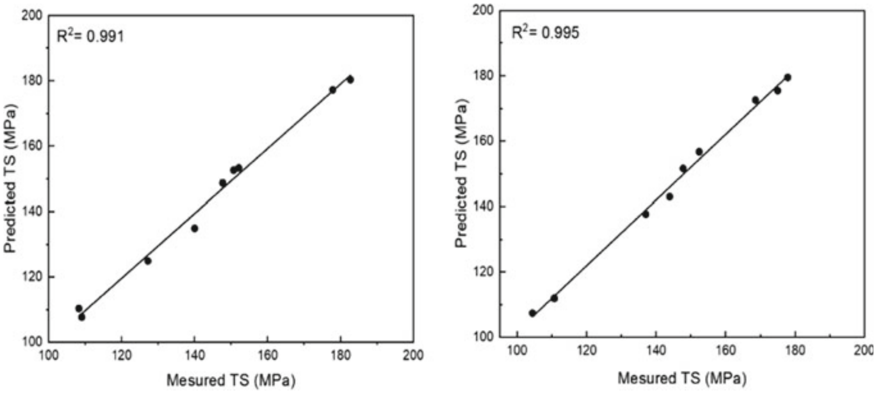


Fig. 13 Scatter plot of measured and predicted tensile strength (TS) for the pattern sinusoidal (on the left) and the triangular pattern (on the right)

patterns. The predicted (TS) values follow the measured values with an overall error of less than 14%.

3.5.1 Influence of Laser Power, Frequency and Amplitude Oscillation for Sinusoidal and Triangular Welding Patterns

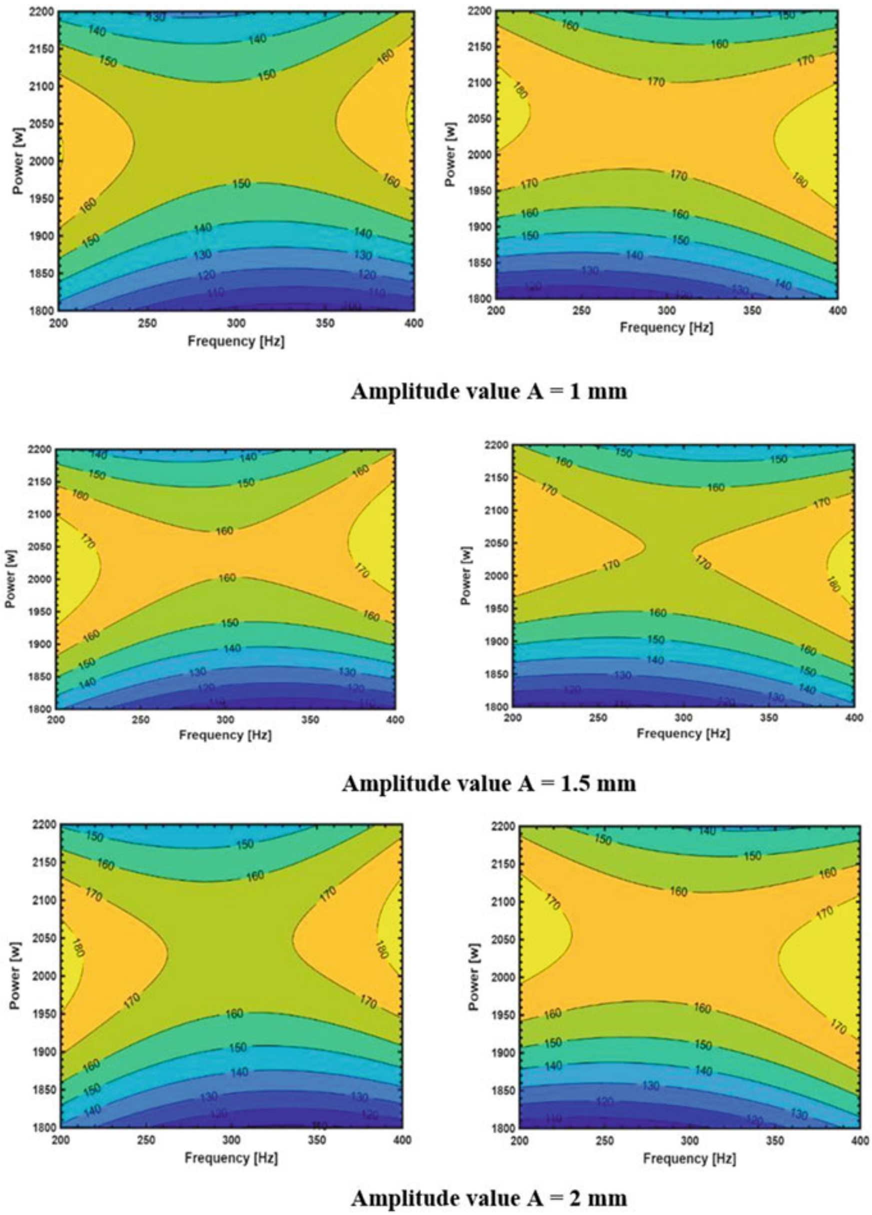
The summary of the ANOVA of the two Tables 7 and 8 which corresponds to the welding by the sinusoidal and triangular sweep shows that for the laser power followed by the frequency of oscillation welding contributes significantly to the tensile strength because their P values are less than 0.05. For the triangular mode, we just find the laser power parameter, which is significant, with a $P < 0.05$. The laser oscillation amplitude with a P value > 0.05 has no significant effect for both welding modes. This is confirmed by the analysis graphic presented in Fig. 7, we notice that the variation of the mode oscillation amplitude sinusoidal when goes from (1 mm) to (1.5 mm) presents a negligible dimension and when it goes from (1.5 mm) to (2 mm), it shows a negligible increase in the mean of (TS), but there is always a factor which has no influence on the mean tensile strength because its variation of (1 mm) to (2 mm) is almost linear, normally from previous Shangren Li results [15], it confirms that the porosity could be removed by laser oscillation with a high frequency of more than 200 Hz and a large diameter greater than (2 mm). Hence from this result we can say that the decrease in porosity automatically reflects an increase in (TS) and the values used of parameter amplitude in this work are less than or equal to 2 mm. Then it is for this reason that the effect of the amplitude is not significant for the two welding modes. These trends are similar for the measured experimental values and the values predicted by the model. Within the range of parameters (i.e. processing window) used in this work, tensile strength increased with increasing welding frequency for both welding modes (sine and triangular) more particularly when the oscillation frequency changes from 300 to 400 Hz, this can be interpreted as follows: normally when the frequency increases the period of oscillation decreases then this generates a compression of the shape (sinusoidal and triangular) and then the laser will interact with a large surface of the sample. Even if the laser power (P) is a significant parameter for both welding modes, we noticed that when the power goes from (2000 W) to (2200 W) we have a decrease in the (TS), this can be explained by the loss of alloying element by vaporization due to the increase in laser power then this is reflected in a decrease in (RT).

3.5.2 Effect of Bidirectional Interaction on Tensile Strength

As shown in Tables 7 and 8, the interactions (P * P) and (F * F) contribute significantly to the tensile strength for the two welding modes, we also notice that the interaction (P * F) for the sinusoidal mode is not significant because its $P > 0.05$. However, this same interaction contributes significantly to the tensile strength for the triangular pattern with a $P < 0.05$. Therefore, only the effect of this interaction (P * F) was analysed. In order to understand the effects of the interaction between welding laser power (P) and oscillation frequency on the tensile strength of laser welded AA 5052-H32 and to prove the suitability of the model, a plot of the contours was generated, as shown in Fig. 8. This is called the Response Surface methodology or RSM which is a set of mathematical and statistical methods that can quickly and efficiently predict the optimal response based on process parameters. Contour lines predicted tensile strength values for varying laser power with different values of oscillation frequency. For example, the sinusoidal pattern tensile strength has increased from 100 to < 170 MPa. When the laser power increased from 1800 to 2050 W for an oscillation frequency of 200 Hz to 400 Hz and a fixed oscillation amplitude of 1 mm, the triangular mode tensile strength increased from 100 to < 180 MPa. Based on the analysis of the different RSM response surfaces of both sinusoidal and triangular welding patterns and to maximize the tensile strength of AA 5052-H32, the range of parameters have to be chosen as the following, a welding power between 1900 W up to 2100 W with oscillation frequencies between 350 to 400 Hz as well as an amplitude greater than 1.5 mm (Fig. 14).

4 Conclusion

1. Two laser beam oscillation patterns, the sinusoidal and triangular pattern, were successfully studied for weldability analysis of AA 5052-H32 within a parameter window (laser power 1800–2000–2200 W; oscillation frequency 200–300–400 Hz; oscillation amplitude 1–1.5–2 mm).
2. The maximum tensile strength is (~183 MPa) for the sinusoidal pattern and (~178 MPa) for the triangular pattern. These two maximum values of tensile strength were found for each pattern at the following same parametric



Amplitude value A = 1 mm

Amplitude value A = 1.5 mm

Amplitude value A = 2 mm

Fig. 14 Contour plots showing the interaction effect of sinusoidal (left) and triangular (right) oscillation frequency versus welding laser power with different amplitude hold values

combination, 2000 W laser power, 400 Hz oscillation frequency and 2 mm amplitude.

3. Two mathematical regression models to predict the tensile strength of laser welding of AA 5052- H32 in the parameter window used in this work have been developed and validated successfully for each swing pattern.
4. Laser power and the oscillation frequency have significant effects on the tensile strength of welds for the sinusoidal pattern. For the triangular pattern, we found that the significant effects are the laser power and the power frequency oscillation interaction.
5. For the two oscillation patterns (sinusoidal and triangular), the tensile strength of the welded parts is more sensitive with laser power values between 1800–2000 W and oscillation frequency values between 300–400 Hz with amplitudes greater than 1 mm.
6. Currently, AA 5052-H32 thin sheets are used for manufacturing lightweight and economical TV screen frames on the backs of seats in airplanes, it would be interesting to extend this study to other kind of patterns with other welding configurations.

Acknowledgements We are very grateful to Professors Sasan Sattapanah Karganroudi and El Ouardi Oualid for their helpful discussions. We are also very grateful to our referees for their helpful suggestions and critical remarks that allowed us to expand the scientific content of our work.

References

1. Mankari K, Acharyya SG (2017) Development of stress corrosion cracking resistant welds of 321 stainless steel by simple surface engineering. Appl Surf Sci 426:944–950
2. Gao XL, Liu J, Zhang LJ (2018) Dissimilar metal welding of Ti6Al4V and Inconel 718 through pulsed laser welding-induced eutectic reaction technology. Int J Adv Manuf Technol 96:1061–1071
3. Windmann M, Röttger A, Kügler H, Theisen W (2016) Laser beam welding of magnesium to coated high-strength steel 22MnB5. Int J Adv Manuf Technol 87(9):3149–3156
4. Cao XJ, Jahazi M, Immarigeon JP, Wallace W (2006) A review of laser welding techniques for magnesium alloys. J Mater Process Technol 171(2):188–204
5. Soltani HM, Tayebi M (2018) Comparative study of AISI 304L to AISI 316L stainless steels joints by TIG and Nd: YAG laser welding. J Alloy Compd 767:112–121
6. Ahn J, Chen L, He E, Dear JP, Davies CM (2018) Optimisation of process parameters and weld shape of high power Yb-fibre laser welded 2024–T3 aluminium alloy. J Manuf Process 34:70–85
7. Chen W, Molian P (2008) Soudage au laser à deux faisceaux d'aluminium AA 5052–H19 ultra-mince. Int J Adv Manuf Technol 39:889–897
8. Bunaziv I, Akselsen OM, Salminen A, Unt A (2016) Soudage hybride laser à fibre—MIG d'un alliage d'aluminium 5 mm 5083. J Mater Process Technol 233:107–114
9. Abioye TE, Farayibi PK, Kinnell P, Clare AT (2015) Microstructures de Ni-Ti à gradation fonctionnelle synthétisées en cours de processus par dépôt direct de métal par laser. Int J Adv Manuf Technol 79(5–8):843–850

10. Popov S (2008) Fibre laser overview and medical applications. Tunable laser applications, 2nd edn. CRC Press, New York, USA, pp 197–226
11. Okon P, Dearden G, Watkins K, Sharp M, French P (2002) Laser welding of aluminium alloy 5083. In: Proceeding of 21st international congress on applications of lasers and electro-optics, Scottsdale, 14–17 Oct, pp 1–10
12. Ayoola WA, Suder WJ, Williams SW (2018) Parameters controlling weld bead profile in conduction laser welding. J Mater Process Technol 249:522–530
13. Al-Sayyad A, Bardon J, Hirchenhahn P, Santos K, Houssiau L, Plapper P (2018) Aluminum pretreatment by a laser ablation process: influence of processing parameters on the joint strength of laser welded aluminum–polyamide assemblies. Procedia CIRP 74:495–499
14. Shanavas S, Dhas JER, Murugan N (2018) Weldability of marine grade AA 5052 aluminum alloy by underwater friction stir welding. Int J Adv Manuf Technol 95:4535–4546
15. Li S, Mi G, Wang C (2020) A study on laser beam oscillating welding characteristics for the 5083-aluminum alloy: Morphology, microstructure and mechanical properties. J Manuf Process 53:12–20
16. Wang Z, Oliveira JP, Zeng Z, Bu X, Peng B, Shao X (2019) Laser beam oscillating welding of 5A06 aluminum alloys: microstructure, porosity and mechanical properties. Opt Laser Technol 111:58–65
17. Wang L, Gao M, Zhang C, Zeng X (2016) Effect of beam oscillating pattern on weld characterization of laser welding of AA6061-T6 aluminum alloy. Mater Des 108:707–717
18. Yu Y, Wang C, Hu X, Wang J, Yu S (2010) Porosity in fiber laser formation of 5A06 aluminum alloy. J Mech Sci Technol 24(5):1077–1082
19. Matsunawa A, Mizutani M, Katayama S, Seto N (2003) Porosity formation mechanism and its prevention in laser welding. Weld Int 17(6):431–437
20. El-Batahgy A, Kutsuna M (2009) Laser beam welding of AA5052, AA5083, and AA6061 aluminum alloys. Adv Mater Sci Eng
21. Katayama S, Nagayama H, Mizutani M, Kawahito Y (2009) Fibre laser welding of aluminium alloy. Weld Int 23(10):744–752
22. Madhusudhana Reddy G, Srinivasa Murthy CV, Viswanathan N, Prasad Rao K (2007) Effects of electron beam oscillation techniques on solidification behaviour and stress rupture properties of Inconel 718 welds. Sci Technol Weld Joining 12(2):106–114
23. Trushnikov DN, Koleva EG, Mladenov GM, Belenkiy VY (2013) Effect of beam deflection oscillations on the weld geometry. J Mater Process Technol 213(9):1623–1634
24. Sun Z, Karppi R (1996) The application of electron beam welding for the joining of dissimilar metals: an overview. J Mater Process Technol 59(3):257–267
25. Zhao D, Wang Y, Liang D, Zhang P (2016) Modeling and process analysis of resistance spot welded DP600 joints based on regression analysis. Mater Des 110:676–684
26. Montgomery DC (2017) Design and analysis of experiments. Wiley
27. Lawson J (2014) Design and analysis of experiments with R, vol 115. CRC press

The Internet of Things Solutions for Transportation

Arunima Sharma and Ramesh Babu Battula

Abstract Before the principal vehicle was created by Henry Ford in 1908, all shipments were dealt with via carriage and ponies. Organizations needed to battle with low security and helpless proficiency utilizing these strategies. At the point when the vehicles began to show up in the business sectors and on streets, individuals began utilizing them for the conveyance of merchandise. It brought about a lesser opportunity to transport items and decreased coordinations costs. Later toward the finish of the twentieth century, coordinations encountered a sensational change that changed the conveyance interaction: coordinations programming. From that point forward, innovation has assumed control over the execution and arranging of different errands, including administrative work and manual work process, while getting weak data. Innovation is constantly upsetting the coordinations area and changing the way how cargo, deals orders, materials, merchandise, creation and stock are overseen. Organizations consistently search for an answer that can carry insight to the coordinations tasks work process and assist with diminishing significant expenses. As speed, insight and proficiency turned into the critical deciding elements; the coordinations area has embraced arising advancements, including AI, IoT and blockchain, to fulfill the rising need and manage complex cycles.

Keywords Internet of Things · Artificial intelligence · Intelligent transportation systems · Vehicles · Logistics

1 Introduction

The Internet of Things (IoT) is transforming standard urban communities into shrewd urban communities. In a smart city, sensors, information and availability consolidate to make computerized innovations and correspondence frameworks, further developing activities across a city [1].

A. Sharma (✉) · R. B. Battula
Malviya National Institute of Technology, Jaipur, India
e-mail: rbbattula.cse@mnit.ac.in

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_15

The present city organizers are attempting to join enormous information applications, 5G organizations, complex security frameworks and more to change residents' lives and work. These complex, consistently developing frameworks guarantee another age of productive, safe and harmless to the ecosystem metropolitan spaces [2].

Urban areas face three significant difficulties in their journey to become brilliant urban communities: overseeing tremendous measures of information, embracing IoT across each city work and getting fundamental foundation against a developing number of dangers [3].

Huge information has fantastic applications in shrewd urban communities—in the event that it tends to be assembled and perceived. The chance to utilize huge information to advise better, more productive activities in smart urban areas is tremendous, however just if that information can be adequately gathered and broke down to drive those upgrades. Productive arrangement is basic to moving and handling huge measures of information from a wide and developing exhibit of gadgets [4].

Enormous IoT is the future, and urban areas need to plan for it today. As the reception of the smart city idea develops, the utilization of gigantic volumes of IoT gadgets and sensors to screen and control each part of city activities and usefulness will grow [5]. Enormous IoT arrangements incorporate somewhere in the range of hundreds to millions of IoT gadgets, with the essential objective of communicating and burning-through modest quantities of information from many sources. The test will fabricate, sending, working and keeping up with the gadgets with physical, programming and administration segments that offer versatility, security, low force and minimal expense to deal with the development of inclusion and usefulness extent of the arrangement. In anticipation of full 5G IoT over NR, outfitting 5G NSA networks with dynamic range sharing (DSS) will future-evidence current LTE-based monstrous IoT norms in organization now.

Urban areas should design now and be ready to incorporate sensors and associated IoT gadgets into their tasks persistently. This constant incorporation is fundamental to guarantee that usefulness and interchanges stay continuous as smart urban areas advance. Hearty cell network plans make the establishment for this development.

Smart city framework requires strategic security at scale. As urban areas become more reliant upon computerized capacities, they likewise become more helpless against cyberattacks. A complex assault on a force matrix could make huge and conceivably unsalvageable damage a city and its occupants. Moral inquiries around the utilization of individual data, like facial acknowledgment information, likewise surface in a smart city climate [6]. Responsibility and client following additionally become issues as client volume increments and more individuals approach touchy information. Incorporating complex security into the whole information venture from the very edge over portable organizations across the tremendous wired web foundation and the cloud is significant in keeping smart urban communities secure.

5G will assume a considerable part in speeding up the smart city change through expanded capacities, for example, the huge scope computerization of vehicle armadas, smart matrix utility administrations and brilliant natural checking. With private 5G and LTE, explicit IoT applications and administrations can be genuinely isolated from the wide-region full scale organization to satisfy the network needs for associations with uplifted security needs like utilities, modern locales and wellbeing focuses. Corporate and instructive grounds and different conditions can acquire consistent wandering with the capacity to switch among public and private organizations, keeping everybody associated and secure.

1.1 IoT Brings Fundamental Changes in the Transit Equation

The Internet of Things (IoT) has the potential to change the vehicle industry by significantly modifying how transportation frameworks accumulate information what's more, data by uniting the significant specialized and business patterns of versatility, robotization what's more, information investigation. IoT alludes to the systems administration of actual articles using implanted sensors, actuators, and different gadgets that can gather and send data about continuous movement in the organization. The information accumulated from these gadgets can then, at that point be dissected by transportation specialists to:

- Improve the voyager experience, with more trustworthy transportation, improved client benefits and better and that's just the beginning precise correspondence and data.
- Increase wellbeing, by better understanding travel framework activities through sensor information that tracks everything from abnormalities in train speeds, street temperatures, airplane part condition, to the quantity of vehicles holding up at a crossing point.
- Reduce clog and energy use, using ongoing information to further develop how authorities scale assets to fulfill need, with the readiness to respond rapidly to quick changing traffic patterns, or to address traffic sway on fuel use, the climate, and local financial intensity.
- Improve functional execution, by proactively checking basic framework and making more effective cycles to decrease working expenses and further develop framework limit.

2 Coordinations and Transportation Management

Coordinations and Transportation Management is a field that needs examination and exactness. It controls the conveyance of products or materials from providers to clients. It is the duty of coordinations experts to zero in on transportation, essentially, the acquisition and arranging of transportation for products and materials [7].

Coordinations and Transportation Management System mostly handles the accompanying parts:

- Pickup and Delivery Request
- Transporter Management
- Pickup Optimization
- Distribution center Management
- Travel
- Conveyance
- BI and Revealing.

2.1 Pickup and Delivery Requests

The job of a client is to appoint pickup and conveyance demands that go to the transporter. Here's the means by which you can foster an interface for clients to send pickup and conveyance demands. You don't have to foster another answer for dealing with the pickup and conveyance of merchandise without any preparation. There are numerous coordinations programming accessible on the lookout, including NetSuite and McLeod, that can be utilized to deal with the tasks successfully.

NetSuite

NetSuite is one of the famous stages for coordinations activities that work with incorporated anticipating and planning, production network and stock, income the executives, client relationship the board and business knowledge.

McLeod

McLeod gives shipping programming and transportation the executives answers for the shipping organizations and permits shipping dispatch activities the board, report imaging, armada the executives, business measure computerization and EDI. You can utilize the above answers for handle the pickup and conveyance demands. At the point when the client sends the solicitations for pickup or conveyance, the solicitation is shipped off the transporter who appoints the shipment to various drivers. It becomes fundamental for transporters to examine if the pickup or conveyance is finished on schedule or how frequently the driver digresses from the arranged course. It very well may be conceivable by carrying out AI and IoT to the coordinations programming.

Gartner has anticipated that half of the huge worldwide organizations will utilize progressed investigation, IoT and AI in their production network tasks.

We should see how arising advancements can guarantee the on-time get and conveyance of merchandise.

2.2 Transporter Management

Utilizing Predictive Analytics calculations, transporters can recognize if the conveyances are done productively or not.

- Prescient Analytics Guaranteeing on Schedule and In-Full Delivery
  Continuous prescient coordinations examination guarantee that armadas show up on schedule, products are gotten and continued on time so shipments are conveyed to clients when they need it. Sensor-empowered resources or IoT gadgets implanted in trucks, prepares or transports feed information like motor execution and speed and send it to the transporters who can demonstrate and anticipate the assessed appearance times and motor disappointments. For instance, telematics information caught from a vehicle can uncover its speed, position, condition and time passed on to arrive at the objective [8].
  The caught information can be utilized to tell recipients of deferrals alongside the heap/dump exercises, trucks making a beeline for a similar objective and item conveyance necessities. Thus, it guarantees limited postponements and satisfied client assumptions. Shrewd coordinations utilizing AI and IoT helps ports, transporting organizations, providers and specialists advance asset use and their timetables.
- Coordinations Demand Forecasting Dependent on Inventory and Orders Data
  With Logistics Demand Forecasting, organizations can expect the interest for shipments and items across the production network. Coordinations organizations need to execute an estimating model to assess limit request dependent on the blend of recorded information, including stock information and request information [9].
  With custom interest guaging models, organizations can accomplish an exact estimate that assists them with understanding the level of extra limit required, decrease the kilometers spent repositioning resources and further develop payload vehicle limit and resource use.
  Since the coordinations organizations can gauge resource and shipment requests precisely, they can expand payload limit and popularity products can be conveyed to clients on schedule.
- Programmed IoT and AI-Driven Shipment Notification
  Carrying out IoT sensors in coordinations makes following of merchandise more open. Sensors are utilized to catch and trade information. IoT permits taking care of information distantly across the organization foundation. For instance, sensors can be inserted in vehicles conveying products starting with one spot then onto the next. Information caught by sensors can be changed over into significant experiences utilizing AI.
  Computer based intelligence empowered investigation work with following of shipments from flight focuses to the last objective and send following reports during the excursion. Continuous observing of products gives data, including:

  1. Altering or burglary during the excursion
  2. Flight and appearance times
  3. Live area of the shipment
  4. Any deviations from the planned course.

  With all the data close by before the shipment arrives at the objective, coordinations organizations can work on the get and conveyance of orders. The information caught during the store network of merchandise can likewise gauge future interest that aides in item advancement and organic market arranging. When the item is reached at the last objective, fabricating organizations can keep on advancing the creation of products.

We should think about the case of prepared food things. Various sorts of sensors can be utilized to screen the food creation express, the temperature under which they are kept and dispatching time. Peril Analysis and Critical Control Points Checklists are utilized all through the assembling, creation and conveyance methodology to guarantee that the nature of food isn't hampered.

IoT sensors send valuable information identified with the food inside the production network, empowering organizations to try sanitation arrangements. It's difficult aides in gathering food handling guidelines yet additionally keeping up with client reliability and trust with complete straightforwardness.

Another job of the transporter is to furnish the advanced courses and collaborate with distribution center proprietors for the coordinated conveyance of products. Advancing the transporter course utilizing brilliant advancements, including AI and IoT, can further develop coordinations tasks and diminish an opportunity to get and convey products. IoT gives continuous bits of knowledge that can be observed and revealed. With far off checking abilities, it becomes simpler to distinguish in case there are any deferrals because of unfriendly climate conditions or upkeep issues with trucks. Consequently, transporters can be observed successfully all through the store network.

We will presently clarify how improved course arranging utilizing AI and IoT can offer advantages to the coordinations organizations.

- Advanced Route Planning Using AI and IoT

  Burdens can be alloted for pickup to drivers dependent on their flow area. Drivers close by the pickup area are relegated assignments for getting products and drivers get warning about the advanced course to arrive at the distribution center. In view of the kind of merchandise to be gotten, for instance, transient or durable items, trucks and stockrooms are chosen for trade or conveyance.

  Trucks outfitted with IoT sensors give continuous investigation of course advancement, guaranteeing dependability and diminishing travel times. Data and cautions caught by sensors are shipped off coordinations specialist co-ops.

  Data given by IoT sensors include:

  1. Sort of Goods
  2. Ongoing area of the transporter
  3. Deviations from the arranged course
  4. Pickup/Delivery to Warehouses
  5. Temperature/Humidity.

  Carrying out AI on the assembled information can assist with anticipating the accompanying elements:

  1. Assessed conveyance time
  2. Driver conduct examination
  3. Nature of things, for instance, transitory products
  4. Contrasts between arranged course and the real course.

Catching and examining the information would help coordinations organizations cut down pointless expenses, work on an opportunity to get and convey products and give the enhanced courses. It becomes conceivable to see how the genuine way is unique in relation to the arranged one and what could be the purposes for deviations from the arranged course. Subsequently, AI and IoT joined gauges more advanced courses for conveyance and pickup later on.

With improved course arranging, it very well may be feasible to design travel and trade. Since stockrooms can get data about the drivers' assessed season of appearance, they can proficiently anticipate the travel of products. For instance, instruments and work needed for the dumping of products at the distribution center can be orchestrated ahead of time so that travel/trade should be possible rapidly.

2.3 Pickup Optimization

It is fundamental to upgrade the pickup of orders with the goal that items are not influenced under any conditions. Advancements, including AI and IoT, can guarantee streamlined pickup.

Trucks are chosen for pickup dependent on the sort of burden to be conveyed. Computer based intelligence learns recorded examples and permits transporters to settle on choices definitely. In light of sorts of products to be moved, an AI-based model predicts the right truck for getting merchandise. For instance, on the off chance that you need to get transitory products or delicate merchandise that require additional consideration, you can design the pickup of merchandise in IoT-empowered trucks.

IoT empowered trucks can accumulate the accompanying information about items:

- The temperature under which merchandise are put away
- Constant area of the item
- Moistness openness during transport
- Truck information, including speed, fuel costs.

The information is imparted to the coordinations organizations and item proprietors with the goal that they could screen the nature of items is kept up with all through the inventory network cycle.

2.4 Stockroom Management

- Truck and Warehouse Collaboration

1. Burden Preparation
   Artificial intelligence and IoT joined can work with transporter and stockroom joint effort by interfacing them. For instance, IoT sensors prepared in a truck

sends the constant area of the truck and its ETA to distribution center supervisors. Utilizing the data given by sensors, distribution center chiefs can keep the necessary space empty and get ready for the dumping of merchandise before time. It will assist distribution centers with dealing with their timetables productively and exactly. Load and dump readiness should be possible ideal with clever expectation utilizing AI. Thus, the sit tight an ideal opportunity for both coordinations organizations and stockroom administrators gets decreased.

2. Entryway Planning
   Sensors introduced around the distribution center region would send the data identified with entryways close to the empty space to transporters dependent on their GPS information and assessed season of appearance. It will assist with shipping drivers lessen an opportunity to track down the right way to get a fast section into the stockroom. In this way, AI and IoT additionally permit the problem free entryway wanting to work with better cooperation between coordinations organizations and distribution centers.

- Security and Compliance

It is fundamental to forestall drive-aways during the dumping/stacking of merchandise from the truck. To guarantee wellbeing, you can utilize IoT-empowered locks that work with interlocking of the trailer's compressed air brakes with the dock entryway. Brilliant locks guarantee that the truck can't withdraw until dumping/stacking gets finished. Trailers can possibly leave the stockroom when the dock entryway is shut. It can protect your hardware and representatives [10].
Another approach to empower the security of the stockroom is by utilizing PC vision. IoT cameras can be introduced in the stockroom region that catches the situation with stacking/dumping and burden move. Edge registering can be applied to the smart cameras that empower the vital activity dependent on any occasion got during load move action. For instance, if any delicate great gets separated during dumping because of the awful conduct of work, stockroom supervisors can get informed and move can be made against that individual.

2.5 Travel

Travel tasks can be changed utilizing IoT-empowered sensors and AI innovation. GPS sensors introduced in trucks give its constant area dependent on which AI model assesses the assessed season of appearance. IoT sensors prepared in trucks can catch data, including crash episodes, the temperature under which merchandise is kept and stickiness. In light of the continuous information assembled by IoT gadgets, coordinations organizations can follow travel tasks progressively.
Executing prescient investigation would assist organizations with understanding the distinction between anticipated time and the real time the truck takes to arrive at the objective. Clients and coordinations organizations can know whether the merchandise are conveyed under the right temperature conditions.

- Burden Exchange Optimization
  In the event that, the truck gets over-burden or meets with a mishap, transporters need to trust that quite a while will discover elective alternatives. In any case, IoT and AI can assist with lessening the stand by time by catching on going information and empowering wise activities. Artificial intelligence based models for load trade streamlining can be carried out to permit the fast trade of burden from one truck to another. For model, if a truck meets with a mishap, IoT sensors would catch this data and another vehicle will be appointed consequently to trade the heap and convey it to the last destination. GPS gadgets introduced in the truck give data, including scope and longitude, constant area of the vehicle and movement of the vehicle. Hence, transporters can rapidly decide the state of a vehicle during a mishap or mis-happening and courses of action for trade can be made likewise. That is the way arising advancements guarantee the consistent travel and conveyance of products.

2.6 Conveyance

The subsequent stage in the process is to get it done for end-clients. The transporter would go to the distribution center to convey products to end clients. Like the above interaction, trucks would be told about the empty entryway for getting products and stockroom administrators would be educated about the heap planning. The heap will be moved to the transporter under the total security utilizing IoT-empowered locks and PC vision-based cameras. When the stacking is done, the transporter would leave and the products will be conveyed to the clients.

2.7 BI and Reporting

When the pickup and conveyance of items are done, organizations would require an exhaustive report that would contain the negative and positive patterns of execution during the stockpile of merchandise.

Since organizations require granular straightforwardness into their transportation costs for overseeing and controlling them viably, the interest for business knowledge inside the coordinations and transportation space is soaring. They need to distinguish underlying drivers and dissect negative patterns in execution and cost to make smart moves. Business Intelligence permits changing over information into important data. Prior, announcing was simply restricted to extricating information, getting it from a framework and bringing it into an accounting page or data set where an organization would attempt to utilize it and convert it into helpful information.

Yet, these days, business knowledge has arrived at a higher level, where organizations can create important reports that exhibit every one of the information about coordinations suppliers in a scorecard design. Elements, remembering for time get and conveyance, limit responsibilities and driver conduct are relegated measurements that assist clients with deciding the exhibition of transporters. Likewise, administrators who require an every day and speedy outline of what's going on can utilize continuous dashboards that give ongoing data and assist clients with tackling issues as they happen. Dashboards offer organizations the upside of speedy response time as clients don't need to trust that somebody will make and send reports.

Organizations utilize BI and answering to show designs found in authentic information that can anticipate openings and future dangers in the store network or transportation organizations. How about we see how transporters can utilize business insight to further develop coordinations and inventory network tasks with a model. Assume one transporter has a 90% on-time conveyance rate for load move reliably, however he needs to go to the base of the issue hauling down the other a modest amount of shipment. Did delays happen in light of the messed up paths? Is there any issue with the hardware or transporter?

With business insight, it becomes conceivable to limit delay-causing factors. For instance, if the transporter came to late because of the clog in that manner, organizations can find the issue with business knowledge and make important moves. With BI information, it becomes simpler for coordinations organizations and their group to settle on functional choices all the more proficiently. Subsequently, executing BI across the shipments can bring about a start to finish pickup and conveyance time improvement.

3 IoT Situations in Transportation

IoT arrangements guarantee to make transportation associations more intelligent and more fruitful at what they do. The IoT is at the center of powers reshaping transportation to give more prominent security, more effective travel, further developed vehicle and airplane upkeep, what's more, more essential traffic the board. Instances of transportation IoT include:

- More productive, less exorbitant mass travel, that utilize organizations of sensors, computerized cameras, and correspondence frameworks to build framework limit furthermore, improve traveler wellbeing and solace while bringing down expenses and dangers.
- Dynamic side of the road message signs for shrewd transportation frameworks, which show constant street status, cost rates, path terminations and travel times consequently transferred from sensors and cameras.
- Autonomous vehicles, with the capacity to detect their climate, foresee conduct, speak with different vehicles and their environmental elements, and respond quickly to genuine interstate situations.

- Video observation arrangements, which include highresolution CCTV cameras to get air terminals and rail stations, including ceaseless observing of identification control designated spots and development of individuals furthermore, swarms. Insightful video investigation programming robotizes early location of dubious conduct furthermore, deserted gear.

3.1 IoT Availability Innovation Necessities in Transportation

Rising worldwide traffic and public vehicle fills development for the principle sections in transportation—associated transport and airplane. In spite of a low number of endpoints, vehicle endpoint arrangements represent most of market development because of the intricacy of individual arrangements.

Associated transportation involves use cases like driver connection, armada following and prescient upkeep planning to lessen vehicle vacation while expanding generally efficiency. Such applications have complex innovation prerequisites—solid inclusion in metropolitan and provincial regions, worldwide network crosscountry and cross-mainland, situating abilities for area following, a great of administration to guarantee that resources don't "go dull", and future-proofness of the innovation.

The last implies that the gadgets have SIMs and modules supporting over-the-air updates to stay away from the requirement for expensive trading out in the field. The headway of utilization cases that predict ongoing responsiveness and investigation will also strain transfer speed necessities. Thus, most of sent shipping arrangements will go to 4G/5G given low value affectability and adaptable energy proficiency necessities.

3.2 Advantages of IoT for Transportation

Some more extensive advantages that apply to the utilization of IoT innovation inside the transportation area include:

1. Improve Customer Experience
   IoT innovations help to furnish clients with more exact, state-of-the-art, continuous information to all the more likely arrangement travels and further develop correspondence.
2. Further Developed Safety
   The capacity to follow things, for example, train speeds, airplane part conditions, street temperatures and the quantity of vehicles at a convergence utilizing IoT empowered innovation would all be able to assist with working on the wellbeing of our travel frameworks around the world.

3. Functional Performance
   Transport Agencies embracing IoT advancements are now beginning to see benefits as far as functional execution. Urban areas can all the more likely screen basic foundations and foster proficient cycles to limit working expenses and further develop framework limit.
4. Ecological Improvements
   By better observing clog, IoT empowered frameworks can respond rapidly to developing traffic examples and return continuous information to assist individuals with arranging their excursions better. Diminishing blockage and energy use emphatically affect the climate.

3.3 Difficulties of IoT Organization

The IoT brings exceptional streams of information, introducing difficulties for organization and information the executives along with expanded security hazards. To address these issues, transportation specialists need to adjust customary network plans to give new levels of organization knowledge, mechanization what's more, security.

Transport associations need a cost effective organization foundation that safely handles immense progressions of information, and is likewise simple to oversee and work. The foundation must:

- Provide a basic, robotized measure for IoT gadget on boarding. Enormous IoT frameworks can contain a large number of gadgets or sensors, what's more, physically provisioning and dealing with these endpoints is intricate and blunder inclined. Computerized on boarding empowers the IoT stage to powerfully perceive gadgets and allot them to the suitable got network.
- Provide a safe climate against cyber attack and information misfortune. Since the numerous arranged gadgets and sensors in transportation IoT networks give a relating plenitude of potential assault vectors, security is basic for relieving dangers of cyber crime. Security is important at different levels, including control of the IoT networks themselves.
- Supply the right organization assets for the IoT framework to run appropriately and proficiently.

Numerous gadgets in the IoT framework convey strategic data that requires a particular degree of QoS. For example, some utilization cases require legitimate data transmission reservations on a superior organization foundation to guarantee administration unwavering quality.

3.4 Cybercrime

The development of IoT in transportation likewise brings a blast of network protection dangers, as the expansion of sensors and associated gadgets extraordinarily extends the organization assault surface. IoT is particularly vulnerable in light of the fact that numerous IoT gadgets are fabricated without security at the top of the priority list, or worked by organizations that try not to comprehend current security prerequisites. Thusly, IoT frameworks are progressively the failure point in transportation network security.

- The dispersed refusal of-administration assault on Dyn in October 2016 that cut down a significant part of the web was executed through hacked arranged gadgets such as surveillance cameras and computerized video recorders.
- Hackers assaulted the San Francisco Muni public travel framework network in November 2016, delivering ticket machines and other figuring foundation inoperable as part of a ransomware scheme.
- In 2016, Chinese security analysts took controller of a Tesla Model S from a distance of 12 miles, meddling with the vehicle's brakes, entryway locks, dashboard PC screen and other electronically controlled highlights in the electric car. The following year, similar gathering of programmers again assumed responsibility for the vehicle, regardless of Tesla's fixing of the underlying vulnerability.

3.5 Building a Safe IoT Network Foundation

Ensuring IoT traffic and gadgets is a test that can't be addressed by any single security innovation. It requires an essential methodology that exploits different security shields. To help associations exploit the advantages and lessen the dangers of IoT, Alcatel Lucent Venture (ALE) gives a staggered security system. Brew's system conveys security at each layer of the framework, from the individual client and gadget out to the organization layer itself.

It additionally gives an IoT regulation procedure to improve on gadget onboarding and convey the right network assets to run the framework appropriately and productively, all in a protected climate to shield transportation frameworks from digital assault.

3.5.1 IoT Control

To empower IOT control, all clients, gadgets and applications inside the ALE network are relegated profiles. These profiles, which characterize jobs, access approvals, QoS levels, and other approach data, are handed-off to all switches and passageways in the organization [11].

- Devices are put in “virtual holders” utilizing network virtualization strategies that permit numerous gadgets and organizations to utilize something very similar actual framework, while staying segregated from the remainder of the organization.
- In these virtual compartments, QoS and security rules are applied.
- By isolating the organization with virtual holders, on the off chance that a break happens in one piece of the virtual network, it doesn’t influence different applications.
- When another IoT gadget is associated, the organization consequently perceives its profile and appoints the gadget to the fitting virtual climate.
- Communication is restricted to the gadgets inside that climate and to the application in the server farm that controls these gadgets.
- Because all clients likewise include profiles inside the ALE network, admittance to the IoT virtual compartments can be restricted to approved people and gatherings.

3.5.2 Top to Bottom Security

Notwithstanding IoT regulation, ALE organizing advances give layered security across various levels of the organization.

- Secure enhanced code shields networks from inborn weaknesses, code misuses, inserted malware, and potential indirect accesses that could think twice about, switches and other strategic center and equipment.
- At the client level, profiles guarantee clients are verified also, approved with the proper access rights.
- At the gadget level, the organization guarantees that gadgets are verified and agreeable with set up security rules.
- At the application level, the organization builds up rules with respect to admittance to explicit applications, counting hindering, restricting transfer speed and controlling who can get to what.
- At the organization level, ALE switches and passageways offer brilliant investigation capacities that give perceivability furthermore, point by point data about the organization, clients, gadgets and applications being utilized on the organization.
- ALE smart investigation likewise give profound bundle assessment capacities, which can identify the kind of information and applications traveling through the organization, making it conceivable to recognize uncommon organization traffic designs furthermore, unapproved movement and organization interruptions.

4 Case Study of Internet of Things Solutions in Transportation

4.1 INRIX

Inrix (Kirkland, Washington) breaks down information from street sensors and vehicles to give constant leaving and traffic data just as experiences that are utilized to all the more securely test and convey self-driving vehicles. Furthermore, the organization's Population Analytics administration utilizes GPS and versatile organization information to respond to inquiries regarding travel propensities and populace thickness.

Inrix as of late added 100,000 new parking areas to its data set. As per a report, the organization currently has information on in excess of 212,000 stopping areas in 15,140 urban communities across 88 nations.

4.2 CHARIOT

Ford-possessed Chariot (Austin, Texas) plans to further develop the public travel insight and simplicity gridlock by giving a superior transportation choice to suburbanites, venture arrangements and contracts. As its statement of purpose proclaims, "When the world sudden spikes in demand for more intelligent courses, lower costs and better ride encounters, we'll all things considered take vehicles off the street and change your twice-day by day disappointment into a piece of your day you really anticipate." Chariot as of late moved its development center to big business clients as opposed to the buyer market.

4.3 CONCIRRUS

Concirrus' (London, England) Quest Motor assists guarantors with seeing and oversee hazard through an information investigating application that gives them continuous bits of knowledge into driver conduct, including factors like speed, slowing down, closely following and recurrence of late evening driving. Other than information separated by the organization's exclusive application, other carefully passed on data—traffic designs, neighborhood climate, impact information—is additionally considered. The organization's Quest Marine answer for transportation dissects things like vessel measurements, developments, nearby climate, apparatus data and more to give new bits of knowledge and rating factors.

Concirrus as of late inked a multi-year marine circulation manage protection specialist and hazard the board organization Marsh to utilize Concirrus' Quest Marine stage.

4.4 Dash

Dash innovation works through versatile application to further develop eco-friendliness, pinpoint important fixes, anticipate support and screen driving propensities. It can even assist with figuring out what precisely that "check motor" ready means and find the right specialist.

Dash (New York) was remembered for Fast Company's rundown of World's Most Innovative Companies 2018.

4.5 FLASHPARKING

Functions of FlashParking's (Austin, Texas) versatile connected innovation for parcels, carports and valet tasks incorporate touch screens and smart stations, alongside cloud-run programming. For example, its valet administration includes an installment and recovery stand so visitors can pay ahead of time and solicitation vehicles, enormous screen screens that communicates vehicle recovery status to visitors, vehicle and staff following to monitor left vehicles and valet staff, key following and the sky is the limit from there. There's likewise an intelligent vehicle outline so visitors can increase any spaces of harm they notice.

FlashParking as of late started a joint test case program with parking space reservation organization SpotHero in Austin, Texas.

4.6 FYBR

Fybr's smart (St. Louis, Missouri) leaving meters utilize subterranean sensors that decide if a vehicle is left in some random spot, alongside that spot's accurate area. The driver would then be able to pay for the spot by means of the organization's cell phone application. In the event that the designated time lapses and isn't broadened, an alarm is shipped off city stopping implementation work force.

Fybr collaborated with SMC Labs to utilize IoT innovation for stopping, water system the executives, air quality and resource the board in two Silicon Valley "advancement zones."

4.7 G.E. TRANSPORTATION

G.E. Transportation rail industry (Chicago, Illinois) administrations incorporate IoT empowered availability, ongoing condition checking, prescient investigation and that's only the tip of the iceberg. The organization's innovation, e.g., "clever journey control", likewise boosts train lengths, further develop taking care of and diminish fuel utilization. BNSF Railway is teaming up with GE Transportation to construct an all-battery electric train.

4.8 MAERSK

Maersk's (Copenhagen, Denmark) far off containing the executives program utilizes interior sensors to assemble and communicate real-time information on everything from temperature and moistness to CO_2 levels. It additionally works with ongoing day in and day out GPS following of compartments, gives programmed notices that keep load proprietors mindful of any deviations in temperature or pull down rates, empowers freight rerouting and further develops security. Maersk as of late presented its own moment booking affirmation administration to make saving a holder much the same as booking a flight ticket.

4.9 MIAMI INTERNATIONAL AIRPORT

The Miami air terminal's (Miami, Florida) immense organization of guides interface with a versatile application so individuals can do everything from get headings and flight updates to shop and sweep tickets. When it started its program in 2014, Miami International was the main air terminal on the planet to have total and open guide sending. After four years the training is significantly more typical at air terminals and different scenes.

4.10 MOTOLINGO

MotoLingo's (Tulsa, Oklahoma) telematics innovation plays out an assortment of security and proficiency capacities, remembering recording driving time for a client's cell phone, changing speed increase and slowing down from constant input, figuring hazards through cell phone GPS and the sky is the limit from there. Discretionary elements incorporate driver scores dependent on factors like speed cutoff points and traffic lights, and symptomatic OBD equipment. The organization likewise offers an application called MotoCarma for youngsters who are currently getting their student's grant.

4.11 NEXT TRUCKING

Connecting drivers with transporters, NEXT's (Los Angeles, California) administrations incorporate ongoing following of shipments and postpone warnings; support from a devoted in-house account supervisor; transporter task by means of portable application; map pinpointing of transporters; coordinating with accessible delivery loads and that's just the beginning. Thanks to a $21 million money imbuement, in the principal half of 2018 NEXT added six new jobs to its chief supervisory crew and is currently filling in excess of 40 different situations in designing, item advancement, advertising and deals.

4.12 SHIPPABO

Shippabo (Los Angeles, California) gives IoT-empowered store network the board arrangements, including request the executives, customs bits of knowledge, shipment following, SKU level item perceivability, mechanized warnings and that's only the tip of the iceberg. According to a Forbes article in mid 2018, Shippabo has shown promising development, raising $2.8 million dollars. In its subsequent year, it significantly increased deals to $3 million and added 17 workers.

4.13 TERBINE

Terbine's (Las Vegas, Nevada) information commercial center incorporates transferred IoT data from an assortment of sensor-prepared areas, including the cultivating, delivery and traffic businesses. It's twirling with computerized insights regarding what's going on in reality, and those subtleties can be bought by invested individuals. The organization intends to make machine-produced information indexable, controllable and quickly pertinent for business employments. Terbine as of late dispatched another IoT Data Exchange to enormously support "the sharing of machine-created information between organizations, public offices and scholarly foundations."

4.14 VENIAM

Veniam's (Mountain View, California) cloud stage handles an enormous deluge of information that empowers a wide range of vehicles to speak with one another in a huge metropolitan biological system of moving things. Per its own depiction, the organization transforms vehicles into Wi-Fi areas of interest and constructs networks that extend remote inclusion and gather heaps of city information. Veniam CTO Rui

Costa as of late revealed to Forbes that by 2025, vehicles will create 10 exabytes [10 million terabytes] of information.

4.15 SEPTA: Positive Train Control (PTC)

The Southern Pennsylvania Transportation Authority (SEPTA) gives light rail, tram and transport administration to more than ten lakhs riders every day in and around Philadelphia. SEPTA was one of the early travel frameworks to assemble their Positive Train Control (PTC) establishment, a complex train-flagging framework intended to forestall accidents, crashes and track laborer wounds coming about because of speed and sign infringement. SEPTA worked with Digi to send the right availability answer for PTC.

4.15.1 The Digi WR44-RR Portable Access Switch

The Digi WR44 RR is the vital correspondences center in all trains and vehicles, handing-off PTC information messages to and from waysides through 220 MHz radio and empowering distant framework support, arrangement and organization the executives over a phone connect. Expanded organization unwavering quality and rail framework perceivability broadens execution past PTC toward Communications-Based Train Control (CBTC), bringing about more effective planning, more noteworthy limit and expanded fuel reserve funds.

4.16 TransData: Passenger Ticketing and Information System

Train passenger Systems integrators in the IoT space have huge freedom today to help the requirements of associations across the vehicle area, from city transport and light rail organizations to shipping organizations in the store network, and significant distance traveller trains. The necessities are developing as these organizations work to meet consistence prerequisites and contend in their commercial centers by offering upgraded types of assistance and further developed security.

TransData is an IoT frameworks integrator that creates applications for public travel, like installment and IDs frameworks, for the Slovak market. TransDatas leader item is a complex arrangement that upholds an expansive scope of public vehicle abilities:

- Secure toll exchanges
- Simple to-utilize electronic card framework improves on traveler encounters
- GPS-followed course direction limits delays
- Show nearby shops, eateries and focal points

- More solid Internet access and rapid traveler Wi-Fi
- Screen traffic action with on-vehicle surveillance cameras
- Course correspondences through a focal warehouse or dispatch.

The applications above are empowered by Digi ConnectCore six super minimal framework on-module (SOM), which upholds TransData network prerequisites at a reasonable value point. TransData tagging and data frameworks require unrivaled video execution, Wi-Fi and Bluetooth, and network to the vehicle information framework and cell modem. IoT applications in transportation additionally require a steady bundle and little structure factor that can withstand tough conditions like outrageous warmth, stickiness and vibration while keeping up with network availability to play out these perplexing undertakings.

4.17 SMART: Public Transit System Computer-Aided Dispatch

The Suburban Mobility Authority for Rapid Transit (SMART) metro transport armada of 330 biodiesel and half and half electric transports covers in excess of 1100 miles and supports 32,000 riders every day. With this broad armada, it is basic to screen the vehicles to guarantee the most elevated levels of traveller wellbeing and on-time execution.

The business issue to tackle included the update of a maturing CAD/AVL (Computer-Aided Dispatch/Automatic Vehicle Location) framework based on an inheritance simple radio organization associated by means of three rented towers. smart originally assessed relocating from simple to advanced signals and expanding the quantity of pinnacles, yet that was cost-restrictive. At last SMART changed to VOIP on cell for CAD, exploiting the bundle need administrations incorporated into the Digi WR44 R versatile cell switch.

With its change to a cell based AVL, SMART can gather and dissect a lot more extensive scope of information and measurements—including vehicle area and speed—progressively. Upkeep information is additionally caught to assist with forestalling breakdowns and speed up fix cycles, to limit vehicle personal time. Information is sent to the travel tasks focus through a profoundly secure VPN burrow, while administrators can speak with Central Dispatch utilizing VoIP handsets.

Because of these updates and upgrades, the SMART initiative appraisals they are saving more than $70,000 every year.

4.18 Macchina: Auto Control Center

Macchina and Digi devicesMacchina worked with Digi to create a moderate 4G LTE arrangement with a little impression. The group picked the Digi XBee Cellular installed modem dependent on its plan—an open source interface for vehicle specialists and experts to program a gadget or administration into the car secondary selling. Our vision is to offer a one-to-many interface, clarified Josh Sharpe, Macchina boss specialized official. In the information base world, you may call this middleware. The gadget creator will actually want to make one gadget with one interface to our board and we handle incorporation to many vehicles.

The item accordingly empowers engineers to Another approach to consider Macchina is that its like a key to open the control focus of the vehicle. When you are in, you can utilize Macchina to make changes and changes to the vehicle. You can do anything from straightforward ventures, such as halting that irritating ding, to more perplexing redesigns, for example, opening more pull or further developing fuel economy.

Macchina basically gives a task layout that empowers improvement and supports development. Designers like the open source stage to compose code, just as a local area of vehicle specialists, aficionados and experts to talk with as they investigate different ways to deal with their own item advancement.

5 AI Based Intelligence and IoT

In the early long stretches of the twentieth century, the presentation of the auto changed society in manners that came to a long ways past the straightforward exchanging of a pony drawn vehicle for a Model T. Recently secluded networks were associated by new streets, makers could get their products to more business sectors, ventures were made, and openings opened up for a recently portable labor force. As feature writer George F. Will put it, “In an evident flash, mankind from what has been known as ‘the abuse of distance’.” Today, mechanical and social powers such electric and self-governing vehicles and shared versatility are merging to introduce an insurgency that is similarly as groundbreaking, if not more so.

The web of things (IoT), with its capacity to interface edge sensors and information authorities to the cloud, is vital to putting self-ruling (self-driving) and associated vehicles in the city. Filled by man-made consciousness (AI), these IoT-driven vehicles support smart urban communities and a reasonable future by lessening costs and ecological effect and further developing wellbeing and traffic stream. They likewise support new transportation-as-a-administration plans of action, for example, the undeniably well known ridesharing and ride-hailing administrations.

New advances are likewise being utilized for network safety in associated vehicles. The consistent progression of information from drivers, vehicles, producers, administration focuses, and local area foundation should be shielded from noxious abuse. Only a couple years prior, for instance, analysts had the option to remotely hack the dashboard PCs in some Chrysler vehicles, taking over the dashboard capacities as well as guiding, transmission, and brakes.

As an ever increasing number of associated vehicles arise onto the public streets, the volumes of information they will produce and burn-through will become monstrous. That is the place where blockchain will join IoT and AI and convey imaginative answers for portability. Secure by plan, blockchain is alter safe and is customized to ensure the huge volumes of information that associated vehicles will deliver, while guaranteeing straightforwardness and exactness. A blockchain arrangement upheld by network safety skill can ensure both vehicle frameworks and individual data from being hacked [12].

The abilities of these new advancements go past improving and securing the driving experience. As numerous producers have learned, AI and IoT can proactively anticipate mechanical disappointments and breakdowns, setting aside cash and forestall mishaps. Indeed, by 2025, McKinsey gauges that prescient upkeep will save makers 240 to 627 billion annually.

These equivalent advantages can apply to vehicle vendors, mechanics, and proprietors too. For instance, an associated, IoT-based vehicle could recognize an issue before quick consideration is required. It could mention to its administrator what it needs, request the part, make an arrangement for itself at the closest assistance place, and afterward—if the vehicle is independent—drive itself there. This situation might be nearer than you may might suspect, and its foundations are as of now grounded in prescient upkeep.

Prescient upkeep arrangements use examination and AI to analyze likely issues before they put vehicles down and out, permitting proprietors and drivers to set aside cash and stay away from the burden of unexpected fixes. Also, AI-based prescient upkeep makes vehicles more secure, making drivers aware of expected issues before they become dangerous.

As the excursion to the eventual fate of portability unfurls, further advances will arise to drive it forward. It will be interesting to observe how networks, organizations, and society explore this parkway to more astute and more secure urban communities.

6 Applications AI-IoT in Transportation

Examiners foresee that by 2020, 75% of new vehicles will include IoT network. The rate increment portrays purchaser applications, yet associated vehicles should earn revenue from different areas like delivery, coordinations, and transportation [13].

Pioneers in these businesses would be insightful to get ready for a future where AI and the IoT change transportation the executives.

Here are five potential applications to consider.

6.1 Smooth Out Decision-Making

David Poulsen, CutCableToday’s IT master, says associated, or self-sufficient, vehicles, are appealing a direct result of the advancements that undergird them. “The Internet of Things (IoT) is one piece of the condition,” Poulsen clarifies. “The other part is computerized reasoning (AI). It goes about as the driver, helping the associated 'thing,' which could be a vehicle or stock framework, settle on more brilliant choices."

As applied to transportation the executives, that mechanized dynamic capacity is basic. Associated vehicles, shipments, and frameworks assist with following and chronicled detailing. Yet, ongoing experiences and reactions happen through manmade reasoning.

At TOPBOTS, an on the web, instructive asset for everything AI, essayist Mariya Yao calls the cycle "transforming store network coordinations into computerized exchanging." She gives a model: Amazon's capacity to convey bundles to an individual's entryway in less than two hours. Simulated intelligence and the IoT smooth out the whole interaction, from request to conveyance, to set aside time and cash and fulfill client need.

6.2 Upgrade Operations

DHL, the worldwide coordinations supplier, sets another utilization of AI and the IoT streamlining. Its 2016 Logistics Trend Radar report recommends that large information and robotized supply chains could prompt beforehand unfathomable degrees of streamlining [14].

In any case, that streamlining isn't disengaged to a solitary part of transportation the executives. Maybe, DHL predicts a world where fabricating, coordinations, warehousing, and conveyances become progressively proficient, useful, and productive. The supplier accepts the pattern will become animated in the following ten years.

DHL could be right. General Electric, for instance, has begun coordinating AI into its trains to improve wellbeing and speed. Daniel Malak at Motionloft offers another utilization for AI and the IoT: enhancing traffic. He says transportation the executives organizations profit with Motionloft by utilizing it to examine traffic designs and "upgrade strategic approaches, for example, conveying police powers just during top busy times, having support groups fix streets that get the most travel, and sending disinfection teams to clean open regions just when required."

6.3 Oversee Warehouses

Tim Young of Vero Solutions shares another way AI and the IoT Could change transportation the executives in his infographic taking a gander at stockrooms. He says AI could affect six spaces of tasks.

"Usefulness levels, stock cycles, and worker compensation are only three fields," clarifies Young, "that are relied upon to be upset and improved by AI innovation in distribution centers in simply a question of years."

The other three regions identify with powerful correspondence, distribution center tasks, and robot laborers. Youthful's illustration of robot laborers includes an organization recently referenced: Amazon. The brand has been trying out robots in its distribution centers to build efficiency and, apparently, quality control.

6.4 Lessening Downtime and Repairs

Transportation organizations additionally use AI and the IoT to alleviate expensive fixes and personal time. Interior diagnostics, for instance, can make clients aware of support issues, which guards travelers—no victories while going not too far off at seventy miles each hour, for instance—and expands the lifetime worth of the vehicle.

Daniel Dombach at Zebra further enlightens the idea, adding that the Internet of Things conveys distant checking capacities. Organizations that utilize them can proactively react to upkeep issues and furthermore survey stock records and parts accessibility.

Dombach additionally makes a substantial statement, saying that AI and the IoT could "decline protection related expenses." Business Insider's The Insurance and the IoT Report finds that safety net providers use vehicle utilization information to advise evaluating on approaches and charges. The report covers customer protection strategies explicitly, yet its discoveries effectively mean business interests.

6.5 Go Driverless

Man-made intelligence and the IoT could affect more than back-end frameworks and cycles. The two could deliver driverless vehicles, a thing apparently the domain of tech monster Google. However, Google isn't the only one in the undertaking. Tesla, Ford, Daimler, and even Uber all case driverless drives.

George Zarkadakis at Willis Towers Watson gets down on the Uber story in his article The Impact of Artificial Intelligence in Transportation, refering to the occurrence as "a reminder." He proceeds, "Computerized reasoning (AI) and AI (ML) might actually prompt the full robotization of truck armadas."

Obviously, Zarkadakis' comment brings up the issue of what befalls the transporters. Goldman Sachs Economics Research gives an answer [15]. The organization reveals to CNBC that driverless trucks could deliver employment misfortunes of 25,000 every month in years and years.

Jack Stewart at WIRED offers a more inspirational outlook; he says customary driver occupations will change once self-governing vehicles become a reality, yet these positions will not really vanish. He likewise adds other constructive outcomes of this change, like reducing expenses and further developing street security.

7 Artificial Intelligence Answers for Intelligent Transportation

The commitment of AI to the field of transport industry has been monstrous and broad. The arrangements incorporate self-governing vehicles, traffic the executives, enhanced directing, and coordinations subsequently giving wellbeing of vehicles and drivers. ITS are fabricated utilizing the information produced from the gadgets introduced in the vehicles through AI advancements.

The current examination centers around four sub-frameworks identified with transportation—to be specific, Intelligent Traffic Management System, Intelligent Public Transport System, Intelligent Safety Management System, and Intelligent Manufacturing and Logistics System. Tables 1, 2, 3 and 4 portray the information hotspots for AI arrangements, related issues in the sub-framework, job of AI and the advantages accomplished.

From Table 1, we see that AI gives answers for transportation issues by recommending elective courses, ongoing following of traffic signals during gridlock. This assists with overseeing traffic in a productive way in the end prompting checking of natural contamination and building manageable urban areas.

From Table 2, we see that AI gives arrangements on foreseeing climate and traffic designs, street the board, ready age to officials on the job. These frameworks help drivers, workers and walkers before the beginning of their excursion. It is important to have the help of innovation to fabricate a proficient public vehicle framework that aides in arranging and the dynamic cycle.

From Table 3, we see that AI has diminished the quantity of mishaps on streets, predicts mishaps dependent out and about conditions, cautioning drivers towards street security and so on An economy runs effectively when the vehicle business is proficient. It is important to accomplish something very similar by building safe vehicle framework with the assistance of AI advancements.

From Table 4, we see that auto industry is profited by AI arrangements during the assembling interaction of vehicles. Sensors, cameras and different advances play had an impact in this industry for better advantages. A portion of the in-assembled AI arrangements in a car have become fundamental segments in traveler vehicle just as business vehicle sections.

8 Simulated Intelligence Achievements in Transportation Across the Globe

As seen in the conversations up until this point, the ability of AI to tackle issues identified with transportation is by all accounts a characteristic fit. However, just like the case with AI in each and every other industry, the reception of these applications fluctuates across associations and topographies. In light of the natural and geographic

Table 1 Smart traffic management systems

Wellspring of information	Issues	Job of AI	Advantage	Past studies
Vehicles with intelligent frameworks	Expanded expense because of gridlock	AI instruments to anticipate traffic stack up	Better fuel saving capacity and lesser contamination to climate	Momentary gridlock expectation by assessing traffic boundaries accomplished utilizing ML models
Information from PDAs	Directing	Elective course ideas	Efficient	Driver conduct checking frameworks through information produced from advanced cells use ML strategies
Astute vehicle frameworks	Unusual gridlock	ID of dirtying substances in air	Controling of ecological contamination	Various air quality files are joined utilizing fluffy rationale alongside recreated strengthening and molecule swarm improvement method to distinguish air contamination
Traffic signals and vehicles	Pinnacle hour traffic the board	Realtime following of blockage and calculations in traffic signals	Control of higher and lower traffic designs	Continuous data assembled from traffic signals are noticed for ideal green-red circulation before AI arrangements are sent for investigation
Information from vehicles	Expansion in the quantity of vehicles out and about	Example distinguishing proof	Better perception and dynamic	The strength of AI strategies, explicitly ANN is sent to foresee gridlock in heterogenous rush hour gridlock conditions

Table 2 Public transport system

Wellspring of data	Issues	Role of AI	Benefit	Previous studies
Developed designs, street surfaces, climate and traffic patterns	Variability in the data	Prediction of varieties in the examples through AI algorithms	Planning and decision making	Short-term gridlock expectation done utilizing traffic volume, thickness, inhabitance, travel time, clog file
Constant information from drivers and passengers	Traffic congestion	Optimization of routes	Shortens the hour of movement	–
Man-made intelligence controlled vehicles for merchandise delivery	Variation in conveyance time, place suggestions to further develop driving patterns	Improved efficiency and further sales	The most ideal conveyance course is shown up at utilizing	The Vehicle Routing Optimization to apply prescient knowledge in street transport
Sensors from brilliant roads	Wear and tear of the road	Automatic ready age to officers	Road management	A maintainable ITS is accomplished with the coordination of sensor innovation with transportation foundation guaranteeing vehicle and traveler security

elements, the applications could be both straight forward and muddled, far off and not far off, positive or plausible.

8.1 Man-Made Intelligence Applications Across Associations

Utilizations of AI in different associations in the transportation area is given in Table 5. US is by all accounts front sprinter in these applications. This is presumably because of lesser populace and better street foundation when contrasted with agricultural nations like India. New businesses which are imaginative get acceptable measure of financing to foster models in created nations. A large portion of

Table 3 Insightful safety management system

Wellspring of information	Issues	Job of AI	Advantage	Past examinations
Sensors from intelligent vehicles	Weakness and sluggishness of drivers	Auto-pilot framework actuation	Stay away from mishaps	Different incorporated sensors in a self-governing vehicle decides the security and achievability
Significant distance trucks	Persistent driving hours and obscure territory	Wellbeing observing of drivers	Forecast of mishaps	Continuous estimation of physiological boundaries of drivers are taken care of to web cloud and examined utilizing AI utilizing smart in-vehicle wellbeing checking frameworks
Self-driving vehicles	Low execution and security issues	Vulnerable side ready, versatile voyage control, progressed driver help frameworks	Saves season of drivers	Self-driving vehicles guarantee less exertion and speculation towards security techniques for drivers
Constant information transmission	Expanded time and cost	Streamlining of courses	Expectation procedures to estimate vehicle volume	Self-sufficient vehicles secure ongoing and exact information on vehicle position and state prompting better vehicle taking care of and wellbeing
Checking through sensors	Fix or refueling	Controller the board	Saving of fuel, further develop mileage	Insightful visual labels introduced on vehicles give portability backing and following system

Table 4 Smart manufacturing and logistics system

Wellspring of data	Issues	Role of AI	Benefit	Previous studies
Clever vehicles	Need for maintenance	Combining information from IoT sensors, upkeep logs expectation models are created	Better forecast and machine failure	Reduced cost and further developed availability to low-class populace through independent vehicles
Associated vehicles	Repairs and maintenance	Connected vehicles booking prescient and preventive maintenance	Empowerment of vehicle observing businesses	Connected traveler vehicles are superior to physically determined vehicles in the event that they work dependably with better UIs
Vehicles fitted with technologies	Increase underway and conveyance cost	Shared information across vehicles and routes	Improved cost reserve funds across the whole inventory network, going from acquisition to research and development	C-ITS—Cooperative ITS give ongoing hand crafted data to explicit drivers
Organization based structure	Large number of solicitations because of manual information entry	AI based frameworks recover information easily from the network	Faster handling of bills, invoices	Smart telephone connected home to vehicle associated vehicles to direct tedious assignments
Solicitations and documents	Anomalies in solicitations, consistence verification	Prediction and handling of extortion detection	High level of exactness	–
Contracts	Extracting information which isn't structured	Natural language handling innovations for understanding of invoices	Extraction of basic data	–

the arrangements are tested during significant distance driving when contrasted with traveler vehicle section.

8.2 Reception of AI by Transport Organizations

AI is probably going to decidedly affect city foundation by giving precise prescient conduct models of person's developments, their inclinations and their objectives. However AI in transportation arranging applications have gotten critical in the new past, there is a worry of security and wellbeing of people identified with information. There is plausible of government and lawful guidelines directing the speed at which advancement and reception happens in this industry because of these moral contemplations.

Without moral agreement on numerous parts of innovations, singular associations who are on an AI venture should factor moral contemplations. However couple of associations use machines to compose code, overall people keep on composing it. Because of this factor inclinations, presumptions, discernments might discover their direction into the calculations being created. Associations should address themselves: What is moral AI? Where do administration and moral AI cross-over? How to wipe out inclination in AI dynamic? and so forth. Because of this perspective, there is a variety in the reception of AI by different governments and city organizations.

Reception of AI by different vehicle enterprises and its advantages are given in Table 6. Because of the impact of the nearby unofficial laws, the reception is by all accounts changing across different urban communities and provinces of India.

9 Conclusion

Transport, Logistics, and Fleet administration organizations use IoT innovation not exclusively to build effectiveness and efficiency of tasks yet in addition to improve the manner in which activities happen. The Internet of Things will in a general sense change the manner in which load and product are followed while eliminating arrangement cost and lessening arrangement cost. The broad utilization of advanced mechanics in huge distribution centers, and to deal with unsafe merchandise is now far reaching, just like the driverless, remote-controlled vehicles.

The Internet of Things doesn't simply illuminate a solitary use case or drive. It impacts each part of transportation and coordinations and can upset the manners by which organizations move payload starting with one spot then onto the next. In spite of the fact that idea chiefs don't have a clue how the IoT will examine 50 years, current investigation proposes that the upcoming delivery and getting cycles will work definitely more productively than they do today.

The transportation area is continually developing to offer more secure, quicker, cleaner and more agreeable drives. The following critical industry shift is within

Table 5 Man-made intelligence achievements across the globe in transportation

AI applications	Organization	Country
A self-driving, intellectual electric transport—Olli, transports travelers to mentioned area and gives ideas on neighborhood touring. Olli is controlled by IBM's Watson Internet of Things (IoT) for Automotive	Local Motors	United States
Surtrac framework was introduced in an organization of nine traffic lights and it anticipated and identify auto collisions and conditions by changing over traffic sensors into astute agents	Rapid stream technologies	Pittsburgh, United States
Otto finished the world's first self-ruling truck conveyance conveying 50,000 jars of Budweiser brew for over a distance of 120 min	Otto (Uber)	San Francisco, United States
TuSimple, a Chinese beginning up finished 200 miles of driverless truck drive. The driving framework was prepared utilizing profound learning techniques	TuSimple	United States
GE's wise cargo trains furnished with sensors distinguishes things close by the track. There is a 25% decrease in train disappointment rates	GE transportation	Germany
In-house AI innovation of Hitachi decreased the force devoured in driving moving stock. Right mix of functional information extricated from the moving stock saw 20% decrease in yearly footing power	Hitachi	Japan
The Department of Transportation expects AI upgraded request and conjecture displaying in street cargo transportation management	DoT	United States
On-time conveyance of individuals and bundles through self-sufficient transports despite non consistency in climate designs, traffic designs, city infrastructure	–	Finland, Singapore, China

Table 6 Reception of AI by transport enterprises

State Transport Corporation	AI application	Benefits
Bangalore Metropolitan Transport Corporation	AI cameras GPS trackers Facial Recognition	To screen driver conduct identified with rest because of exhaust and speeding
Karnataka State Transport Corporation	Sensors fitted at the front guard of the transport where the driver waves at it each 3–4 min	Sensor cuts off gas pedal if the driver doesn't wave at the sensor significant distance extravagance transports
Metropolitan Transport Corporation (Chennai)	Intelligent Traffic the board system	Automatic number plate acknowledgment cameras fueled by OCR peruses criminal traffic offenses. Programmed age of challan for installment of a fine which is shipped off the violator
Uttar Pradesh State Transport Corporation	Anti-impact system	Continuous observing of driver for objects inside 180 m reach by signaling
Maharashtra Transport Corporation	IVADO and Next AI Canadian companies	Set up AI bunches for different ventures incorporating transportation—investment in R and D, Technologies for transport
Telangana Transport Corporation	Chatbots for client support	AI addresses numerous inquiries. Troublesome inquiry is sent to higher specialists
West Bengal Transport Corporation	Patha Disha—AI app	Availability of seats on explicit transports, assessed appearance season of transports. Following conduct—commuter criticism and conduct
Toronto Transit Commission	Self-driving travel shuttle	Supervised by human drivers at first. A drive to settle last mile network to public vehicle
French National Railway Company	Chatbots for travel passengers	Helps explorers plan their every day trip and explore across the city in case of inescapable postponements
Street and Transport Authority, Dubai	Smart and maintainable transportation utilizing AI—Automated Bus Track Control System, Smart passerby signal system	Monitoring the state of transports—driver weariness, reconnaissance cameras across vehicles. Improvement of transport proficiency, sensors to apportion person on foot signal intersection
Service of Transport, Singapore	nuTonomy—a self-driving vehicle organization banding together with grab to make self-governing taxi	Self-driving transports and cargo vehicles to affect public vehicle
Transport for London	Sopra Steria gives admittance to data	Road traffic, transport execution, climate and street attempts to decrease clog and street the executives

reach, and IoT is driving the charge. The capability of IoT is prodding a rush of shrewd vehicles and associated framework.

The worldwide savvy transportation market could reach $262 billion by 2025, because of the worth of IoT in vehicles. In any case, the advantages don't end at monetary accomplishment for vehicle makers. It's improving practically every part of the business.

Most travel industry enhancement issues advantage from profound bits of knowledge, and the precept that unrivaled information drives better travel results is in no way, shape or form antagonistic. For instance, the mind-boggling larger part of urban areas request inside and out testing of independent vehicles prior to allowing them to travel their roads. Public transportation following further develops administration by diminishing stand by times at tram stages and transport terminals to build ridership. Armada directors track everything from fuel utilization to routinely arranged support calls to get the leap on hardware that will unavoidably separate.

Utilization of computerized drives assumes a significant part in the profoundly cutthroat coordinations industry. Web of Things innovation further develops effectiveness and straightforwardness over the entire transportation cycle, keeping the activities run as expected and consistently. Progressed armada the executives, stock following, natural checking, just as danger avoidance are the most productive IoT arrangements in Logistics Tech.

Also, today, the coordinations business encounters an inescapable pattern towards uberization of transportation administrations. "Uber for Heavy Equipment" permits to effectively organize transportation of large equipment and, with an IoT framework, track transportation cycles progressively.

Acknowledgements This work was sponsored by SERB, the service of the Department of Science and Technology (DST), Government of India, Project "SWARD -Secure next-generation Wireless Access RaDio technology for smart cities in India" (award number EEQ/2018/001482).

References

1. Iyer LS (2021) AI enabled applications towards intelligent transportation. Transp Eng 100083
2. Jwaid MF, Juboori HKS (2021) Vehicles for open-pit mining with smart scheduling system for transportation based on 5G. Turk J Comput Math Educ (TURCOMAT) 12(5):827–835
3. Derawi M, Dalveren Y, Cheikh FA (2020) Internet-of-things-based smart transportation systems for safer roads. In: 2020 IEEE 6th world forum on internet of things (WF-IoT). IEEE
4. Fernando WKAUK, Samarakkody RM, Halgamuge MN (2020) Smart transportation tracking systems based on the internet of things vision. In: Connected vehicles in the internet of things. Springer, Cham, pp 143–166
5. Yang H et al (2021) Supporting transportation system management and operations using internet of things technology
6. Suresh A, Udendhran R, Balamurugan M (2020) Internet of things based solutions and applications for urban planning and smart city transportation. In: Internet of things in smart technologies for sustainable urban development. Springer, Cham, pp 43–62

7. Velasco-Hernandez G et al (2020) Intersection management systems and internet of things: a review. In: 2020 IEEE 16th international conference on intelligent computer communication and processing (ICCP). IEEE
8. Ekanayake LJ et al (2019) Smart protector: a real-time theft prevention system for transportation management. In: 2019 14th conference on industrial and information systems (ICIIS). IEEE
9. Manoj Kumar N, Dash A (2017) Internet of things: an opportunity for transportation and logistics. In: Proceedings of the international conference on inventive computing and informatics (ICICI 2017)
10. Mohanta BK et al (2020) Survey on IoT security: challenges and solution using machine learning, artificial intelligence and blockchain technology. Internet Things 11:100227
11. Alcatel Lucent Enterprise (2018) The internet of things in transportation
12. Bhushan B et al (2021) Unification of blockchain and internet of things (BIoT): requirements, working model, challenges and future directions. Wireless Netw 27(1):55–90
13. Milić DC, Tolić IH, Peko M (2020) Internet of things (IoT) solutions in smart transportation management. Bus Logist Mod Manag
14. Sankar KM, Booba B (2020) The usage of internet of things in transportation and logistic industry. In: Intelligent computing and innovation on data science. Springer, Singapore, pp 431–438
15. Husniah H et al (2020) The economic aspects of internet of things for a public transport system. In: PROCEEDING MICEB (Mulawarman international conference on economics and business), vol 2

A Novel GAN-Based System for Time Series Generation: Application to Autonomous Vehicles Scenarios Generation

Samy Kerboua-Benlarbi, Mallek Mziou-Sallami, and Abdelkrim Doufene

Abstract Adversarial models have been widely used for image automatic generation, with several recent models taking into account the real data manifold coverage. However, there are still remaining challenges to generate time series data due to the complexity of their invariant characteristics. In this work, we propose a novel GAN-based system for time series generation. We design a novel representation of multivariate time series, that enables the use of image-based Generative Adversarial Networks. To assess the feasibility of our method, we apply it to generate various autonomous driving scenarios, towards a fully-automatic framework of self-driving testing. To quantitatively evaluate its efficiency, we conduct an empirical study on different GAN architectures. For each model, we compare the manifold of generated data with the one from real data, using a coverage metric based on persistent homology. The comparison results underline the great interest of gradient penalty and the consistency term in the case of WGANs and prove the ability to generate realistic driving scenarios using the proposed representation of multivariate time series.

1 Introduction

Time series analysis and study constantly benefit from the advancements of machine learning, and especially deep learning. In many contexts, data diversity is an asset during training and more. Indeed, capturing interesting patterns may help to make models much more robust to the task for which they were designed. A first bottleneck is to be found in the fact that these data have to describe as many situations as possible,

S. Kerboua-Benlarbi · A. Doufene
IRT SystemX, Palaiseau, France
e-mail: samy@kerben.fr

A. Doufene
e-mail: abdelkrim.doufene@irt-systemx.fr

M. Mziou-Sallami (✉)
CEA, Evry, France
e-mail: Mallek.mziou@cea.fr; Mallek.mziou-sallami@irt-systemx.fr

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_16

but it’s practically impossible to gather all considered situations. In order to overstep this problem, one may generate new time series, corresponding to what concerned models would need to achieve their goals. In fact, the choice of a generation method directly depends on what definition could be used to represent those time series. In this work, time series will be represented as channels of an image that constitute a description of each mentioned situations. Our work and future ones will benefit from it as it is both precise and general enough to consider all possible models and all possible learning paradigms.

The search for a representation that effectively helps time series maintain the meaning of the situations they describe is essential. A well-defined coverage of all situations would imply an understanding of the relevant data and an effective way to generate them. Thus, the notion of manifold is appropriate because we theoretically acknowledge that high dimensional data distributions, such as multivariate time series ones, are lying on lower dimension manifolds [1]. Computing the real shape of a manifold is not trivial, but we can model its behavior thanks to a specific family of neural networks frameworks, called Generative Adversarial Networks. So, the objective is to use a model from the former and generate data implicitly along the manifold, on which every realistic data points are closely lying. Once a generation process is build, a proper evaluation of its capacities needs to be done. Topological Data Analysis and persistent homology will then be used to identify quantitatively how efficient our method is.

One field in particular could use this idea: autonomous driving. Indeed, autonomous vehicles need various and numerous data in order to capture interesting behaviors, or patterns in a more general way, that will help them drive safely and efficiently. An autonomous driving scenario, e.g. a situation in which a driving system would be engaged, could be several numerical sequences, turned into channels of an image. Each pixel would describe a parameter of the environment at each time step, such as speed or steering. Here, it is important to notice that a scenario is not a function of the common perception that is usually described for autonomous systems inputs (cameras, ...). A model would then use such depiction as inputs for training, evaluation or even certification at the same time.

Using these ideas, our first contribution consists in proposing a novel generic representation of time series as multi-channel images. Secondly, this new representation was integrated in a GAN-based protocol for self-driving scenarios generation. Such system is a first step towards a fully-automatic autonomous vehicle testing system. Despite the fact that the generation of driving scenarios is not common to the best of our knowledge, it is not new [2, 3]. These previous studies tend to focus on the formulation and generation of test scenarios from a formal context. GANs then take on their importance since they have been used in the context of autonomous cars, either as a driving training tool [4, 5], for data augmentation via domain adaptation and transfer [6] or even in the context of behavioral cloning and reinforcement learning [7]. The generation of scenarios via Bayesian optimization and expected improvement [8] is also an example of related work. However, most of these methods have had the idea of using natural images (identifiable in the real world), while this is not what we will working with. Indeed, we are not using perception as in [9] but indicators that are not images. Most of all, these methods have been used for perception, while this work focuses on the process of decision making in [10]. GANs then remained an interesting choice with regard to the state of the art. The remaining of this paper is organized as follows. In Sect. 2, a brief survey of Generative adversarial networks is given with a focus on time series generation and evaluation metrics. After that, we will present our approach for scenarios processing and time series encoding as multi-channel images. Section 4 is devoted to the experimental setup. In Sect. 5, we will analyze our generation results independently, first in terms of learning dynamics from as far as stability and equilibrium are concerned, and then in terms of driving data generation in a quantitative and qualitative manner. Finally, we conclude and present some perspectives for future work in Sect. 6.

2 Related Work

2.1 Generative Adversarial Networks

Generative adversarial networks or *GANs* [11] emerge from game theory. Indeed, they are composed of two entities: a generator G and a discriminator D. The former tries to fool D by generating realistic samples, while the latter always tries to be able to distinguish G's creations from real examples (Fig. 1).

Given $\mathbb{P}_r$ the real data distribution and $\mathbb{P}_g$ the generated one, this *min-max* game attempts to find a Nash Equilibrium by alternated optimization, given the below expression as a cost function:

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{z\sim\mathbb{P}_g}[\log(1 - D(z))]$$

Another formulation was introduced to overcome gradient issues arising from the initial one. The generator's update rule then went from $\mathbb{E}_{z\sim\mathbb{P}_g}[\log(1 - D(z))]$ to $-\mathbb{E}_{z\sim\mathbb{P}_g}[\log(D(z))]$. Nevertheless, numerous problems remain [12], such as the *mode collapse* or the theoretical assured existence of a "perfect" discriminator which could separate the supports of the two data distributions. Several works brought methods to

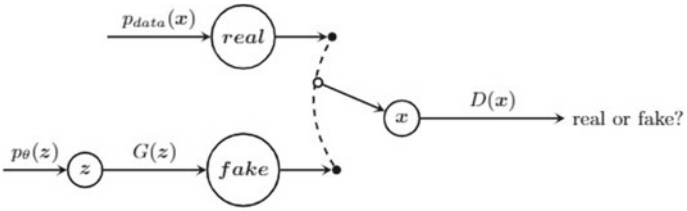


Fig. 1 The adversarial framework for data generation

improve the learning process of GANs [13, 14]. A big enhancement was made with the use of the Wasserstein distance which gave the Wasserstein GANs or *WGANs* [15], thanks to the Kantorovich-Rubinstein duality [16]. The cost function is then similar to the one observable in vanilla GANs, except that we do not talk about D as a discriminator anymore, but as a "critic". In fact, a WGAN does not minimize a Jensen-Shannon divergence [11], but learns a real function instead. It is worth noticing that it is often needed to iterate several times over the discriminator before iterating over the generator, in order to ensure an efficient sharing of information between the two parts of the framework. It holds to the below expression as a cost function, with $\mathbb{D}$ the set of all 1-Lipschitz functions and $\mathbb{P}_g$ the generated distribution:

$$\min_{G} \max_{D \in \mathbb{D}} \mathbb{E}_{x \sim \mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{z \sim \mathbb{P}_g}[\log(1 - D(z))]$$

To make WGANs operational, we need to enforce 1-Lipschitz continuity on the discriminator. Consequently, the first works clip the gradients at each learning iteration. Yet, this process is considered hazardous, which steered research towards the development of a gradient penalty [17]. Indeed, this penalty allows the enforcement of the 1-Lipschitz continuity by interpolating points on lines between the real data distribution and the generated data one. However, this term takes in account really specific examples, leaving several zones of the real data manifold unexplored. In addition, interpolated data might be far from both manifolds during the first learning iterations and nothing insures that the two distributions will come up to each other enough so that the 1-Lipschitz continuity exists. A consistency term [18] was proposed, as an extension of the gradient penalty discussed above. This term leads to a better exploration of the real data manifold, and enforces Lipschitz continuity to the real data surroundings: these regularizations theoretically improve the stability and convergence of GANs, and also the generation itself. It is worth notice that gradient penalty has not been used only for WGANs [19].

There are many different cost functions for many different models including Bayesian formulations [8] or a third network [20], using a genetic approach [21] or even using Variational auto-encoders [22]. However, a recent work [23] has shown that the choice of a cost function does not matter compared to optimizing hyperparameters. Empirical studies have validated these elements and gone further by showing that beyond the choice of model parameters, regularization and normalization play a real role in the stability of learning and the quality of the images generated [24]. By regularization, we mean any constraint placed on the cost function or any formulation that regulates the standard of gradients [17, 18, 25, 26]; normalization concerns the weights of the network [27–29]. Other studies have brought a new formalization of GANs to show their difficulty in converging regardless of the chosen cost function [30], or have precisely proven interesting convergence properties by applying the gradient penalty to a classic GAN [19], but also by analyzing the gradient descent in GANs [31]. Several very recent models have surpassed all the others thanks to the attention mechanism [32] or by scaling GANs [33].

2.1.1 GANs with Time Series

From recurrent neural networks and its enhancements, sequences can then be generated: indeed, since the prediction of the next action is at the heart of these networks, it is possible in the case of natural language for example, to sample the beginning of a sentence and to ensure that words are predicted one by one. However, this task is not trivial and can lead to disappointing results [34]. It would then be interesting to consider GAN as a "framework" before being a model: by this we mean that a GAN consists of a generator and a discriminator which can be convolution networks, but theoretically any other type of architecture, continuing whatever it manages to explore the said manifolds. With the help of previous recurrent networks, it is possible to design GANs capable of generating sequences, taking care to adapt the principle of recurrence to the cost functions stated so far. The non-differentiability of the choice of each word after the softmax poses nevertheless a problem with regard to the backpropagation required to train the discriminator. Several studies have tried to overcome this problem, using methods from reinforcement learning [35], using a different formulation of the softmax function [36], or using the capabilities of auto-encoders to sample not words but entire sentences on the explored manifold [37]. Some work, particularly in the generation of continuous sequences (music, medical field, ...) attest the possibility of using recurrent networks as they stand, by using linear layers at the output of a recurrent layer to be able to apply a softmax more easily [38, 39].

Thus, RNNs and other variants can be applied to GANs [40], but it should be noted that it is currently theoretically impossible to use Wasserstein GANs directly [39], due to the latter's formulation (the conditions on the formulation of the Critic no longer work). We will also note the use of recurrent GANs in word sequence generation without pre-training [41] which is an interesting idea, since the discriminator in these models generally tends to become too strong too quickly [38]. The handling of missing values in time series also uses GANs on some applications [42] and there exist several recent works in the use of GANs for time series modeling [43], financial strategies fine-tuning [44]. A recent work [45] uses temporal convolutional networks in GANs for time series generation.

2.2 Evaluation of GANs

Many metrics exist for the evaluation of GANs. More specifically, since most of the work concerns the generation of natural images, the metrics seek to attest to both the quality of the images generated and their diversity. For example, we can note the *Inception Score* (IS) [14] which aims to verify if the images are varied and if they look like something real using a pre-trained classifier on a natural image dataset. The *Fréchet Inception Distance* [19] aims to improve the Inception Score by calculating the distance between two multivariate Gaussian configured by the outputs of the Inception Network on real and generated data. Other methods use a trained auxiliary classifier to distinguish generated images from false ones [46, 47]. We can also train a network on real data and measure its performance on generated data, or the opposite [39].

Other methods do not involve training from another network, such as Maximum Mean Discrepancy (MMD) [48] which assesses the dissimilarity between two data distributions. Some variants exist, using kernels [49]. One should notice that the discrepancies that GANs use can also be used as metrics [50]. Choosing a metric depends on what you want to measure, because there are no metrics that can detect all the phenomena inherent in the generation. The characteristics most often sought are over-fitting, the quality of the data generated or sensitivity to spatial distortions. Some works list all the metrics with their strengths and weaknesses [51], while others try to understand which ones not to use, such as Parzen estimators or log-likelihood on GANs [52]. There is still a need in a metric that can approximate the intrinsic geometry of the data to think in terms of coverage.

2.3 Evaluation by Simplicial Homology

Simplicial homology establishes its origins in topological data analysis. Its strength lies in the possibility of analyzing multidimensional data, which corresponds perfectly to our case. Topological data analysis or *TDA* seeks to understand the structure and the geometric nature of data. In this sense, it is able to capture invariant characteristics that reflect the reality of datasets. The principle of persistent homology is then an algebraic method to extract information about the overall structure of data, such as “holes”, that could indicate in which areas they are concentrated. In other words, it makes it possible to know the areas of space where little or no data is present. This approximation is done by connecting each group of close points to represent them as multidimensional graphs called *simplicial complexes*. The filtration operation allows to build a set of simplicial complexes, by adjusting an ϵ meant to be the radius of each ball produced around each sample. Let’s assume $\mathbb{X}$ a set and a metric d. The different filters are based on the use of d to test whether several points belong to the same component. For each value that ϵ can take, we define the homology of the scatter plot on which we work, where the number of holes evolves according to the evolution of ϵ. Once filtration is obtained, persistent homology can be used to attest to the “lifetime” of observed holes: it is the calculation of the *persistence bars* [53]. Thus, we have a measure of the evolution of the approximate shape of manifolds (Fig. 2).

There are several ways to build simple complexes such as *witness filtration* or *Rips Filtration*. The *Geometry score* [55] is a metric that uses results of persistent homology and more specifically persistence barcodes. By looking at the statistics on its components, we can know how long each of them lived over the period of the topological study: by this, we mean that each number of holes is weighed by the maximum value of ϵ, which gives the relative lifetime of this number of holes. This is a confidence measure since a number of holes that lasts for a while (in terms of ϵ evolution) indicates that this number reflects the shape of the data manifold. Its com-

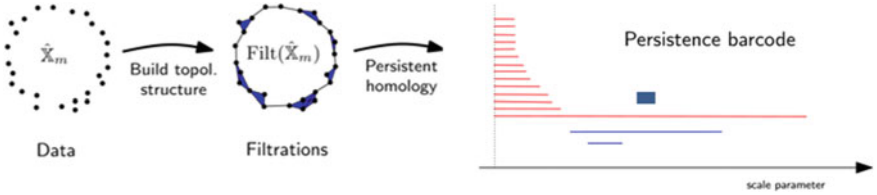


Fig. 2 "Pipeline" for the study of simplicial complexes [54]

putation requires repeating the experiment several times, and the values obtained for each relative lifetime are averaged, giving what is call their $Mean\ Relative\ Living\ Time$, or $MRLT$. For two datasets X and Z, the geometry score is then calculated using the quadratic error between the two $MRLTs$ of each distribution. So the lower the geometry score is, the more it reflects the proximity to the manifold. Indeed, this would mean that the number of holes does not decrease. In short, this metric and the TDA make it possible to evaluate GAN-based system in terms of generation coverage and model collapse, regardless of the size and dimension of the data used.

3 Approach

3.1 Image-Based Generation

The usefulness of time series in the context of autonomous driving is clearly apparent. Indeed, the definition of a driving scenario implies taking into account the environment around the autonomous vehicle called *ego* over a certain period of time; therefore, our data are time-dependent signals. It is a succession of points, multivariate series (as many variables as the input of the model), in a space of several dimensions. For example, a point would characterize the curvature of the road, the speed and the steering angle of the vehicle. Each point is dependent on previous ones, as it could be in the case of trajectories, or financial values. We therefore need to build generative adversarial models that can handle temporal sequences, multivariate in our case.

Recurrent neural networks already showed their ability to generate sequences, but our goal is to analyze the behavior of generative adversarial networks when it comes to captures the intrinsic information of multivariate time series. In other words, we answered specific questions: Do image-based adversarial generative models are able to generate high dimensional and complex time series? And if so, are they able to reach a sensible coverage of the situations they describe? Although some GAN formulations [35–37] have allowed the generation of sequences, while others have succeeded in integrating recurrent neural networks as they stand in a classic GAN-type model [38, 39], theoretical gaps were observed making their exploration uncertain. However, "image-based" gan generation, including DCGANs (convolutions) [56] had many improvements studied by empirical [24] and theoretical [19, 57] studies, that have made their learning easier (stability, convergence, etc.). The state of the art had already highlighted the greatest importance of normalization processes (batch, per layer or *layer*, spectral or *spectral*) and regularization processes (*Gradient Penalty* with [17, 19] and *Consistency Term* with [18]), when faced with the choice of a specific cost function (GAN, WGAN, LSGAN, ...) [23, 24].

Using image-based GANs with our representation, that would be obtained by converting time series data into images, was therefore a required step. By doing so, we would then be able to use proven image-based methods for time series generation. Applied to autonomous driving, It would produce a novel protocol for efficient generation, with a good coverage.

In our empirical study, many elements were taken into account (normalization of networks, regularization of cost functions, hyperparameters, use of the *Two Times Update Rule* or TTUR). It should be noted that the TTUR is a rule which is intended to accelerate the convergence of models, and reduce the difference between the number of discriminator and generator iterations (From 10 to 1, to 1 to 1 for example). In fact, it is done by using two different learning steps for the model networks [57]. This would give an overview of the devised models and allow to conclude on the effectiveness and usefulness of our protocol for driving data generation.

3.2 Time Series Processing

Once each component of scenarios were known as univariate time series, it was therefore necessary to consider a modelling that could be appropriate for the use of GANs. The theoretical foundations of some transformations [58] could lead to a loss of information. This problem would have made it tedious to achieve the objectives of this work. To generalize the idea of a fine discretization of periodic univariate series to grayscale images [59], it was decided to propose a new method for non-periodic multivariate ones. A similar case consists in the expansion of periodic functions with the Fourier series which was generalized for non-periodic functions. Such generalization is possible in the distribution theory. The process is not afflicted by the absence of periodicity, and it raises a question on GANs capacity to handle even more complex signals. The question raised by curves to surface representation of an object leads to a second similar problem. In fact, different types of representations for curves and surfaces are common in computer graphics and geometric design. One can convert images from Cartesian to polar coordinates [60, 61]. It is possible to express planar curves with explicit representation $y = f(x)$ or parametric representation $x = x(t)$ and $y = y(t)$. The latter is typically the image of the planar curves [62]. Note that the information of the object form remains invariant. Hence, the idea of multivariate series coding as an image is relevant. Usually, images are encoded on 3 channels (RGB) to represent the color spectrum. Some representations exist where more channels can be reached (multiple satellites, ...). It is interesting to consider one channel of a future image for each univariate series of a scenario. Computing the average over all channels and converting the resulting values into a single greyscale image was not an option. Indeed, this would not only imply a huge loss of information across all sequences of the scenario, but also destabilize the reversal process. Instead, a transformation of each univariate series into an image is performed, and each representation is concatenated to the others, to form a multi-channel image. For example, if a driving scenario is composed of four different series (speed, road orientation, ...), each of them would be transformed into a greyscale image that would then constitute a channel in our final representation, leading to a 4-channel image. On the other hand, RGB can be used, which then gives a 12-channel image. The precision retained after the transformation depends only on the representation we choose: indeed, the finer the representation is, the more efficient the discretization is, but the higher the storage costs. It is moreover the RGB representation that was chosen guaranteeing an efficient and relatively inexpensive storage of information.

Series discretization corresponds to a division of the amplitude range into 2^{24} intervals, between the minimum and maximum values of each series over the entire dataset to prevent bias, assigning to each point the index of the interval in which it's lying, i.e. a value ranging from 0 to 16777215, which is converted to 24 bits binary integers. The previous operation, called quantization, then allows building triplets of values t, where $\forall v \in t, v \in [0, 256]$. This is done by shifting and isolating the 24 bits in groups of 8. The last step of the conversion is a simple *resizing* of the encoded series to images of size 64×64 (Fig. 3).

One may consider the fact that this resizing might induce a loss of information; but intuitively, we're going from the time series domain to the images one, which turns the temporal correlations to spatial ones. Indeed, a small loss could be observed, but since the information is kept spatially between the "end" (last pixel) and the

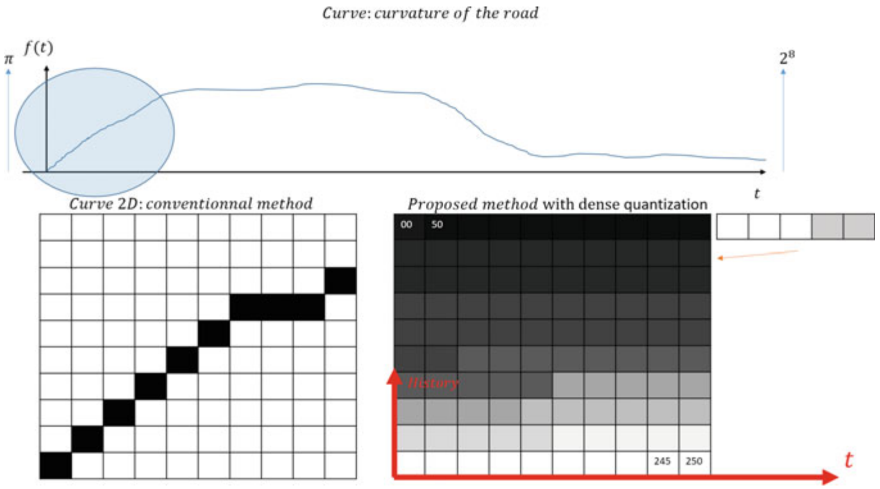

Curve: curvature of the road
π
f(t)
2^8
t
Curve 2D: conventionnal method
Proposed method with dense quantization
00
50
History
245
250
t

Fig. 3 Within our method, we go beyond the classical view that considers the portion of each univariate series as 2D curve in binary image. We are able to have RGB or grayscale image in which each line represents a window on the series, the latter being represented entirely as the segments stack up. Such representation preserves more details

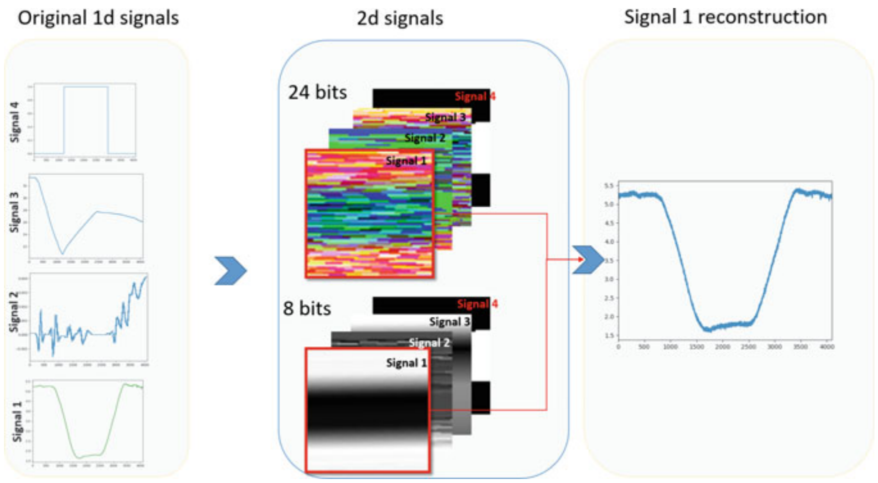


Fig. 4 Complete process of time series to image conversion

"beginning" (first pixel) of each consecutive lines on the image, we assume that it is not going to reach a problematic threshold. Moreover, our scenarios are series whose periodicity is non-existent, unlike the method in [59]. Indeed, if we applied convolution without a sliding window, we would loop back to the beginning of the series and have redundancy of information. Several filters were tested (such as Hamming or Blackman), taking care to effectively study which ones could have the least impact in terms of information loss, with minimal loss being inevitable. Finally, despite convincing results, a Fast Fourier Transform was used, taking advantage of a threshold in the frequency domain. Figure 4 and Algorithm 1 summarize the process for the RGB representation.

Algorithm 1 Encode a driving scenario

```
procedure PROCEDURE ENCODE_SCENARIO
Input: $S \in [A, B]^k$ A, B min and max for each feature, $k \in \mathbb{N}^*$; $n \geq 8$ bits
    $S_T$ : List
    for $f \in S$ do
        Discretize $[A, B]$ in $2^n$ intervals $I_t$, $t \in [0, 2^n]$
        Quantify : $\forall i \in f, i = t$ where $i \in I_t$
        if $n > 8$ then
            Cut in groups of 8 bits to obtain a value for each channel
        Reshape each remained vector $V_f$ as a squared image
        $S_T$ : Append $V_f$
    Return $S_T$
```

One shall convert scenarios to different image sizes by using another sampling rate, the duration of scenarios enforcing a specific size. Images generated by GANs are usually noisy. Removing noise from generated images as it stands would cause us to lose information. So, the images were first reconstructed in series, and denoised using a low-pass filter.

3.3 Data Collection

All the experiments were conducted using data from a driving simulator provided by the project partner *Groupe PSA*. The latter was designed to offer a working tool for anyone interested in the theme of autonomous cars, since it is a realistic simulator including data collection tools and a well provided native API.

The simulator was then able to collect data in line with our definition of a scenario in a supervised context. Then, a study of the generation of these scenarios was possible. The frequency at which the data were collected would induce univariate series of the same length. In our simulator, the data recovery frequency was about 256 Hz, over 16-seconds scenarios, with 4096 recovered values ($4096 = 64 \times 64$). Because of background functions calls in the simulations, this frequency rate was not fixed and small errors intervals were observed, even if it is not a real issue. All in all, 4096 values were needed to create an image of size 64×64, and 5375 scenarios were then gathered for training.

4 Experimental Setup

An empirical study was conducted to evaluate the generation of driving scenarios. A DCGAN architecture was maintained throughout the study to allow an effective comparison of cost functions, but also because it is the architecture known to be the best at the image generation level (CNN for discriminator D and generator G), better than multi-layer perceptrons. The trained models correspond to the GANs and WGANs, to which the same regularization procedures with specific harshness (gradient penalty and consistency term) and the same forms of normalization (batch, layer, spectral) were applied. Based on theoretical explorations, normalizations were applied either to both networks, or put at a specific position in the case of the spectral one. The ADAM optimizer was the one of all models except the WGAN which uses RMSProp, and all hyperparameters are described in Tables 1 and 2. All inputs are normalized between -1 and 1 to prevent gradient problems. A nomenclature was designed to effectively identify all 108 models for the future. The latter corresponds to a sequence of several acronyms:

Table 1 Hyperparameters for models not using the Two Times Update Rule

	GAN	GAN-GP	GAN-GP-CT	WGAN	WGAN-GP	WGAN-GP-CT
α_g	0.0002	0.0001	0.0002	0.00005	0.00001	0.0002
α_d	0.0002	0.0001	0.0002	0.00005	0.00001	0.0002
β_1	0.5	0.5	0.5	×	0.5	0.5
β_2	0.999	0.999	0.999	×	0.999	0.999
n_{disc_iters}	1:1	1:1	1:1	5:1	5:1	5:1
λ_1	×	10	10	×	10	10
λ_2	×	×	2	×	2	2
M	×	×	0.2	×	×	0.2

Table 2 Hyperparameters for models using the *Two Times Update Rule*

	GAN	GAN-GP	GAN-GP-CT	WGAN	WGAN-GP	WGAN-GP-CT
α_g	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
α_d	0.0004	0.0002	0.0002	0.0002	0.0002	0.0002
β_1	0.5	0.5	0.5	×	0.5	0.5
β_2	0.999	0.999	0.999	×	0.999	0.999
n_{disc_iters}	1:1	1:1	1:1	2:1	2:1	2:1
λ_1	×	10	10	×	10	10
λ_2	×	×	2	×	×	2
M	×	×	0.2	×	×	0.2

1. the type of GAN, i.e. 6 possible acronyms (GAN, GANGP for *Gradient Penalty*, GANGPCT for *Gradient Penalty* and *Consistency term*, WGAN, WGANGP, WGANGPCT)
2. the type of normalization (if dropout, always at 0.2):
   - **R** for a normalisation batch in D and G, **Rd** for a dropout addition in D
   - **LN** for a normalisation layer in D and G, **LnD** for a dropout addition in D
   - **RSN** for spectral normalisation only in D (batch normalisation in G), **RSnD** for dropout addition in D
   - **SN** for a spectral normalisation in D and G, **SnD** for a dropout addition in *D*, **SnDB** for a batch normalisation addition after the spectral normalisation in G [32].
3. **FC** for "Fully convolutional"
4. **2D** for the image data format
5. **TTUR** if two times update rule added (two different learning steps for D and G optimizers).

The hyperparameters defined for each model are described in these two tables. α and β correspond to the parameters of the generator and the discriminator/critic optimizers. n_{disc_iters} is represented as n:1 where n is the number of iterations of the discriminator before updating the generator. λ_1 is the parameter of the gradient penalty, λ_2 and M are those of the consistency term.

5 Results

In this section, we will analyze our results, first in terms of learning dynamics, and then in terms of driving data generation, to study the adequacy of the proposed representation for the image-based GAN. By learning dynamics, we refer to learning stability like learning curve magnitudes, and we consider a model convergence when a state of equilibrium is reached i.e. when both learned curves stabilize in zones of fixed magnitudes.

5.1 Stability and Equilibrium

We need to qualitatively analyze the impact of each type of normalization on learning. Our main objectives focus on which normalization methods improve stability and convergence, and whether one of them exceeds the use of regularizations.

We can see that the learning dynamics of GANs, like WGANs, are far from stable in all cases (Figs. 5, 6, 7 and 8). By applying only Batch Normalization (BN), stability is far from being achieved with oscillations the amplitudes of which are extremely high. This is also the case when applying layer normalization (LN). The interest of this method lies in the formulation of the gradient penalty which, in order to preserve the properties on which it is based, prefers its use in the discriminator/critic to batch normalization. GANs do not benefit from easier convergence and the equilibrium state is then less distinct, this phenomenon being observed just as much for WGANs. The layer normalization even seems to increase the oscillations of the generator and the discriminator, compared to the batch normalization. However, this last point must be qualified because the respective oscillations follow generally the same order of magnitude. Using LN either on GANs and WGANs is not theoretically expected to bring about major changes, since the WGAN's formulation, although different from the GAN's one, is not subject to any condition on global normalization. In other words, for the unregularized GANs and WGANs, the batch normalization as well as the layer normalization should not be considered as a source of learning stability, even if an equilibrium seems to be reached (for WGAN, we have a divergence at the beginning that reaches a balance with strong spaced oscillations).

By applying a gradient penalty (GP), learning becomes more stable. If we look at the two previous normalizations, the learning of GANGP is not optimized by a BN. On the other hand, the LN allows a clearer balance to be achieved (the two curves are

Fig. 5 GAN with BN in D and G, TTUR

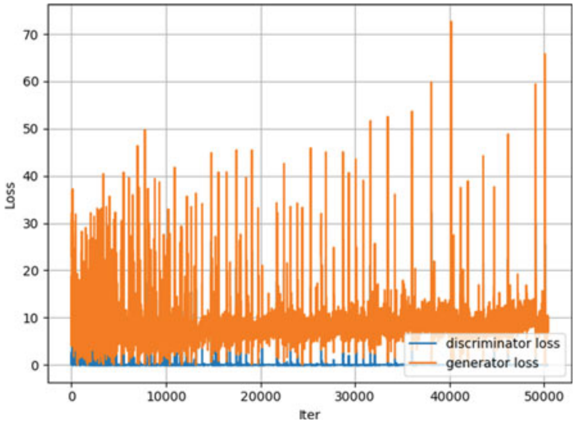


Fig. 6 GAN with LN in D and G, TTUR

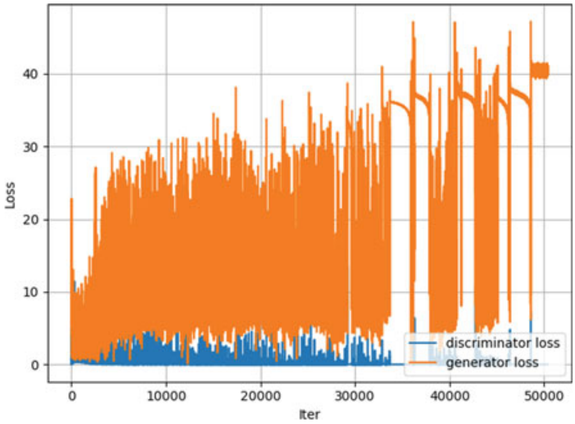
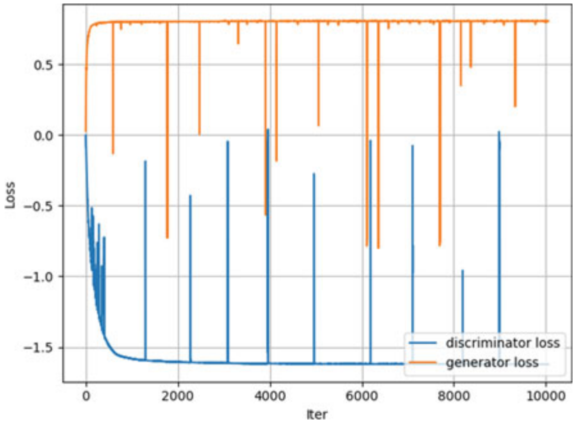


Fig. 7 WGAN with BN in D and G



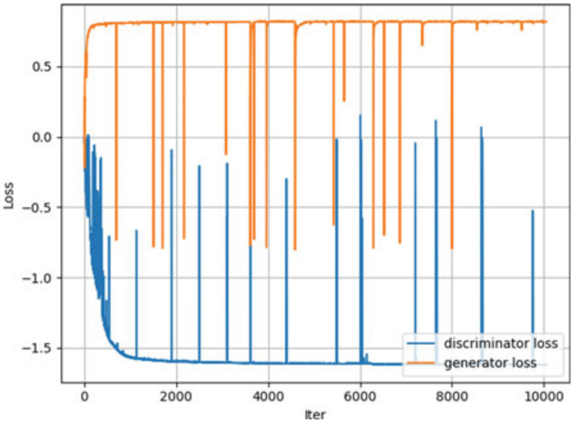


Fig. 8 WGAN with LN in D and G

“one on top of the other”) despite the oscillations, which is logical given its usefulness in this situation. For the WGANGP, the results do not support this assumption, since the curves for the BN and the LN are almost identical. The hypothesis would be that convergence improvements come only from the GP and that LN produces a better generation since it guarantees the validity of the interpolation performed by GP. It should be noted that for LN and BN, the WGAN losses reach very high values. However, the loss of the latter is supposed to be correlated to the quality of the elements generated [15]: this would mean that the model should produce incorrect elements with such behavior. We will see that this conjecture is not proven in our case. Thus, BN and LN do not have a significant impact, and it is worth noting that the GP provides an effective first response to the issue of stability and convergence, especially for the WGANGP. If we apply the consistency term (CT) in addition to the GP, to both GANS and WGANs, we observe that the LN produces the same results as for the GP alone. Knowing that by cross-checking the learning results of a GP with BN, the consistency term is more helpful in achieving a state of equilibrium. Stability appears to be the same as for GP and GPCT models. In this sense, the CT does not seem to bring more than GP to the point of learning itself. GPs and CTs therefore have an impact on the convergence process and stability, but their real appeal would probably lie in the quality of the elements generated.

The most interesting results are achieved by the use of spectral normalization (SN) (Figs. 9, 10, 11 and 12). In terms of stability and convergence, the latter greatly stabilizes the entire process and makes it possible to achieve a clear equilibrium, regardless of the model chosen and regardless of the placement of the normalization (in D and/or in G, followed by a batch normalization in the generator, ...). It should also be noted that coupled with a WGANGP with dropout, it provides exceptional stability.

Let us note two other interesting results, the addition of dropout almost always seems to reduce oscillations of the curves, thus improving stability. The TTUR proves what it was introduced for, by facilitating convergence and, above all, by accelerating

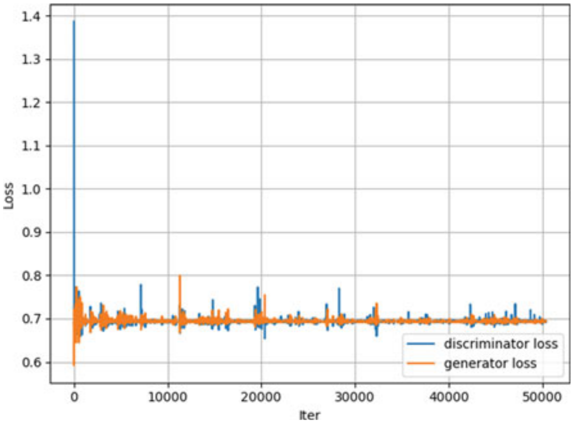


Fig. 9 GAN with SN in D and G

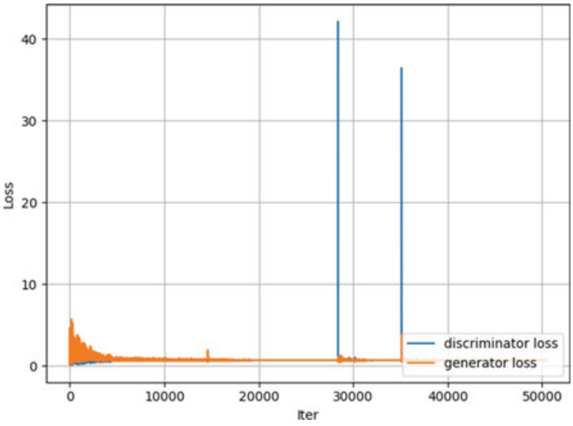


Fig. 10 GANGP with SN in D and G

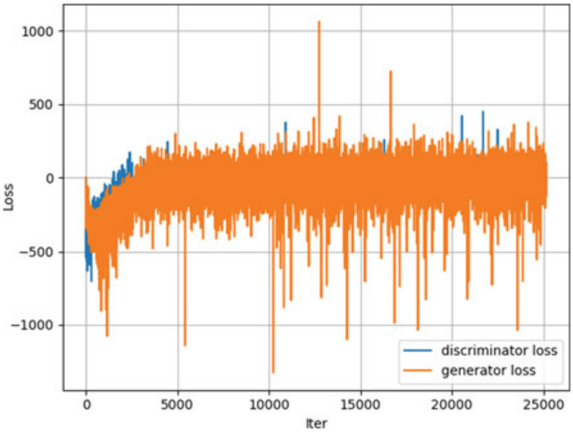


Fig. 11 WGAN with SN in D, and TTUR

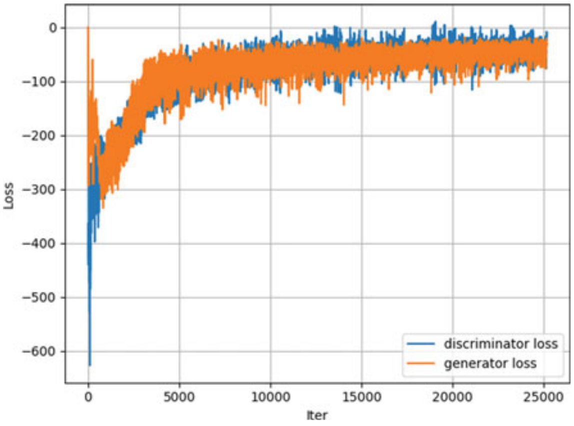


Fig. 12 WGANGPCT with SN, dropout in D, and TTUR

it. In conclusion, it is certain that spectral normalization exceeds both regularizations, and this is even truer when combined with the latter. But in addition to the powerful aspect of SN, it would be interesting to observe GP and CT as terms that influence generation quality more than stability and convergence, again given our data. It is of course necessary to look at the generated series and the geometry score to be able to really determine if spectral normalization brings a real plus when it is alone, and if there is a link between adversarial learning dynamics and generated data in the context of scenarios generation.

5.2 Generation Quality

The BN and the LN do not bring anything significant in terms of convergence and stability. We observe empirically the same phenomenon regarding the quality of generated data. Indeed, the series derived from GAN models that use these normalizations do not succeed in correctly capturing the entire data distribution and some series are therefore closer to hazardous oscillations than those of interest (Figs. 13, 14, 15 and 16). It should be noted, however, that the series generated by GANs using the LN or the BN are sometimes realistic. In a way, specific measures could be considered as simpler univariate series, leading to a much better understanding of them by GANs; more complex series are then too difficult to capture. The same idea can be found for the WGANs, which are supposed to be better: not all series are correctly captured, but some are extremely well captured. One might think that this is due to the relative simplicity of these series, when compared to the others forming the scenarios.

When applying GP, the results differ. For GANs, the series correspond to sine waves in most cases, indicating that the model failed to capture the data distribution

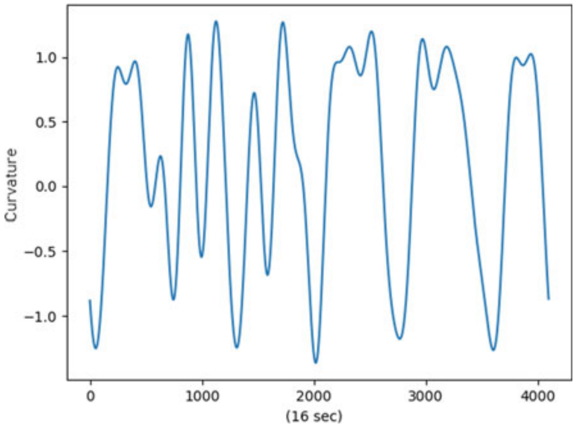


Fig. 13 Generated curve of the road's curvature using GAN

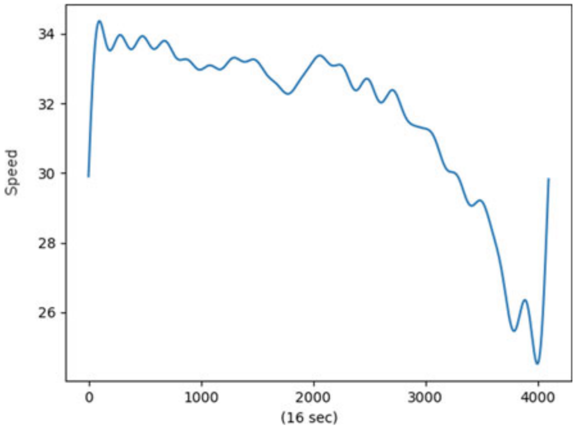


Fig. 14 Generated curve of the absolute speed using GAN

(this behavior is found for any type of normalization in GANGP). For WGANs, the series are immediately improved (Figs. 17, 18, 19 and 20): this is the logical result of this regularization since the exploration of the two data distributions, or at least areas close to them, is carried out and promotes the Lipschitz continuity necessary for the proper functioning of WGANs. If we apply the CT in addition, we find exactly the same behavior for GANs: the models failed to capture the real data distribution. For WGANs, the generation is further improved, as the exploration of space becomes finer, always in accordance with the formulation of the Critic: indeed, we also observe correlations between the different series. For example, if the curvature of the road is increasing, the speed of the ego vehicle is decreasing. Several noise artifacts are sometimes observed with the gradient penalty, and their presence slowly disappears if we reduce the harshness of the penalty. The same phenomenon occurs when the CT is added, with the same reaction towards changes in its harshness parameter.

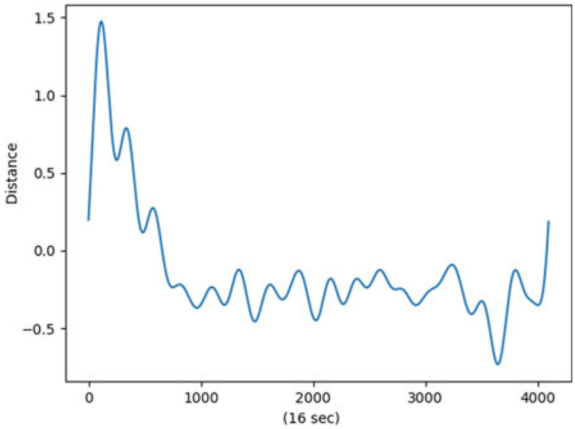


Fig. 15 Generated distance from the road center using GAN

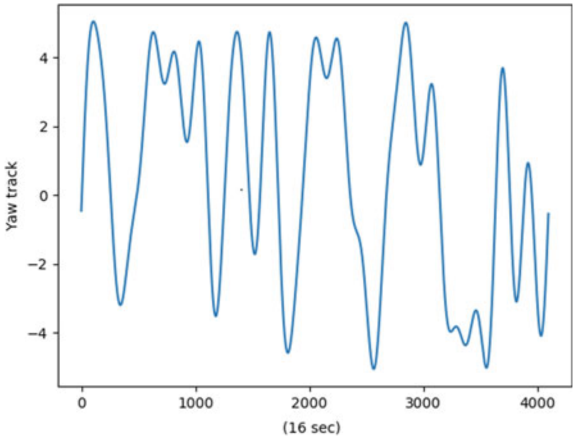


Fig. 16 Generated curve of Yaw track using GAN

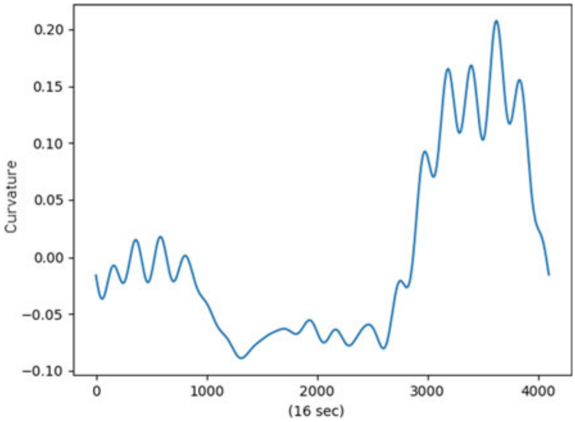


Fig. 17 Generated curve of the road’s curvature using WGAN

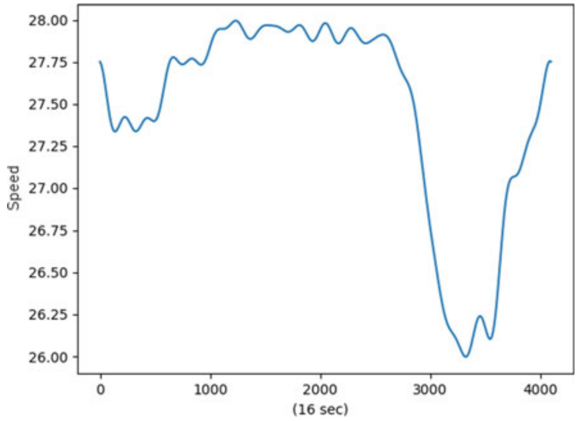


Fig. 18 Generated curve of the absolute speed using WGAN

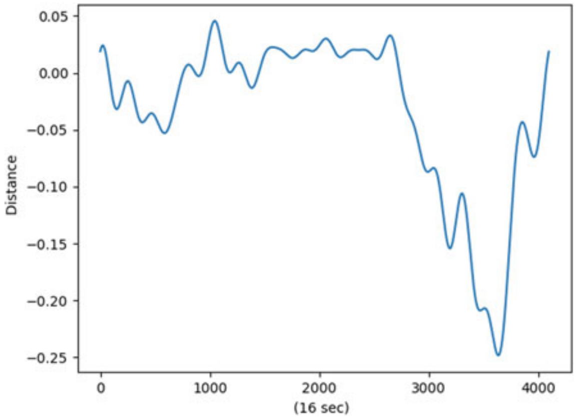


Fig. 19 Generated distance from the road center using WGAN

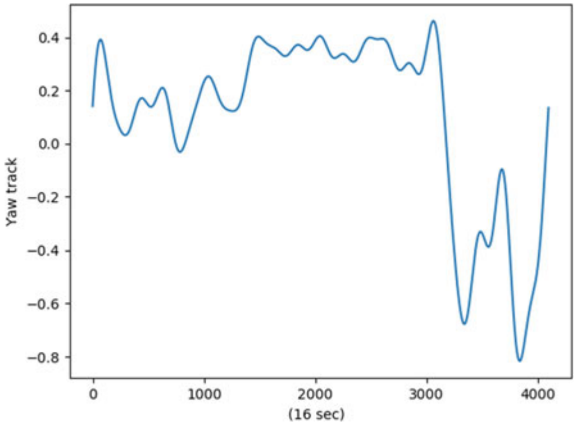


Fig. 20 Generated curve of Yaw track using WGAN

However, it should be noted that spectral normalization does not necessarily bring a benefit in terms of generation quality. To be more precise, it should be pointed out that for a GAN without regularization, the SN does not cause any significant change in the quality of generation; similarly, for the GANGP and GANGPCT, SN did not rectify model failures. On the other hand, examples generated from WGAN models that have used SN alone or with regularization tend to be generally more realistic than those generated from models that do not use it. This can be explained simply by the fact that the formulation of spectral normalization favors the WGAN Critic, so learning is more effective, and the addition of regularization to this normalization only improves the exploration of spaces as stated above. The results of the gradient penalty and the consistency term on GANs can be explained by the fact that the latter are calculated from the softmax outputs of the discriminator the inputs of which are normalized between -1 and 1: in theory, the scale of distances would make gradient penalty extremely harsh and would encourage the discriminator to focus only on minimizing the penalty without worrying about its primary task, not letting the consistency term bring his added value. In other words, spectral normalization enhances the quality of generation only for WGAN type formulations. In fact, the main observed fact of spectral normalization is its ability to promote the generation of specific scenarios. In the case of traditional GANs, we observe the generation of straight lines, which are very atypical scenarios in the dataset (Figs. 21, 22, 23 and 24), even if these straight lines could also indicate that the model overfitted. In the case of WGANs with or without regularization, the same observation applies to atypical scenarios that could be described as *critical*; by this, we mean close to an erratic behavior of the model that would make decisions and then change its mind very quickly on several occasions. Similarly, it also appears to slightly reduce the presence of noise artifacts in some complex series within generated scenarios.

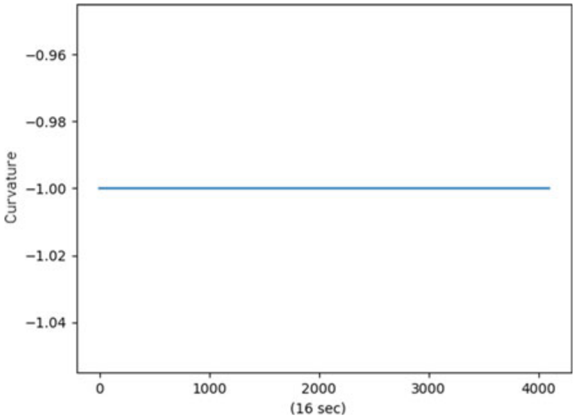


Fig. 21 Generated curve of the road's curvature using GAN

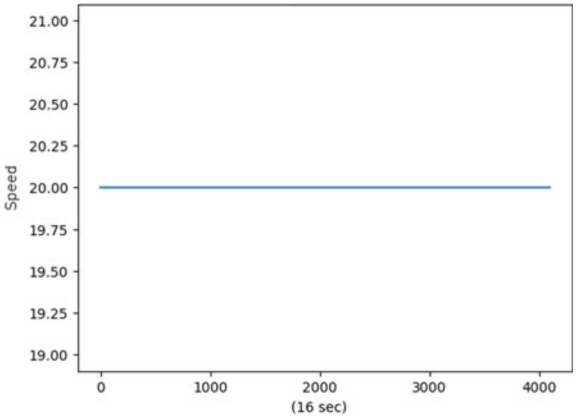


Fig. 22 Generated curve of the absolute speed using GAN

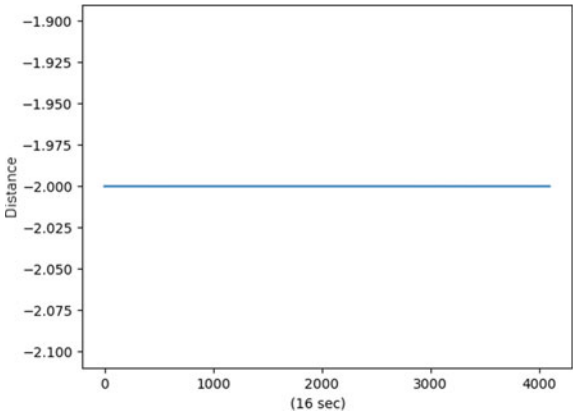


Fig. 23 Generated distance from the road center using GAN

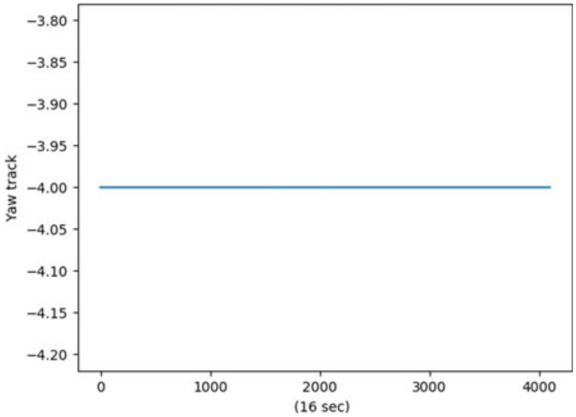


Fig. 24 Generated curve of Yaw track using GAN

5.3 Geometry Score

The results of the geometry score evaluation conclude this experimental study. We remind you that the lower this score is, the more it reflects the success of GANs learning. Indeed, it corresponds to the “distance” between the approximation of the shape of two simplicial complexes over time (variation of the said parameter ϵ). A low score therefore implies an adversarial generative network that has captured the shape of the manifold on which real data is based and has therefore learned effectively to reproduce the spatial elements of our input data. At the coverage level, this implies that GANs cover the actual data manifold correctly and that the generation coverage level is maximized; indeed, by calculating MRLTs, the shape of the manifold of generated data approaches the actual data one (number of holes in particular). It is interesting here to start by studying results of spectral normalization. The latter gives good geometry scores, but not the best in all types of formulations:

Table 3 Geometry scores for the GAN formulation

Model	Score	Model	Score	Model	Score
GAN Rd FC 2D	0.306	GANGP RSN FC 2D TTUR	0.353	GANGPCT R FC 2D TTUR	0.342
GAN SN FC 2D	0.079	GANGP SN FC 2D	0.340	GANGPCT SN FC 2D TTUR	0.296
GAN LnD FC 2D TTUR	0.14	GANGP Rd FC 2D TTUR	0.297	GANGPCT SnDB FC 2D	0.312
GAN R FC 2D TTUR	0.48	GANGP SN FC 2D TTUR	0.346	GANGPCT SN FC 2D	0.313
GAN SN FC 2D TTUR	0.079	GANGP LnD FC 2D	0.323	GANGPCT Rd FC 2D	0.308
GAN LN FC 2D TTUR	0.077	GANGP SnDB FC 2D TTUR	0.294	GANGPCT LnD FC 2D	0.333
GAN RSnD FC 2D	0.214	GANGP Rd FC 2D	0.344	GANGPCT LN FC 2D TTUR	0.305
GAN R FC 2D	0.075	GANGP LN FC 2D TTUR	0.311	GANGPCT LN FC 2D	0.349
GAN SnD FC 2D TTUR	0.095	GANGP RSnD FC 2D TTUR	0.349	GANGPCT RSnD FC 2D_TTUR	0.322
GAN SnD FC 2D	0.141	GANGP RSN FC 2D	0.355	GANGPCT RSnD FC 2D	0.304
GAN Rd FC 2D TTUR	0.223	GANGP SnDB FC 2D	0.326	GANGPCT SnD FC 2D TTUR	0.307
GAN LnD FC 2D	0.292	GANGP R FC_2D TTUR	0.359	GANGPCT RSN FC 2D TTUR	0.331
GAN RSnD FC 2D TTUR	0.574	GANGP RSnD FC 2D	0.380	GANGPCT SnD FC 2D	0.336
GAN RSN FC 2D TTUR	0.574	GANGP SnD FC 2D	0.327	GANGPCT SnDB FC 2D TTUR	0.345
GAN SnDB FC 2D	0.081	GANGP LN FC 2D	0.252	GANGPCT R FC 2D	0.353
GAN RSN FC 2D	0.57	GANGP SnD FC 2D_TTUR	0.296	GANGPCT RSN FC 2D	0.338
GAN LN FC 2D	0.123	GANGP R FC 2D	0.363	GANGPCT Rd FC 2D TTUR	0.279
GAN SnDB FC 2D TTUR	0.495	GANGP LnD FC 2D TTUR	0.378	GANGPCT LnD FC 2D TTUR	0.350

in this sense, the SN improves the stability and convergence of the model, as well as the quality of the generation because it values 1-Lipschitz continuity, and this valorization is found in the geometry score of all models that use it. In such a case, the latter is generally low, except for the GANGP and the GANGPCT. The qualitative study of the scenarios generated by the latter revealed the failure of the learning: the geometry score is therefore necessarily much higher for all the models concerned (Table 3) (values higher than 0.3, compared to values lower than 0.1 for many other models). Empirically, it is fairly easy to conclude that, on average, models using the Wasserstein formulation are more able to address the coverage problem. Indeed, the Earth Mover distance seems more suitable (it is used in topological data analysis). The corresponding geometry scores are all low (Table 4) and the use of GP and CT

Table 4 Geometry scores for the WGAN formulation

Model	Score	Model	Score	Model	Score
WGAN RSN FC 2D	0.082	WGANGP SnD FC 2D	0.093	WGANGPCT LnD FC 2D	0.118
WGAN SN FC 2D	0.074	WGANGP SN FC 2D TTUR	0.113	WGANGPCT SnDB FC 2D TTUR	0.116
WGAN RSnD FC 2D TTUR	0.164	WGANGP SnDB FC 2D TTUR	0.075	WGANGPCT LnD FC 2D TTUR	0.077
WGAN LnD FC 2D	0.083	WGANGP R FC 2D	0.092	WGANGPCT SN FC 2D TTUR	0.142
WGAN LnD FC 2D TTUR	0.071	WGANGP RSnD FC 2D	0.079	WGANGPCT LN FC 2D	0.081
WGAN SnDB FC 2D	0.064	WGANGP LnD FC 2D	0.09	WGANGPCT SnDB FC 2D	0.082
WGAN Rd FC 2D	0.061	WGANGP Rd FC 2D TTUR	0.068	WGANGPCT R FC 2D	0.098
WGAN R FC 2D TTUR	0.076	WGANGP R FC 2D TTUR	0.093	WGANGPCT Rd FC 2D	0.065
WGAN RSnD FC 2D	0.144	WGANGP RSnD FC 2D TTUR	0.075	WGANGPCT Rd FC 2D TTUR	0.074
WGAN SN FC 2D TTUR	0.07	WGANGP LN FC 2D TTUR	0.123	WGANGPCT SnD FC 2D	0.077
WGAN LN FC 2D TTUR	0.085	WGANGP RSN FC 2D	0.069	WGANGPCT SN FC 2D	0.081
WGAN RSN FC 2D TTUR	0.080	WGANGP Rd FC 2D	0.079	WGANGPCT SnD FC 2D TTUR	0.068
WGAN Rd FC 2D TTUR	0.124	WGANGP SnD FC 2D TTUR	0.067	WGANGPCT RSN FC 2D	0.131
WGAN SnD FC 2D TTUR	0.170	WGANGP LnD FC 2D TTUR	0.068	WGANGPCT RSN FC 2D TTUR	0.124
WGAN SnDB FC 2D TTUR	0.096	WGANGP LN FC 2D	0.106	WGANGPCT RSnD FC 2D	0.121
WGAN R FC 2D	0.065	WGANGP SN FC 2D	0.109	WGANGPCT R FC 2D TTUR	0.065
WGAN LN FC 2D	0.071	WGANGP SnDB FC 2D	0.067	WGANGPCT RSnD FC 2D TTUR	0.071
WGAN SnD FC 2D	0.102	WGANGP RSN FC 2D TTUR	0.067	WGANGPCT LN FC 2D TTUR	0.071

is really important: by exploring manifolds more effectively, these regularizations ensure that the number of areas of space that were used to learn distributions is maximized, and that the results of the persistence diagrams coincide over time. Thus, all WGANGPs and WGANGPCTs achieve a similar level of coverage, without focusing on any particular normalization. The latter once again addresses the problem of convergence and stability. We can therefore say that the level of coverage becomes better thanks to regularization.

6 Conclusions and Future Works

Generating realistic data here comes from developing a novel method to represent multivariate time series. This representation made it possible to study the dynamics of generative adversarial networks in a rigorous way, and in the context of autonomous driving. The results using the chosen representation showed the ability of WGANGP and WGANGPCT to generate diverse and varied realistic scenarios. In the case of autonomous driving scenarios, the generation using multi-channel images is accurate. And according to the geometry score, the elements produced are realistic with regard to learning data.

The geometry score makes it possible to assess data coverage based on the actual data manifold. However, the data may be on a manifold that does not take into account all possible scenarios. Data collection biases support this hypothesis. Creating a new score that would measure the level of coverage beyond the manifold, potentially on its external borders for unknown scenarios, would better address the issue of evaluating levels of coverage achieved by the generation.

Acknowledgements The authors wish to express their thanks for the financial support of the Institute for Technological Research (IRT) Systemx.

References

1. Goodfellow I, Bengio Y, Courville A (2017) Deep learning. MIT Press, Cambridge
2. Tamilarasan S, Jung D, Guvenc L (2018) Drive scenario generation based on metrics for evaluating an autonomous vehicle controller. In: WCX World congress experience, SAE International
3. Feng G, Jianli D, Yingdong H, Zilong W (2019) A test scenario automatic generation strategy for intelligent driving systems. Math Probl Eng 19
4. Ghosh A, Bhattacharya B, Chowdhury SBR (2016) SAD-GAN: synthetic autonomous driving using generative adversarial networks. arXiv:1611.08788
5. Yang L, Liang X, Wang T, Xing E (2018) Real-to-virtual domain unification for end-to-end autonomous driving. In: Proceedings of the European conference on computer vision (ECCV), pp 530–545
6. Uricar M, Krizek P, Hurych D, Sobh I, Yogamani S, Denny P (2019) Yes, we GAN: applying adversarial techniques for autonomous driving. arXiv:1902.03442
7. Fabbri C, Sharma J (2018) D-GAN: autonomous driving using generative adversarial networks. https://cameronfabbri.github.io/papers/gtav.pdf

8. Saatci Y, Wilson AG (2017) Bayesian GAN. In: Advances in neural information processing systems, vol 30. Curran Associates, Inc., pp 3622–3631
9. Mziou Sallami M, Ibn Khedher M, Trabelsi A, Kerboua-Benlarbi S, Bettebghor D (2019) Safety and robustness of deep neural networks object recognition under generic attacks. In: Gedeon T, Wong K, Lee M (eds) Neural information processing. ICONIP 2019. Communications in computer and information science, vol 1142. Springer, Cham
10. Khedher MI, Mziou-Sallami M, Hadji M (2021) Improving decision-making-process for robot navigation under uncertainty. In: ICAART (2)
11. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
12. Arjovsky M, Bottou L (2017) Towards principled methods for training generative adversarial networks. arXiv e-prints
13. Metz L, Poole B, Pfau D, Dickstein JS (2016) Unrolled generative adversarial networks. CoRR. arXiv:1611.02163
14. Salimans T, Goodfellow IJ, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training GANs. CoRR. arXiv:1606.03498
15. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein GAN. arXiv e-prints
16. Villani C (2008) Optimal transport: old and new. Grundlehren der mathematischen Wissenschaften, Springer, Berlin, Heidelberg
17. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of Wasserstein GANs. In: Advances in neural information processing systems, pp 5767–5777
18. Wei X, Gong B, Liu Z, Lu W, Wang L (2018) Improving the improved training of Wasserstein GANs: a consistency term and its dual effect. arXiv:1803.01541
19. Fedus W, Rosca M, Lakshminarayanan B, Dai AM, Mohamed S, Goodfellow I (2017) Many paths to equilibrium: GANs do not need to decrease a divergence at every step. arXiv:1710.08446
20. Chongxuan L, Xu T, Zhu J, Zhang B (2017) Triple generative adversarial nets. In: Advances in neural information processing systems, pp 4088–4098
21. Chaoyue W, Chang X, Xin Y, Dacheng T (2018) Evolutionary generative adversarial networks. CoRR. arXiv:1803.00657
22. Anders Boesen Lindbo L, Søren Kaae S, Ole W (2015) Autoencoding beyond pixels using a learned similarity metric. CoRR. arXiv:1512.09300
23. Lucic M, Kurach K, Michalski M, Gelly S, Bousquet O (2018) Are GANs created equal? A large-scale study. In: Advances in neural information processing systems, vol 31. Curran Associates, Inc., pp 700–709
24. Karol K, Mario L, Xiaohua Z, Marcin M, Sylvain G (2018) The GAN landscape: losses, architectures, regularization, and normalization. CoRR. arXiv:1807.04720
25. Roth K, Lucchi A, Nowozin S, Hofmann T (2017) Stabilizing training of generative adversarial networks through regularization. In: Advances in neural information processing systems, pp 2018–2028
26. Naveen K, Jacob DA, James H, Zsolt K (2017) How to train your DRAGAN. CoRR. arXiv:1705.07215
27. Takeru M, Toshiki K, Masanori K, Yuichi Y (2018) Spectral normalization for generative adversarial networks. CoRR. arXiv:1802.05957
28. Ba JL, Kiros JR, Hinton EG (2016) Layer normalization. arXiv:1607.06450
29. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR. arXiv:1502.03167
30. Mescheder L, Geiger A, Nowozin S (2018) Which training methods for GANs do actually converge? arXiv:1801.04406
31. Nagarajan V, Kolter JZ (2017) Gradient descent GAN optimization is locally stable. In: Advances in neural information processing systems, pp 5585–5595
32. Zhang H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks. arXiv:1805.08318

33. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. arXiv:1809.11096
34. Ilya S, Oriol V, Quoc VL (2014) Sequence to sequence learning with neural networks. In: NIPS
35. Lantao Y, Weinan Z, Jun W, Yong Y (2016) SeqGAN: sequence generative adversarial nets with policy gradient. CoRR. arXiv:1609.05473
36. Kusner MJ, Hernández-Lobato JM (2016) GANS for sequences of discrete elements with the Gumbel-softmax distribution. CoRR
37. Donahue D, Rumshisky A (2018) Adversarial text generation without reinforcement learning. CoRR. arXiv:1810.06640
38. Esteban C, Hyland SL, Ratsch G (2017) Real-valued (medical) time series generation with recurrent conditional GANs. arXiv:1706.02633
39. Mogren O (2016) C-RNN-GAN: continuous recurrent neural networks with adversarial training. CoRR. arXiv:1611.09904
40. Arnelid H, Zec EL, Mohammadiha N (2019) Recurrent conditional generative adversarial networks for autonomous driving sensor modelling. In: 2019 IEEE Intelligent transportation systems conference (ITSC), pp 1613–1618
41. Press O, Bar A, Bogin B, Berant J, Wolf L (2017) Language generation with recurrent generative adversarial networks without pre-training. CoRR. arXiv:1706.01399
42. Luo Y, Cai X, Zhang Y, Xu J, Xiaojie Y (2018) Multivariate time series imputation with generative adversarial networks. In: Advances in neural information processing systems, pp 1596–1607
43. Takahashi S, Chen Y, Tanaka-Ishii K (2019) Modeling financial time-series with generative adversarial networks. Phys. A Stat. Mech. Appl. 527
44. Soares Koshiyama A, Firoozye N, Treleaven PC (2019) Generative adversarial networks for financial trading strategies fine-tuning and combination. CoRR. arXiv:1901.01751
45. Wiese M, Knobloch R, Korn R, Kretschmer P (2019) Quant GANs: deep generation of financial time series. arXiv e-prints
46. Lopez-Paz D, Oquab M (2016) Revisiting classifier two-sample tests. arXiv:1610.06545
47. Shmelkov K, Schmid C, Alahari K (2018) How good is my GAN ? In: Proceedings of the European conference on computer vision (ECCV), pp 213–229
48. Fortet R, Mourier E (1953) Convergence de la répartition empirique vers la répartition théorique. Ann Sci l'École Normale Supérieure 70:267–285
49. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. J Mach Learn Res 13:723–773
50. Im DJ, Ma H, Taylor G, Branson K (2018) Quantitatively evaluating GANs with divergences proposed for training. arXiv:1803.01045
51. Borji A (2019) Pros and cons of GAN evaluation measures. Comput Vis Image Underst 179:41–65
52. Theis L, Van den Oord A, Bethge M (2015) A note on the evaluation of generative models. arXiv:1511.01844
53. Ghrist R (2008) Barcodes: the persistent topology of data. Bull Am Math Soc 45:61–75
54. Chazal F, Michel B, An introduction to topological data analysis: fundamental and practical aspects for data scientists. Front Artif Intell. https://doi.org/10.3389/frai.2021.667963. https://arxiv.org/pdf/1710.04019.pdf
55. Khrulkov V, Oseledets IV (2018) Geometry score: a method for comparing generative adversarial networks. CoRR. arXiv:1802.02664
56. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434
57. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local NASH equilibrium. In: Advances in neural information processing systems, pp 6626–6637
58. Wang Z, Oates T (2015) Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence

59. Brophy E, Wang Z, Ward TE (2019) Quick and easy time series generation with established image-based GANs. CoRR. arXiv:1902.05624
60. Sellami M, Ghorbel F (2012) An invariant similarity registration algorithm based on the analytical Fourier-Mellin transform. In: 2012 Proceedings of the 20th European signal processing conference (EUSIPCO), pp 390–394
61. Saidani M, Malek S, Faouzi G (2016) Geometric invariance in digital imaging for the preservation of cultural heritage in Tunisia. Digit Appl Archaeol Cult Heritage 3(4)
62. Ron G (2003) Pyramid algorithms, Chapter 1. In: The Morgan Kaufmann series in computer graphics, San Francisco, pp 1–43

Fuzzy Set Theory-Based Approach for Mining Spatial Association Rules: Road Accident as a Case Study

Addi Ait-Mlouk, Mohamed Ait-Mlouk, Fatima-Zahra El Mazouri, Arindam Dey, and Tarik Agouti

Abstract Despite intensive research over the last decades in spatial data mining, a considerable amount of collected data brings new challenges to the extraction of spatial association rules. Notwithstanding, association rules are examining the numerical and categorical data accurately; however, it is not directly extensible to efficiently analyze real-world spatial data. Analyzing spatial data exposes significant difficulties, including particular representations of spatial information like geometrical measures, topological measures, and spatial object dependencies. Most current studies ignore the integration of geographical information in the association rules mining process, which is not beneficial for the decision-makers to extract relevant association rules. To deal with this issue, we proposed a fuzzy set theory-based approach for mining spatial association rules; This proposed approach would be especially valuable for decision-makers suffering from imprecision and non-quality of extracted association rules. An analysis of road accidents using spatial data shows that the use of fuzzy set theory improves the quality and the precision of extracted association rules. Overall, the proposed approach contributes to a better understanding of spatial association rules and provides meaningful information that can aid decision-makers to improve the performances and precision of spatial association rules.

A. Ait-Mlouk (✉)
Department of Computing Science, Uppsala University, Uppsala, Sweden
e-mail: addi.ait-mlouk@it.uu.se

M. Ait-Mlouk
Department of Geology, Cadi Ayyad University, Marrakech, Morocco
e-mail: mohamed.aitmlouk@ced.uca.ma

F.-Z. El Mazouri
IETR laboratory UMR 6164, Polytech de Nantes university, Nantes, France
e-mail: fatima-zahra.elmazouri@univ-nantes.fr

A. Dey
Department of Computing Science, Saroj Mohan Institute of Technology, Kolkata, India
e-mail: arindamdey@bbit.edu.in

T. Agouti
Department of Computing Science, Cadi Ayyad University, Marrakech, Morocco
e-mail: t.agouti@uca.ac.ma

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_17

Keywords Data mining · Spatial association rules · Spatial predicates · Fuzzy logic · Road accidents · Geographical data

1 Introduction

Road accidents caused a major public health problem worldwide; this requires enormous efforts to identify all indicators contributing to this problem. The World Health Organization (WHO) [1] reported that road accidents caused over 1.24 millions deaths and as many as 50 millions casualties worldwide, without forgetting the material damages. To study this problem, data mining (non-trivial extraction of implicit, previously unknown, and potentially useful information from data [2]) techniques have driven important gains in other areas and should also be adopted to enhance road safety. The use of sensors, VANET, GPS, and RFID technologies generates a huge amount of data; this seems an application area of spatial data mining; however, to date, there have been a few approaches to improve road safety using spatial data. The existing optimization methods have long been computerized, but do not provide insights that are the purpose of spatial data mining [3]. In the transportation sector, companies are confronted with a variety of risks resulting from shipping safety, transportation costs, and a high rate of accidents. Therefore, it is necessary to adopt adequate measures that take into account all these challenges. The presented paper introduces an approach to spatial association rules based on the fuzzy set theory. The main purpose of this approach is to produce insights and enough knowledge that enables decision-makers to provide the best decision that can permit avoiding risky routes using spatial data mining and fuzzy set theory.

Previous work in spatial data mining [4–11], lead to a set of interesting methods among them, the association rules analysis. This method has become an important technique for discovering useful knowledge designating certain association relationships among a set of spatial and non-spatial predicates. Notwithstanding, the huge amount of real-world data make it a promising source for association rules, and mining this data could uncover meaningful insights. Moreover, it examines conditional interaction among input datasets and produces association rules of type IF-THEN. For instance, the following rule, “most gas stations are close to the highway,” is a spatial rule of the form: $is\_a(X, gas\_station) \rightarrow close\_to(X, highway)$.

Association rules techniques have been applied in several fields, including business planning, diagnostic support, medical research, and telecommunications. The final step of the rules validation will let the user face the principal difficulty of the most important rules. Therefore, it is necessary to assist the user in the validation task by implementing a pre-processing task. The pre-processing task aims to prepare the spatial context by using the fuzzy set theory [12]. This task must consider the decision-makers’ preferences, spatial predicates, and metric distance between the objects of different thematic layers.

Our proposed approach aims to generate insights and sufficient knowledge from spatial data to allow decision-makers to make the right decision, thus enabling safety optimization and reducing road accidents. Besides, the integration of fuzzy logic within spatial association rules contributes to a better understanding of the dynamics of road accidents. It provides more insightful details that can guide stakeholders better optimize transport quality and safety. Our approach has the following great advantages:

1. Mining interesting spatial association rules;
2. Improve the precision and the quality of spatial association rules by using fuzzy set theory;
3. Improve road safety by analyzing historical data and identify routes that may be dangerous.

The remaining of this paper is structured as follows: Sect. 2 looks into the latest literature related to spatial association rules and data mining. Section 3 illustrates the proposed methodology for mining spatial association rules by using the fuzzy set theory. Section 3 shows the results and discussion. The last section will summarize the work done and highlight its contributions.

2 Related Works

Big data brought the light on association rules as a significant technique that appears in a broad range of real-world scenarios. Association rules offer an improved understanding of what is behind the data. It also uncovers meaningful insights. Thus, it can benefit plenty of useful applications and can assist a lot of applications such as basket analysis [13], chemo-informatics [14], biomedical [15], images [16], and in particular the problem of road safety [17–21]. This broad spectrum of applications allowed association rules to be applied to various datasets, including sequential, spatial, and graph-based data.

We applied fuzzy set theory as a means to deal with both quantitative and qualitative variables. The fuzzy set theory provides a number of tools able to handle the vagueness precision in the natural language and the knowledge itself. The literature has provided several attempts to develop fuzzy pattern mining algorithms able to deal with this kind of data and to discover relationships between them. One of the most popular approaches is Fuzzy Frequent Pattern Mining where the extracted rules can be subject to new rules based on fuzzy set theory. The first fuzzy association rule mining consists of two main steps: Finding the frequent itemsets, and then, generating fuzzy rules based on the previously extracted frequent itemsets. Apriori is an approach that several researchers have tried; among them, Chan and Au [22] who this algorithm to convert data into linguistic terms using the fuzzy set theory. Kuok et al. [23] proposed an approach to handle quantitative databases for generating fuzzy association rules. Kuok et al. [24] gave another algorithm able to converts quantitative data into linguistic terms. Moreover, Lin et al. [25] proposed an algorithm

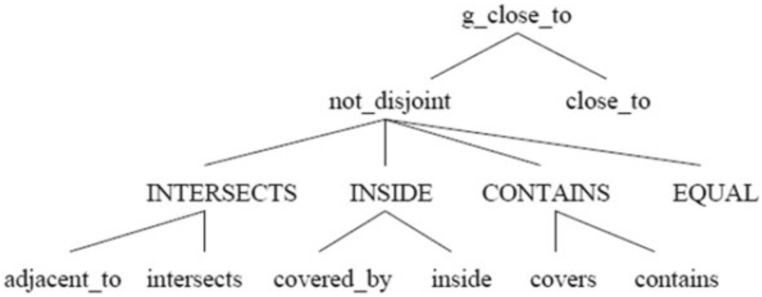


Fig. 1 The hierarchy of topological relations

(UBFFPT) that measures the upper bound membership values of frequent itemsets. This algorithm requires many databases passes to build the tree; this can be long and directly impact response time.

Spatial association rules mining is a technique able to discover relationships between spatial and non-spatial data. Several researchers used this approach for knowledge discovery purposes [26–30]. However, most of these works were on standard databases and a very few of them cover knowledge discovery in spatial databases. In this context, Koperski et al. [3] introduced the extraction of spatial association rules by extending the technique of mining association rules in transaction databases by using several types of spatial predicates, the Fig. 1 presents the concept hierarchy of topological relations.

The algorithm proposed by Koperski et al. takes into input a spatial database, a query, minimum support, and minimum confidence. The output of this algorithm consists of multi-level spatial association rules for all spatial objects and spatial relations. To improve the performance of Koperski et al. method, Salleb [11] proposed the ARGIS algorithm as an iterative algorithm based on the concept of a link table and defined two categories of thematic layers, reference layers, and descriptive layers. The proposed algorithm takes into input a spatial database, reference layers, descriptive layers, spatial and non-spatial predicates, minimum support, and minimum confidence. The output of this algorithm consists of multi-level spatial association rules for all spatial objects and spatial relations between reference and descriptive layers. The basic principle of ARGIS is to consider each time two and only two thematic layers to find spatial associations. Hence, it is not possible to extract rules regrouping more than two thematic layers. To tackle this problem, Marghoubi et al. [31] proposed a codification system of spatial objects to identify a thematic layer of any spatial object; their algorithm used Galois Lattice [32] for mining spatial association rules. Koperski et al. approach is based on the concept hierarchy, fixing a distance according to the user's needs. Similarly, in Salleb and Marghoubi et al. works, one must first fix a distance for the determination of spatial predicates. This can be described by the following spatial association rule R: $is\_a(X, gas\_station) \rightarrow close\_to(X, highway)$. If the distance between "gas stations" and "highway" is less than 50 m then one can say that the gas station is close to "highway", and if the distance between "gas stations" and "highway" is higher than 50 m then one can say that the gas station is "not close" to "highway". From a linguistic point of view, a distance equal to 52 m can be the subject of new rules R1 and R2. Besides, the result does not permit decision-makers to specify the most pertinent rules. This imprecision can be improved by using fuzzy logic [12]; more details about the proposed approach will be given in the next section.

3 Methodology

In this section, we explain the different steps constructing our proposed approach. starting with the association rules mining:

3.1 Association Rules Mining

Association rule mining is a rule-based machine learning method for identifying interesting associations between objects in large databases. It is an implication of the form $A \rightarrow B$ such as $A, B \subset I$, and $A \cap B = \emptyset$. Each rule comprises two different sets of items A and B, where A is called antecedent or left-hand-side (LHS) and B is called consequent or right-hand-side (RHS). For example, $Driver \rightarrow Vehicle$, suggests that a strong relationship exists between two items Driver and Vehicle. However, the application of the association rule technique in large-scale data produces a huge amount of the extracted rules, which can contain weak rules. Thus, multiple constraints on various measures of significance and interest are utilized to filter the interesting rules. Practically, the best-known constraints are minimum thresholds on support and confidence as well as the lift and certainty factor. The support is defined as the proportion of transactions in the database which contain the items A, the formula is given by the Eq. 1:

$$Supp(A \rightarrow B) = Supp(A \cup B) = \frac{|t(A \cup B)|}{t(A)} \tag{1}$$

While the confidence defines how frequently items in B appear in a transaction that contains A, the formula is given in Eq. 2:

$$Conf(A \rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)} \tag{2}$$

The lift computes the ratio between rule's confidence and the support of itemset in the rule consequent, the formula is given in Eq. 3:

$$Lift(A \rightarrow B) = \frac{Conf(A \cup B)}{Supp(B)} \tag{3}$$

The certainty factor (CF) measures the variation of the probability that B is in a transaction when only considering transactions with A. An increase of CF indicates a decrease in the probability that B is not in a transaction that A contains. We determine the CF of a given association rule $(A \rightarrow B)$. The CF formula is given in Eq. 4:

$$CF(A \rightarrow B) = \begin{cases} \frac{C(A \rightarrow B) - S(B)}{1 - S(B)} & if \;\; C(A \rightarrow B) > S(B) \\ & \frac{C(A \rightarrow B)}{S(B)} \; otherwise \end{cases} \tag{4}$$

3.2 Spatial Association Rules

Spatial association rules are a set of associations among a set of spatial and possibly non-spatial attributes of geographical objects. These relationships, called spatial predicates, can describe topological relationships between spatial objects, such as disjoint, intersects, adjacent to, etc. They can also provide intel regarding distance, orientation, possible intersections, etc. [3].

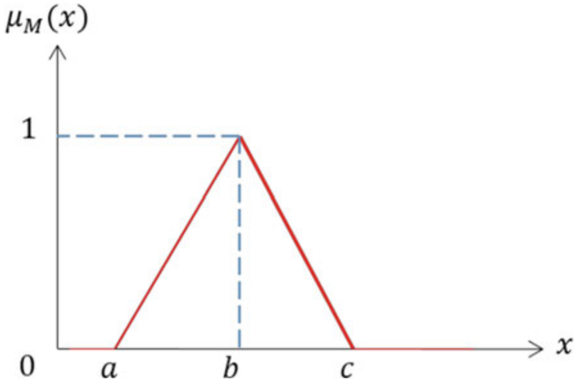
A spatial association rule can be described as an $X \rightarrow Y$ form where where X and Y are sets of predicates, where some can be spatial. Like their classic counterparts, spatial association rules come with a minimum support and a minimum confidence. Spatial association rules mining that can generate all possible combinations of items (spatial or not) that will provide the best support and confidence.

3.3 Fuzzy Set Theory

In many real cases, the linguistic assessments of human perception can be inconsistent, incomplete, vague, and even imprecise, representing it by numerical value may prove unrealistic. However, the use of interval judgment would be practically preferable than fixed values judgment [33]. Besides, the evaluation of spatial predicates provides intrinsic complexity linked to the metric distance between objects of thematic layers. In such a situation, Fuzzy Set Theory (FST) [12, 34] presents an effective approach to improve the uncertainty of human preferences. FST provides mathematical representations or dealing with the intrinsic imprecision of real-life problems. A fuzzy set A of a universe of discourse X is designated by a membership function μ_A, $\forall x \in \mu_A \in [0, 1]$. The set A is defined by:

$$A = (x, \mu_A(x))|x \in X$$

Fig. 2 The triangular fuzzy numbers (TFNs)



The Triangular Fuzzy Numbers (TFNs) is among the most common shapes of fuzzy numbers. It is represented with triplets (a, b, c) such that $a \preceq b \preceq c$ as shown in Fig. 2.

The mathematical formula of the membership function is defined as follows:

$$\mu_M(x) = \begin{cases} 0\, if \;\; x \preceq a \\ (x-a)/(b-a) \;\; if \;\; a \preceq x \preceq b \\ (x-b)/(c-b) \;\; if \;\; b \preceq x \preceq c \end{cases} \tag{5}$$

The term b represents the possible value, a and c represent the lower and upper limits respectively used to reflect the fuzziness of the related preferences.

By using FST, the concept of linguistic expression can be quantified by using fuzzy numbers. In this proposed approach, linguistic variables were considered to express experts' assignments, and the positive Triangular Fuzzy Numbers (TFNs) were used to quantify linguistic variables for the computation of the metric distance between objects of thematic layers (Table 1).

In this paper, we extend Salleb and Margoubi et al. approach [11, 31]. Indeed, we aim to extract spatial association rules by obtaining relevant spatial predicates using fuzzy logic. Our proposal is based on three main modules (Fig. 3). The first is the preparation of the spatial context, the second aims to handle the uncertainty of human preference, such as the metric of the distance between objects of thematic

Table 1 Fuzzy predicates table

Predicate Id	Name	Triangular fuzzy numbers
P01	Inside	(0,10,15)
P02	Close	(10,15,20)
P03	Intersect	(15,20,25)

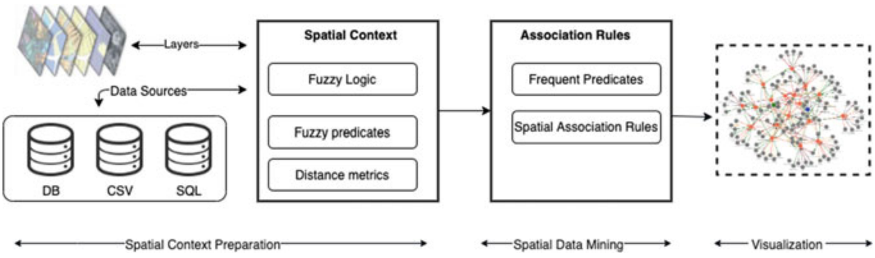


Fig. 3 Overview of the proposed approach: (1) Spatial context preparation, (2) Fuzzy Spatial association rules, and (3) Visualization

layers, and the third is the extraction of spatial association rules. More details about the approach will be given in the next subsections.

The preparation of spatial context It is the result of computing the spatial relations of each thematic layer object with the other spatial objects while considering non-spatial attributes. In our approach, the decision-makers determine spatial predicates (Table 1), the positive triangular fuzzy numbers (TFN) were used to quantify linguistic variables.

Spatial predicates and metric distance computation The determination of spatial predicates for a different thematic layer is computed by allowing the user to determine the distances between different thematic layers and assign a corresponding spatial predicate using fuzzy logic. For example: If the distance between the spatial objects *"gas station"* and *"highway"* is 20 m, then the *"gas station"* is close to *"highway"*. To compute the distance between two points p and q, we employed the Euclidean distance that is the length of the line segment connecting them; the formula is given in the Eq. 6.

$$dist(p, q) = dist(q, p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{6}$$

Spatial association rule mining algorithm In the initial step, the recommended algorithm is based on the preparation of spatial context through four main stages (Algorithm 1). From line 1 to line 5, the expert choose a thematic layer according to the extraction objectives, from line 6 to line 9 the algorithm fusion the objects of different layers previously defined into one table, from line 10 to line 16, the algorithm defines the spatial relationship between different layers and determines the appropriate spatial predicates. From line 17 to line 21, the algorithm computes a distance for each spatial object compared to the other spatial objects of the same table and affects the appropriate predicate based on distance thresholds using fuzzy logic. Finally, line 22 constructs a spatial context by making the union of all spatial objects.

In the secondary step, (Algorithm 2), the Apriori algorithm takes into input the spatial context previously prepared (Algorithm 1) and minimum support to extract frequent predicates set. The algorithm generates a frequent predicate set (from line

1 to line 3) and then generates frequent spatial predicates (from line 4 to line 12). Finally, extract spatial association rules from frequent spatial predicate by satisfying the minimum confidence threshold. The result is a set of spatial association rules that considering a different correlation between a set of thematic layer objects.

Algorithm 1: Spatial context preparation

```
   Input  : Thematic Layer Tables (TLT), Spatial Predicates Table, Distance table between
            thematic layers
   Output: Spatial Context SC)
1  $Layerlist \leftarrow \emptyset$
2  foreach $(i = 1; TLT! = \emptyset; i++)$ do
3      $Layerlist \leftarrow AddNewLayer(L_i)$
4  end
5  return Layeralist
6  foreach $(j = 1; Layerlist! = \emptyset; j++)$ do
7      $SpatialObjects \leftarrow DataIntegration(LayerList(L_i)$
8  end
9  return SpatialObjects
10 foreach $(k = 1; Layerlist! = \emptyset; j++)$ do
11     foreach $(l = k; Layerlist >= 0; l++)$ do
12         $SpatialPredicateTable \leftarrow$
           $Addpredicate(LayerList(k)), LayerList(l), distance(LayerList(k)$
13         $, LayerList(l), Operator, SpatialPredicate, Percentage)$
14     end
15 end
16 return SpatialPredicateTable
17 foreach $(l = 1; SpatialPredicateTable! = \emptyset; l++)$ do
18     foreach $(m = l + 1; Size > 0; m++)$ do
19         $Distance \leftarrow$
           $CalculDistance(SpatialPredicateTable(l)), SpatialPredicateTable(m)$
20         $Predicate(SpatialObjectTable(l)) \leftarrow$
           $PredicateAffectation(Distance, SpatialPredicateTable)$
21     end
22     $SC \leftarrow$
       $SpatialObjectTable(l), SpatialObjectTable(m), Distance, Predicate(Spatial$
       $ObjectTable(l))$
23 end
24 return $SC$
```

3.4 Research Area and Data Sources

To verify this approach's practicability, we used spatial association rules mining, in which we present an approach to extract frequent predicates from geographic

Algorithm 2: Frequent predicate set mining

```
Input  : Spatial Context, Minimum support
Output: FrequentPredicateSet ($L_k$)
$L_k \leftarrow \emptyset$
foreach ($k = 1; L_{k-1} ! = \emptyset; k + +$) do
    $C_k \leftarrow Apriori - gen(L_{k-1})$
    foreach $t \in SC$ do
        $C_t \leftarrow Subset(C_k, t)$
    end
    foreach $c \in C_t$ do
        $C.count + +$
    end
    $L_k \leftarrow \{condidatein(C_k)/c.count >= supmin\}$
end
return $L_k$
```

data related to road accidents. The data were inferred from reports provided by the Ministry of Equipment, Transport, and, Logistics of Morocco (METL) [35, 36]. To identify the main factors that contribute to road crashes, five factors and nineteen attributes were used [37], see Table 2 for more details.

The factors describe attributes related to the accident (Type, Cause, Time, Number of deaths, Number of injuries), a driver (Age, Experience), vehicle, environment (Light condition, Road geometry, weather conditions). In this study, we defined three thematic layers, Road, Territory (regions), and Institutions (Fig. 4, Table 3), which contains the spatial objects we focused on. We defined a TFN according to the fuzzy predicate in Table 1 by giving the choice to users to define the distance between different objects of layer and affect the appropriate predicate (Table 4).

For example, OR001 is a spatial object that belongs to the thematic layer 'Road', etc. After the algorithm computes the distance between different objects of thematic layers (Fig. 5) using Euclidean distance, we affect the appropriate predicate to the related spatial object (Table 4). Each thematic layer is described by a set of spatial and non-spatial attributes (Table 5 for road, Table 6 for territory, Table 7 for institution). Once the distances are computed, the algorithm links each thematic layer with others by affecting the appropriate predicate according to the fuzzy distances (F.distance).

3.5 Evaluation and Results Analysis

Once the predicates are associated with each spatial object, we use the proposed algorithm to extract frequent predicate set (see Algorithm 2). This algorithm illustrates the predicate by frequency (Fig. 6). Moreover, we used arulesViz [38] to visualize the extracted rules. Besides, the results have shown sensitivity towards the minimum

Table 2 Factors contributing to road accidents

Attribute name	Values	Description
Accident_ID	Integer	Accident ID
Accident_Type	Fatal, Injury, Property damage	Accident type
Driver_Age	$\prec 20$, [21–27], [28–60] $\succ 61$	Driver age
Driver_Sex	M, F	Driver sex
Driver_Experience	$\prec 1$, [2–4], $\succ 5$	Driver experience
Vehicle_Age	[1–2], [3–4], [5–6] $\succ 7$	Service year of the vehicle
Light_Condition	Daylight, Twilight, Public lighting, Night	Light condition
Weather_Condition	Rain, Fog, Wind, Snow	Weather condition
Road_Condition	Highway, Ice Road, Collapse Road, Unpaved Road	Road condition
Road_Geometry	Horizontal, Alignment, Bridge, Tunnel	Road geometry
Road_Age	[1–2], [3–5], [6–10], [11–20] $\succ 20$	Road age
Time	[00–6], [6–12], [12–18], [18–00]	Accident time
City	Marrakesh, Casablanca, Rabat...	City
Particular_Area	School, Market, shops...	Particular area
Season	Autumn, Spring, Summer, Winter	Seasons of year
Accident_Causes	Alcohol effects, Fatigue, loss of Control, Speed, Brake Failure ...	Accident causes
Number_of_injuries	1, [2–5], [6–10] $\succ 10$	Injuries number
Number_of_death	1, [2–5], [6–10] $\succ 10$	Number of deaths
Victim_Age	$\prec 1$, [1–2], [3–5] $\succ 5$	Victim Age

support we have introduced in the first step of the algorithm *(minsupp = 0.2)*. The second step is to extract spatial association rules by using the minimum confidence *(minconf = 0.5)*; the result is given in Table 8.

In the previous study [3, 11, 31], if the distance between two spatial objects *OI001* and *OR001* is equal 10 m, we can say that the accident in road *OR001* is inside to institution *OI001*. Nevertheless, if the distance equals 12 m, we cannot say that the accident is inside *OI001*, this imprecision can be handled by using fuzzy logic. As explained, the use of fuzzy logic can produce two different rules: If the accident occurs in the distance $d = 12$ meters between the object *OI001* and the object *OR001*, we can extract two spatial association rules: The accident is close to *OI001* with the percentage of 10%, and the accident is inside *OI001* with the percentage of 90%".

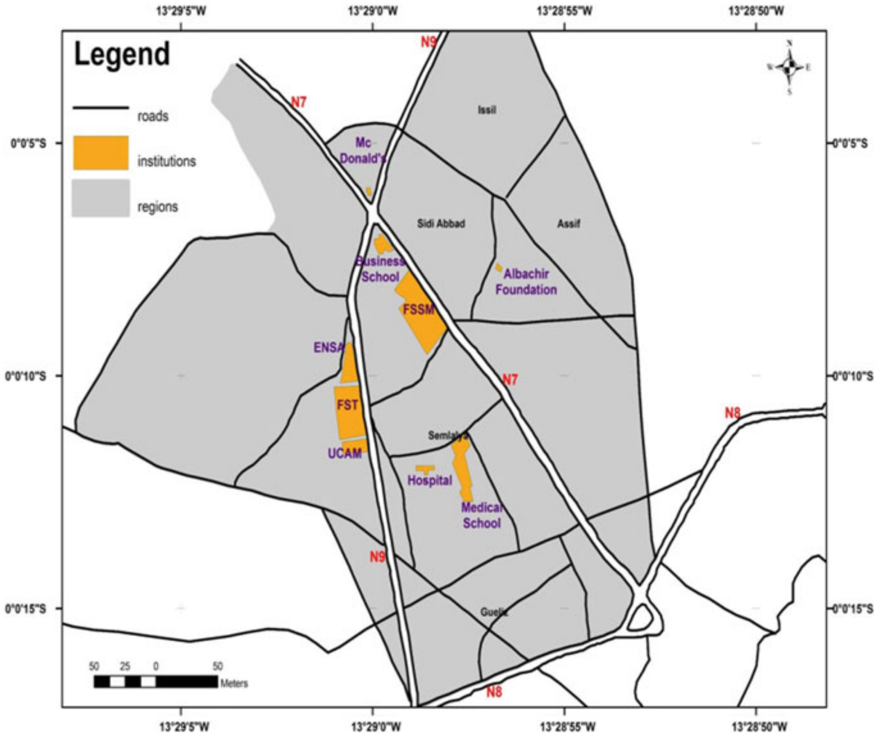


Fig. 4 Geographical map with thematic layers (Road, Territory, Institution)

To evaluate spatial association rules, Support (S), Confidence (C), and Lift (L) have been used. In the association rules extraction the accuracy measure does not make sense and does not provide any evaluation. However, it is possible to use a certainty factor (CF) [39] to evaluate rules. It is an alternative of accuracy measure that can be adapted to spatial association rules. The certainty factor is a measure of the variation of the probability that B is in a transaction when only considering transactions with A. The formula of CF is given in 5:

$$CF(A \rightarrow B) = \begin{cases} \frac{C(A\rightarrow B)-S(B)}{1-S(B)} \; if \; C(A \rightarrow B) > S(B) \\ \frac{C(A\rightarrow B)}{S(B)} \; otherwise \end{cases} \tag{7}$$

We consider an association rule $A \rightarrow B$ as strong when its support and CF are greater than thresholds $minsupp$ and $minCF$, respectively. The evaluation of top extracted rules according to the decision-makers preferences are given in Table 8.

The literature presents a number of research projects that study road accidents and their main causes. Chong et al. [31] introduced a decision tree based approach to analyze the severity of road accidents. They found that the main causes behind these accidents were alcohol, seat belts, and lack of lighting. Sohn et al. [33] com-

Table 3 Table of thematic layers and spatial objects

Predicate ID	Predicate	Triangular fuzzy numbers
P01	Inside	(0,10,15)
P02	Close	(10,15,20)
P03	Intersect	(15,20,25)

Table 4 Definition of predicates based on fuzzy distance

Road	Territory	Institution	Predicate	F. distance
OR001	OT001	OI001	P01	(0,10,15)
OR001	OT002	OI002	P02	(10,15,20)
OR001	OT002	OI001	P03	(15,20,30)
OR002	OT003	OI002	P01	(15,20,30)

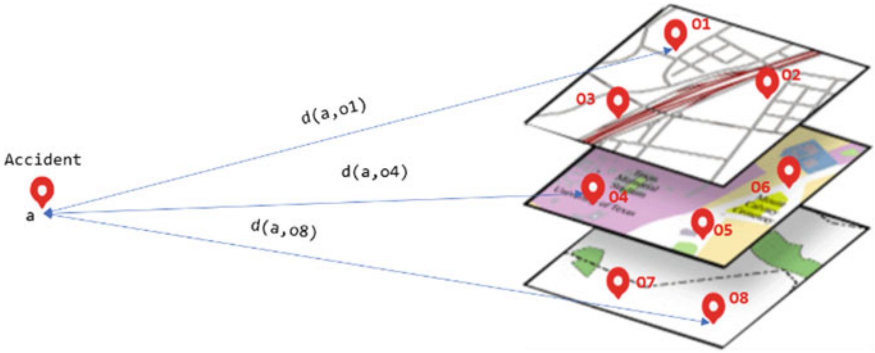


Fig. 5 Computation of metric distance between objects

Table 5 Road thematic layer objects

Object ID	Form	Geometry	Age
OR001	Line	Horizontal	[6–10]
OR002	Line	Alignment	[3–5]
OR003	Line	Bridge	[6–10]

Table 6 Territory thematic layer objects

Object ID	Form	Population
OT001	Polygon	5000
OT002	Polygon	10,000
OT003	Polygon	15,000

Table 7 Institution thematic layer objects

Object ID	Form	Type	Age
OI001	Polygon	School	[6–10]
OI002	Polygon	Bank	[3–5]
OI003	Polygon	Faculty	[10–20]

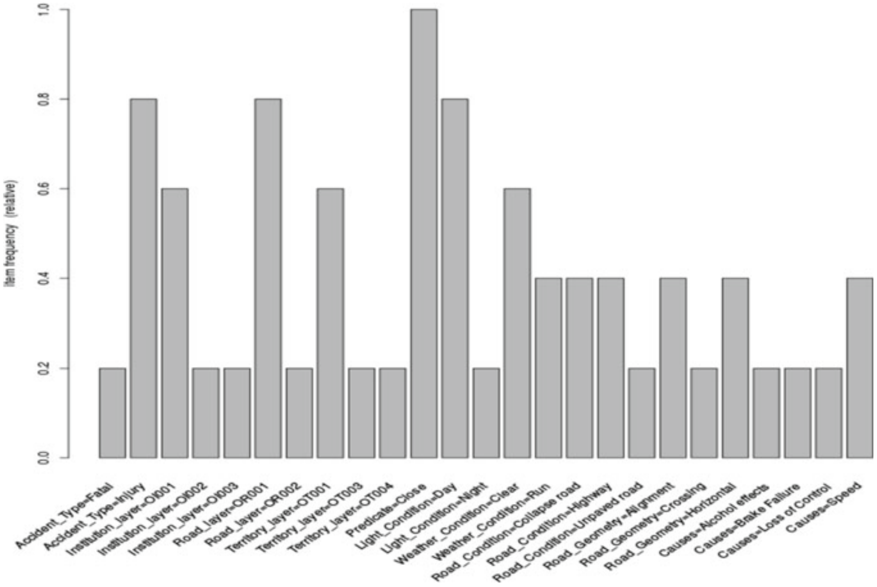


Fig. 6 Frequent predicates using Aprori algorithm

pared three algorithms (Neural Networks, Logistic Regression, and Decision Trees) to study the severity of road traffic in Korea. Chang and Wang [12] analyzed the severity of road accidents by using non-parametric classification tree techniques. Wong and Chang [34] studied the main factors behind the severity of accidents and found that the most dangerous cases come from the lack of driving experience, as well as drinking. Ait-Mlouk et al. [37, 40, 41] proposes an approach to association rule mining-based Multiple Criteria Analysis for a road accident. This list of work is not exhaustive, road accidents have become a real tragedy, and many researchers contribute to finding solid solutions. In [21] the obtained association rules indicate a close correlation between the terms 'Normal Surface of Road' and 'Normal Atmospheric Condition', 'Driver' and 'Pedestrian', as well as 'Pavement' and 'Pedestrian'. This correlation can explain that dangerous accident can come from excess speed and drivers' carelessness. As good as these results can be, there's still a vagueness into them. This comes from the lack of information related to the distance between the accident location and other spatial objects from different thematic layers.

Table 8 Top 20 extracted spatial association rules

Id	Spatial rule	S	C	L
1	$\{Institution\_layer = OI001 \rightarrow Light\_Condition = Day\}$	0.6	1.0	1.25
2	$\{Territory\_layer = OT001 \rightarrow Road\_Layer = OR001\}$	0.6	1.0	1.25
3	$\{Institution\_layer = OI001, Predicate = Close \rightarrow Light\_Condition = Day\}$	0.6	1.0	1.25
4	$\{Territory\_layer = OI001, Predicate = Close \rightarrow Road\_Layer = OR001\}$	0.6	1.0	1.25
5	$\{Predicate = Close, Weather\_Condition = Clear \rightarrow Road\_Layer = OR001\}$	0.6	1.0	1.25
6	$\{Road\_Layer = OR001 \rightarrow Predicate = Close\}$	0.6	1.0	1.25
7	$\{Predicate = Close \rightarrow Road\_Layer = OR001\}$	0.8	0.8	1.00
8	$\{Light\_Condition = Day \rightarrow Predicate = Close\}$	0.8	1.0	1.00
9	$\{Predicate = Close \rightarrow Light\_Condition = Day\}$	0.8	0.8	1.00
10	$\{Accident\_type = Injury \rightarrow Predicate = Close\}$	0.8	1.0	1.00
11	$\{Predicate = Close \rightarrow Accident\_type = Injury\}$	0.8	0.8	1.00
12	$\{Institution\_layer = OI001 \rightarrow Predicate = Close\}$	0.8	1.0	1.00
13	$\{Territory\_Layer = OT001 \rightarrow Predicate = Close\}$	0.8	1.0	1.00
14	$\{Weather\_Condition = Clear \rightarrow Predicate = Close\}$	0.8	1.0	1.00
15	$\{Institution\_layer = OI001, Light\_Condition = Day \rightarrow Predicate = Close\}$	0.6	1.0	1.00
16	$\{Road\_layer = OR001, Territory\_Layer = OT001 \rightarrow Predicate = Close\}$	0.6	1.0	1.00
17	$\{Road\_layer = OR001, Weather\_Condition = Clear \rightarrow Predicate = Close\}$	0.6	1.0	1.00
18	$\{Road\_layer = OR001, Light\_Condition = Day \rightarrow Predicate = Close\}$	0.6	1.0	1.00
19	$\{Accident\_Type = Injury, Road\_layer = OR001 \rightarrow Predicate = Close\}$	0.6	1.0	1.00
20	$\{Accident\_Type = Injury, Light\_Condition = Day \rightarrow Predicate = Close\}$	0.6	1.0	1.00

In literature, the primary issue of spatial association rules is an insufficient integrated solution, and similar systems are focused on classical association rules and ignored spatial data. This can lead to inconsistent data analysis that involves spatial components. To overcome this issue, we proposed this approach by exploiting the potential of fuzzy set theory and association rules. The literature is rich with studies highlighting the correlation between weather and lighting conditions, drivers' behavior, and accident severity in the context of road accident analysis and association rules. However, the use of fixed values judgment and human preferences to compute the distance between the local where the accident occurs and spatial objects of different thematic layers can be vague and imprecise. Our proposed approach confirms an association between different variables, but the integration of fuzzy logic for spatial predicates computation improves the precision and the quality of spatial association rules by using fuzzy set theory to handle the uncertainty of human preferences. The results of our approach can help to provide accurate results for best decision making that can help to save lives and reduce economic losses.

4 Conclusion

This paper was an introduction to new approach to mine the strongest and most meaningful spatial association rules. Our algorithm uses fuzzy set theory to improve associations rules mining. We applied this algorithm on road accident data to confirm its performance and scale up abilities. The results were encouraging as the algorithm was able to provide proper spacial association rules mining that can help stakeholders improve road safety.

This work can be further improved with different future ideas. For instance, it would be useful to introduce a multi-criteria decision analysis to select only the relevant rules according to the stakeholder preferences. Besides, it would be interesting to extend our approach to a big data structure able to counter issues such as response time and data storage, especially when it comes to spatial data.It can also be interesting to combine this approach with machine learning, natural language processing, and computer vision in order to understand driver behaviors and better optimize road safety.

Data Availability

The data used to support the findings of this study were inferred from reports provided by the ministry of equipment and transportation.

Acknowledgements The authors would like to thank Lili Jiang for her support. Furthermore, we are grateful for ISCL and Uppsala University in Sweden.

References

1. The World Health Organization (WHO). https://www.who.int/home. Accessed on July 2019
2. Frawley W, Piatetsky-Shapiro G, Matheus C (1992) Knowledge discovery in databases an overview. AI Mag 13:57
3. Krzysztof K, Jiawei H (1995) Discovery of spatial association rules in geographic information databases. In: Egenhofer MJ, Herring JR (eds) Proceedings of the 4th international symposium on advances in spatial databases (SSD '95). Springer, London, pp 47-66
4. Han J, Kamber M, Tung AKH (2001) Spatial clustering methods in data mining: a survey. In Miller HJ, Han J (eds) Geographic data mining and knowledge discovery. Taylor and Francis, London, pp 33–50

5. Mennis J, Liu JW (2005) Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. Trans GIS 9(1):5–17
6. Miller H, Han J (2009) Geographic data mining and knowledge discovery: an overview. Geographic data mining and knowledge discovery. CRC Press, Taylor and Francis Group, pp 1–26
7. Ester M, Kriegel HP, Sander J (1997) Spatial data mining: a database approach. Advances in spatial databases. Springer, Berlin, pp 47–66
8. Koperski K, Han J, Stefanovic N (1998) An efficient two-step method for classification of spatial data. In: 1998 international symposium on spatial data handling SDH'98, Vancouver, Canada, pp 45–54
9. Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26:1481–1496
10. Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F (2005) A spacetime permutation scan statistic for the early detection of disease outbreaks. PLoS Med 2:216–224
11. Salleb A (2003) Recherche de motifs fréquents pour l'extraction de règles d'association et de caractérisation. Ph.D. thesis, Orléans University, France
12. Zadeh LA (1965) Fuzzy Set Inf Control 8(3):338e353
13. Aguinis Herman, Forcum Lura E, Joo Harry (2013) Using market basket analysis in management research. J Manag 39:1799–1824
14. Auer Jens, Bajorath Jürgen (2006) Emerging chemical patterns: a new methodology for molecular classification and compound selection. J Chem Inf Model 46:2502–2514
15. Jothi Neesha, Husain Wahidah et al (2015) Data mining in healthcare—a review. Proc Comput Sci 72:306–313
16. Jothi Neesha, Husain Wahidah et al (2016) Steganalysis based on steganography pattern discovery. J Inf Secur Appli 30:3–14
17. Cipriani Ernesto, Nigro Marialisa, Fusco Gaetano, Colombaroni Chiara (2014) Steganalysis based on steganography pattern discovery. Eur Transp Res Rev 6:139–148
18. Sanmiquel Lluís, Rossell Josep M, Vintró Carla (2015) Study of Spanish mining accidents using data mining techniques. Saf Sci 75:49–55
19. Mirabadi Ahmad, Sharifian Shabnam (2015) Application of association rules in Iranian Railways (RAI) accident data analysis. Saf Sci 48:1427–1435
20. Kumar Sachin, Toshniwal Durga (2015) A data mining framework to analyze road accident data. J Big Data 2:1–26
21. Mazouri El, Zahra Fatima, Abounaima Mohammed Chaouki, Zenkouar Khalid (2019) Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France. J Big Data 6:1–5
22. Chan KCC, Au WH (1998) An effective algorithm for discovering fuzzy rules in relational databases. In: Proceedings of the 1998 IEEE international conference on fuzzy systems, pp 1314–1319
23. Hong TP, Kuo CS, Chi SC (1999) Mining association rules from quantitative data. Intell Data Anal 3:363–376
24. Kuok CM, Fu A, Wong MH (1998) Mining fuzzy association rules in databases. SIGMOD Rec 27(1):41–46
25. Lin CW, Hong TP, Lu WH (2010) A two-phase fuzzy mining approach. In: International conference on fuzzy systems (2010)
26. Chun-Hsien C, Li Pheng K, Yih Tng C, Xiao Feng Y (2014) Knowledge discovery using genetic algorithm for maritime situational awareness. Exp Syst Appl 41(6)
27. Faisal S, Sohail A (2015) A fuzzy based scheme for sanitizing sensitive sequential patterns. Int Arab J Inf Technol 12(1)
28. Shweta M, Kanwal G (2013) Mining efficient association rules through apriori algorithm using attributes and comparative analysis of various association rule algorithms. Int J Adv Res Comput Sci Softw Eng 3(6). ISSN: 2277 128X
29. Yasuhiko M (2010) Co-location pattern mining for unevenly distributed data: algorithm, experiments and applications. Int J Comput Sci Eng 5(3/4):185–196
30. Dai C, Chen L (2016) An algorithm for mining frequent closed itemsets with density from data streams. Int J Comput Sci Eng 12(2/3):146–154

31. Marghoubi RA, Boulmakoul A, Zeitouni K (2005) Spatial mining with the Galois lattice for information technologies. In: International conference on modeling and simulation, ICMS05 Marrakech, Morocco, 22–24 Nov 2005, p 86
32. Andor Csaba, Joó András, Mérö László (1985) Galois-lattices: a possible representation of knowledge structures. Eval Educ 9(2):207–215. https://doi.org/10.1016/0191-765X(85)90015-1
33. Chan FTS, Kumar, N (2007) Global supplier development considering risk factors using fuzzy extended AHP-based approach. Omega 35:417e431
34. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8:199e249
35. The Ministry of Equipment Transport and logistics of Morocco (METL). http://www.equipement.gov.ma/en/Pages/home.aspx. Accessed on Jan 2019
36. Ait-Mlouk A, Agouti T, Gharnati F (2017) Mining and prioritization of association rules for big data: multi-criteria decision analysis approach. J Big Data 4:42. https://doi.org/10.1186/s40537-017-0105-4
37. Ait-Mlouk A, Gharnati F, Agouti T (2017) An improved approach for association rules mining using multi-criteria decision support system: A case study in road safety. Eur Transp Res Rev 9(3):1–13. https://doi.org/10.1007/s12544-017-0257-5
38. Hahsler M, Chelluboina S (2011) Visualizing association rules: introduction to the R-extension package arulesViz. R project module
39. Sánchez Daniel, Serrano Jose, Blanco Ignacio, Martin-Bautista Maria, Vila M (2019) Using association rules to mine for strong approximate dependencies. Data Min Knowl Discov 16:313–348
40. Ait-Mlouk A, Agouti T (2019) DM-MCDA: a web-based platform for data mining and multiple criteria decision analysis: a case study on road accident, SoftwareX, vol 10, p 100323, ISSN 2352-7110
41. Ait-Mlouk A, Agouti T, Gharnati F (2017) Mining and prioritization of association rules for big data: multi-criteria decision analysis approach. J Big Data 4:42. https://doi.org/10.1186/s40537-017-0105-4

A Mobile Application for Real-Time Detection of Road Traffic Violations

Tulsi Pawan Fowdur and Girish Luckhun

Abstract The most dominant cause of road accidents is traffic law infringement. The exponential increase in the number of vehicles in several countries has increased the number of traffic rule violations and consequently the number of accidents. Several techniques have therefore been proposed to detect such violations using sensors and cameras. In this chapter, a mobile application has been proposed which warns users about the speed limitations as well as overtaking and parking regulations in a specific region. The mobile application detects the traffic violations pertaining to speeding, unauthorized overtaking and parking in real-time. The application also sends the violation pertaining to the specific driver to a centralized database whereby the authorities can take corrective actions. The mobile app uses the GPS tracker to obtain the exact GPS coordinates of the vehicle. It then sends this location to a cloud database which has all the information pertaining to the road traffic regulations for that location. Once the app receives the road traffic information from the database, it detects the speed of the vehicle using its accelerometer and runs an algorithm to detect overtaking and parking in illegal zones by using the GPS coordinates of the vehicle. After that the app warns the driver of any traffic violations and sends the violations to the cloud if required. The system was tested on a given road in Mauritius and it proved to be very accurate in detecting and recording violations such as exceeding the speed limit on the road, parking in a yellow line zone and overtaking on solid white lines.

Keywords Traffic violations · Road · Real-time · Detection · Cloud · GPS · Mobile

T. P. Fowdur (✉) · G. Luckhun
Department of Electrical and Electronic Engineering, University of Mauritius, Réduit, Mauritius
e-mail: p.fowdur@uom.ac.mu

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_18

1 Introduction

Due to the exponential growth of traffic globally, road traffic management has become a consequential problem in today's society. With the increase in traffic, traffic jams and accidents have considerably increased. This significantly affects the economy of a country [1]. In fact with the rapid development of technologies, vehicles can reach speeds that are considered very unsafe and have caused a lot of harm and death. For this particular reason, different speed limits have been set across the road to reduce damage. Unfortunately, many drivers usually do not take these speed limits and other road traffic regulations seriously. As a result several devices and IoT based techniques have been introduced over the years to improve traffic infringement detection, documentation and prosecution. An overview of some research in this direction is given next.

The authors of [2] developed a system to calculate the speed of a vehicle using a computer and a mounted camera. The proposed system detected the corner of the vehicle's license plate across multiple frames, based on a background subtraction sensitive to the color of the vehicle's license plate. It also extracted the elapsed time from the frame rate and then calculated the speed of the moving vehicles. The vehicle tracking was carried out by classifying the earlier segmented region as a license plate using the Support Vector Machines (SVM) classifier and by extracting the license number using a three-layer Artificial Neural Network (ANN) classifier. To monitor incoming traffic, the camera was placed on the side of the road surface. A speed correction factor was also introduced based on the height of the vehicle above the road surface. The speed detection varied in proportion to the ground speed of the GPS speed meter. Another speed detection system for vehicles was presented in [3]. The system used a video camera, a computer and MatLab software as a low-cost alternative and a more genuine system for radar monitoring systems. Background evaluation was carried out employing the average filter method and subtraction using the combined saturation and value method. The binarized image was applied to edge detection and morphological operations to segment the vehicles in the image and find their bounding boxes. The location of the centroids of the moving vehicles was then traced and the vehicle speeds were computed with the image frame rate. The system could attain correlation speeds with an average error of ±7 km/h.

The work in [4] described the framework and components of the Advanced Driver Assistance System (ADAS) experimental platform. It provided drivers with feedback on the traffic violations they committed during their driving. The test bed consisted mainly of two parts: A computer vision subsystem for traffic sign detection and recognition, which operates day and night, and an Event Data Recorder (EDR) for the recording of data relating to certain traffic violations. A module for traffic sign detection and recognition, which operates both day and night, can detect vertical signs along the road. This module is intended to alert drivers in hazardous situations such as "speeding," "no passing zone," "intersections," "stop signs," "yield give way signs," "dangerous turns," "steep slopes" or "road works." In such situations, the system sends warnings through the vehicle loudspeakers in the form of acoustic messages.

Moreover in [5], the authors proposed a framework for safer driving in Mauritius that uses an on-board car diagnostic module (OBDII) for capturing data such as the average speed of a vehicle, revolution and acceleration of the engine. This module sends the data to a cloud platform where the data is analysed using an adaptive algorithm and performs driver behaviour prediction in real-time. Mobile alerts such as messages, voice commands or beeps could be sent to drivers depending on their behaviour. A survey was also carried out to evaluate the acceptance rate of such a framework by people of different age groups in Mauritius.

Another research by [6] proposed a method for regulating infringements of parking using computer vision. A camera mounted on enforcement vehicles obtained a still color image of the parked vehicle. The acquired image was pre-processed and binarized by a morphological algorithm. The shadow area was extracted from the image, using brightness information, and the shadow of the vehicle was determined using the region's properties. Whether the vehicle was parked illegally or not was determined by the lane type and the comparison of the vehicle's shadow and lane relative position. The result showed that the yellow solid line detection ratio was 95.8% when 47 images were included.

In [7], the system proposed used both hardware (microcontroller as embedded system) and a software rule-based system for smart decision as well as a database for driver violation recording. The system monitored the speed of the cars and would first advise the driver to slow down. However, if the speed of the car is not reduced, it forces the driver to slow down. The system contains an embedded computer that can process the data sent by a server. The embedded computer records violations and immediately sends them back to the traffic police department. The system can be implemented in all types of vehicles (old, modern, private or commercial vehicles, taxis, trucks), by installing a sensor system that receives digital readings from the analog and mechanical meters of the vehicle, and then sends the data to be processed by the mini computer.

The authors of [8] proposed a system that detects and tracks vehicles advancing through the surveillance region. The proposed method uses a single camera to identify, track and withdraw the vehicle speed estimation parameter. The camera is placed on the traffic signal pole about 10 m above the middle of the road. The maximum tracking accuracy achieved by the suggested technique was up to 98.3% in the afternoon period with an average tracking accuracy of around 92.2%.

In [9], the system employed techniques for real-time information detection and warning to detect infringement of traffic speed, as well as to inform police and the car owner of the breach committed. Furthermore, the system could detect whether or not the driver is drunk and ensure that the passengers are wearing seat belts. The system sends a check message to the car that examines the parameters and documentation of the car status. If anything is found to be inappropriate then the vehicle is seized. In conclusion, it was found that the system overcomes the majority of the drawbacks of existing solutions and thus distributes low cost and low maintenance.

In [10], the system used the car movement characteristics, stop lines and traffic signal status, to analyse red signals and also lane changing violation. Once the system detects the violation, it seizes a snapshot of the infringing car and pinpoints the car's license plate number. An android application was developed that provides details of the violation when inputting the vehicle's license plate number by an authorized officer. The violation detection system achieved good performance for lane changing and detecting red signals violation, in terms of accuracy.

The mobile application developed by [11], contained a database that stored the picture, name, address, phone number, bank account number and also the records of previous infringements committed by drivers. All the information are displayed on the traffic police's smart phone. If a driver commits an offence, the traffic police can use the app to automatically acquire the fine from the account of the bank of the driver and post the notification to his/her phone app. However, if the same vehicle was found to be confronted again, a specific action will was taken.

Other related works includes the one in [12] in which the authors developed a real-time cloud based bus tracking system using an Arduino microcontroller with a GPS/GSM shield to record and relay the position of a vehicle being tracked to a cloud server. A control station could then obtain the vehicle position in real-time and applied prediction algorithms to the data to predict the arrival time of the vehicle to a specific destination. Additionally, in [13] the authors proposed two adaptive real-time congestion prediction approaches. These techniques choose between five different prediction algorithms using the Root Mean Square Error model selection criterion. The implementation consisted of a Global Positioning System based transmitter connected to an Arduino board with a Global System for Mobile/General Packet Radio Service shield that relays the vehicle's position to a cloud server. A control station then accesses the vehicle's position in real-time, computes its speed. Based on the calculated speed, it estimated the congestion level and it applied the prediction algorithms to the congestion level to predict the congestion for future time intervals. Finally in [14] the authors analysed the driving styles in both the real world and the simulated world by extracting them via the nonnegative matrix factorization method on the collected Vehicle Sensor Data (VSD). This analysis allowed the driving style differences to be quantitatively interpreted. It was observed that the drivers tend to be more erratic when driving the simulator and the deviation in the perception of temporal gap in the two circumstances was revealed. These findings can be used to calibrate the driving simulator and construct more reliable driving behavior models.

Although many techniques have been developed for the detection of a moving vehicle's speed, tracking of an overtaking car and detection of the yellow line, most of them have used hardware (e.g. camera, sensors) to detect the speed of a car. To the best of our knowledge, to date, no previous works have considered full software-based solution to detect traffic violations. In this paper therefore a mobile application which detects road traffic violations such as over speeding, illegal parking and overtaking, has been developed. The mobile app uses the GPS tracker to first obtain the exact GPS coordinates of the vehicle. The app then sends this location to a cloud database which has all the information pertaining to the road traffic regulations for that location. Once the app receives the road traffic information from the database, it detects the speed of the vehicle using its accelerometer and an runs an algorithm to detect overtaking and parking in illegal zones by using the GPS coordinates of the vehicle. After that the app is able to warn the driver of any traffic violations and send the violations to the cloud if required.

The rest of the chapter proceeds as follows. Section 2 describes the proposed system and its implementation details. Section 3 demonstrates how the system is deployed, tested and discusses the results obtained. Section 4 concludes the chapter.

2 System Model

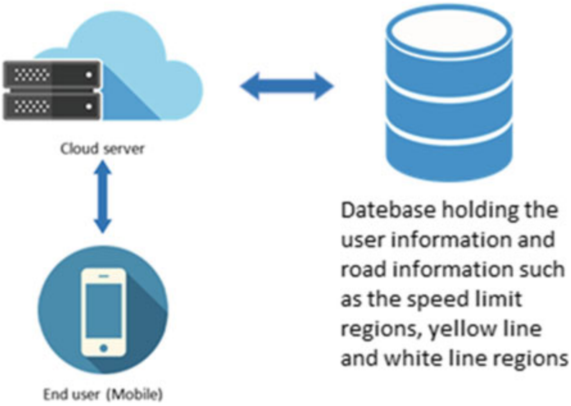
Figure 1 shows the complete system model used for detecting road traffic violations. The mobile app is used for capturing the current location and average speed of the user. This data is then transmitted to a cloud server. If the current location of the user is within a region stored in the database, a text will be displayed on the user's phone indicating that region's speed limit. Furthermore, if the user exceeds the speed limit, the latter is notified and the data is stored in a database that enters the location, the average speed, the start time and the end time of the infringement. In addition to the speed limit, the mobile app will also check whether the user is in a yellow line region or on a single white lane road. If so, a message will be sent to the mobile app to warn the user.

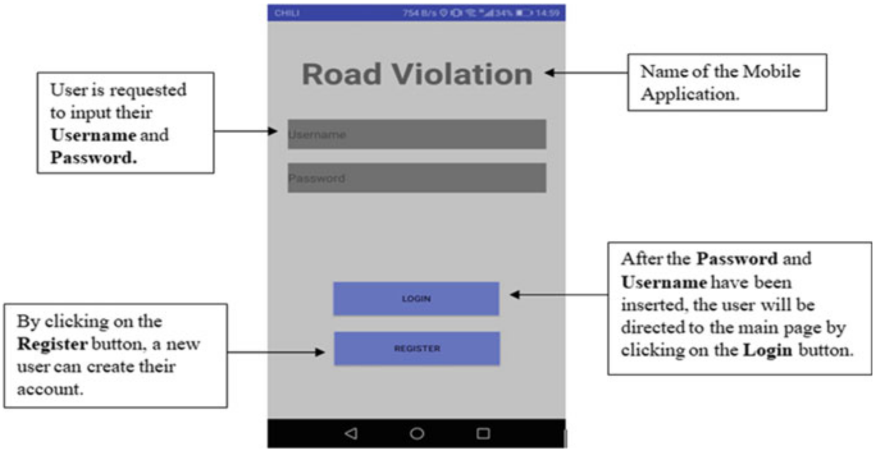
A. The Mobile Application

The mobile application consists of three main layouts. The first layout pertains to the login page whereby the user enters his username and password before accessing the application as shown in Fig. 2.

The database is the one shown in Fig. 1 and more details on its structure will be provided in the next sub-section. The second layout is for users who have not

Fig. 1 Proposed system





Name of the Mobile Application.

User is requested to input their **Username** and **Password.**

After the **Password** and **Username** have been inserted, the user will be directed to the main page by clicking on the **Login** button.

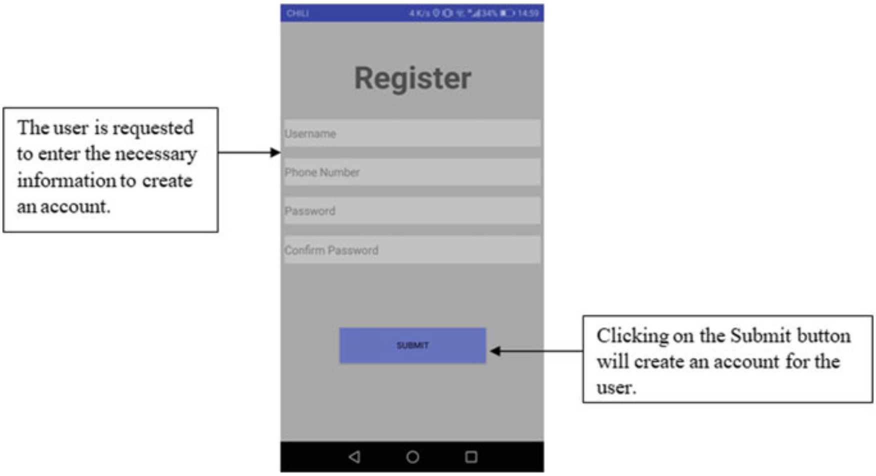
By clicking on the **Register** button, a new user can create their account.

Fig. 2 Login layout

yet registered. When the Register button is pressed, the screen shown in Fig. 3 is displayed.

The cloud server is the one shown in Fig. 1 and more details on its functions will be provided in the next sub-section. The third layout is the Main Application GUI which is triggered once a user has logged in successfully as shown in Fig. 4.

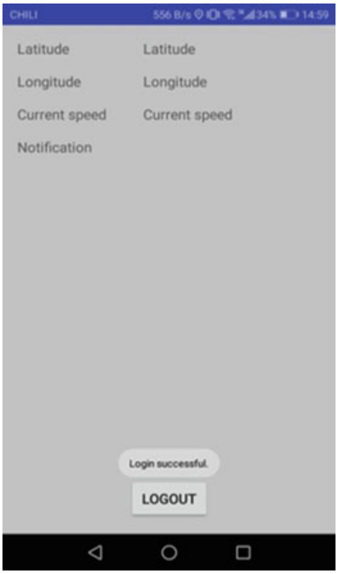
The pseudocode for the main GUI is as follows:



The user is requested to enter the necessary information to create an account.

Clicking on the Submit button will create an account for the user.

Fig. 3 Register GUI



Login successful.

Fig. 4 Main application GUI

```
Send user location to server
Send user current speed to server
Get response from server
if (violation code from server is 0)
        display message for unknown region
else
        display message for road information
if (violation code from server is 1)
        display message for no violation
else if (violation code form server = 2)
        display message for exceeding speed limit
else if (violation code from server is 3)
        display message indicating user stopped in a yellow line region
```

```
else if (violation code from server is 4)
        display message indicating user stopped in a yellow line region with  the time spend in the region
else if (violation code from server is 5)
        display message indicating user has committed a yellow line violation
end if
```

B. The Cloud Server

The cloud server is based on the 00webhost server and is a free and open source cross-platform web server solution stack package, consisting mainly of the Apache HTTP Server, MySQL database, and interpreters for scripts written in PHP. PHP programming is used to manipulate data in the MySQL database. The components of the cloud server are shown in Fig. 5.

B.1 Database Structure

The following six database tables are used in this system:

- Speed Limit and Road Markings table
- User table
- User Violation table
- Yellow Line Violation table

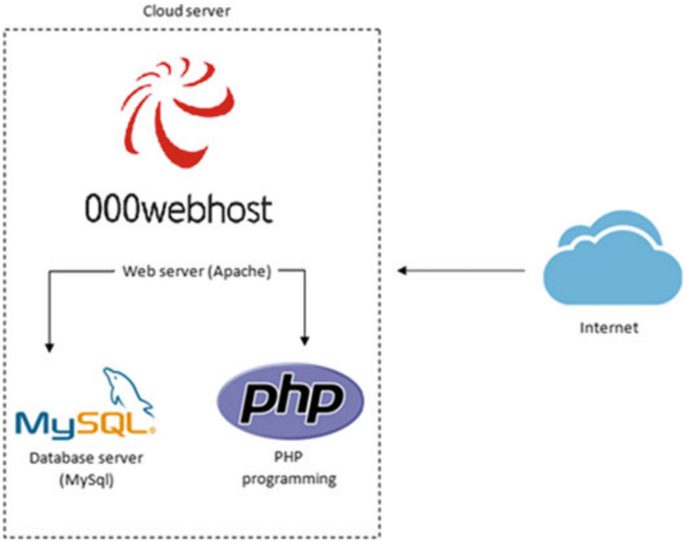


Fig. 5 Cloud server architecture

Table 1 Speed limit and road markings

Field	Type	Description
ROAD_LOCATION_ID(PK)	Integer	Primary key
SPEED_LIMIT	Double	Speed limit of a certain region
LATITUDE_BOTTOM	Double	Latitude of point 1
LONGITUDE_BOTTOM	Double	Longitude of point 1
LATITUDE_TOP	Double	Latitude of point 2
LONGITUDE_TOP	Double	Longitude of point 2
YELLOW_LINE_Y_N	Integer	Yellow line on road
WHITE_LINE_Y_N	Integer	White line on road

- Non overtaking zone location table
- Overtaking violation table

Table 1 is used to store the speed limits for a given region as well as information on the types of road markings present. The ROAD_LOCATION_ID field is the primary key that provides a unique identification number for each record. SPEED_LIMIT is the speed limit of a certain region. The LATITUDE_BOTTOM, LONGITUDE_BOTTOM, LATITUDE_TOP and LONGITUDE_TOP fields store the GPS coordinates of the region. The YELLOW_LINE_Y_N field indicates whether or not there is a yellow line in that region. The WHITE_LINE_Y_N field shows whether or not there is a single white line in the area.

Table 2 stores the user's details. The user must enter his/her information via the mobile app. USER_ID is the primary key that provides each record with a unique identification number. PHONE_NUMBER and EMAIL are the user's mobile number and email respectively. PASSWORD is a character string that provides access to the service.

Table 3 is used to record the Speed violation done by a user. It contains a foreign key to the User information table (Table 2) as well as the Speed Limit and Road Marking table (Table 3) so as identify the region where the violation occurred. VIOLATION_ID is the primary key, AVG_SPEED is the user's average speed, START_TIME is the time the violation occurred, STOP_TIME is the time the violation ended. To determine whether a new record must be added or updated, the FINAL field is used.

Table 4 shows the yellow line violation table. It is similar to the speed violation table except that it is used to record yellow line violations.

Table 2 User information table

Field	Type	Description
USER_ID(PK)	Integer	Primary key
PHONE_NUMBER	Integer	Phone number of the user
PASSWORD	String	Password of the user

Table 3 Speed violation table

Field	Type	Description
VIOLATION_ID	Integer	Primary key
USER_ID	Integer	User_ID of the person
ROAD_LOCATION_ID	Integer	Location of the committed violation
AVG_SPEED	Double	Average speed of the person
START_TIME	Datetime	Start time of the violation
STOP_TIME	Datetime	End time of the violation
FINAL	Boolean	Check if new record needs to be added or updated

Table 4 Yellow line violation table

Field	Type	Description
VIOLATION_ID	Integer	Primary key
USER_ID	Integer	User_ID of the person
ROAD_LOCATION_ID	Integer	Location of the committed violation
LATITUDE	Double	Latitude of yellow line
LONGITUDE	Double	Longitude of yellow line
START_TIME	Datetime	Start time of the violation
STOP_TIME	Datetime	End time of the violation
FINAL	Boolean	Check if new record needs to be added or updated

Table 5 is used to record non-overtaking zones in a given region i.e. a region with solid white lines in the middle of the road. ROAD_LOCATION_OVERTAKING_ID is the primary key, ARR_LATITUDE records all the latitudes in an array and these correspond to latitudes A, B, C and D shown in Fig. 6. ARR_LONGITUDE registers all the longitudes in an array corresponding to longitudes A, B, C and D in Fig. 6. LEFT_Y_N determines whether the vehicle is moving in the left side of the road.

Figure 6 demonstrates how the road is divided into two sections. The first row (row 1) indicates that the vehicle is in the lane 1, row 2 indicates that the vehicle is in the lane 0. The road area is bounded by four corners. Each latitude and longitude pertaining to a given segment of the road is saved in the database using Table 5 to identify non overtaking zones.

Table 5 Non overtaking zone location

Field	Type	Description
ROAD_LOCATION_OVERTAKING_ID	Integer	Primary key
ARR_LATITUDE	Varchar	Array of latitude
ARR_LONGITUDE	Varchar	Array of longitude
LEFT_Y_N	Integer	Direction of the moving car

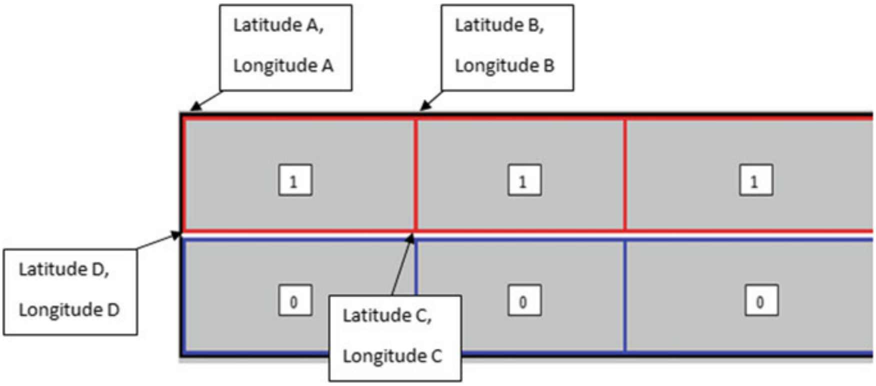


Fig. 6 Extracting coordinates from the road

When the user overtakes in a non-overtaking region, Table 6 is used to record the violation committed. **VIOLATION_ID** is the primary key of the table. **USER_ID** shows which user has committed the violation. **ROAD_LOCATION_OVERTAKING_ID_START** records the time when the violation started and **ROAD_LOCATION_OVERTAKING_ID_END** records the time when the violation stops. **LEFT_Y_N** indicates in which lane the vehicle was moving. **OVERTOOK Y_N** shows that an overtaking violation has occurred and **FINAL** indicates the end of the violation.

B.2 PHP Server Functions Implementation

The PHP server functions were implemented using the following files:

(i) **dbConnection.php**

Table 6 Overtaking violation

Field	Type	Description
VIOLATION_ID	Integer	Primary key
USER_ID	Integer	User_ID of the person
ROAD_LOCATION_OVERTAKING_ID_START	Integer	Start of overtaking violation
ROAD_LOCATION_OVERTAKING_ID_END	Integer	End of overtaking violation
LEFT_Y_N	Integer	Direction of the moving car
OVERTOOK_Y_N	Double	Overtaking taking place
FINAL	Boolean	Check if new record needs to be added or updated

This function is used to establish a connection to the server so that data from the MySQL database can be accessed. To create the connection, the username, password and database name must be entered. If a connection error occurs, an error message will be sent to the client. The pseudocode for this function is as follows:

```
Define the host of the database
Define the username of the database
Define the password of the database
Define the name of the database
Create the Connection
If Connection is invalid
    Display Error message:  "Connection error!"
end if
```

(ii) **login.php**

To authenticate the mobile user, the login.php script is used. The script runs after the user login button is pressed. The pseudocode for this script is as follows:

```
Get username typed by the user
Get the password typed by the user
if username and password do not match
    Print out invalid login unsuccessful
end if
if username matches but password does not match
    Print out Incorrect password
end if
if both match
    Start the session
end if
```

(iii) **register.php**

When an account is created by a user from the mobile app, this script checks if the entered username is unique and then sends it to the database. The pseudocode for this script is as follows:

```
Get username typed by the user
Get Phone number typed by the user
Get the password typed by the user
if username already exists in the database
    Print out Username already exists.
else
    Print out Username has been added correctly.
end if
```

(iv) **model.php**

All queries are saved in this script.

(v) **server.php**

This script contains the main application logic which determines the interaction between the server and the mobile application. For instance, when the server receives data from the mobile app, it checks whether the user is in a speed limit region. If so, it sends the user a message. The following pseudocode describes the functions of this script.

```
Get user speed from client
Get user id from client
Get user's last speed violation record from database
Get user's last yellow line violation record from database
Get recorded regions from database
if user's location is within recorded region from database then
    if user is committing a speed violation then
        if user's speed > speed limit in region then
            Update violation record in database - update time
            Send violation 2
        else
            Update violation record in database - update time
            and indicate to stop recording current violation
            if user's speed is 0 and is in a yellow line region then
            Insert new yellow line violation
            violation code 3
        else
            violation code 1
        end if
    end if
else
    if user speed > speed limit in region then
        Insert new speed limit violation record
        if user was committing a yellow line violation then
            if user stopped in region for more than 10 mins then
                Save violation in database
            else
                delete violation from database
            end if
        end if
        Violation record 2
    else if user speed == 0 then
      if user is committing a yellow line violation then
        if user is in same location since yellow line started then
                Update violation record - set new time
          if user is committing violation for more than 10 mins then
                    violation code 5
                else
                    violation code 4
                end if
            else
            if user stopped in region for more than 10 mins then
                    save violation in database
                else
                    delete violation
                end if
                violation code 1
            end if
        else
            if current region is in a yellow line region
                 Insert a new yellow line violation record
               in database to start a new violation
                violation code 3
            else
                violation code 1
            end if
        end if
    else
        if user was committing a yellow line violation then
         if user committed the violation for more than 10 min then
```

```
                    insert same violation in database
                else
                    delete violation
                end if
            end if
                violation code 1
        end if
    end if
else
    if user if committing a speed limit violation then
        save violation in database
    end if
    if user was committing a yellow line violation then
        if user committed violation for more than 10 mins then
            save violation in database
        else
            delete violation in database
        end if
    end if
    violation code 0
end if
```

(vi) **Overtaking.php**

The logic of this script can be illustrated and explained using Figs. 7, 8 and 9. When the vehicle enters a region of solid white line (non-overtaking zone) as shown in Fig. 7, the mobile app will indicate that the user is in a non-overtaking area. If the user overtakes another car i.e. it crosses the white lines as shown in Fig. 8 and enters the opposite lane; the mobile app will flag a violation send the necessary information to the database. Ultimately, the mobile app will stop the violation when the user re-enters its correct lane as shown in Fig. 9.

The pseudocode for the overtaking.php script is as follows:

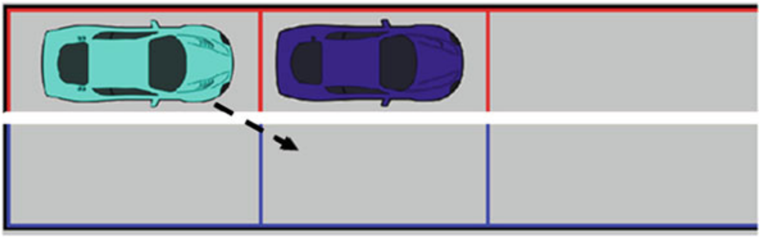


Fig. 7 Cars in respective lane

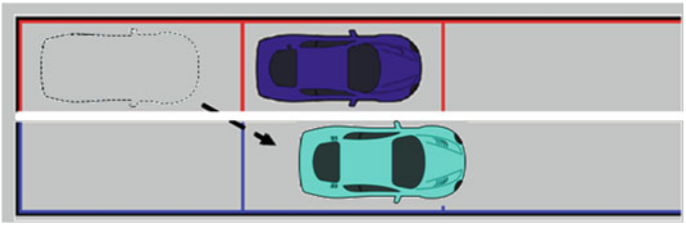


Fig. 8 Car committing a violation

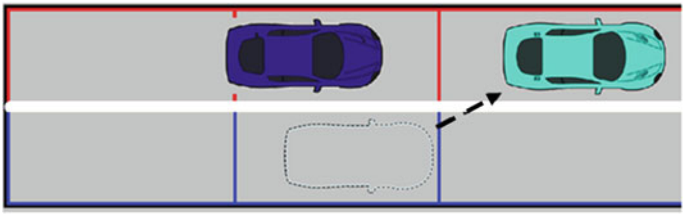


Fig. 9 Car returning to respective lane

```
Get user’s current latitude
Get user’s current longitude
Get user’s current speed
Get user’s id
Get user’s last record in database
if (current speed > 0) then
    Get region recorded in database
    if user’s current location is within a recorded region)
        if (user’s location was being recorded)
            if (user changes lane)
                Update existing location indicating user is committing a violation in database.
                Send violation code 1 to client
            else
                Update existing location
                Send violation code 0 to client
            end if
        else
            Insert a new location record to record the lane in which the user is
            Send violation code 0 to client
        end if
    else
        Send violation code 0 to client
    end if
else
    send violation code 0 to client
end if
```

3 System Testing and Results

The system is first tested for its ability to detect the three road traffic violations defined previously. After that some validation testing is performed to determine the reliability of the system.

3.1 Testing the System's Operation

A. *The Main Application Interface*

Figure 10 shows the mobile application's main layout when it is lanunched. The user's Latitude and Longitude are shown, followed by the user's Current Speed. The speed limit of the region is also displayed. In addition, if it is a yellow line or white line region, the application will let the user know.

If the mobile app is not connected to the internet, a message will be displayed as shown in Fig. 11.

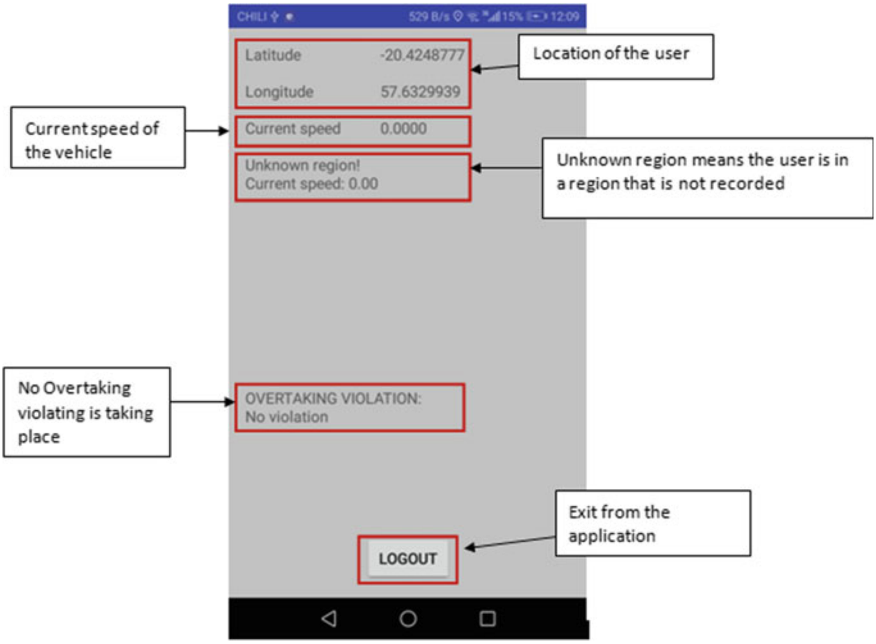
B. *Speed Limit Violation*

When the user enters a region with a speed limit, the speed limit of the area, speed of the user as well as an indication whether a speed limit violation is being committed by the user will be shown as in Fig. 12.

For the sake of testing, the speed limit of the given area has been set to 0 km/h. Figure 13 shows the speed of the area that was inserted in the database which is 0 km/h. This speed limit has been used only for testing purposes and can be set to any realistic value.

Figure 14 shows the area in which the speed limit is being violated.

When the user commits a speed limit violation, the Speed Violation table records data as shown in Fig. 15.



Latitude -20.4248777
Longitude 57.6329939
Current speed 0.0000
Unknown region!
Current speed: 0.00
OVERTAKING VIOLATION:
No violation

Location of the user

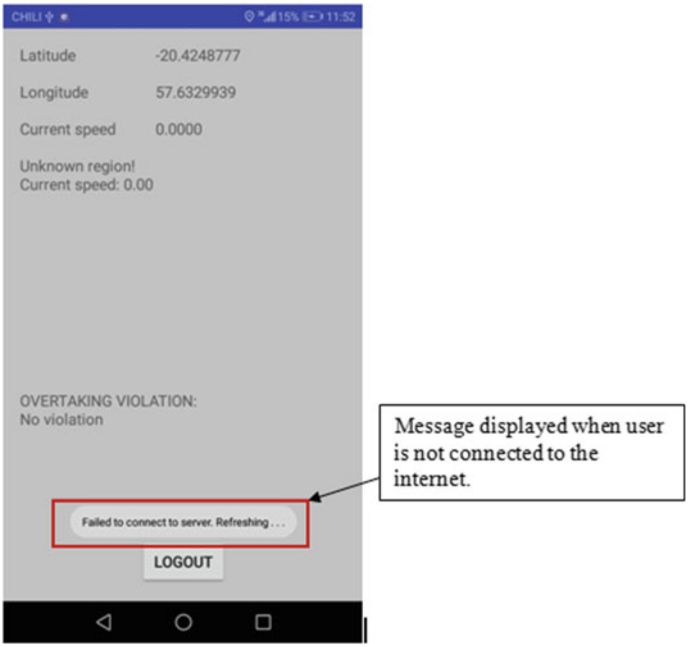
Current speed of the vehicle

Unknown region means the user is in a region that is not recorded

No Overtaking violating is taking place

Exit from the application

Fig. 10 Main UI of the application



Latitude -20.4248777

Longitude 57.6329939

Current speed 0.0000

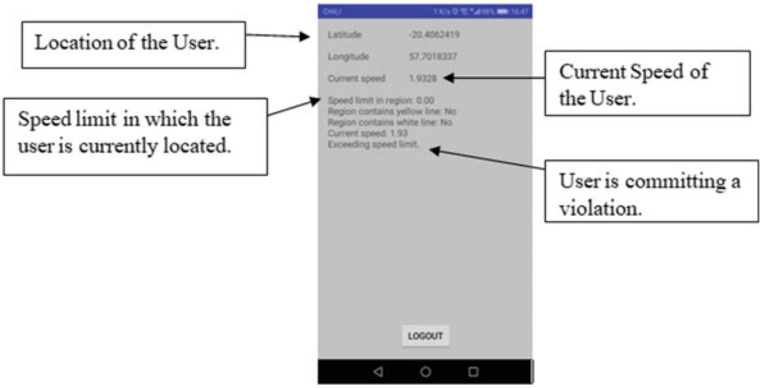
Unknown region!
Current speed: 0.00

OVERTAKING VIOLATION:
No violation

Failed to connect to server. Refreshing . . .

Message displayed when user is not connected to the internet.

Fig. 11 Error message when not connected to the internet


Location of the User.

Speed limit in which the user is currently located.

Latitude -20.4062419

Longitude 57.7018337

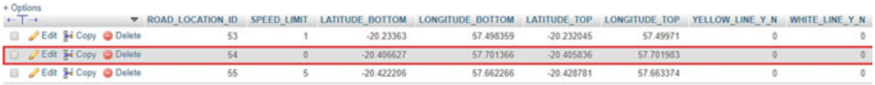
Current speed 1.9328

Speed limit in region: 0.00
Region contains yellow line: No
Region contains white line: No
Current speed: 1.93
Exceeding speed limit.

Current Speed of the User.

User is committing a violation.

Fig. 12 User committing a speed limit violation



	ROAD_LOCATION_ID	SPEED_LIMIT	LATITUDE_BOTTOM	LONGITUDE_BOTTOM	LATITUDE_TOP	LONGITUDE_TOP	YELLOW_LINE_Y_N	WHITE_LINE_Y_N
□ Edit Copy Delete	53	1	-20.23363	57.498359	-20.232045	57.49971	0	0
□ Edit Copy Delete	54	0	-20.406627	57.701366	-20.405836	57.701983	0	0
□ Edit Copy Delete	55	5	-20.422206	57.662266	-20.428781	57.663374	0	0

Fig. 13 Database of the speed limit location

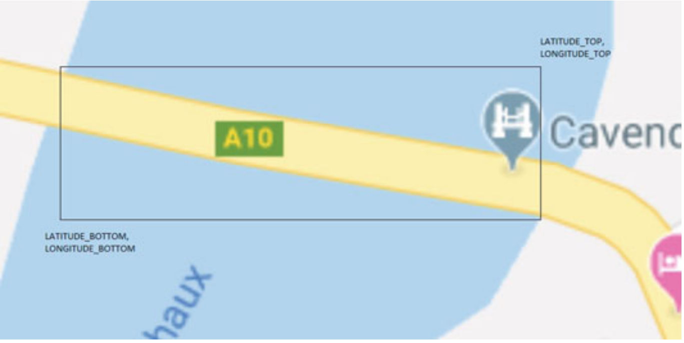
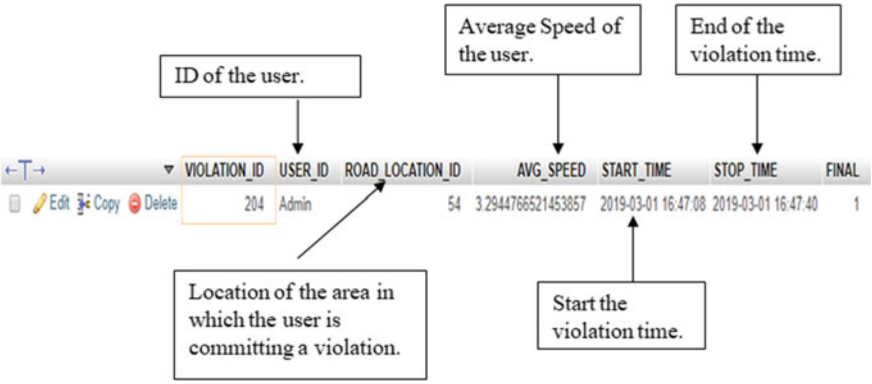


Fig. 14 Location of the speed limit


ID of the user.
Average Speed of the user.
End of the violation time.
Location of the area in which the user is committing a violation.
Start the violation time.
Edit
Copy
Delete

	VIOLATION_ID	USER_ID	ROAD_LOCATION_ID	AVG_SPEED	START_TIME	STOP_TIME	FINAL
	204	Admin	54	3.2944766521453857	2019-03-01 16:47:08	2019-03-01 16:47:40	1

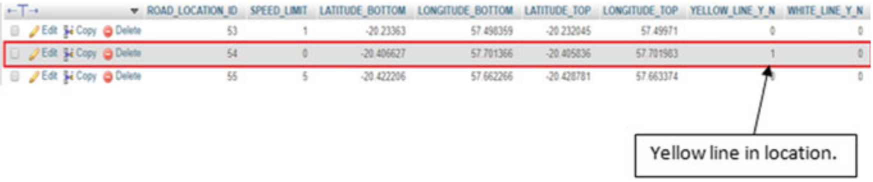
Fig. 15 Data saved when a speed violation has occurred

C. *Yellow Line Violation*

When the user stops on a yellow line, the application warns the user that the area is a non-parking area. If the user has not moved after 10 min, a violation will occur. Figure 16 shows the location of the yellow line.

Figure 17 shows a car in non-parking zone.

Figure 18 shows the application when the user commits a yellow line violation.



ROAD_LOCATION_ID	SPEED_LIMIT	LATITUDE_BOTTOM	LONGITUDE_BOTTOM	LATITUDE_TOP	LONGITUDE_TOP	YELLOW_LINE_Y_N	WHITE_LINE_Y_N
53	1	-20.23363	57.498359	-20.232045	57.49971	0	0
54	0	-20.406627	57.701366	-20.405836	57.701983	1	0
55	5	-20.422206	57.662266	-20.428781	57.663374	0	0

Fig. 16 Yellow line locations in database

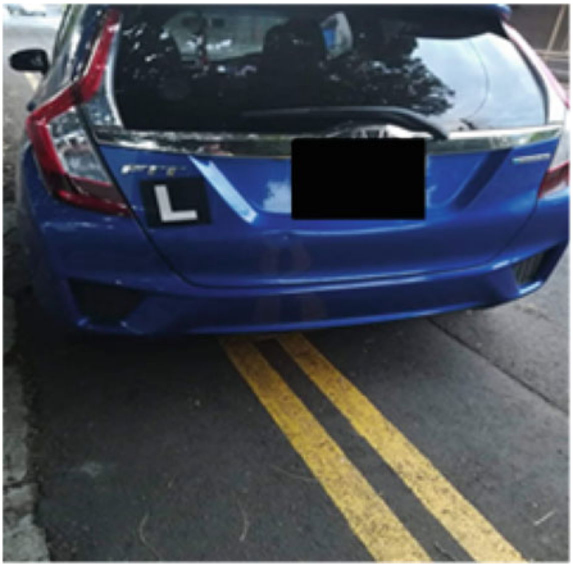
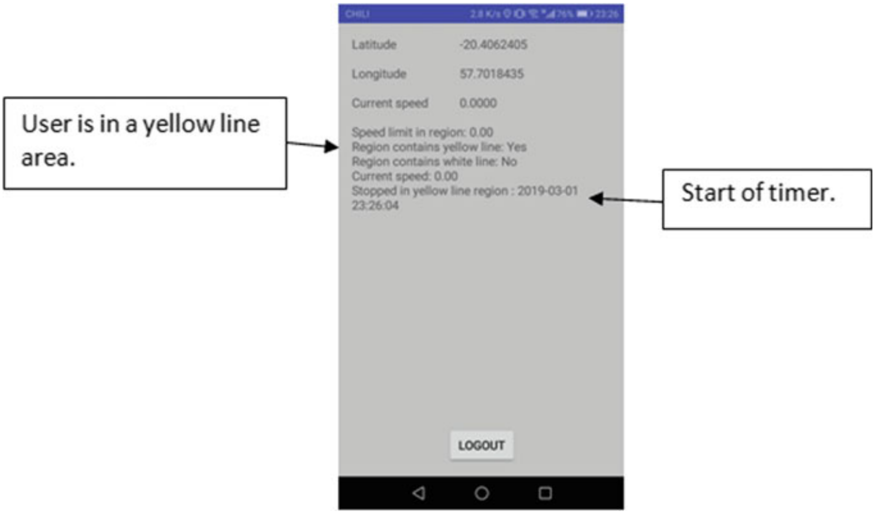

[redacted]

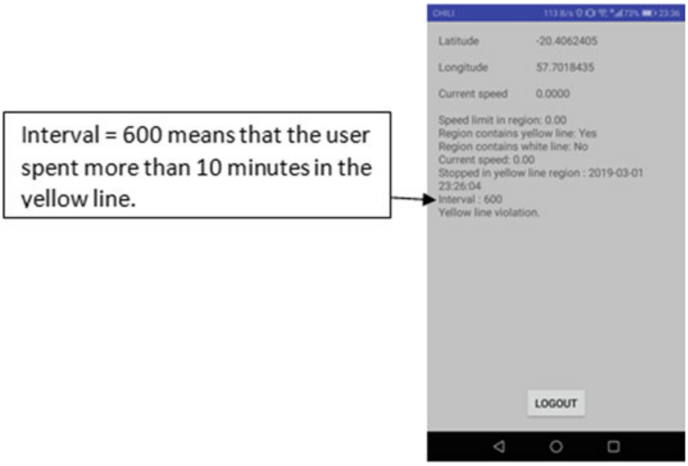
Fig. 17 Car in a yellow line location



```
Latitude          -20.4062405

Longitude         57.7018435

Current speed     0.0000

Speed limit in region: 0.00
Region contains yellow line: Yes
Region contains white line: No
Current speed: 0.00
Stopped in yellow line region : 2019-03-01
23:26:04
```

Fig. 18 User in a yellow line zone

When the interval reaches 600 s (10 min), the application warns the user that a yellow line violation has occurred as shown in Fig. 19, and the data is saved in the database.



Interval = 600 means that the user spent more than 10 minutes in the yellow line.

Latitude -20.4062405
Longitude 57.7018435
Current speed 0.0000

Speed limit in region: 0.00
Region contains yellow line: Yes
Region contains white line: No
Current speed: 0.00
Stopped in yellow line region : 2019-03-01 23:26:04
Interval : 600
Yellow line violation.

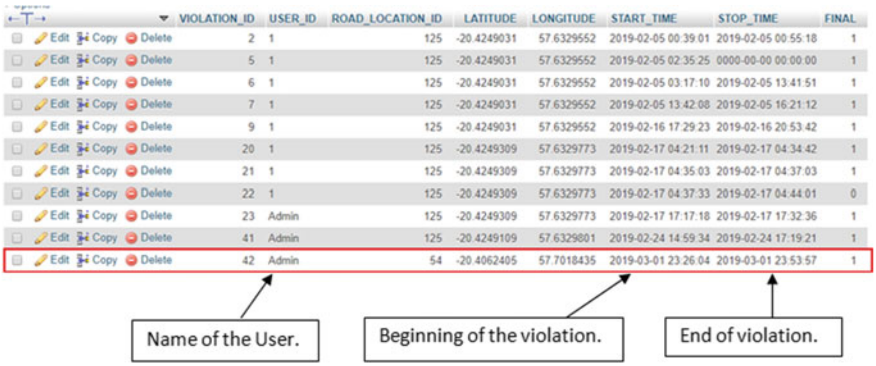
Fig. 19 User committing a yellow line violation

Figure 20 shows how the yellow line violation is recorded in the database, which includes the user's name, the location of the violation and the time.

D. *Overtaking on White Line*

Figure 21 shows a particular region where there is a solid line in the middle of the road indicating that overtaking is not allowed. The road is divided into two rectangular or square sections so as to get the latitude and longitude of the corners. The latitude and the longitude are saved in the database respectively as shown in Fig. 22.

When the user enters a solid white line region, the application will inform the user as shown in Fig. 23. When the user crosses the white line to overtake a vehicle, the user is warned by the application that there has been an overtaking violation as shown in Fig. 24.



←T→	VIOLATION_ID	USER_ID	ROAD_LOCATION_ID	LATITUDE	LONGITUDE	START_TIME	STOP_TIME	FINAL
□ Edit Copy Delete	2	1	125	-20.4249031	57.6329552	2019-02-05 00:39:01	2019-02-05 00:55:18	1
□ Edit Copy Delete	5	1	125	-20.4249031	57.6329552	2019-02-05 02:35:25	0000-00-00 00:00:00	1
□ Edit Copy Delete	6	1	125	-20.4249031	57.6329552	2019-02-05 03:17:10	2019-02-05 13:41:51	1
□ Edit Copy Delete	7	1	125	-20.4249031	57.6329552	2019-02-05 13:42:08	2019-02-05 16:21:12	1
□ Edit Copy Delete	9	1	125	-20.4249031	57.6329552	2019-02-16 17:29:23	2019-02-16 20:53:42	1
□ Edit Copy Delete	20	1	125	-20.4249309	57.6329773	2019-02-17 04:21:11	2019-02-17 04:34:42	1
□ Edit Copy Delete	21	1	125	-20.4249309	57.6329773	2019-02-17 04:35:03	2019-02-17 04:37:03	1
□ Edit Copy Delete	22	1	125	-20.4249309	57.6329773	2019-02-17 04:37:33	2019-02-17 04:44:01	0
□ Edit Copy Delete	23	Admin	125	-20.4249309	57.6329773	2019-02-17 17:17:18	2019-02-17 17:32:36	1
□ Edit Copy Delete	41	Admin	125	-20.4249109	57.6329801	2019-02-24 14:59:34	2019-02-24 17:19:21	1
□ Edit Copy Delete	42	Admin	54	-20.4062405	57.7018435	2019-03-01 23:26:04	2019-03-01 23:53:57	1

Name of the User.

Beginning of the violation.

End of violation.

Fig. 20 Data is stored in the database when user commits a yellow line violation

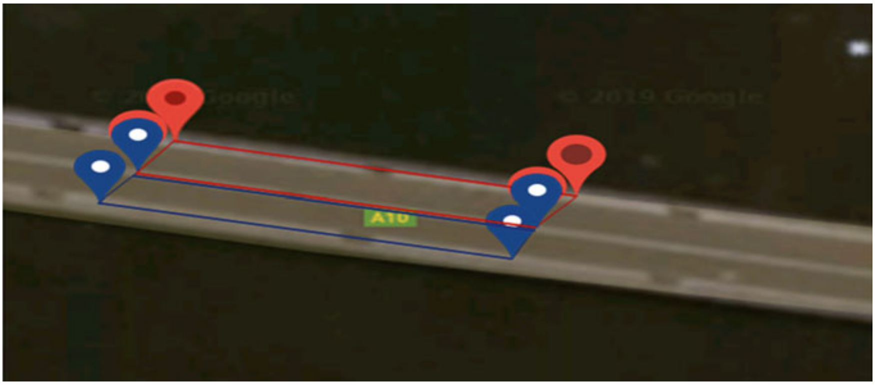
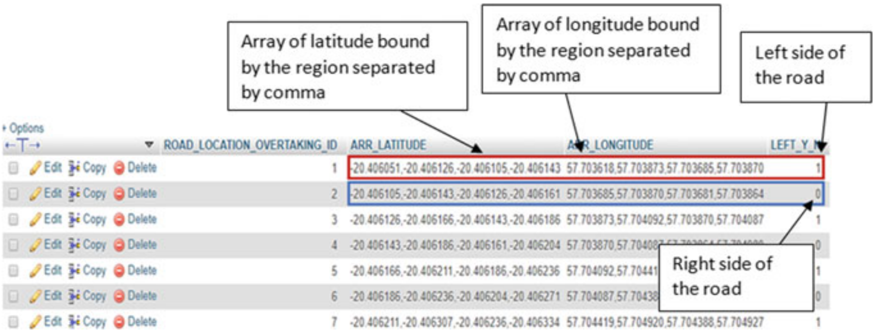


Fig. 21 Retrieving coordinates from road


Array of latitude bound by the region separated by comma

Array of longitude bound by the region separated by comma

Left side of the road

Right side of the road

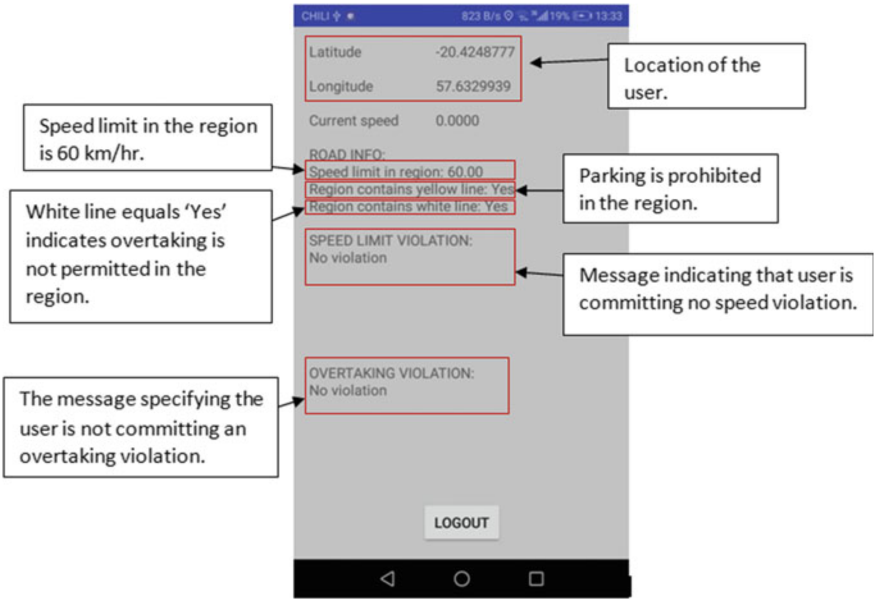
ROAD_LOCATION_OVERTAKING_ID	ARR_LATITUDE	ARR_LONGITUDE	LEFT_Y_N
1	-20.406051,-20.406126,-20.406105,-20.406143	57.703618,57.703873,57.703685,57.703870	1
2	-20.406105,-20.406143,-20.406126,-20.406161	57.703685,57.703870,57.703681,57.703864	0
3	-20.406126,-20.406166,-20.406143,-20.406186	57.703873,57.704092,57.703870,57.704087	1
4	-20.406143,-20.406186,-20.406161,-20.406204	57.703870,57.70408	0
5	-20.406166,-20.406211,-20.406186,-20.406236	57.704092,57.70441	1
6	-20.406186,-20.406236,-20.406204,-20.406271	57.704087,57.70438	0
7	-20.406211,-20.406307,-20.406236,-20.406334	57.704419,57.704920,57.704388,57.704927	1

Fig. 22 Storing the Location of solid white lines on the road

This violation is then recorded in the database as shown in Fig. 25.

3.2 *Testing the Application's Performance in Detecting Speed Limit and Overtaking Violations*

The mobile application was tested in 50 different cases whereby the speed limit was both respected and violated. In all 50 cases, the application successfully detected violation and non-violation of the speed limit for the specific region. Similarly for overtaking, 50 different cases which consisted of instances where overtaking was allowed and also not allowed were tested. The application again proved to be able to correctly detect overtaking violations and non-violations, as long as there was no interruption in the connectivity and GPS service.



Latitude -20.4248777

Longitude 57.6329939

Current speed 0.0000

ROAD INFO:
Speed limit in region: 60.00
Region contains yellow line: Yes
Region contains white line: Yes

SPEED LIMIT VIOLATION:
No violation

OVERTAKING VIOLATION:
No violation

Location of the user.

Speed limit in the region is 60 km/hr.

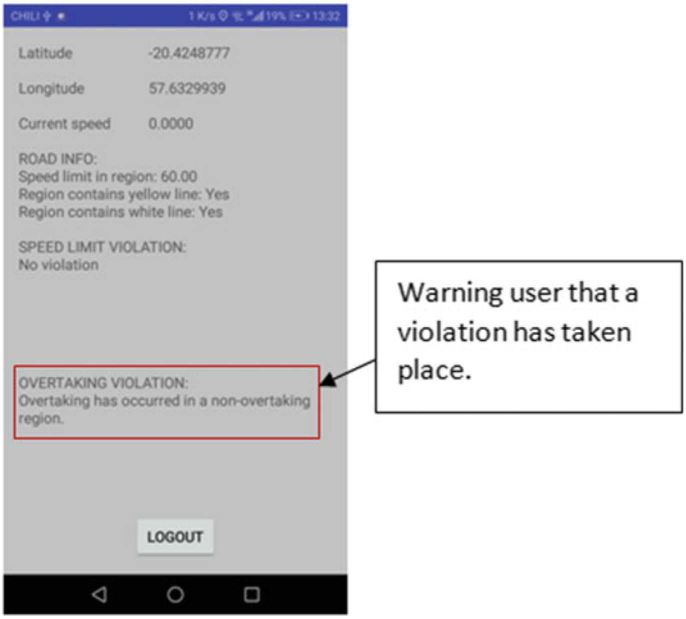
Parking is prohibited in the region.

White line equals 'Yes' indicates overtaking is not permitted in the region.

Message indicating that user is committing no speed violation.

The message specifying the user is not committing an overtaking violation.

Fig. 23 User in a non-overtaking region



Latitude -20.4248777

Longitude 57.6329939

Current speed 0.0000

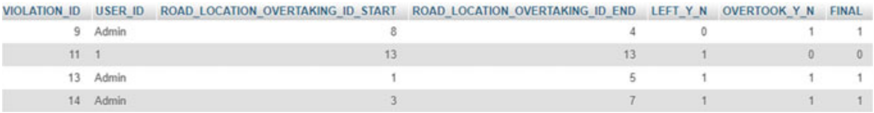
ROAD INFO:
Speed limit in region: 60.00
Region contains yellow line: Yes
Region contains white line: Yes

SPEED LIMIT VIOLATION:
No violation

OVERTAKING VIOLATION:
Overtaking has occurred in a non-overtaking region.

Warning user that a violation has taken place.

Fig. 24 User committing an overtaking violation in a solid white line region



VIOLATION_ID	USER_ID	ROAD_LOCATION_OVERTAKING_ID_START	ROAD_LOCATION_OVERTAKING_ID_END	LEFT_Y_N	OVERTOOK_Y_N	FINAL
9	Admin	8	4	0	1	1
11	1	13	13	1	0	0
13	Admin	1	5	1	1	1
14	Admin	3	7	1	1	1

Fig. 25 Information saved when an overtaking violation has occurred

4 Conclusion

The aim of this work was to design a system that could detect traffic violations such as exceeding the speed limit on the road, parking in a yellow line zone and overtaking on solid white lines. The system consisted of a mobile application and a cloud server. The mobile application was used to monitor and transmit the speed and GPS locations of the vehicle as well as receiving road traffic information from the server. The server has a database which included several tables to provide the mobile application with information on traffic regulations in a particular area depending on the GPS location of the vehicle. The system could also be used to assist drivers by warning them of the road traffic rules in a specific region. The warnings are issued in the form of messages in the mobile application and give a predefined time limit to the driver to react to the current traffic situation. If the driver does not respond to the warning, a traffic breach is committed, and is recorded in a database. The system was tested in a real life scenario on a given road in Mauritius and it proved to be very accurate in detecting and recording these three violations. A viable future work could be to extend the capabilities of this system to detect other traffic violations as well as drivers with unsafe driving states such as drowsiness.

Acknowledgements The authors are thankful to the University of Mauritius for providing the necessary facilities to conduct this research.

References

1. Atubi AO, Gbadamosi KT (2015) Global positioning and socio-economic impact of road traffic accidents in Nigeria: matters arising. Am Int J Contemp Res 5(5)
2. Ginzburg C, Raphael A, Weinshall D. A cheap system for vehicle speed detection. https://arxiv.org/abs/1501.06751
3. Gholami A, Dehghani A, Karim M (2010) Vehicle speed detection in video image sequences using CVS method. Int J Phys Sci 5(17):2555–2563. http://www.academicjournals.org/IJPS
4. Aliane N, Fernandez J, Mata M, Bemposta S (2014) A system for traffic violation detection. Sensors 14(11):22113–22127
5. Bassoo V, Hurbungs V, Ramnarain-Seetohul V, Fowdur TP, Beeharry Y (2017) A framework for safer driving in Mauritius. Fut Comput Inf J 2(2):125–132
6. Kim AR, Rhee SY, Jang HW (2016) Lane detection for parking violation assessments. Int J Fuzzy Logic Intell Syst 16(1):13–20
7. Elmahalawy AM (2014) A car monitoring system for self recording. J Traffic Log Eng 2(3):164–171

8. Kumar T, Singh Kushwaha D (2017) Traffic surveillance and speed limit violation detection system. J Intell Fuzzy Syst 3761–3773
9. Angel MT, Aishwarya M, Freny V, Livya G (2017) Traffic violation detection system. J Emerg Technol Innov Res 4(03)
10. Chaudhari P, Yawle R, Chaudhari P (2016) Traffic violation detection and penalty generation system at a street intersection. In: Proceedings of the international conference on data engineering and communication technology, ICDECT 2016, vol 1, pp 799–807
11. Komal K, Labhesh G, Punam D, Pooja S, Agrawal D (2017) Android based traffic rules violation detection system "Vehitrack." Int J Comput Appl 163(8):1–4
12. Fowdur TP, Khodabacchus MN (2017) Performance analysis of a real-time cloud based bus tracking system with adaptive prediction. IJMEC 7(25):3454–3473
13. Nadeem MK, Fowdur TP (2018) Performance analysis of a real-time adaptive prediction algorithm for traffic congestion. J Inf Commun Technol 17(3):493–511
14. Qi G, Guan W, Li X, Hounsell N, Stanton NA (2019) Vehicle sensor data-based analysis on the driving style differences between operating indoor simulator and on-road instrumented vehicle. J Intell Transp Syst Technol Plann Oper 23:144–160

A Vision Towards an Artificial Intelligence of Medical Things

IoT Based Health Monitoring System and Its Challenges and Opportunities

Mohammad Nuruzzaman Bhuiyan, Md. Masum Billah, Dipanita Saha, Md. Mahbubur Rahman, and Mohammed Kaosar

Abstract With the incarnation of novel COVID-19, health care is getting more preference in each country. IoT-based health monitoring systems might be the best option to monitor infected patients and be helpful for elderly population. In this paper, analyzed different IoT-based health monitoring systems and their challenges. Searched through established journal and conference databases using specific keywords to find scholarly works to conduct the analysis. Investigated unique articles related to this analysis. The selected papers were then sifted through to understand their contributions/research focus. Then tried to find their research gap and challenges, created them into opportunities and proposed a GSM-based offline health monitoring system that will conduct with the healthcare providers through communication networks. Hopefully, this model will work as an absolute pathway for the researchers to establish a sustainable IoT-based health monitoring system for humankind.

Keywords IoT · Health monitoring · GSM module · Sensors · Analytical model · Medical services

M. N. Bhuiyan (✉) · D. Saha
Noakhali Science and Technology University, Noakhali 3814, Bangladesh
e-mail: nuruzzaman.iit@nstu.edu.bd

D. Saha
e-mail: dipanita.iit@nstu.edu.bd

Md. Masum Billah
American International University Bangladesh (AIUB), 408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
e-mail: billah.masumcu@aiub.edu

Md. Mahbubur Rahman
Islamic University, Kushtia 7003, Bangladesh

M. Kaosar
Murdoch University, 90 South Street, Murdoch, WA 6150, Australia
e-mail: mohammed.kaosar@murdoch.edu.au

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_19

1 Introduction

We all know health is wealth. Like all developed countries the whole world is trying day by day to facilitate a sustainable healthcare system. But whenever any pandemic situation like Covid-19 arises in the world, all our achievements face losses. Thus, the world fell into a great crisis. The technological development of IoT can protect the human world in this disaster a lot. But still the world didn't accept this IoT technology for various reasons. Doctors, nurses, hospitals, and the related stockholders are prescribing their patients in an old style even in a crisis moment. As a result, they themselves are affected with these diseases. In it, many frontline fighters gave their lives in this COVID-19.

IoT enabled health Monitoring System can analyze and appropriate for rural areas over a long distance and reduce quality service costs. The use of IoT reduces hospital service time and travel time of patients with chronic diseases; hence, it makes the service more efficient and affordable for both the patients and clinical professionals. We describe how healthcare technologies can provide for the nation's different societal levels. m-health interaction methods promote great potential practice for public health and clinical.

According to [1], "In most of the rural and urban areas the healthcare system is not good enough. However, a substantial dimension of the common and lower-middle classes become inadequate healthcare resources in urban areas." Health care facility is not in hand reach distance for all. So, in general the patients visit the physician's office for his treatment. In other cases, some doctors give treatment to their patients online. In both cases doctors and patients face few problems. Internet facility is one of the big problems. Even if you have the internet facility, bandwidth is very low. For real time monitoring Bandwidth is very important. It is challenging to intend a healthcare monitoring system, including numerous network design, embedded design, data analysis design, modeling, verification, and monitoring. Another thing is all the systems are not user friendly. Security and trust are another issue.

A Health monitoring system is a human–machine interaction that is the entirety regarding some most suitable applicants as measurement. To simplify healthcare, stretch to the limit, raise complexity as a barrier to faster decision-making, reduce variability, and integrate care delivery into emerging technologies. Revolve around the ability to monitor patients and provide the best care possible.

Rural health experience vocalized data access limitations, including time, privacy, inadequate access, lack of facilities, skills, costs, and insufficient Internet infrastructure. Among those developing communication tools and media, the patient is active over a messaging tool and can accept the messages non-availability internet to achieve information access.

To solve these problems, many researchers, doctors, nurses, and stockholders have proposed various models to monitor their patients in real-time, though these monitoring systems have some imperfections. Best of our knowledge, most of the system is not smart enough. Even most of the systems have their own challenges. So, in this paper, we have analyzed most of the relevant health monitoring systems and tried to solve their problems and proposed a new model which will work both in online and offline mode which will really help the human. The rest of this paper is organized as follows. "Related works" section describes different health Monitoring Systems. "Challenges and Opportunities" section discussed the challenges they themself faced to develop or proposed the system. "Proposed Model" section we have designed our model and explained how researchers can utilize our system. "Conclusion" Section concludes the paper with its future works.

2 Related Works

Today many countries are growing significant experience with a healthcare monitoring system in hospitals and health centers, and the use of portable healthcare monitoring systems with emerging technologies are becoming a worldwide issue. We can ensure face-to-face consulting to telemedicine facilitates in healthcare systems by the apparition of the Internet of Things (IoT) technologies. IoT uses sensors and networks to connect computers to the internet. For health monitoring, the sensors then forward the information to distant locations like M2M, which are machinery for computers, machines for people, handheld devices, or smartphones [2] that is a flexible and more convenient way of tracking and optimizing care to any health problem. To ensure a modern life for the human being, the modern system provides a flexible interface [3], assistant devices [4], and mental health management [5].

For continuous monitoring of human health, the two most significant indicators are heart rate and body temperature. To facilitate the patients, it is necessary to ensure good room conditions in the hospitals that include room humidity, level of all gases like CO and CO_2, and that can be able to determine the quality of room environment [6]. It is suggested that the room humidity should be between 30 and 65% and free of toxic gases. To establish a sustainable healthcare system many researchers, contribute a lot. Frequency based S-band was developed by [7]. Non-invasive breathing monitoring system was developed for diabetics patients [8].

A smart framework and automatic evaluation for elderly people's dependencies was proposed by [9]. Intelligent medical box was proposed by [10], IoT based health monitoring system for soldiers were proposed by [11], context aware mobility supported health monitoring system were proposed by [12], for IoT devices a suitable blockchain based novel framework were modified by [13], For security reason a blockchain model was proposed by [14], To reduce security task a collaborative security model proposed by [15], the components of health monitoring like smart cities design presents by [16]. Internet based health monitoring systems a general outline for challenges and opportunities were reviewed by [17], CO and CO_2 sensor-based health monitoring systems were developed by [18]. By using IoT and RFID tags an effective healthcare monitoring system was designed by [19]. Basic symptoms for health monitoring systems were developed by [20]. Health monitoring kit for the IoT environment was developed by [21].

Based on non-invasive technique pulse rate detection was proposed by [22]. Smart phone-based health monitoring systems were introduced by [23]. A fully functional cardiovascular disease sensing system was developed for smartphones by [24]. IoT based adaptive safety monitoring devices were developed by [25].

3 Challenges and Opportunities

3.1 Technical Challenges and Opportunities

3.1.1 Device Reliability

Results of health monitoring depend on the system design of the device. The device needs to be designed with time management in mind. So, all the companies that manufacture health devices need to build machines capable of delivering their healthcare.

3.1.2 Internet Connectivity

The Internet connection simultaneously promises to improve quality and access to health care while reducing its cost. For this purpose, the internet connection plays a notable role in the monitoring system of healthcare. Many technical and non-technical limitations available to applications. For example, the Internet needs to be improved to provide bandwidth and latency in video counselling and remote surgery. Internet connection must be connected to protect against the internet from being disrupted due to a lack of connection.

3.1.3 Security and Privacy

The highly personal nature of health information and the detrimental effects of improper disclosure of such information can affect social status. The level of protection required for some health information is extremely high. This type of protection must be provided through the relevant protocols embedded in the network as well as the computers on the network and the computers connected to the Internet. The rules governing the proper disclosure of information as it will be a technical security system such as encryption and decryption methods. Health data is quite expensive in the world because, with all this data, doctors or physicians can research any disease. You must pay attention to the system information's privacy so that no one can enter by an authorized user.

3.1.4 Data Management

Data analysis and data management are intertwined. Physicians analyze the conducted data and data management needs to do for data analysis advanced healthcare services; data management tools can be used to manage the data.

3.1.5 Data Connectivity

Bandwidth speeds, artificial intelligence, and various API usage need to be increased, and data transmission limits need to be reduced to transfer health information accurately.

3.1.6 Internet Disruptions

In IoT, during the experiment with the effectiveness of the medical IoT software, the required concern is an internet connection if we want to enable it, including the deal with loads, network bandwidth, latency, and metrics in different applications. In healthcare providers, internet access through several wireless technologies such as 3G/4G/5G WiMAX. Internet Disruption (ID) supports killing the systems that can control the unexpected loads, and users get the best efficiency for accessing the enabled IoT healthcare systems.

3.1.7 Bandwidth

In biomedical analysis, the bandwidth requirement is highly considered for teleconferencing, transferring the high-resolution of real-time images. Internet of Things healthcare (IoTh) monitoring system that transmits data and precise patterns is towards enhancing the reliability on the internet. It is difficult to foretell the long-term demands in the healthcare monitoring system. Many medical data can be collected at very high rates throughout the world, which is combined with traffic increases into individual research centers for collecting large-scale data from medical system organism's communities [26].

3.1.8 System Design

System infrastructure is an integral part of healthcare systems. It is challenging to intend a healthcare monitoring system, including numerous network design, embedded design, data analysis design, modeling, verification, and monitoring. Healthcare challenges now face a clinical organization's reach to remove complexity as a barrier to more effective decision-making, reduce variability, and care delivery integrates emerging technologies with IoT systems.

System design tools to transmit and other embedded monitors share the same interface to overcome complexity in a comprehensive patient monitoring solution that works with the current IoT system. It is a challenge to investment, while risk-based defense in-depth security protects the privacy of patients and the integrity of data works by customizing and standardizing the patient to monitor the solution to the hospital with protocols. Health monitoring is a total solution designed around our challenges and support to help achieve the monitoring goals confidently to deliver quality care everywhere.

3.1.9 Security, Privacy, and Trust Issues

Security, privacy, and trust issues of the IoT systems are important as systems consist of various types of body sensors to store patients' and physicians' personal information. In [27], a serious risk to patients' health and their private information can occur due to the unauthorized access of IoT sensor devices. As it increases the threat to users' security and privacy. Cloud is used to capture, aggregate, process, and transfer medical information with the help of medical and mobile devices and the layer of these devices is vulnerable to tag cloning, spoofing, RF jamming, and cloud polling.

In cloud polling, a device directly accepts the command from the traffic through a man-in-the-middle attack. A common barrier to Denial of service (DoS) attacks is a redundancy (the use of multiple devices on the network), which can affect healthcare systems and patient safety in health systems. IoT systems need large numbers of complex software and hardware, and the fast detection of potential security threats remains a challenge. In [28], the wearable devices and available powerful search engines like Shodan enable locating Internet-connected devices despite the lack of security standards of these devices. These devices are now vulnerable to attacks in [29]. In [28], we need to enforce security and privacy to protect the centralized datasets of personal and family medical records and genomic data from hackers and malicious software.

3.1.10 Data Management

The importance of healthcare data management is increasing day by day as the rising of an aging population and chronic patient conditions require new ways of monitoring and managing patient healthcare data. As IoT combines a network of intelligent devices with big data analytics, healthcare providers and patients will contend with new data management issues. IoT also mobilizes patient data on the premises and puts available data into the hands of both patients and healthcare providers. The relationship between big data analytics and IoT was explained in [30] where Various IoT data analytic types like Real-time analytics, Memory-level analytics, BI analytics, massive analytics, Offline analytics were discussed. Connectivity, Storage, Quality of services, Real-time analytics, and Benchmark are needed for data analytics in IoT as suggested in [31].

A functional framework that identifies the acquisition, management, processing, and mining areas of IoT big data is introduced in [32]. Data Acquisition and Integration Module, Data Storage Module, Data Management Module, Data Processing Module, Data Mining Module, Application Optimization Module were discussed. And the new concerns are
Who will own the data? The patient/The provider/The healthcare institution/The data-collection device developer? Who will ensure its accuracy? And how and where will data be stored and who will pay for that service?

3.1.11 Data Connectivity

Health-care definitions are complicated, and metrics are constantly changing in the health-care sector. Clean, formatted, and precise data collected from the healthcare system is facing challenges every day [33]. They are facing a problem related to the data connection. For instance, an important financial measure called length of stay (LOS) was introduced in [34] that was also used by clinicians. If users do not know the explanation of the metric that was indicated, and which metric needs to use then the rating of LOS may vary, and thought may be skewed. How long a patient physically stays in bed is determined by clinicians to calculate LOS. But a financial company calculates LOS on a 24-h scale that ends at midnight.

As a result, an interruption can occur during the data recording. Integration of big data in the health-care system is a challenging issue due to the complexity of data. When we need to update the frequent information such as the health variables of a patient, we will not need to update some passive information like geographic location and contact information. It's necessary to maintain data integrity while updating information because improper data control may pose a risk to data integrity. According to HIPAA regulations rules in [35], maintaining these databases is now challenging because of maintenance costs.

3.2 Social Challenges and Opportunities

3.2.1 Financial

If the patient must take IoT healthcare service, he must bear a little more cost rather than the physical system. Everyone cannot bear this cost to provide IoT healthcare service; among all, device cost must be reduced. It is necessary to have the latest and most advanced equipment in the health sector. But finding the capital is often a struggle. In healthcare, new technology solutions may improve patient outcomes, increase operating efficiencies, and add new service lines. More and more, it's about offering hospitals managed solutions. Advanced electronic medical records (EMRs) can provide better cost-effectiveness and a great profit margin by decreasing the health system's risk, improving treatment decisions, continuity of online care, and decreasing patient admissions is introduced in [36]. A pre-paid, value-based medicine model is more effective, as opposed to the more common fee-for-service model. In [37], suggested that the economic challenges healthcare providers face right now are:

- Managing investment in a capital-constrained environment.
- Accessing digital innovation and technological transformation.
- Adapting to market forces and regulation and compliance requirements.

In [38], the hospital chief financial officer announced that reducing the costs while improving quality and maintaining a strong workforce is a constant pressure for hospitals and health systems. As payment levels continue to change and healthcare becomes more consumer-centric so comes an array of financial challenges.

3.2.2 Human Resource

Nowadays health care systems are facing an increasing number of problems with the management of human resources. In health care, human resources are defined as the clinical and non-clinical staff who are responsible for public and individuals' health interventions. In [39], the performance and the benefits of the health system largely depend on the knowledge, skills, and motivation of those individuals for delivering health care services. For enhancing HR management, [40] suggested some ideas like:

- Reducing the number of staff, altering the skill mix of staff, and arrangement of more adaptive employment.
- Improving training curricula.
- Developing staff performance by introducing lucrative compensations.

Challenges faced by the HR professionals are introduced in [41–43]. Staff shortage (the reasons are massive workload for medical professionals, aged generation are beginning to retire and leave the workforce in droves), turnover rates (rate for healthcare workers over-all is double than other jobs due to long term high stress, financial shortage, improper work-life balance), employee burnout (is tangled up in the issues of staff shortages and turnover. These rates are 70% among nurses, while 50% for doctors and nurse practitioners. There's a significant correlation between burnout rates and increases in patient infection rates. They report lower satisfaction rates in facilities where burnout rates are higher), training and development (ongoing training, licensure, and development are critical in the healthcare sector as there exists a shortage of knowledge for handling modern devices and the internet) are the challenges for the healthcare system [44].

In [45], the Hospital chief financial officer announced that reducing the costs while improving quality and maintaining a strong workforce is a constant pressure for hospitals and health systems. As payment levels continue to change and healthcare becomes more consumer-centric so comes an array of financial challenges.

3.2.3 Standardization

Several vendors are manufacturing a wide range of goods and devices in the healthcare sense, and new vendors continue to join this innovative technological race [25]. Though, standard rules and regulations for coherent interfaces and protocols for devices were not followed as result interoperability issues are rising. Prompt efforts are necessary if we want to address the device variety. Standardization of IoT-based healthcare technologies should consider a variety of issues e.g., communication layers and protocols, media access control (MAC) layers, physical (PHY) layers, interfaces of devices, and data aggregation and gateway interfaces.

Electronic health records, Heartbeat records, Pressure records are standardization issues that are considered as management of different value-added amenities. We should also consider access management and healthcare professional management. Different organizations related to mHealth and eHealth and IoT researchers can make working groups for ensuring the standardization of IoT-based healthcare services where standardization groups like the Information Technology and Innovation Foundation (IETF), the Internet Protocol for Smart Objects (IPSO), and the European Telecommunications Standards Institute (ETSI) can give their efforts patient and expert. This model describes the GSM Module, which transmits SMS to the medical expert by integrating several sensors (e.g., body temperature sensor, heartbeat, ECG, blood pressure, blood glucose, breathing sensor). Patients can save their lives with the help of GSM if they face any emergency. GSM protocol will be followed by the equipment to detect different healthcare sensors and then begin the rescue assistance operation by the expert. The GSM module function gets information configured with a communications protocol named User Datagram Protocol (UDP) that facilitates the exchange of messages between computing devices in a network [46].

4 Proposed Model

Discussed in several challenges of IoT Healthcare, and accusations are counteracting health monitoring and getting the suitable output. We believe some challenging solutions are available in different countries [47]. With the growing challenges of IoT healthcare, its solutions are also progressing. Several solutions address health monitoring systems more effectively, and those solutions are discussed below.

The proposed model occurs; patient monitoring is more than a network of bedside and mobile devices sensitive temperament together and combined to a central station that fits securely into the hospital's environments to help drive clinical and economic outcomes. It takes consecutive current patient data then feeds it automatically to the patient records. Patient monitoring online and offline solution taps the power of advanced decision support tools to help detect the onset and offset of critical events like body temperature and heart rate. So, doctors can make learned care decisions, and it gives on-the-move caregivers virtually anywhere, anytime visibility to a patient's vital signs and alarms. They can communicate and collaborate to decide on the best course of care.

The offline mode of the proposed model measures the different parameters of the body externally determined by the stored data in the database. The model represents to identify the person to whom the data belongs. Historical stored data set to assess the status of the patient's health condition. It connects to a mobile phone with a radio component to the database for subscriber parameters like ID, location, authentication key, etc. It contains a representation of the data stored in the Home Location Register (HLR). This database includes all patient data regarding authorized subscribers using the global mobile communication (GSM) center network. The approach of numerous straightforward models to make the liveliness is to do it offline. The proposed a model to address the IoT Health Monitoring System to lead the possible control from everywhere. During the COVID-19 pandemic, the IoT based health monitoring system is growing rapidly in various sectors. The essence of information retrieval and Natural Language Processing (NLP) can filter the user data to be sent to the medical expert using Machine Learning [48].

The model concentrates on the health monitoring system as a real-time and offline activity with medical devices that expert doctors monitor. Patient data transmitted through multiple connected devices are linked with network protocols and stored in the Server, Health IoT Could, or database. Systems are as signed to persistent data stored by the system that access control and security of the system in phases of an access model. This model also defines security issues, such as the authentication tool and encryption keys algorithms.

The model illustrated sensor based IoT devices connected with user data. Application of IoT devices sends their respective data to IoT networks using the Internet of Things Health (IoTh) network is mainly used to analyze and process it for access and storage of data like server/IoT, cloud or database. It analyzed and stored the independent data in an appropriate database [45] in real-time. The networking protocols give the purpose of transferring data among the devices proceeding a network being the Internet. A real-time monitoring system can continuously be monitored. IoT devices collect the data from devices that are delivered to the IoT platform through IoT network protocols. The IoT Networks implements for data performances, and here presents a solution to the problems mentioned earlier. A real-time activity needs to use IoT networks to store the data to whom the data belongs.

The proposed model aims to improve the existing real-time monitoring environment through offline activity and concentrate on the offline health monitoring system that will work with cellular IoT without a Wi-Fi connection and Wi-Fi module. A GSM connection will use cellular networks to communicate with the A GSM connection will use cellular networks to communicate with the patient and expert. This model describes the GSM Module, which transmits SMS to the medical expert by integrating several sensors (e.g., body temperature sensor, heartbeat, ECG, blood pressure, blood glucose, breathing sensor).

Patients can save their lives with the help of GSM if they face any emergency. GSM protocol will be followed by the equipment to detect different healthcare sensors and then begin the rescue assistance operation by the expert. The GSM module function

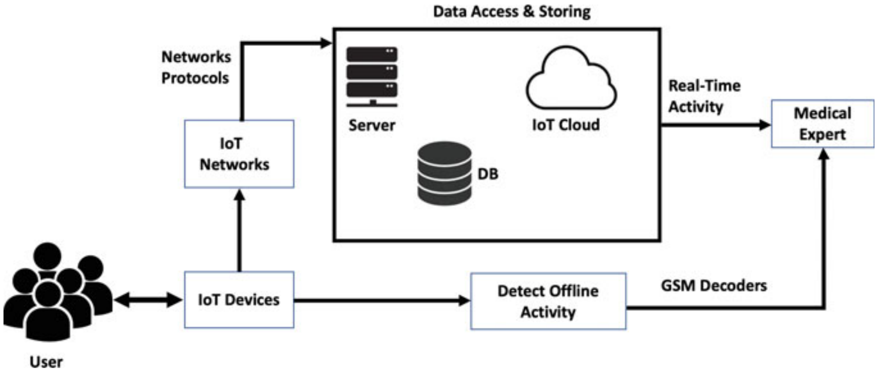


Fig. 1 Offline based IoT health monitoring system

gets information configured with a communications protocol named User Datagram Protocol (UDP) that facilitates the exchange of messages between computing devices in a network [46]. Finally, the sensor data is transferred to a telephone network, and at the receiver side, the signal transmission occurs in the reverse direction. Figure 1 shows the proposed model of Offline based IoT Health Monitoring System.

5 Conclusion

These research work will give a new opportunity that has excellent prospects for researchers and relevant stockholders to manufacture a smart health monitoring system. Our aim was to address specific challenges to deploy IoT health monitoring systems successfully. Authorized doctors and nurses can monitor and have performed a literature survey to highlight various challenges in IoT-based health monitoring systems in third-world countries. Proposed a solution (offline-based) to resolve some issues specifically related to connectivity and bandwidth in underdeveloped countries properly work on it, hopefully they can help the world to monitor their patients from anywhere and can reduce the infection like COVID-19.

References

1. Viswanathan H, Lee EK, Pompili D (2012) Mobile grid computing for data- and patient-centric ubiquitous healthcare. In: 2012 The first IEEE workshop on enabling technologies for smartphone and internet of things (ETSIoT). https://doi.org/10.1109/ETSIoT.2012.6311263
2. Sun Q, Li H (2011) Research and application of a UDP-based reliable data transfer protocol in wireless data transmission. In: 2011 International conference on computer science and service system (CSSS). IEEE
3. Duggal R (2003) Urban healthcare-issues and challenges. Management 15:31

4. https://advances.massgeneral.orgcardiovasclar/vieo.aspx?id=1061
5. Andriopoulou F, Dagiuklas T, Orphanoudakis T (2017) Integrating IoT and fog computing for healthcare service delivery. In: Keramidas G, Voros N, Hübner M (eds) Components and services for IoT platforms. Springer, Cham
6. Kahn JG, Yang JS, Kahn JS (2010) 'Mobile' health needs and opportunities in developing countries. Health Aff (Millwood) 29(2):252–258
7. https://www.healthcaremaiing.com/healthcare/medical-executives-email-list.html
8. https://www.amerilist.com/medhealth-healthcare-professionals-mailing-list
9. Yang X (2018) Wandering pattern sensing at S-band. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2017.2787595
10. Yang X, Fan D, Ren A, Zhao N, Alam M (2019) 5G-based user-centric sensing at C-band. IEEE Trans Ind Inform. https://doi.org/10.1109/TII.2019.2891738
11. Lemlouma T, Laborie SR, Philippe R, Abderrezak A, Kenza (2014) A study of mobility support in wearable health monitoring systems: design framework. CRC Press/Taylor Francis
12. Srinivas M, Durgaprasadarao P, Raj VNP (2018) Intelligent medicine box for medication management using IoT. In: 2nd International conference on inventive systems and control (ICISC). https://doi.org/10.1109/ICISC.2018.8399097
13. Pallavi K, Tripti K et al (2019) Secure health monitoring of soldiers with tracking system using IoT: a survey. Int J Trend Sci Res Dev (IJTSRD) 3(4)
14. Amine B, Abderrezak R, Badache N (2015) A study of mobility support in wearable health monitoring systems: design framework. In: 2015 IEEE/ACS 12th International conference of computer systems and applications (AICCSA). https://doi.org/10.1109/AICCSA.2015.7507158
15. Dwivedi AD, Srivastava G, Dhar S, Singh R (2019) A decentralized privacy-preserving healthcare blockchain for IoT. Sensors. https://doi.org/10.3390/s19020326
16. Bhuiyan MN, Rahman MM, Billah MM, Saha D (2021) Internet of things (IoT): a review of its enabling technologies in healthcare applications. Standards protocols, security, and market opportunities. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2021.3062630
17. Divya S, Tripathi RC (2020) Performance of internet of things (IOT) based healthcare secure services and its importance: issue and challenges. Available at SSRN 3565782
18. Ngankam HK et al (2019) An IoT architecture of microservices for ambient assisted living environments to promote aging in smart cities. In: Pagán J, Mokhtari M, Aloulou H, Abdulrazak B, Cabrera M (eds) How AI impacts urban living and public health. Lecture notes in computer science, vol 11862. Springer, Cham
19. Kadhim KT, Alsahlany AM, Wadi SM, Kadhum HT (2020) An overview of patient's health status monitoring system based on internet of things (IoT). Wirel Pers Commun 1–28
20. Islam MM, Rahaman A, Islam MR (2020) Development of smart healthcare monitoring system in IoT environment. SN Comput Sci 1:185
21. Khan SF (2017) Health care monitoring system in internet of things (IoT) by using RFID. In: 2017 6th International conference on industrial technology and management (ICITM), pp 198–204. https://doi.org/10.1109/ICITM.2017.7917920
22. Tamilselvi V, Sribalaji S, Vigneshwaran P, Vinu P, Geetha Ramani J (2020) IoT based health monitoring system. In: 2020 6th International conference on advanced computing and communication systems (ICACCS). IEEE, pp 386–389
23. Acharya AD, Patil S (2020) IoT based health care monitoring kit. In: 2020 Fourth international conference on computing methodologies and communication (ICCMC), pp 363–368
24. Banerjee S, Roy S (2016) Design of a photo plethysmography based pulse rate detector. Int J Rec Trends Eng Res
25. Gregoski MJ, Mueller M, Vertegel A, Shaporev A, Jackson BB, Frenzel RM, Sprehn SM, Treiber FA (2012) Development and validation of a smartphone heart rate acquisition application for health promotion and wellness telehealth applications. Int J Telemed Appl. https://doi.org/10.1155/2012/696324
26. Williams P, McCauley V (2016) Always connected: the security challenges of the healthcare internet of things. In: 2016 IEEE 3rd World forum on internet of things (WF-IoT), pp 30–35

27. Oresko JJ (2010) A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. IEEE Trans Inf Technol Biomed. https://doi.org/10.1109/TITB.2010.2047865
28. Kumar SP, Samson VRR, Sai UB, Rao PLSDM, Eswar KK (2017) Smart health monitoring system of patient through IoT. In: 2017 International conference on I-SMAC (IoT in Social, mobile, analytics and cloud) (I-SMAC). https://doi.org/10.1109/I-SMAC.2017.8058240
29. Zeadally S, Isaac JT, Baig Z (2016) Security attacks and solutions in electronic health (E-health) systems. J Med Syst. https://doi.org/10.1007/s10916-016-0597-z
30. Das AK, Zeadally S, He D (2018) Taxonomy and analysis of security protocols for internet of things. Future Gener Comput Syst 110–125
31. Nambiar AR, Reddy N, Dutta D (2017) Connected health: opportunities and challenges. In: 2017 IEEE International conference on big data (big data). https://doi.org/10.1109/BigData.2017.8258102
32. Marjani M, Fariza N, Abdullah G, Ahmad K, Ibrahim HTA, Aisha S, Ibrar Y (2017) Big IoT data analytics: architecture, opportunities, and open research challenges. IEEE Access 5247–5261
33. Ahmed E, Ibrar Y, Ibrahim HTA, Imran K, Abdel I, Muhammad I, Athanasios V (2017) The role of big data analytics in internet of things. Comput Netw 459–471
34. Sharma SK, Wang X (2017) Live data analytics with collaborative edge and cloud processing in wireless IoT networks. IEEE Access. https://doi.org/10.1109/ACCESS.2017.2682640
35. Anagnostopoulos I, Zeadally S, Exposito E (2016) Handling big data: research challenges and future directions. J Supercomput 72:1494–1516
36. Burke J (2015) Is that data valid? Getting accurate financial data in health-care. Health Catalyst. Available via: www.healthcatalyst.com/financial-data-in-healthcare-edw
37. Health systems: improving. Performance, Geneva. http://www.who.int.proxy.lib.uwo.ca:2048/
38. Hipaa, Health information privacy. www.hhs.gov/hipaa/index.html
39. Javier M, Liz C, Tim M (1999) A review of human resource issues in the health sector. DFID Health Systems Resource Centre, P-Pub, London
40. Nicole H, 5 Challenges facing HR professionals in the healthcare industry. Available via: https://www.accesscorp.com/blog/5-challenges-hr-healthcare/
41. Jay M, 5 Challenges facing human resources management in healthcare. Available via: https://www.excelforce.com/insights/5-challenges-facing-human-resources-management-in-healthcare
42. Tori F, 4 Big challenges facing HR professionals in the healthcare industry. Available via https://www.bamboohr.com/blog/challenges-facing-hr-professionals-healthcare-industry/
43. Healthcare's leading financial challenges and opportunities in 2019. Available via: https://www.jpmorgan.com/commercial-banking/insights/healthcare-financial-challenges-2019
44. Jess W, Top financial challenges in health care. Available via:https://www.healthcarebusinesstech.com/financial-challenges/
45. Islam SMR, Kwak D, Kabir MH, Hossain M, Kwak K (2015) The internet of things for health care: a comprehensive survey. IEEE Access. https://doi.org/10.1109/ACCESS.2015.2437951
46. Billah MM, Bhuiyan MN, Akterujjaman M (2021) Unsupervised method of clustering and labeling of the online product based on reviews. Int J Model Simul Sci Comput
47. Hopalı E, Vayvay Ö (2018) Internet of things (IoT) and its challenges for usability in developing countries. Int J Innov Eng Sci Res
48. Saber Mahani A, Godarzi G, Baroni M, KHakiyan M (2010) Estimation of technical efficiency of general hospitals of Kerman University of Medical Sciences by data envelopment analysis (DEA) method in 2007. J Kerman Univ Med Sci 16(1):59–67

Wireless Body Sensor Networks: Applications, Challenges, Patient Monitoring, Decision Making, and Machine Learning in Medical Applications

Alaa Shawqi Jaber and Ali Kadhum Idrees

Abstract The power of the networks of Wireless Body Sensor are too huge to ignore. WBSNs have promised to enable their innovative, vast, simultaneous, and accurate monitoring applications in a variety of fields including health care, fitness and sport training, social interactions, and the monitoring of industrial workers. The objective of this paper is to lend some understanding on the scientific background of WBSNs and presenting recent advances in this field especially applications focus on remote monitoring for elderly and chronically diseases patients. In order to fulfillment the scientific concept of WBSN, a comprehensive study involving WBSNs architecture, challenges, healthcare applications and their requirements. Following, discussing the most important characteristics of the WBSN includes data collecting, fusion, risk evaluation and decision making. Moreover, shedding lights on machine learning techniques and their role in medical application. Finally, the paper recommends that the awareness of relevant issues and future development of WBSNs are regarded as a perfect solution to monitor the patient's life.

Keywords Wireless body sensor networks · Patient monitoring · Multisensor data fusion · Decision making · Machine learning · Energy efficiency

1 Introduction

Health is the prime challenge for humanity around the globe. Potentially, there is a need for a cost-effective, miniature and reliable health monitoring systems which would be efficient and easy to use without any literacy boundaries. Smart healthcare systems by Wireless Body Sensor Networks (WBSNs) was designed and developed

A. S. Jaber
Information Networks Department, University of Babylon, Babylon, Iraq
e-mail: alaa.shawqi@uobabylon.edu.iq

A. K. Idrees (✉)
Department of Computer Science, University of Babylon, Babylon, Iraq
e-mail: ali.idrees@uobabylon.edu.iq

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_20

to fulfillment these essentials [1]. BSN is the boon of deep research in WSN. WBSN healthcare system was used to predict and treat the disorders and diseases at the earlier stage before it becomes difficult to handle or maybe a threat to human life. Above and beyond, various advantages of such smart healthcare systems covered health quality enhancement, reducing the rates of death due to delay in treatment, discovering sudden diseases strokes or accidents and finally decrease the cost of healthcare services. In contrast, traditional ways cost time, expenses or make hospitalizing is out of reach for many people. Thus they neglect their health issues [2].

BSN consist of homogeneous and/or heterogeneous sensing nodes comprising of wearable and biocompatible sensors. It integrates the technology of sensing, intelligent information processing, pervasive computing and communication. WBSN has the capability of gathering the parametric physiological data like body accelerations, blood pressure, pulse rate and bio-impedance. Meanwhile the objective of fusion data is to process information from multiple, heterogeneous sources to estimate near accurate view of physiological, behavioral, health or emotional state of a person [3].

However, WBSNs have some challenges with regard to the programming of applications used with WBSNs and their performance, especially their weariness, battery, computational work, and data storage. Usually, WBSNs applications require high sampling rates that might have an effect on the real-time data processing with the transmitting efficiency. This is because of the computation power and the bandwidth provided through the infrastructure of the WBSN which is in general evaluated as inadequate. The limited resources require appropriate allocation on the nodes in terms of the huge raw data collected, memory, and energy, in addition to the processing. This point is very precarious in signal processing systems [4].

Gaining a realistic understanding of foremost research trends and revealing WBSN applications that promulgated by researchers is the primary aim of this work. This literature reviewed reputed journals and provided an insight into contemporary status of research. Consequently, the following questions have been answered through this work:

1. How technologies and applications have been evolved in WBSNs?
2. What are the onerousness in WBSN areas that researchers confronted?
3. What are the realistic demands the researchers should add?
4. How researchers investigate patient health risk evaluation and invent accurate decision making approaches using WBSNs technologies?
5. Which areas could be the probable future of WBSN research?

Rest of this chapter is structured in twelve sections. Section 2 discuss the previous work of researchers. Sections 3 and 4 elaborate Wireless Body Sensor Networks and their applications. Sections 5 and 6 discuss challenges in WBSNs and requirements for healthcare applications. Section 7 details about biosensors data collecting. Illness score concept demonstrates in Sect. 8. Sections 9, 10 and 11 provide discussions about data fusion, risk evaluation with decision making and machine learning in medical applications. Final section reveals conclusion.

2 Related Work

A systematic literature review was done to investigate issues such as consumption of energy by mean of data reduction, adaptation of sampling rates and fusion of sensed data. These functions are essential to understand the heterogeneity of data and come out with essential information that contributes in make the medical decisions. Table 1 covers all the aforementioned aspects. Models such as NEEF, digital CS based single-spot Bluetooth ECG node and EEG Fractals lossy compression are proposed by authors in [5–7] respectively. They investigated approaches to minimize the collection of medical data and eventually consumed less energy .

Other researchers in [8–12] concerned with adapting the sampling rate techniques. Some of them use statistics tool like ANOVA with a quadratic Bezier curve as a behavior (BV) function, adjusted Fisher test, Spline interpolation. Others use machine learning to forecast the sampling rate such as LSTM, heuristic methods or even training models regarding the problem of class imbalanced big data.

Fusion system is the integration of sensor data and information to produce valuable information. Multisensor fusion in WBSN is still a nontrivial task due to directly impact the performance of vital application. Authors divulged various schemes to fuse the sensitive physiological parameters. For example, functioning a supervised learning method for detecting heartbeats and classifying arrhythmias is one of these schemes adopted by [14]. Other researchers in [16] proposed an architecture for the distribution of a hierarchical data fusion. Consequently, the survey in this literary section exposed diverse approaches for bio-physical measurements to be processed and managed on WBSNs. Table 1 shed lights on researches covering topics like energy efficient, adaptive sampling, data reduction, multisensor data fusion and decision making.

Table 1 Survey on main characteristics of patient health related approaches in the literature

Proposed approach	Mutlisensor fusion	Energy-efficient	Adaptive sampling	Data reduction	Decision making
Mehrani et al. [10]		✓	✓	✓	
Rendon et al. [11]		✓	✓	✓	
Al-Nassrawy et al. [7]		✓		✓	
Malathy et al. [5]		✓			
Johnson and Khoshgoftaar [12]		✓	✓	✓	
Vitabile et al. [13]	✓				✓
Sciré et al. [14]		✓			
Dautov et al. [15]	✓				✓
Luo et al. [6]		✓			
Azar et al. [8]		✓	✓	✓	
Navarro et al. [16]	✓				✓
Habib et al. [9]		✓	✓	✓	✓

3 Wireless Body Sensor Networks (WBSNs)

WBSNs or WBANs (Wireless Body Area Networks) are networks contain set of heterogeneous biosensors and a coordinator node Fig. 1. These tiny-sized biosensors could be either invasive (i.e. implantable sensors) or non-invasive when attached to human skin (i.e. wearable sensors) [17]. Each biosensor has its particular criteria and functions in different missions. Assessing variations in a patient's vital sign rates, as well as detecting patient's status or emotions (fear, stress, happiness, etc.) are some examples of these sensors functions. The sensors are communicated to a specific coordinator node, represented by a portable device carried by the patient, smartphone, or patient data aggregator (PDA), as these have better processing capabilities and relatively less energy constrains. The function of this device is to send biologically significant signals about the patient's condition to the medical staff in order to create an actual medical diagnosis, allowing them to make appropriate decisions [18]. Different types of biomedical sensors can be applied to collect different vital data from the patient, such as body temperature, ECG (ElectroCardioGram), EEG (ElectroEncephaloGram), heartbeat, and blood pressure. In fact, these biosensors cannot only collect the health data but also analyze them. For the above mentioned reasons, this technology minimizes the health care cost by providing the medical teams instantly with the vital data for patients inside their homes or away from the hospital. In addition, the general patient's condition monitored for a long time remotely and continuously will change over time. It is subjected to several health situations that can be severe or even constant. This thing has a direct impact on the acquisition of data and finally on the consumed energy of the WBSN and the initial discovery of risky situations [19]. Table 2 reports some of the commercially available sensor nodes. Most of those biosensors are mainly attentive in continuous monitoring of physiological parameters especially vital signs.

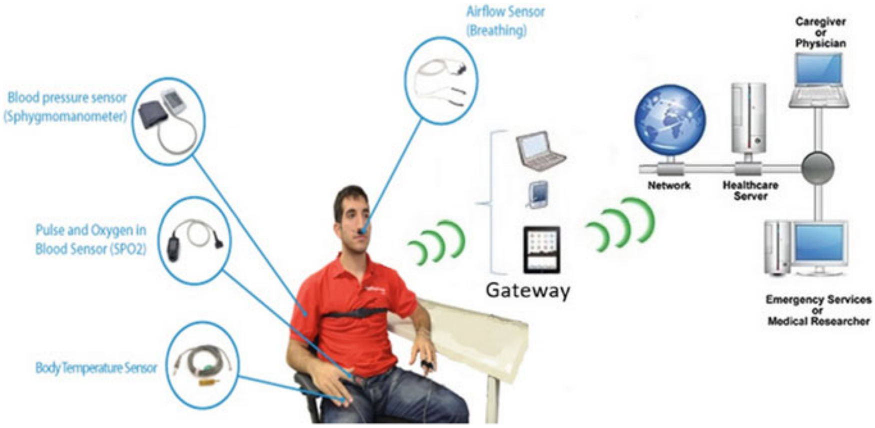

Airflow Sensor (Breathing)
Blood pressure sensor (Sphygmomanometer)
Pulse and Oxygen in Blood Sensor (SPO2)
Body Temperature Sensor
Gateway
Network
Healthcare Server
Caregiver or Physician
Emergency Services or Medical Researcher

Fig. 1 Topology of WBSN system

Table 2 Commercially available sensor nodes on the market

Model	Company	What measures	Key features
Vital Tracer [20]	VTLAB	Blood pressure, oxygen level in the blood, SpO_2, heart rate, temperature, gyroscope, step count, ECG and respiratory rate	Smartwatch wearable device collects raw bio-signals. Accessed to patient's data in real-time and kept on the cloud or can be moved from the watch to phone or PC utilizing Bluetooth/USB link to be customized to research or patient needs
CamNtech [21]	Actiheart5	ECG, heart rate, HRV, energy expenditure and activity recorder	Attaches to the chest/plug-in device, programmable: row data output-open data format, connected via direct USB/Bluetooth
Hillrom Extended Care Solution [22]	Welch Allyn	BP, Blood pressure, plus oral/axillary thermometry, SpO_2 technology measures blood oxygen saturation	Capture vital signs at home, send measurements via Bluetooth, then send to the care team data-driven decision-making
BiPS Medical [23]	BiPSMed	Blood pressure (diastolic and systolic), heart, respiration rates, and blood oxygen level	Wearable wireless device, the sensor system programmed to record and transmit data to electronic medical records (EMRs) on demand (from continuous to intermittent, to meet the needs of the patient)
HEXOSKIN Health Sensors and AI [24]	HEXOSKIN	Continuous cardiac, respiratory, activity, stress, cognitive, mental disorders, neurology and sleep monitoring	Smart wearable shirt for in home rehabilitation. Bluetooth connectivity with iphone, ipad and android devices. Open Data API allows to collect and download raw data and use own analytics software for health monitoring
Shimmer 3IMU, 3ECG, 3GSR+ [25]	Shimmer	ECG signal to detect heart rate and respiration rate, blood pressure and inertial sensing via accelerometer, gyroscope and magnetometer	Wearable wireless device, automated data upload via cellular and WIFI communications. Configurable participant interface to full analysis. Integrates into a full featured Clinical Trial Management Software platform

4 WBSN Applications

The WBSNs are initially applied for improving the systems of health in order to increase its efficiency in detecting diseases in an early stage, as well as monitoring, assistance, and post-surgical feedback. Alternative uses of WBSNs are found in other medical, such as nonmedical fields. WBSNs medical devices are divided into three classes according to their functions and roles: Wearable, Implant, and Remote-controlled WBSNs [17, 26], the following subsections will describe the aforementioned classes.

4.1 Wearable WBSNs

This type of biosensors are in place of observing and checking the patient's health progress. Wearable sensors term came from its direct contact with patient. They have objects called wearable objects in charge of converting physical signals into electrical signals [27]. Many real life application concerning wearable sensors have been arise nowadays. Some examples are, ECG monitor evaluation [28, 29], Human activity monitoring [30, 31], sleep monitoring [32, 33], fall detection [34, 35], asthma monitoring [36, 37].

4.2 Wearable WBSNs

Known also in-body sensors, any movement in the organ whether affected or transplanted can be detect by using implantable sensors. Applications in the body contain monitoring and modifying programs to adjust biosensors embedded in the body [26, 27]. Moreover, the implantable devices have the ability to perform wireless communication, as well as analyzing and delivering warnings to support human life [17]. For instance, ICD (Implanted Cardiac Defibrillators) [38], pacemakers [39], drug and baclofen pumps [40], neuro stimulators [41], blood glucose-level sensors [42, 43], cancer detection [44].

4.3 Remote-Controlled WBSNs

The last type is remotely controlled medical sensor devices which are concerned with three different aspects. AAL (Ambient Assisted Living), telemedicine systems and patient monitoring. All these types contribute to the development of the elderly's lifestyle where technological advances add a sense of independence to their self-care system [17]. Next subsections will illustrate the three aspects sequentially.

4.3.1 Ambient Assisted Living

The AAL system was found for assisting people in their Activities of Daily Living (ADL), it is applicable in smart flats and houses, and long-term caring hospitals [45]. Gingras et al. [46] proposed a highly modular architecture that consists of four layers to conduct health analyses of elderly. In addition, they proposed a new automated process to select an appropriate algorithms for a given task. Wang and Cook [47] present three multi-resident tracking algorithms, namely the nearest neighbor with sensor graph (NN-SG), global nearest neighbor with sensor graph (GNN-SG), and multiresident tracking with sensor vectorization (sMRT), for solving the issue of data association among sensor events and residents within smart environment.

4.3.2 Telemedicine Systems

Telemedicine systems are mainly to be applied in the field of tele-medication mhealth (mobile health) applications, it diagnoses and treats patients remotely via telecommunications technology [48]. According of institute of medicine in US [49] the definition of telemedicine can be stated as using electronic information and telecommunication technology for supporting and promoting clinical health care in case the participants are distantly separated. The authors in [50] developed an Artificial Medical Intelligence (AMI) for automated analysis of fluorograms through a cloud service design. This achievement includes telemedicine systems with AMI based modern telecommunications and different AMIs for mammography, fluorography and cardiograph. Nasri and Mtibaa [51] suggested a structure for a smart mobile IoT health care system that monitors the risk of a patient by means of a smartphone and 5G. The data is received through an android interface from a WBSN using Wi-Fi, ZigBee, or Bluetooth, which requires the user to permit sending data over the Web. The system is to advise and alert the medical team so that preventive measures can be taken.

4.3.3 Patient Monitoring

The final aspect is patient monitoring which responsible for careful observation of the patient's physical interaction linked to health status and at the same time enables internetworking among several networks and devices. Authors in [52] could develop a comprehensive model to monitor patients regularly over an interconnected network that link the medical healthcare staff to the patients. The main purpose behind this model is to minimize the workload of the medical staff, reduce the chance of doctors and nurses to be infected by the COVID19 disease, and to increase the total effectiveness of patient monitoring within hospitals.

5 Challenges in WBSNs

Despite the fact that WBSNs are a distinct kind of WSNs, there are a number of challenges that characterize the WSBN, and several additional issues need a more sufficient solution. The following challenges are stated:

5.1 Energy Consumption

In general, the WBSN devices are powered by batteries. Since WBSNs make use of a limited number of nodes that are relatively small (as compared to alternative WSNs), a larger constrain is added to the energy consumed during communications. The type of applications determine the amount of energy to be consumed by the WBAN nodes. For example, most implanted sensors must operate for several years, whereby the operations are performed for a long time without any replacement of the battery [53] while others are organized and deployed over the human body. Therefore the design of an in/on-sensor energy efficient methods is still an open research issue challenged by reducing the duty cycle, managing the power usage of these sensors and maximizing the battery life [54]. Researchers in [55] proposed Fast Compressive Electrocardiography (FCE) technique by using Compressed Sensing (CS). CS is considered to be a less complex ECG data compression scheme for wearable wireless biosensor devices. Despite, they tend to be highly complex on a computational level. The CS decoding in FCE depends on Weighted Regularized Least-Squares (WRLS) to tackle this problem. Other researchers in [56] proposed a new WBSN architecture using a pyramid interconnection to decrease the amount of energy consumed and the delay in data collection, and to increase the resiliency.

5.2 Signal Processing

Commercializing miniaturized WBAN devices leads to an increased rate of WBAN users, which in turn demands network resource managements and usages because of the increased traffic of data. This remarkable data volume demands new WBAN criteria that are required for supporting such high data rates using the limited network resources within the transmitting channel. Other cases include abnormal alterations in the data configurations and frame structures may also contribute to excessive network resource requirements [57]. In order to monitor real time health, the communication of streaming videos (as in endoscopic capsule) and ECG trace require an increased bandwidth and transmitting energy because of the highly complicated issues of data acquisition noises, data volumes, and transmitting channel errors. These criteria could cause performance degradation problems that require mitigating measures of high costs [58].

5.3 Heterogeneity of Devices

The concept of heterogeneity in WSNs can be defined differently based on the capabilities of sensor nodes in terms of computation or energy. Computational heterogeneity implies that nodes have differing memories, transceivers and processing powers, whereas energy heterogeneity refers to the varying energy levels of nodes. Alternatively, heterogeneity could be defined based on the traffic whereby various intensities, packet length distributions or bandwidth requirements are associated with the traffic generated by the nodes [59]. The modern multisensor nodes come with nonintrusive sensors placed upon one radio board, and could thereby generate heterogeneous traffic, as is the case in multisensor nodes which monitor the patients' vitals in a patient monitoring system. Collecting signals through several sensors allow more reliable diagnoses in heterogeneous WBANs as compared to homogeneous ones. Following the developments of heterogeneous WBANs, the Wireless Electro-encephalography Sensor Networks (WESNs) [60] have been studied using distributed signal processing. The design of such a network follows two hierarchies, namely the Hierarchical Fully-Connected Topology (HFCT) and Ad-Hoc Nearest-Neighbor Topology (ANNT) for improving how energy-efficient it is through the use of the distributed Multi-channel Weighted Wiener Filter design (MW2F).

5.4 Data Anomalies

Anomalies can be described as atypical features that are found in data or device operations, which display a deviation from the actual attributes. As for big data scenarios, anomalies may arise from the devices of ageing or malfunctioning, power scarcities, adversarial intrusions, noise, erroneous calibration, electro-magnetic interferences and coexistence, moisture on sensor contacts, and insertion forged data [57]. The limitation found in WBANs in terms of computation leads to anomalous data which in turn expands the intricacies when operating them. However, these may affect the patients fatally. Therefore, a number of researches suggest data analyses, ML tools, appropriate error correction and interference-avoidance methods for detecting anomalies in WBANs. The work in [61] proposed a double level lightweight and adaptive anomaly detection approach for discarding inaccurate and faulty measures, so that alarms are only raised whenever patients actually seem to be in an emergency situation.

5.5 Path Loss and Environmental Challenges

Path loss or attenuation leads to drop in energy density of electro-magnetic waves when propagating throughout space. There are a number of causes that lead to path loss in WBANs, including postural body movements, the node mobility, obstacles in the environment, and body tissue absorption. The positioning of WBAN devices could be in, on, or off-body. The latter two types of positioning could witness less path losses than in-body devices; this is because the profound effect of various di-electric features of the body tissue layers on the propagation of radio signals. Therefore, the transmission of the radio signals faces extreme attenuation [62]. Maheswar et al. [56] proposed an algorithm to choose a more suitable next hop input parameters, such as mobility, load and energy levels so as to avoid drop packets and delivers higher throughput.

6 Healthcare Application Requirements

The development of WBANs forms a challenge due to the wide variety of requirements demanded in their applications. The most important user requirements include safety, ease of use, mobility, data rate and reliability. The next subsections will cover these requirements clearly.

6.1 Safety for the Human Body and Bio-compatibility

The International Commission on Non-Ionizing Radiation Protection (ICNIRP) has specified a number of general limitations that need to be taken into consideration for guaranteeing health safety when exposing the body different timely electro-magnetic fields [63]. Considering the frequency ranging between 100 kHz and 10 GHz, these limitations are based on the Specific Absorption Rate (SAR). Due to the factor of sensors overheating, WBAN sensors require a direct attachment to the human skin. This usually comes from the radiation of antenna when transmitting data, thereby causing damage to the heat sensitive body tissues [64]. SAR is a representation of the mass normalized speed of coupling Radio Frequency (RF) energy to bio-tissues, using the unit of watts per kilogram (W/kg). Therefore, it is suggested that WBANs should be complying with either global [63] or local SAR restrictions and limitations, like the ones stated by the European Union in Europe [65] and by the FCC for USA [66]. By the same token, bio-compatibility is another aspect of WSNs requirements. Bio-compatibility has a significant influence on the biofouling. Biofouling can be defined as a reaction of the body tissues to the implantation of sensors beneath the human skin or other body parts, causing proteins, cells and other undesired biomaterials to accumulate upon the body surface. This process is considered to be one of the reasons that cause the sensor current to degrade and eventually fail.

6.2 *Mobility Support*

There are two main benefits associated with WBSNs, namely monitoring portability and independent locations, which are considered to be essential factors in expanding the application venues of WBSNs. However, there are several limitations involved, such as patient mobility [67]. In WBSNs, mobility and patient movement form critical issues in changing the sensor node locations, and eventually disconnect sensor nodes from their coordinator node (CN). As a consequence, this will lead to the increase in packet drop rate. Selem et al. [68] proposed mobTHE protocol to handle seamless mobile communication with on-body nodes. It deals with the disconnecting issues that result from sensor node mobility beside the network sustainability like maximal packet throughput, extended node life span, and lower temperatures.

6.3 *Reliability*

Reliability in WBSNs is highly necessary as it has a direct influence upon the monitoring quality of patients. Life-threatening situations that are not detected in time could end up being fatal for the patients. Therefore the data obtained during monitoring should receive correctly by health professionals. There are three essential factors necessary for a reliable network that fits the user's criteria, namely fault tolerance, QoS, and security [69].

6.3.1 Fault Tolerance

WBANs may often be subjected to interference, body fading, and a decreased reliability and throughput. Therefore, it is important to design protocols that cover unexpected issues which in turn will represent a promising solution for enhancing its reliability and fault-tolerant communication. This will be of use during the consistent patient monitoring as well as in obtaining the necessary data for making diagnoses in time [70]. To address these issues, authors in [71] introduced a fault-tolerant scheme of high energy efficiency for improving the network reliability in WBSNs. This mainly involves adopting the cooperative communicating and network coding strategy for minimizing the chances of channel impairments as well as the body fading effects, eventually reducing ensued fault, rates of bit error, and the amount of energy consumed.

6.3.2 Quality of Service (QoS)

There are many difficulties faced when meeting the quality criteria in e-health applications, particularly considering its energy-efficient features, quality of sensed data, the amount of network resources consumed, and its overall latency. A break, pause, delay or packet loss might end up fatally whenever an emergency case takes place. The body sensor quality indicates how accurate and sensitive the measured data is as obtained through the sensors [72]. There are a number of parameters for determining the quality of BSNs, including in the following points:

- The delay factor, especially in data collection/acquisition, processing and transmission at the level of sensor nodes.
- Bandwidth, capacity and throughput, which refer to the capacity of sensor networks for sending data through a link during a particular timespan.
- The sensor observation accuracy to detect all critical events.
- Trust-worthiness, some data obtained from two or more sensors might overlap as they cover the same area.

As an example of involving a QoS, Samanta et al. [73] designed a scheme of cost-effective heuristic packet scheduling for providing a high network throughput and sufficient QoS for a WBSN. Besides, the consideration of delay-constraint in achieving the optimized packet transmitting delays and managing heavy traffic load formed in an ideal alignment with a higher priority of medical emergent patients.

6.3.3 Security

Personal health information should be treated with high confidentiality, integrity, availability and authentication for avoiding any forms of data breaching. Leakages in patient health information will cause problems in terms of law and loyalty. It is therefore important to take several security measures into consideration when protecting users from such risks. An efficient, scalable, and usable performance is necessary when creating security architectures in WBANs [64].

6.4 Data or Bit Rate

The data rate is the speed of transferring data from the source to its destinations and vice versa, whose measurements is based on the network's speed (megabits per second). In term of medical uses, the network reliability could be increased through measuring data rates, whereby high data rate devices tend to be associated with low bit error rates (BER) [74]. The bit rate requirement varies in light of their applications, as well as the data type in transmission. It ranges between less than 1 kbps (as in temperature monitoring) to 10 Mbps (as in video streaming) [75]. Bit rates may involve on or more than one link, whereby several devices take part in transmitting/receiving information from or to a single coordinator simultaneously. To cover the mentioned aspects, authors in [76] introduced the RDTDRO scheme whereby they ensure a minimal amount of energy to be consumed through the optimization of data transmitting rate in light of a particular transmission distance and number of receiving antennas, whereby the BER criteria are still covered. The authors in [77] considered two issues in their proposing method: increasing the data rate for medical emergency applications with minimal transmitting power for which limited delay and reliability are necessary, as well as the increase in transmitting power (TP) for maintaining the data transmitting rates.

6.5 Ease of Use and Hardware Design

The ease of use of sensor devices is an essential and highly important criterion that enables WBSNs to be adopted widely in daily healthcare applications. Considering the patient's point of view, the hardware design for the bio-sensors that are either implanted in or attached onto the human body have to be small, easy to put on, few in number, unobtrusive, ergonomic, and sometimes even stylish [78].

7 Data Collection

Collecting of bio-physical measurements from human body is an essential and mandatory task in BSN applications. The system accuracy is determined by how accurate the recorded measurement values are. They are based upon three aspects, namely (i) the type of sensors employed (ii) their placement or positioning upon the patient's body (iii) skills [1]. In most healthcare applications whereby wearable sensors are used, it is favorable for sensors to be able to monitor the physiology of patients in a simultaneous manner. The assessment of physiology involves measuring biological, chemical or physical phenomena. The technical difficulty lies in the maintenance of consistent contact throughout a longer time span and under various conditions. On the other hand, the difficulties that are related to the healthcare aspect of these applications include achieving highly sensitive, accurate, and specific data to detect any anomalies simultaneously [75]. Following this line, real medical readings obtained from Multiple Intelligent Monitoring in Intensive Care (MIMIC I and II) databases of PhysioNet are utilized in this study [79]. The database resource consists of three interdependent entities: the PhysioBank, the PhysioToolkit, and the PhysioNet. The PhysioBank is a huge and extended archive of digital recordings that include physiological signals and other relevant data to be used in the field of biomedicine, such as critical arrhythmia, congestive heart failures, sleep apnea, neurological disorder, and aging. The PhysioToolkit represents a library of open-source software that can be of use in physiological signal processing and analyses. Finally, the PhysioNet is an on-line forum for exchanging and disseminating recorded biomedical signals and open-source software to be analyzed. Several patient records and their vital signs such as heart rate (HR), respiration rate (RESP), systolic blood pressure (ABPsys), blood temperature (BLOODT), and oxygen saturation (SpO2)) are used in this dissertation. Using such records is to ensure an actual and dynamic evaluation in instantaneous time, where the risk value of the vital signs indicate how severe the patient's health status and the vital signs themselves are.

8 Illness Score Concept

It is necessary to score the rate of illness or wellness so as to visualize the patient's health status, as it represents a practical method for remote healthcare staff for observing the patient's health condition and making suitable decisions accordingly within healthcare operation. In Emergency Departments (EDs) of hospitals, the triage system is widely used. Triage system is a system that involves assigning grades of urgency to injuries or illnesses for deciding upon the priority when treating multiple patients [71]. Various researches shows the importance of this system. Hook and Acharya [80] produced model electronic triage application based on the current Emergency Severity Index (ESI) and analyzed the result by the decision tree. The authors in [81] proposed an out-patient triage system determined by the hybrid Dynamic Uncertain Causality Graph (hybrid DUCG) for reducing the misdiagnosis result from out-patient triage error, and eventually helping triage nurses in improving the accuracy of the triage procedure. The method in [82] involves classifying the patient morbidity into four colored groups (green, yellow, orange and red) based on how severe the patient's condition is. The design of the color-coding system was recommended by the WHO.

9 Multisensor Data Fusion

The fusion of multi-sensor data can be defined as the process whereby observations from multiple sensors are combined for providing a robust and comprehensive description of particular environments or processes [83]. The attention on the multisensor fusion in WBSN has increased because it presents several benefits within a network suffering from various constraints, including losing data, inconsistent status, and affected sensor samples. It is important to decrease risk through altering the trust value of the fused data and the reasoned decisions, in addition to improve the extent to which healthcare applications are robust [84]. Evaluating the health criteria of a critically diseased or acutely-ill patient requires collecting several vital signs continuously so as to form an overall view of the patient's condition and eventually present an appropriate health evaluation. More clearly, the vital measurements transmitted from several biosensor nodes to reach the coordinator, whose role is fusing the received data intended for obtaining significant information and have appropriate decision concerning the patient's status.

10 Risk Evaluation and Decision Making

There are a number of works that attempt to design a technology to monitor the vital signs of patients. Bio-sensor healthcare systems involve placing biosensor nodes upon the patient's body for screening their health conditions in a simultaneous timing. Traditional patient-monitoring systems tend to be of lower accuracy. Therefore it is necessary to propose new systems for improving its accuracy through the use of a number of measurements in assessing global risks. The efficacy of healthcare systems could be enhanced by adding real-time decision-making systems to predict potential risks with higher accuracy. First, the WBSN nodes send the obtained data from several biosensor nodes to the coordinator to fuse the received data. This simultaneous evaluating measurement though N biosensors helps predicting the patient's risk levels when making decisions. The decision mostly represents a piece of advice involving a prediction or correction of the patient's case and may trigger or alert particular actions. Therefore, WBSNs can be used to improve healthcare monitoring systems and to recognize the levels of risk that indicate how severe the patient's status is in case of emergencies [85].

11 Machine Learning for Medical Applications

Machine learning (ML) is the field of research that studies the ability of making machines capable of learning. ML involves applying particular computer algorithms onto sets of data known to the event results, and it involves learning the training data and predicting new data according to the learning results. It tends to be inductive rather than deductive [86]. Nowadays, healthcare issues are becoming the center of attention, and the current developments mainly include the application of ML. Previous works tend to shed light on the application of ML algorithms onto data that has been gathered through the use of sensors so as to predict diseases, monitor health conditions, and eventually take part in the decision-making process. Many smart algorithms have been introduced using ML techniques for predicting diabetes [87], thyroid [88], heart diseases [89], AMD disease [90], stress detection [91], monitor health conditions [92] and human activity recognition [93]. ML can be classified into supervised, semi-supervised, and unsupervised forms of learning, as is the case with certain algorithms including the decision tree, nearest neighbor classifier, Naïve Bayes classifier, K-means, ANN (Artificial Neural Network) and SVM (Support Vector Machine) [86]. Besides these algorithms, regression tree is one of the fundamental supervised learning techniques that address multiple regression problems, as will be explained in the subsequent sub-sections.

11.1 Regression Tree Scheme

Regression trees provide a tree-based approximation with good promise for research in many academic fields. CARTs (Classical And Regression Trees) are analytic tree methods which can induce trees for predicting both categorical and real-valued targets. Depending on the measurement nature of the dependent variable, trees can divided into two classification and regression trees. Classification trees used on nominal (or categorical) dependent measures, whereas regression trees used on interval (or continuous) dependent measures as well as when there is not an obviously linear relationship or any mathematical relationship between the independent variables (predictors) [94]. However, they are share common characteristics like statistical principles, simplicity in understanding and interpreting, easiness in predicting new response values. Moreover, there are no special requirements for the distribution and specific attributes of the variables; and determines the main predictors automatically. Nevertheless some shortcomings are: requests observations from 50 to 100 with carefully selects the predictor; which mean when significant predictors or observations are absent, it cannot provide a sufficiently precise or stable models [95]. Least Square Regression Tree is a special type and most widely recognized scheme to build a regression model. It depends on a sample of an unknown regression surface for obtaining the parameters of model which minimizing the criterion of LS error [96]. In other words, regression tree utilizes least square and chooses an intuitively split in order to diminish the residual sum of squares among the observations and average in every node to find the optimum split. One more important point to be mentioned, in machine learning lingo is the prune. Trees cannot apply pruning after creation, means they have no bias but potentially large variance allows partially over-fitting to the trees own data sample. In order to avoid over fitting, the decision to divide which feature is limited to a random size n subset of the complete feature set [97].

12 Conclusion

To recap, this chapter presents many aspects regarding WBSNs. In this type of network, the sensed vital signs are gathered by biosensor devices then transmitted to the coordinator for further processing and fusion. Several WBSN applications like wearable, implantable and remote controlled have been presented, besides, main WBSNs challenges and requirements. The data collection and the illness score concept have been further illustrated. So as the received data for obtaining significant information to be evaluated for risk and fused to have appropriate decision concerning the patient’s status, this chapter highlights these concepts. Furthermore, the machine learning for medical application has examined carefully.

References

1. Gandhi V, Singh J (2020) An automated review of body sensor networks research patterns and trends. J Ind Inf Integr 18:100132
2. Gandhi V, Singh J (2020) WBSN based safe lifestyle: a case study of heartrate monitoring system. Int J Electr Comput Eng 10(3):2296
3. Rahmani AM, Gia TN, Negash B, Anzanpour A, Azimi I, Jiang M, Liljeberg P (2018) Exploiting smart e-health gateways at the edge of healthcare internet-of-things: a fog computing approach. Future Gener Comput Syst 78:641–658
4. Idrees AK, Al-Yaseen WL, Taam MA, Zahwe O (2018) Distributed data aggregation based modified k-means technique for energy conservation in periodic wireless sensor networks. In: 2018 IEEE Middle East and North Africa communications conference (MENACOMM). IEEE, pp 1–6
5. Malathy S, Rastogi R, Maheswar R, Kanagachidambaresan GR, Sundararajan TVP, Vigneswaran D (2019) A novel energy-efficient framework (NEEF) for the wireless body sensor network. J Supercomput 1–16
6. Luo K, Cai Z, Du K, Zou F, Zhang X, Li J (2018) A digital compressed sensing-based energy-efficient single-spot Bluetooth ECG node. J Healthc Eng
7. Al-Nassrawy KK, Al-Shammary D, Idrees AK (2020) High performance fractal compression for EEG health network traffic. Procedia Comput Sci 167:1240–1249
8. Azar J, Habib C, Darazi R, Makhoul A, Demerjian J (2018) Using adaptive sampling and DWT lifting scheme for efficient data reduction in wireless body sensor networks. In: 2018 14th International conference on wireless and mobile computing, networking and communications (WiMob). IEEE, pp 1–8
9. Habib C, Makhoul A, Darazi R, Couturier R (2017) Real-time sampling rate adaptation based on continuous risk level evaluation in wireless body sensor networks. In: 2017 IEEE 13th International conference on wireless and mobile computing, networking and communications (WiMob). IEEE, pp 1–8
10. Mehrani M, Attarzadeh I, Hosseinzadeh M (2020) Sampling rate prediction of biosensors in wireless body area networks using deep-learning methods. Simul Model Pract Theory 105:102101
11. Rendon E, Alejo R, Castorena C, Isidro-Ortega FJ, Granda-Gutierrez EE (2020) Data sampling methods to deal with the big data multi-class imbalance problem. Appl Sci 10(4):1276
12. Johnson JM, Khoshgoftaar TM (2019) Deep learning and data sampling with imbalanced big data. In: 2019 IEEE 20th International conference on information reuse and integration for data science (IRI). IEEE, pp 175–183
13. Vitabile S, Marks M, Stojanovic D, Pllana S, Molina JM, Krzyszton M, Sikora A, Jarynowski A, Hosseinpour F, Jakobik A et al (2019) Medical data processing and analysis for remote health and activities monitoring. In: High-performance modelling and simulation for big data applications, pp 186–220. Springer, Cham
14. Scirè A, Tropeano F, Anagnostopoulos A, Chatzigiannakis I (2019) Fog-computing-based heartbeat detection and arrhythmia classification using machine learning. Algorithms 12(2):32
15. Dautov R, Distefano S, Buyya R (2019) Hierarchical data fusion for smart healthcare. J Big Data 6(1):1–23
16. Navarro J, Vidaña-Vila E, Alsina-Pagès RM, Hervás M (2018) Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios. Sensors 18(8):2492
17. Khan RA, Pathan A-SK (2018) The state-of-the-art wireless body area sensor networks: a survey. Int J Distrib Sens Netw 14(4):1550147718768994
18. Jaber AS, Idrees AK (2021) Energy-saving multisensor data sampling and fusion with decision-making for monitoring health risk using WBSNs. Softw Pract Exp 51(2):271–293
19. Jaber AS, Idrees AK (2020) Adaptive rate energy-saving data collecting technique for health monitoring in wireless body sensor networks. Int J Commun Syst 33(17):e4589

20. VTLAB (2021) Vitaltracer. https://vitaltracer.com. Accessed 16 Apr 2021
21. CamNtech (2020) Camntech. https://www.camntech.com/about-us. Accessed 20 Apr 2021
22. Hillrom Extended Care Solution (2021) Hillrom extended care solution. https://www.hillrom.com/en/products/hillrom-extended-care-solution/. Accessed 20 Apr 2021
23. BiPS Medical (2018) Bips medical. https://www.bipsmed.com/. Accessed 16 Apr 2021
24. Hexoskin (2021) The Hexoskin smart clothing monitor. https://www.hexoskin.com/. Accessed 19 Apr 2021
25. Shimmer (2021) Shimmer wearable technology. http://www.shimmersensing.com/. Accessed 18 Apr 2021
26. Yazdi FR, Hosseinzadeh M, Jabbehdari S (2017) A review of state-of-the-art on wireless body area networks. Int J Adv Comput Sci Appl 11:443–455
27. Shokeen S, Parkash D (2019) A systematic review of wireless body area network. In: 2019 International conference on automation, computational and technology management (ICACTM). IEEE, pp 58–62
28. Abualsaud K, Chowdhury MEH, Gehani A, Yaacoub E, Khattab T, Hammad J (2020) A new wearable ECG monitor evaluation and experimental analysis: proof of concept. In: 2020 International wireless communications and mobile computing (IWCMC). IEEE, pp 1885–1890
29. Almusallam M, Soudani A (2017) Feature-based ECG sensing scheme for energy efficiency in WBSN. In: 2017 International conference on informatics, health & technology (ICIHT). IEEE, pp 1–6
30. Ascioglu G, Senol Y (2020) Design of a wearable wireless multi-sensor monitoring system and application for activity recognition using deep learning. IEEE Access 8:169183–169195
31. Wang H, Yan W, Liu S (2019) Physical activity recognition using multi-sensor fusion and extreme learning machines. In: 2019 International joint conference on neural networks (IJCNN). IEEE, pp 1–7
32. Yildiz S, Opel RA, Elliott JE, Kaye J, Cao H, Lim MM (2019) Categorizing sleep in older adults with wireless activity monitors using LSTM neural networks. In: 2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 3368–3372
33. Arulvallal S, Snekhalatha U, Rajalakshmi T (2019) Design and development of wearable device for continuous monitoring of sleep apnea disorder. In: 2019 International conference on communication and signal processing (ICCSP). IEEE, pp 0050–0053
34. Saadeh W, Butt SA, Bin Altaf MA (2019) A patient-specific single sensor IoT-based wearable fall prediction and detection system. IEEE Trans Neural Syst Rehabil Eng 27(5):995–1003
35. Desai K, Mane P, Dsilva M, Zare A, Shingala P, Ambawade D (2020) A novel machine learning based wearable belt for fall detection. In: 2020 IEEE International conference on computing, power and communication technologies (GUCON). IEEE, pp 502–505
36. Ghosh A, Rahman N, Awadalla N, Sagahyroon A, Aloul F, Dhou S (2020) Asthma diagnosis using neuro-fuzzy techniques. In: 2020 Advances in science and engineering technology international conferences (ASET). IEEE, pp 1–4
37. Tsang KCH, Pinnock H, Wilson AM, Shah SA (2020) Application of machine learning to support self-management of asthma with mhealth. In: 2020 42nd Annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE, pp 5673–5677
38. Hata R, Kato T, Yaku H, Morimoto T, Kawase Y, Yamamoto E, Inuzuka Y, Tamaki Y, Ozasa N, Yoshikawa Y et al (2021) Implantable cardioverter defibrillator therapy in patients with acute decompensated heart failure with reduced ejection fraction: an observation from the KCHF registry. J Cardiol 77(3):292–299
39. Hasan RR, Rahman MdA, Sinha S, Uddin MdN, Niloy T-SR (2019) In body antenna for monitoring pacemaker. In: 2019 International conference on automation, computational and technology management (ICACTM). IEEE, pp 99–102
40. El Kheshen H, Deni I, Baalbaky A, Dib M, Hamawy L, Ali MA (2018) Semi-automated self-monitore-syringe infusion pump. In: 2018 International conference on computer and applications (ICCA). IEEE, pp 331–335

41. Reza Pazhouhandeh M, Chang M, Valiante TA, Genov R (2020) Track-and-zoom neural analog-to-digital converter with blind stimulation artifact rejection. IEEE J Solid-State Circ 55(7):1984–1997
42. Islam MdM, Maniur SM (2019) Design and implementation of a wearable system for non-invasive glucose level monitoring. In: 2019 IEEE International conference on biomedical engineering, computer and information technology for health (BECITHCON), pp 29–32
43. Verner A, Butvinik D (2017) A machine learning approach to detecting sensor data modification intrusions in WBANs. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 161–169
44. Welikala RA, Remagnino P, Lim JH, Chan CS, Rajendran S, Kallarakkal TG, Zain RB, Jayasinghe RD, Rimal J, Kerr AR et al (2020) Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. IEEE Access 8:132677–132693
45. Abtoy A, Touhafi A, Tahiri A et al (2020) Ambient assisted living system's models and architectures: a survey of the state of the art. J King Saud Univ Comput Inf Sci 32(1):1–10
46. Gingras G, Adda M, Bouzouane A, Ibrahim H, Dallaire C (2020) IoT ambient assisted living: scalable analytics architecture and flexible process. Procedia Comput Sci 177:396–404
47. Wang T, Cook DJ (2020) Toward unsupervised multiresident tracking in ambient assisted living: methods and performance metrics. In: Assistive technology for the elderly. Elsevier, Amsterdam, pp 249–280
48. Kruse CS, Williams K, Bohls J, Shamsi W (2021) Telemedicine and health policy: a systematic review. Health Policy Technol 10(1):209–229
49. Field MJ et al (1996) Telemedicine: a guide to assessing telecommunications for health care
50. Garichev S, Klassen V, Natenzon M, Safin A, Sergeev S (2019) Mobile telemedicine systems with artificial medical intelligence. In: 2019 International conference on artificial intelligence: applications and innovations (IC-AIAI). IEEE, pp 8–83
51. Nasri F, Mtibaa A (2017) Smart mobile healthcare system based on WBSN and 5G. Int J Adv Comput Sci Appl 8(10):147–156
52. Amin R, Saha TS, Hassan MdFB, Anjum M, Tahmid MdI (2020) IoT based medical assistant for efficient monitoring of patients in response to covid-19. In: 2020 2nd International conference on advanced information and communication technology (ICAICT). IEEE, pp 83–87
53. Qureshi KN, Tayyab MQ, Rehman SU, Jeon G (2020) An interference aware energy efficient data transmission approach for smart cities healthcare systems. Sustain Cities Soc 62:102392
54. Almajed HN, Almogren AS, Altameem A (2019) A resilient smart body sensor network through pyramid interconnection. IEEE Access 7:51039–51046
55. Rateb AM (2020) A fast compressed sensing decoding technique for remote ECG monitoring systems. IEEE Access 8:197124–197133
56. Maheswar R, Maria AR, Sheriff N, Mahima V, Kanagachidambaresan GR, Lakshmi M (2019) Mobility aware next hop selection algorithm (MANSA) for wireless body sensor network. In: 2019 10th International conference on computing, communication and networking technologies (ICCCNT). IEEE, pp 1–5
57. Liu Q, Mkongwa KG, Zhang C (2021) Performance issues in wireless body area networks for the healthcare application: a survey and future prospects. SN Appl Sci 3(2):1–1
58. Cho Y, Shin H, Kang K (2018) Scalable coding and prioritized transmission of ECG for low-latency cardiac monitoring over cellular M2M networks. IEEE Access 6:8189–8200
59. Ali HQ, Ghani S (2020) Multi-sensor based mk/hyperk/1/m queuing model for heterogeneous traffic. Comput Netw 181:107512
60. Manojprabu M, Sarma Dhulipala VR (2020) Improved energy efficient design in software defined wireless electroencephalography sensor networks (WESN) using distributed architecture to remove artifact. Comput Commun 152:266–271
61. Arfaoui A, Kribeche A, Senouci SM, Hamdi M (2019) Game-based adaptive anomaly detection in wireless body area networks. Comput Netw 163:106870
62. Sari A, Alzubi A (2018) Path loss algorithms for data resilience in wireless body area networks for healthcare framework. In: Security and resilience in intelligent data-centric systems and communication networks. Elsevier, Amsterdam, pp 285–313

63. International Commission on Non-ionizing Radiation Protection et al (2020) Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 gHz). Health Phys 118(5):483–524
64. Hasan K, Biswas K, Ahmed K, Nafi NS, Islam MdS (2019) A comprehensive review of wireless body area network. J Netw Comput Appl 143:178–198
65. Karaboytcheva M (2020) Effects of 5g wireless communication on human health. Eur Parliam Res Serv PE 646:172
66. Federal Communications Commission (2019) Radio frequency safety. https://www.fcc.gov/general/radio-frequency-safety-0. Accessed 2 Apr 2021
67. Asam M, Ajaz A, Jamal T, Adeel M, Hassan A, Butt SA, Gulzar M (2019) Challenges in wireless body area network. Proc Int J Adv Comput Sci Appl 10(11)
68. Selem E, Fatehy M, El-Kader SMA (2021) mobTHE (mobile temperature heterogeneity energy) aware routing protocol for WBAN IoT health application. IEEE Access 9:18692–18705
69. Salayma M, Al-Dubai A, Romdhani I, Nasser Y (2017) Wireless body area network (WBAN) a survey on reliability, fault tolerance, and technologies coexistence. ACM Comput Surv (CSUR) 50(1):1–38
70. El Salamouny MY (2018) Fault tolerance in WBAN applications
71. Georgopoulos VC, Stylios CD (2017) Fuzzy cognitive maps for decision making in triage of non-critical elderly patients. In: 2017 International conference on intelligent informatics and biomedical sciences (ICIIBMS). IEEE, pp 225–228
72. Dhanvijay MM, Patil SC (2019) Internet of things: a survey of enabling technologies in healthcare and its applications. Comput Netw 153:113–131
73. Samanta A, Li Y, Chen S (2018) QoS-aware heuristic scheduling with delay-constraint for WBSNs. In: 2018 IEEE International conference on communications (ICC). IEEE, pp 1–7
74. Abidi B, Jilbab A, Mohamed EH (2020) Wireless body area networks: a comprehensive survey. J Med Eng Technol 44(3):97–107
75. Kim S, Iravantchi Y, Gajos K (2019) SwellFit: developing a wearable sensor for monitoring peripheral edema. In: Proceedings of the 52nd Hawaii international conference on system sciences
76. Senthil Kumar K, Amutha R, Palanivelan M, Gururaj D, Richard Jebasingh S, Anitha Mary M, Anitha S, Savitha V, Priyanka R, Balachandran A et al (2018) Receive diversity based transmission data rate optimization for improved network lifetime and delay efficiency of wireless body area networks. Plos One 13(10):e0206027
77. Sodhro AH, Chen L, Sekhari A, Ouzrout Y, Wu W (2018) Energy efficiency comparison between data rate control and transmission power control algorithms for wireless body sensor networks. Int J Distrib Sens Netw 14(1):1550147717750030
78. Velez FJ, Miyandoab FD (2019) Wearable technologies and wireless body sensor networks for healthcare. Institution of Engineering and Technology
79. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Eugene Stanley H (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220
80. Hook-Podhorniak G, Acharya S (2019) Effectual emergency severity adaptation for improved triage care operations. In: 2019 IEEE International conference on bioinformatics and biomedicine (BIBM). IEEE, pp 2628–2633
81. Bu X, Lu L, Zhan Z, Qin Z, Yan Z (2020) A general outpatient triage system based on dynamic uncertain causality graph. IEEE Access 8:93249–93263
82. Khan TR, Hossein KM, Maruf KRI, Fukuda A, Ahmed A (2017) Measurement of illness and wellness score of non-communicable disease patients. In: TENCON 2017-2017 IEEE Region 10 conference. IEEE, pp 2253–2257
83. Aileni RM, Valderrama AC, Strungaru R (2017) Wearable electronics for elderly health monitoring and active living. In: Ambient assisted living and enhanced living environments. Elsevier, pp 247–269

84. Gravina R, Alinia P, Ghasemzadeh H, Fortino G (2017) Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. Inf Fusion 35:68–80
85. Alameen A, Gupta A (2020) Optimization driven deep learning approach for health monitoring and risk assessment in wireless body sensor networks. Int J Bus Data Commun Netw (IJBDCN) 16(1):70–93
86. Dou H (2019) Applications of machine learning in the field of medical care. In: 2019 34rd Youth academic annual conference of Chinese association of automation (YAC). IEEE, pp 176–179
87. Jashwanth Reddy D, Mounika B, Sindhu S, Pranayteja Reddy T, Sagar Reddy N, Jyothsna Sri G, Swaraja K, Meenakshi K, Kora P (2020) Predictive machine learning model for early detection and analysis of diabetes. Mater Today Proc
88. Ouyang F, Guo B, Ouyang L, Liu Z, Lin S, Meng W, Huang X, Chen H, Qiu-Gen H, Yang S (2019) Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. Eur J Radiol 113:251–257
89. Katarya R, Srinivas P (2020) Predicting heart disease at early stages using machine learning: a survey. In: 2020 International conference on electronics and sustainable communication systems (ICESC). IEEE, pp 302–305
90. Hao S, Junye B, Liu H, Wang L, Liu T, Lin C, Luo X, Gao J, Zhao J, Li H et al (2020) Comparison of machine learning tools for the prediction of AMD based on genetic, age, and diabetes-related variables in the Chinese population. Regenerative Therapy 15:180–186
91. Kumar P, Garg S, Garg A (2020) Assessment of anxiety, depression and stress using machine learning models. Procedia Comput Sci 171:1989–1998
92. Gnana Sheela K, Varghese AR (2020) Machine learning based health monitoring system. Mater Today Proc 24:1788–1794
93. Li X, Zhao P, Wu M, Chen Z, Zhang L (2021) Deep learning for human activity recognition. Neurocomputing 444:214–216
94. Ma X (2018) Using classification and regression trees: a practical primer. IAP
95. Gocheva-Ilieva SG, Voynikova DS, Stoimenova MP, Ivanov AV, Iliev IP (2019) Regression trees modeling of time series for air pollution analysis and forecasting. Neural Comput Appl 31(12):9023–9039
96. Matondo SB, Owolawi PA (2019) FSO rain attenuation prediction using non-linear least square regression. In: 2019 International multidisciplinary information technology and engineering conference (IMITEC). IEEE, pp 1–5
97. Torgo L (2017) Regression trees

A Novel Lossless EEG Compression Model Using Fractal Combined with Fixed-Length Encoding Technique

Kahlaa K. Al-Nassrawy, Ali Kadhum Idrees, and Dhiah Al-Shammary

Abstract The aging population in advanced countries and the expensive costs of healthcare system has led to development of intelligent technology which is Wireless Body Sensor Network (WBSN). This intelligent healthcare system requires large amount of medical data produced from various types of biomedical sensors node to be collected, sent, and treated. This fact has created high latency and has increased network traffic. Therefore, health networks will suffer from congestions and bottlenecks. There is a need to minimize health network traffic volume and reduce latency especially in emergency condition where very short response time is required and improve networks performance by reducing the size of the transmitted data. This paper proposed a lossless Fractals compression approach to decrease the sent EEG from the gateway (Patient Data Aggregator (PDA)) cloud. The suggested approach improve the data communication in WBSNs by reducing the size of data traffic over the network. This approach is evaluated and compared with some existing methods and the results show that the introduced approach outperformed the other methods.

Keywords EEG signal · Network traffic · Lossless fractal · Compression model · Fixed-length · Encoding technique

1 Introduction

The WBSN is a significant branch of wireless sensor networks (WSNs) that is designed specifically to achieve a certain practical application [1]. Some scientists use the term Body Area Sensor Network (BASN) or brief Body Sensor Network (BSN) when referring to WBSN where every node consists of medical sensor with

K. K. Al-Nassrawy · A. K. Idrees (✉)
Department of Computer Science, University of Babylon, Babylon, Iraq
e-mail: ali.idrees@uobabylon.edu.iq

D. Al-Shammary
College of Computer Science and Information Technology, University of Al-Qadisiyah, Al Diwaniyah, Iraq

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_21

a sensing unit, these health networks are very close to each other and have similar characteristics, applications and difficulties [2].

WBSN is an evolving technology that can be used in E-Health systems and it consists of sensors positioned in, around or on the human body to monitor different physiological and biochemical parameters such as temperature, blood pressure, blood glucose, Electroencephalogram (EEG), Electrocardiography (ECG), Electromyography (EMG), etc. WBSN's participation began with defense when used in their work, from then on, it has developed in human existence and is now being used in every sector of life. WBSN is made up of tiny devices attached to the patient's body and in charge of sensing, processing and interacting with medical server. These biosensors are intended for particular aim to satisfy the demands of end uses [3].

For instance, an ECG sensor was designed to monitor cardiac operations. The problems that need instant attention are high prices for healthcare, the health of the infant, and the increase in the age group population. Although medical facilities are effectively supplied by physicians, some instances such as cancer need continuous patients monitoring and if therapy begins at a later point, then health enhancement will be hard and that may lead to patient death. Most chronic illnesses such as cancer, obesity, cardiovascular and diabetics need that the patient's condition to be monitored continuously, however this can't be done by stabilizing the patient in one place. When utilizing a WBSN, the patients get higher physical mobility and are no longer required to remain in the hospital [4, 5].

WBSN contains a key node acting as master and set of tiny nodes located in, on, or around the patient body acting as slave nodes. These sensor nodes can communicate with a specific node called network coordinator (sink), e.g. a smartphone, Personal/Patient Data Aggregator (PDA), robots and so on, this is usually less energy-constrained and has more capacity for processing. It is accountable for sending the patient's biological signal to the physician to provide medical diagnosis in real time and enable him to make the correct choices [6].

EEG plays a significant role in epileptic illness diagnosis, brain death, tumors, stroke, understand/manage brain signals and multiple brain diseases. This sensor is utilized to measure the brain's electrical activity by putting electrodes on the human scalp at different locations [7].

The medical systems based on the EEG data require a different number of electrodes and this depends on the type of the application. This number takes the range from 4 to 256 electrodes. For example, the adults require a greater number of electrodes than the children [8].

1.1 Motivation

EEG-based healthcare systems particularly in distant and continuous monitoring applications usually need very big quantities of information to be recorded, transmitted and processed. For example, high-density EEG devices that consist of up to 100 electrodes, each with a sampling rate of 1000 samples/s and every sample is represented by two bytes. Thus, generated data rate can be 1.6 Mbps for one patient. In addition, medical information must be reported to the Mobile-Health Cloud (MHC) every five minutes in normal cases while in emergencies in which highly intensive monitoring is required the collected medical data must be reported every ten seconds [9]. The wireless transmitting of these large amount of health data generates high network traffic, errors and increases congestion, consumed energy, and the latency. Hence, to increase the performance of the network and the smart-healthcare system, it is essential to reduce the huge EEG data using an efficient data reduction method before sending it over the network. Data compression is the effort to reduce the large amount of EEG health data. This aspect in particular is the main focus of this paper.

1.2 Contribution

Lossless Fractal Combined with Fixed Length Encoding (FCFLE) is proposed as a compression approach to extend Fractal compression concept to be a lossless model. EEG signal is kept intact at 100% accuracy by encoding EEG error signal and attached with compressed file to finally transmitted to the destination. The proposed compression model aims to cut down data transmitting delays, reduce storage and enhance the performance of WBSN. The network architecture is explained in Fig. 1. The suggested method is implemented at the PDA.

To verify the efficiency of the proposed approach, a comparison has made with lossless technique recently published [10] and it has found that the proposed model has outperformed it completely.

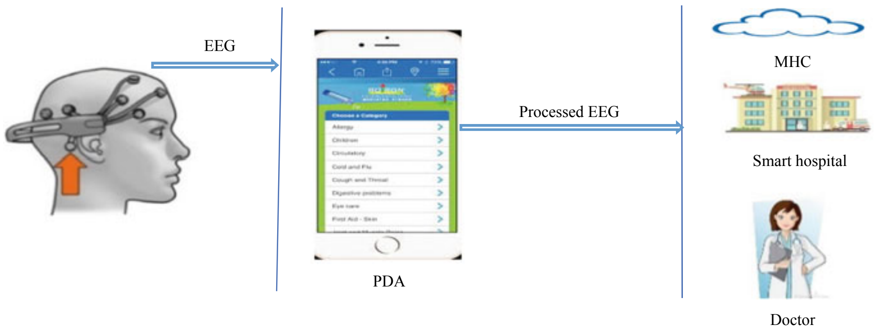


Fig. 1 Proposed system scenario

1.3 Paper Layout

The rest of the paper has the following arrangement: Sect. 2 declares the Related work. Section 3 explains the proposed model: Lossless Fractal combined with Fixed Length Encoding for EEG. In addition, the algorithms of the proposed technique for EEG compression, decompression, encoding are discussed. Section 4 describes and shows the practical work to test the proposed technique, the experimental results are compared to other previous related work in this section. Section 5 presents the conclusions and future work.

2 Related Works

Several approaches have been explored to reduce the volume of transmitted health data and keep device battery survive longer. In the conditions where information should remain 100% intact compared to the original data, lossless data compression is chosen. However, in lossless compression methods a lower data reduction ratio can be achieved. It exploits the similarity of data; no extra information is added and the original data don't lose any information when these parts of data are deleted.

An interesting efficient and simple lossless compression technique has been proposed by Hejrati et al. [10]. Both inter-channel and intra-channel correlation are exploited. At first stage, a preprocessing step is taken to extract intra-channel correlation by utilizing differential pulse code modulation technique. Channels are gathered in distinct clusters, the centroid of every cluster is computed and encoded by using arithmetic coding. At second stage, in every cluster the difference between the centroid and other channels data is computed and encoded by arithmetic coding.

Another study achieved by Hejrati et al. [11]. They proposed a learning-based adaptive transform. They have combined Discrete Cosine Transform (DCT) with neural network method.

On the other hand, lossy reduction attempts to reduce the large amount of data with a minimal loss in information as possible. Higgins et al. [12] have developed a quantization technique that requires little power to reduce data size. The significant reduction bit level can lead to loss of much information. Compression approaches playing factors by changing Discrete Wavelet Transform coefficients quantization level in SPIHT. SPIHT is the art signal compression state based on DWT. Increasing quantization and using SPIHT as an entropy encoder have been shown.

Hussein et al. [13] applied the wavelet transform to compress the EEG data and saving energy. They decreased the distortion in a distinct channel using an optimization approach. Lossy compression approach based on Discrete Wavelet Transform and Adaptive Arithmetic Encoding for EEG under epilepsy disease has been proposed by Nguyen et al. [14]. They exploited the features of epileptic EEG signal to enhance CR. EEG-based seizure detection system used the reconstructed EEG signal to estimate the performance of lossy compression method and high results are achieved for compression ratio by this lossy technique. Ben Said et al. [15] have presented single and multiple modality compression approach based on deep learning scheme at PDA layer. Exploited inter and intra correlation to improve performance. Stacked Auto-Encoders SAE are used for mobile-health system. They analyzed and compared their proposal with other compression methods such CS, DCT and DWT.

3 Lossless Fractal Model Combined with Fixed-Length Encoding Technique

Fractal is defined as an entity consisting of smaller objects, similar to the original object. The fractal descriptions that represent the following characteristics are [16]:

- Fractal is good organized object and it is possible to see the details for any size.
- Fractal offers a geometric explanation for irregular structures which cannot be represented by other mathematical models.
- Fractal represents certain properties of self-similarity.

In the proposed EEG compression approach, the fractal's self-similarity feature has been used.

3.1 Fractal Similarity Measurement

Fractasis definition of selfsimilarity that seeks similarity of the same entity with different scaling and offsetting factors [16]. Basic mathematical Fractal coefficients are Scale (Sc) and Offset (Of). Range (r) is the original numerical object that is divided into several blocks. By using many mathematical ways, the original object generates the Fractal domain object. The Down Sampling approach is applied to introduce this domain. It is smaller than the range (domain = 1/2 range). The blocks in the range are compared with the blocks in the domain to check the similarities and introduce the best match. The Fractals Root Mean Square Error is computed to every match to find the minimum one. The decompression approach will apply the factors of scale (Sc) and offset (Of) to regenerate the domain. These factors are computed as in Eqs. (1) and (2).

$$Sc = \frac{b\sum_{i=1}^{b} d(E_i)r(E_i) - \sum_{i=1}^{b} d(E_i)\sum_{i=1}^{b} r(E_i)}{b\sum_{i=1}^{b} d(E_i)^2 - (\sum_{i=1}^{b} d(E_i))^2} \tag{1}$$

$$Of = \frac{1}{b}\left(\sum_{i=1}^{b} r(E_i) - Sc\sum_{i=1}^{b} d(E_i)\right) \tag{2}$$

Table 1 Four instances of affine transform

No.	Case of transformation
0	Without change in domain block
1	Reflection points in the domain block
2	Exchange between second half and first half of block in the domain
3	Substitution and reflection first half with the second half in domain block

where b represents the size of the block, r(Ei) refer to the *i*th data sample in the block of the range, d(Ei) refers to the *i*th data sample in the block of the domain. The minimum RMS value shows the best similarity to the block in the domain. By using Eq. (3), the value of RMS is calculated [16].

$$RMS = \sqrt{\frac{1}{b}\left[\begin{array}{l}\sum_{i=1}^{b} r(E_i)^2 + Sc\left(Sc\sum_{i=1}^{b} d(E_i)^2 - 2\sum_{i=1}^{b} d(E_i)r(E_i) + 2Of\sum_{i=1}^{b} d(E_i)\right)\\ + Of\left(bOf - 2\sum_{i=1}^{b} r(E_i)\right)\end{array}\right]} \tag{3}$$

To ensure the high similarity between the two blocks in the domain and range, the Affine transform is applied. Four instances are applied by the proposed approach as affine transformations (see Table 1).

3.2 Fractal Compression for EEG Signal

The main idea of fractal compression is to fill the object of range directly from the data of EEG. Then, the vector of domain is created by from the vector of range using down sampling method. Every two consecutive EEG points are replaced by their average value. Dividing the original range and computed domain into non-overlapping blocks with specific size. Fractal independent measurements in the equations are computed in advance to increase the performance of fractal compression technique. After that, a particular transform is applied and every range block is matched with all other non-overlapped domain blocks. Fractals jump step (JS) factor gives the location of next matching block in the domain and determine the overlapping of them, in the proposed method JS size is equal to BS size. Rather than the original EEG blocks in the output file, the transformation coefficients (generally termed fractal coefficients) are stored. As a result of this, only fractal coefficients plus indexes are quantized to the minimum binary representation and included in the compressed file. The scale factor is quantized by Eq. 4.

$$Sc_q = round\left(\frac{Sc}{Sc_{Max}} \times \frac{2^{nb} - 1}{2}\right) \tag{4}$$

where nb represents the number of bits that specified to encode scale factor, Sc_{Max} refers to the maximum value for Sc coefficient. Sc_q represents the quantized scale. The offset coefficient is quantized by Eq. 5.

$$Of_q = round\left(\frac{Of - Of_{Min}}{Of_{Max} - Of_{Min}} \times \left(2^b - 1\right)\right) \tag{5}$$

where b represents the number of bits that allocated to encode offset parameter, Of_{Max} is the maximum value for offset and Of_{Min} refers to minimum permitted offset value. Of_q represents the quantized offset. Equation (6) is used to quantize the domain block index.

$$LOCq = \frac{LOC}{JS} \tag{6}$$

where LOC is the index of domain block, JS is jump step value (i.e. JS defines the distance between each consecutive blocks of domain vector) and LOC_Q is the quantized domain block index.

The fractals compression approach is displayed in Fig. 2. Algorithm 1 is responsible for providing domain and range. Algorithm 2 is utilized to provide Sc, Of, and RMS for any block in the Range similar to a block in the domain.

Fractals independent measurements are:

- $\sum r(E_i)$: sums EEG points inside range block.
- $\sum r(E_i)^2$: sums of squared EEG points in range block.
- $\sum d(E_i)$: sums of points in the domain block.
- $\sum d(E_i)^2$: sums of squared frequencies in the domain block.
- $(\sum d(E_i))^2$: squared sums of each domain block.

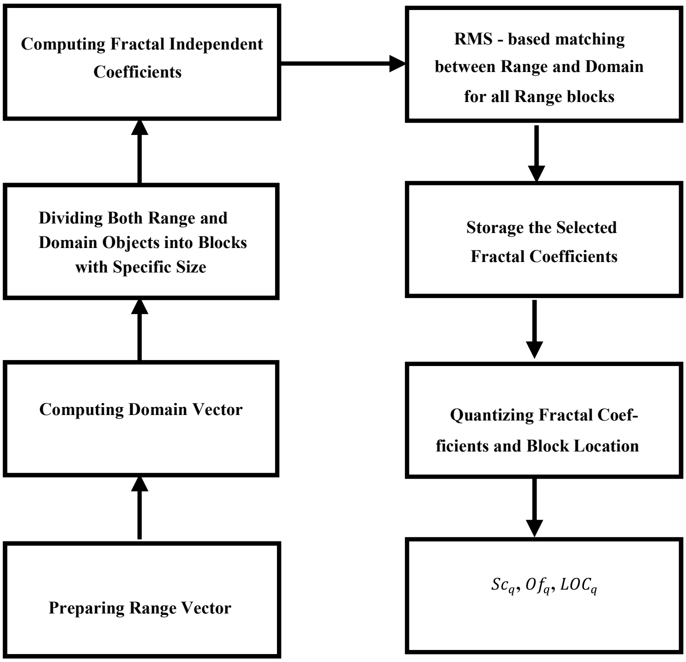

Computing Fractal Independent Coefficients
RMS - based matching between Range and Domain for all Range blocks
Dividing Both Range and Domain Objects into Blocks with Specific Size
Storage the Selected Fractal Coefficients
Computing Domain Vector
Quantizing Fractal Coefficients and Block Location
Preparing Range Vector
Sc_q, Of_q, LOC_q

Fig. 2 Fractal compression model

Algorithm 1. Preparing Range and Computing Domain

```
Inputs:
Reading // hold multiple streams of EEG
P // one packet in Reading
Outputs:
Range // Range vector converted from Reading
Domain // Domain search vector computed from Range by down sampling
for $each\ P$ in Reading do
    $Range_{P} \leftarrow Reading_P$;
    P $\leftarrow P + 1$;
end for
NTOTR // hold size of Range
k ← 0;
for $i \leftarrow 0\ to\ NTOTR$ in Range do
    $Domain_k \leftarrow (Range_i + Range_{i+1})\ /2$
    k←k+1
    i←i+2 ;
end for
$return$ Range, Domain
```

3.3 Fractal Decompression for EEG Signal

Fractal decompression is basically a reverse operation to compression process. EEG Fractal decompression approach is shown in Algorithm (3). The decompression algorithm only gets the quantized coefficients of Fractals (Sc_q, Of_q, LOC_q, AFF) and the determined block size (BS). De-quantization process is required in order to reconstruct Fractal coefficients [16]. The reconstructed scale (Sc_r) coefficient is found by Eq. 7 which is derived from Eq. 4.

Algorithm 2. EEG Fractal Compression

```
**Inputs:**
R // Range vector
D // Domain vector
B // Size of block in Range and Domain objects
NTOTR   // Range object size
NTOTD   // Domain object size
J ← B   // represents jump step in Domain search object
NR ← $NTOTR/B$ //   number of Blocks in Range vector
ND ← $NTOTD/B$ //  number of Blocks in Domain vector
RMSI ←10000 // initialization Fractal RMS with big value
**Outputs:**
SC //  array for storing Scale values
OF //  array for storing Offset  values
LOC //  array to store locations of the selected Domain block
AFF //  array  for storing affine transform case
for $i \leftarrow 0\ to\ NR$ do
    $Rs_i \leftarrow 0$; // sums EEG points inside range block
    $Rsseq_i \leftarrow 0$; // sums of squared EEG points in range block
    for $j \leftarrow (i * B)\ to\ (i * B + B)$ do
    $Rs_i \leftarrow Rs_i + R_j$;
```

```
        $Rsseq_i \leftarrow Rsseq_i + (R_j)^2$;
        i ← i+1, j ← j+1;
    end for
end for
for $i \leftarrow 0\ to\ ND$ do
    $Ds_i$ ← 0, // sums of points in the domain block
    $Dsseq_i$ ← 0; // sums of squared frequencies in the domain block
    for $j \leftarrow (i * B)\ to(\ i * B + B)$ do
        $Ds_i \leftarrow Ds_i + D_j$;
        : $Dsseq_i$ ← : $Dsseq_i + (D_j)^2$;
        $Dsallseq_i \leftarrow (Ds_i)^2$;
        i←i+1;
        j←j+1;
    end for
end for
for $i \leftarrow 0\ to\ NR$ do
    $S_i \leftarrow (B * RDs_i - (Ds_i * Rs_i))/(B * Dsseq_i - Dsallseq_i)$;
    $O_i \leftarrow (Rs_i - S_i * Ds_i)/B$;
    $RMSN_i \leftarrow \sqrt{\frac{(Rsseq_i + S_i * (S_i * Dsseq_i - 2 * RDs_i + 2 * O_i * Ds_i) + O_i * (B * O_i - 2 * Rs_i))}{B}}$;
    If $(RMSN_i < RMSI_i)$
        $RMSI_i \leftarrow RMSN_i$;
        $LOC_i$ ← i;
        $SC_i$ ← SSi;
        $OF_i \leftarrow O_i$;
        $AFF_i \leftarrow AC_i$; //Hold affine transformed domain block
    end if
end for
$\boldsymbol{return}$ SC, OF, LOC, AF
```

```
Algorithm 3. EEG Fractal Decompression
Inputs:
B, SC, OF, LOC, AFF
Outputs
DeRange // decompressed EEG signal
Set $DeDomain_k$ to 0, for every k, k ← 1,…, $NTOTD$
for $i \leftarrow 1\ to\ 4$ do
  for $ind \leftarrow 0\ to\ NR$ do
    S← $SC_{ind}$ // scale of one block
    O ← $OF_{ind}$ // offset of one block
    j ← 0;
    base ← ind * B;
    DeLoc ← $LOC_{index}$;
    $AffineD$ ← AffDeTransformation (DeLoc, $AFF_{index}$);
    for $k \leftarrow base\ to\ (base + B)$ do
      DeRange$_k$ ← $AffineD_j$ * S + O;
      k ← $k + 1$;
      j ← $j + 1$;
      ind← $ind + 1$;
    end for
  end for
  Set both m & k to 1;
  While ( m≤ $NTOTR$ ) do
    $DeDomain_n$ ← ($DeRange_m$ + $DeRange_{m+1}$)/ 2;
    n ←n+1;
    m ← m+2;
  end while
  i← i+1
end for
return DeRange
```

.

$$Sc_r = Sc_q \times Sc_{Max} \times \frac{2}{2^{nb} - 1} \tag{7}$$

where *nb* represents the number of bits that utilized for scale factor quantization, Sc_{Max} is the maximum permitted scale value. Equation 8 which is concluded from Eq. 5 is used to de-quantize offset coefficient Of_r

$$Of_r = Of_q \times \frac{Of_{Max} - Of_{Min}}{2^b - 1} + Of_{Min} \tag{8}$$

where *b* is the number of bits utilized to quantize offset factor, Of_{Max} and Of_{Min} are the maximum and minimum permitted value of offset coefficient respectively. Domain block index can be de-quantized by utilizing Eq. 9 that derived from Eq. 6.

$$LOC_r = Sc_q \times JS \tag{9}$$

where *JS* is the value of jump size.

To reconstruct the block of range, Eq. (10) is used. The vector of the domain is set to 0. The coefficients of dequantized Fractals is used with the reconstructed blocks of the domain to reconstruct the vector of range. Algorithm 3 shows the Fractals decompression.

$$r = Sc \times d + Of \tag{10}$$

where r and d refer to the blocks of domain and range.

3.4 Fixed Length Encoding for EEG

Fixed-Length Encoder is used to encode the error EEG signal and then the encoded EEG error signal is sent along with compressed file. In theory, Fixed length technique encodes all the samples in the error EEG signal with the same length. The length of bit mainly depends on the number of bits of maximum difference in EEG error signal. Equation (11) is used to compute the number of bits required to encode the maximum difference value [17].

$$No_B = round(\log(m) + 0.5) \tag{11}$$

where m represents the maximum difference in EEG error signal and (0.5) is added to the logarithm result to ensure that the needed number of encoding bits are covered [17]. The lossless Fractal combined with fixed-length encoding is explained in the following steps. First, using Fractal for EEG Compression/Decompression model. Then, the differences between the original and decompressed EEG signal are calculated to generate EEG error signal, the maximum difference value in error signal is computed and the number of bits that required to encode the maximum difference is computed also by Eq. 11. Finally, all the samples in EEG error signal are encoded based on the max diff. code.

4 Experiments and Results

This section presents the results of testing the proposed compression model: Lossless Fractal Combined with Fixed-Length Encoding model and evaluating its performance by implementing multiple experiments utilizing real vital signs based on different metrics of performance. In addition, to verify the efficiency of the proposed approach, a comparison has made with various techniques recently published and it has found that the proposed model has outperformed them completely.

The proposed model is tested by using public Bonn University dataset [18]. This EEG dataset includes five groups indicated A-E. Every EEG set contains 100-single channel of duration 23.6 s. Group A contains EEG records of five awake volunteers with eyes open. Group B includes EEG records of five awake persons but with eyes closed. C and D groups include EEG signals for patients with epilepsy recorded during seizure free periods and the group E contains only seizure activities.

This paper utilized some performance metrics such as PRD, processing time, and Compression Ratio to evaluate the proposed compression approach. The results are shown in Tables 2, 3, 4, 5, 6 and 7. The increase in BS specifically has resulted in a higher Compression Ratio but it is important to notice that EEG datasets and BS in this lossless compression model affect the resulting compression ratio. If the maximum difference between the original EEG and decompressed EEG signal is significant, it will require more bits to encode the differences resulting in an increase in compressed file size thus reducing Compression Ratio. As evidenced in Table 3, when using dataset Z the proposed model has achieved CR 3.63 for block size of 50 and decreased to 3.40 for block size of 100.

Finally, the proposed lossless Fractal compression model is compared with recent lossless model [19] as shown in Table 7. It has resulted in potential Compression Ratio and greatly outperformed other compression mode.

Table 2 CR, PRD, processing time values of dataset Z for several block size

BS	CR	PRD	Processing time	
			Compression time	Decompression time
10	3.20	0	0.056	0.029
50	3.63	0	0.038	0.024
100	3.40	0	0.025	0.020
150	3.45	0	0.019	0.013
200	3.47	0	0.013	0.009

Table 3 CR, PRD, processing time values of dataset O for several block size

BS	CR	PRD	Processing time	
			Compression time	Decompression time
10	2.66	0	0.056	0.029
50	3.26	0	0.038	0.024
100	3.40	0	0.025	0.020
150	3.45	0	0.019	0.013
200	3.47	0	0.013	0.009

Table 4 CR, PRD, processing time values of dataset N for several block size

BS	CR	PRD	Processing time	
			Compression time	Decompression time
10	2.46	0	0.056	0.029
50	3.26	0	0.038	0.024
100	3.40	0	0.025	0.020
150	3.45	0	0.019	0.013
200	3.47	0	0.013	0.009

Table 5 CR, PRD, processing time values of dataset F for several block size

BS	CR	PRD	Processing time	
			Compression time	Decompression time
10	2.91	0	0.056	0.029
50	3.63	0	0.038	0.024
100	3.81	0	0.025	0.020
150	3.87	0	0.019	0.013
200	3.90	0	0.013	0.009

Table 6 CR, PRD, processing time values of dataset S for several block size

BS	CR	PRD	Processing time	
			Compression time	Decompression time
10	1.88	0	0.056	0.029
50	2.50	0	0.038	0.024
100	2.58	0	0.025	0.020
150	2.60	0	0.019	0.013
200	2.62	0	0.013	0.009

Table 7 CR of proposed lossless fractal model for five datasets (Z, O, N, F, S)

Proposed lossless model	BS	CR				
		Z	O	N	F	S
	10	3.20	2.66	2.46	2.91	1.88
	50	3.63	3.26	3.26	3.63	2.50
	100	3.40	3.40	3.40	3.81	2.58
	150	3.45	3.45	3.45	3.87	2.60
	200	3.47	3.47	3.47	3.90	2.62
Another model [19]		Z	O	N	F	S
		2.01	1.84	2.23	2.19	1.44

5 Conclusion and Future Work

In this paper, a novel compression approach has been implemented. Lossless Fractal combined with Fixed Length Encoding approaches for compressing EEG data in WBSNs. The main purpose is to explore the possibilities of improving the performance of WBSN via decreasing the EEG data over the network. The PDA is operated by battery; therefore, it is important to conserve its transmission power. The proposed approach has exploited redundancies and correlation inside EEG signal itself. The Compression Ratio, PRD metrics and processing are used for compression model testing. The suggested approach enhances the performance of the WBSN and the results show that the proposed approach introduced a high-quality result and identical to the original EEG signal is regenerated in comparison with other methods. For future work, variable-length encodings like Huffman method for lossless Fractal compression to obtain even better compression ratios can be investigated.

References

1. Gravina R, Alinia P, Ghasemzadeh H, Fortino G (2017) Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. Inf Fusion 35:68–80
2. Fatima M, Kiani AK, Baig A (2013) Medical body area network, architectural design and challenges: a survey. Wireless sensor networks for developing countries. Springer, Berlin, Heidelberg, pp 60–72
3. Mavinkattimath SG, Khanai R, Torse DA (2019) A survey on secured wireless body sensor networks. In: 2019 international conference on communication and signal processing (ICCSP). IEEE, pp 0872–0875
4. Movassaghi S, Abolhasan M, Lipman J, Smith D, Jamalipour A (2014) Wireless body area networks: a survey. IEEE Commun Surv Tutorials 16(3):1658–1686
5. Jaber AS, Idrees AK (2021) Energy-saving multisensor data sampling and fusion with decision-making for monitoring health risk using WBSNs. Softw Pract Experience 51(2):271–293
6. Hayajneh T, Almashaqbeh G, Ullah S, Vasilakos AV (2014) A survey of wireless technologies coexistence in WBAN: analysis and open research issues. Wirel Netw 20(8):2165–2199
7. Hu B et al (2015) Signal quality assessment model for wearable EEG sensor on prediction of mental stress. IEEE Trans Nanobiosci 14(5):553–561
8. Seeck M et al (2017) The standardized EEG electrode array of the IFCN. Clin Neurophysiol 128(10):2070–2077
9. Abdellatif AA et al (2019) Edge-based compression and classification for smart healthcare systems: concept, implementation and evaluation. Expert Syst Appl 117:1–14
10. Hejrati B, Fathi A, Abdali-Mohammadi F (2017) Efficient lossless multi-channel EEG compression based on channel clustering. Biomed Signal Process Control 31:295–300
11. Hejrati B, Fathi A, Abdali-Mohammadi F (2017) A new near-lossless EEG compression method using ANN-based reconstruction technique. Comput Biol Med 87:87–94
12. Higgins G, McGinley B, Jones E, Glavin M (2013) An evaluation of the effects of wavelet coefficient quantisation in transform based EEG compression. Comput Biol Med 43(6):661–669
13. Hussein R, Mohamed A, Alghoniemy M (2015) Scalable real-time energy-efficient EEG compression scheme for wireless body area sensor network. Biomed Signal Process Control 19:122–129

14. Nguyen B, Ma W, Tran D (2018) A study of combined lossy compression and seizure detection on epileptic EEG signals. Procedia Comput Sci 126:156–165
15. Al-Sa'D MF et al (2018) A deep learning approach for vital signs compression and energy efficient delivery in mHealth systems. IEEE Access 6:33727–33739
16. Ibaida A, Al-Shammary D, Khalil I (2014) Cloud enabled fractal based ECG compression in wireless body sensor networks. Future Gener Comput Syst 35:91–101
17. Al-Shammary D, Khalil I (2012) Redundancy-aware SOAP messages compression and aggregation for enhanced performance. J Netw Comput Appl 35(1):365–381
18. Andrzejak RG et al (2001) Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys Rev E 64(6):061907
19. Srinivasan K, Dauwels J, Reddy MR (2011) A two-dimensional approach for lossless EEG compression. Biomed Signal Process Control 6(4):387–394

Securing the Hyperconnected Healthcare Ecosystem

Ramon Sanchez-Iborra and Antonio Skarmeta

Abstract Digitalization across the healthcare industry is improving the quality and personalization of the services provided to patients. Currently, we are observing great advances in aspects such as doctor-patient coordination, real-time health monitoring, communication between different specialist, or administrative tasks automation, among others. However, this hyperconnectivity also brings important challenges related to the cyber-security of the critical eHealth infrastructures. Privacy concerns, intrusion detection, secure data exchange, etc. are crucial factors that should be addressed when designing digital healthcare platforms. In this chapter, a wide overview of the cyber-security landscape in the eHealth sector is presented, emphasizing the biggest challenges to be faced during the next years. Then, as the main contribution of this work, it is proposed a holistic cyber-security platform that tackles privacy and security risks in an automated fashion to foster the development of innovative applications within the healthcare ecosystem.

Keywords eHealth · Cyber-security · Digitalization · Security platform

1 Introduction

The recent acceleration of digital technology and online connectivity within the healthcare, so called eHealth, has led to significant improvements in care delivery for citizens worldwide. However, the downside is increased exposure of these sectors' infrastructures and data to cyber-threats and attacks of growing complexity that may negatively impact on the privacy and safety of citizens, e.g., patients, caregivers and relatives, health professionals, etc. Similar developments can also be observed in

R. Sanchez-Iborra (✉)
University Center of Defense, General Air Force Academy, San Javier, Spain
e-mail: ramon.sanchez@cud.upct.es

A. Skarmeta
Computer Science Faculty, University of Murcia, Murcia, Spain
e-mail: skarmeta@um.es

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_22

other domains such as the energy sector, where, for example, smart and sustainable energy systems have been developed, but at the same time the surface for cyber-attacks has grown [1]. In this line, international initiatives such as the European Directive on security of Network and Information Systems (NIS Directive)[1] identifies such sectors as critical and calls for a better alignment of these domains' needs with the supply side of cyber-technologies. The latter must incorporate security, privacy, and personal data protection by design and by default, as dictated by the European General Data Protection Regulation (GDPR).[2] However, there is still much to be accomplished so as to:

- Enable the practical implementation of relevant international legislation in complex ecosystems.
- Align the cyber-security technologies with the specific domain needs, thus linking the demand and supply sides for cyber-security solutions.
- Build security, privacy, and personal data protection by design and by default in the cyber-security technologies offered to these sectors, so as to ensure the highest levels of protection, data integrity, and confidentiality.
- Demonstrate the impact generated by deploying system prototypes in relevant operational environments and proving the cross-domain applicability of all developed solutions.

Indeed, healthcare is more digitally connected than ever before and dependent on cutting-edge technologies, such as Big Data, Internet of Things (IoT), Artificial Intelligence (AI), cloud computing, etc. [2]. Apart from physical facilities, Healthcare IT (HIT) systems now include medical devices, medical Internet of Things (mIoT), wearables, telemetry equipment, a new breed of healthcare applications, and more [3], as illustrated in Fig. 1. In addition, healthcare delivery is moving beyond the hospital perimeter to focus on the patient in less expensive environments that facilitate care management, i.e., to home care and community care settings. Thereby, a new model of integrated care is basically emerging, demanding collaboration among various stakeholders, numerous interconnected assets and resources, high flexibility requirements, and high levels of security and privacy of sensitive information, very often on a cross-border basis. However, with exponentially increasing health data set sizes and digital connectivity, the cyber-attack canvas and associated cyber-threat models in healthcare are also expanding [4]. Due to the significant tangible and intangible assets at stake, e.g., sensitive personal information, which can be sold to third parties, financial resources, but also patients' "lives", ransomware, malware and phishing attacks as well as data breaches have soared in the healthcare sector [5] with growing concern about (i) attacks targeted at connected devices (for which cooperation of medical device manufacturers is crucial); (ii) systemic cyber-attacks that would affect an entire health ecosystem, e.g., hospitals, care centers, pharmacies, labs, etc., possibly on a cross-border basis; and (iii) human errors, e.g., ICT systems'

[1] https://digital-strategy.ec.europa.eu/en/policies/nis-directive.

[2] https://ec.europa.eu/info/law/law-topic/data-protection_en.

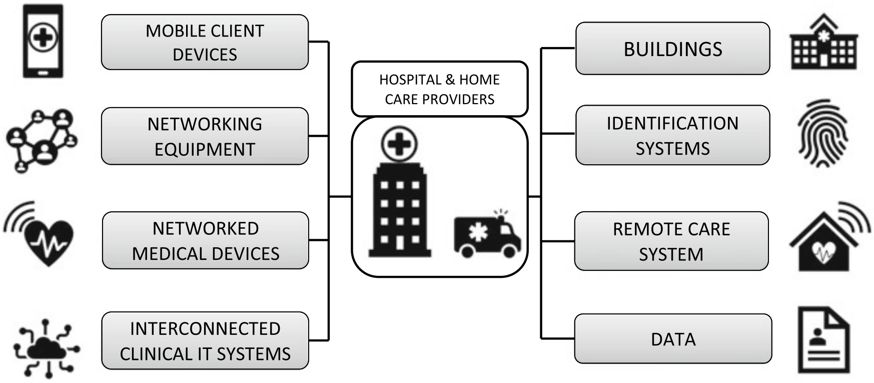


Fig. 1 Typical assets of eHealth ecosystem

configuration errors, unauthorized access control, non-compliance, that create vulnerabilities potentially exploitable by attackers.

In this chapter, it is provided a wide overview of the cyber-security challenges that the eHealth sector will face during the next years as its digitalization process comes to fruition. Besides, a holistic security platform for the critical sector of healthcare is proposed with the aim of covering the entire cyber-security, data protection, and privacy risk handling lifecycle. The rest of the document is organized as follows. Section 2 discusses and provides the necessary background of the pillar technologies for enabling the development of secure eHealth platforms. The most relevant healthcare application scenarios are discussed in Sect. 3. Section 4 presents an holistic cyber-security-aware eHealth architecture. Finally, Sect. 5 closes this chapter, summarizing the most important aspects and drawing future research lines.

2 Background: Security Challenges in e-Health Cyber-Infrastructures

Cyber-security-aware eHealth platforms should introduce new collaborative and holistic approaches for digital security, privacy, data protection, and accountability in healthcare. The focus should be not only on smart hospitals infrastructure and data, but also on home care and community care environments, with a constant emphasis on patient's privacy and safety. Besides, novel security assets for identifying and assessing vulnerabilities and support information-sharing between all involved stakeholders should be developed. Additional reliable and user-friendly tools for enhancing cooperation in response and recovery as well as for providing specialized training to non-technical users are also necessary to close the loop in this complex scenario. Aligned with this diagnosis we discuss the next steps to be taken by separating the discussion in four different pillars.

2.1 Cryptography

Developing provably-secure cryptographic technologies that allow for sharing aggregated data or computations on sensitive data without exposing more information than intended is of utmost importance to encourage data owners to actively participate in data exchange [6]. Such secure aggregation mechanisms would encourage patients to securely share their data with other stakeholders within but also outside the healthcare ecosystem, e.g., research institutions or open data platforms. Companies would also be encouraged to voluntarily provide information about their incident data to national and international cyber-security workgroups such as the NIS Cooperation Group[3] in an integrity-preserving way.

The most common model is the one where individual data producers are not actively involved in the aggregation, but simply send their data to an aggregator, which collects all the data and then performs the computation. Other approaches that involve the end-devices into the aggregation are either too inflexible or too inefficient for resource-constraint data producers: They either require frequent communication between the data sources, distribute the aggregation to a set of aggregators, or even require a number of communication rounds among all data producers as well as the data producers and the aggregator. However, the first steps are being taken towards the optimization of this process by the integration of intelligence within constrained devices [7].

In the aforementioned model, one can make different assumptions about the aggregator, i.e., whether it is fully trusted, semi-honest or even fully malicious. They also differ in the types of aggregation/computations required to be performed by the aggregator, i.e., only simple functions such as summation, statistical moments such as mean or variance or more complex computations. Depending on the desired security guarantees with respect to the involved parties, different cryptographic tools are required. A very common, but inflexible setting is to assume that the aggregator is honest-but-curious, i.e., computes the summation of all the single inputs and only learns the aggregation results but none of the single inputs. A cryptographic primitive to securely achieve such a functionality is aggregator-oblivious encryption (AOE) [8]. In AOE the aggregator learns the aggregation results. Preventing the aggregator from learning the output and making it only available to certain receivers can be achieved by means of homomorphic proxy re-encryption [9]. Orthogonal to the confidentiality guarantees are authenticity guarantees of the involved data. In particular, that (i) the aggregator can be sure that it operates on authentic inputs and that the party receiving the result can be sure that (ii) the computation was performed correctly and (iii) the computation has been performed on authentic inputs. If one does not require confidentiality, it can be achieved by means of verifiable computing techniques [10]. While these techniques are very powerful and can handle very large classes of computations, despite all the approaches to apply them to real world scenarios, their costs, especially for the aggregator, are usually still prohibitive. An alternative approach to simultaneously achieve the key points mentioned above can be achieved by means

[3] https://digital-strategy.ec.europa.eu/en/policies/nis-cooperation-group.

of homomorphic authenticators [10]. Besides, promising techniques to verifiably compute on encrypted data from homomorphic hashing and a combination of homomorphic encryption and authenticators are being developed [11].

2.2 Intrusion Detection

In an ever changing and highly sophisticated threat landscape, healthcare organizations must have the right capabilities in place in order to detect, respond, and mitigate cyber-attacks. However, most of the events/behaviours occurring on a network are not malicious, and most of the system behave as expected. As a result, malicious activity must be isolated by closely looking at systems, users or applications that deviate from the expected behaviour [12].

Currently, these capabilities comprise a number of signature-based security technologies designed to deliver real-time detection. These include antivirus software, intrusion prevention systems, firewalls, content filtering proxies, system-based security layers, web application firewalls, and others [13]. Information supplied by those systems is captured, normalized, aggregated, and correlated in a Security Information Event Management (SIEM) platform, that provides a single view of all security events being triggered, and enables real-time monitoring [14]. While this approach remains valid, it has become clear that it is not enough to defeat advanced attacks against complex organizations and, more importantly, identify previously not-seen malicious behaviour. Traditional Intrusion Detection Systems (IDSs) and SIEMs are usually based on signature-based threat/incident detection, which are not suitable to cover complex attacks. This is because signatures exhibit two limitations: First, they can only detect conditions that have been seen before, and so they cannot identify zero-day attacks or custom attacks that are not replicated across organizations; second, they cannot account for variants of a given attack that do not fit into the signature definition. With the kill chain becoming multi-stage and multi-vector, relying on real-time monitoring to defeat attacks against healthcare systems that can span for weeks is not an option. Moreover, the correlation time window of a SIEM is limited; as such, the possibility to define complex use cases that integrate and make meaningful conclusions about events occurring in long timeframes is not available [15]. Advanced attacks can often be detected only if a deviation from a trend or a typical or expected behaviour is assessed, which is not available in traditional SIEMs. They do not leverage statistical analysis and modelling techniques, however proven to be useful to gather insights not available with traditional approaches [16]. Since SIEMs and classic security technologies are not data-centric, time-to-insights is high, which hinders investigations to deliver timely results.

In the light of the previous discussion, it is desirable the development of analysis tools able to correlate behaviour from multiple sources to catch multi-vector attacks and digests data from multiple sources to fully track attacker activities, including data source from user devices. Thus, the global risk assessment and fraud/intrusion detection should be done considering the device context, involved app, exchanged data, and specific risks depending on the use case, environment, and user. The objective is to obtain risk assessment and management tools capable of (i) having low-dependency on signatures, and relying more on statistical analysis, data mining, and modelling instead; (ii) enabling long-term analysis of security data, allowing to detect advanced attacks that span across long timeframe; and (iii) enabling complex use cases where past information can be leveraged to assess current behaviour. To this end, the User and Entity Behaviour Analytics (UEBA) are state-of-the-art techniques capable of performing such advanced analysis [17]. Besides, the improvement of SIEMs with the help of Machine Learning (ML)-related techniques is a hot and promising field of study [18].

2.3 Cyber-Threat Information Sharing

Sharing alerts is a standard practice in cyber-security systems. Alerts are either shared for free or in the framework of agreed-upon exchange framework or bought in from cyber-security intelligence companies. Sharing incident data however involves more important challenges due to trust and privacy issues [19].

Different threat streams/feeds are a standard base technology of cyber-security systems: Some of the streams originate from local installations of IDS and SIEM software, some are subscribed to in order to receive professionally created warnings from external cyber-security analysis companies and organizations. However, filtering relevant alerts from the information shared by others is still a challenging task. Furthermore, vulnerability and incident handling information is without a doubt sensitive information which parties might be reluctant to share [20]. However, it is crucial that this information is made available for analysis for different parties. Hence, tools for sharing the information so that the privacy of the data owners is preserved, but the information can still be used, are needed. Cyber-security threat intelligence (CTI) sharing has become a hot topic in information security because it is expected that the sharing of the collective wisdom of peer organizations with respect to the evolving threat landscape could enable more efficient and effective incident response [21].

Sharing mechanisms that have an international outreach include European Information Sharing and Alerting (EISAS),[4] information sharing and formatting standards developed by MITRE (TAXII, STIX) [22], and Cyber-investigation Analysis Standard Expression (CASE).[5] The value of these standards comes when there is a need to exchange CTI across different administrative domains since they provide a well-documented format and structure. Representation of CTI information has been standardized using STIX.[6] Additionally, there are numerous open source and com-

[4] https://www.enisa.europa.eu/publications/eisas-deployment-feasibility-study/at_download/fullReport.

[5] https://github.com/casework/case.

[6] https://stixproject.github.io/supporters/.

mercial CTI sharing tools available. For example, the Malware Information Sharing Platform (MISP)[7] is an open source platform supported by the Computer Incident Response Centre of Luxembourg that is widely used for incident report exchange by European Computer Security Incident Response Teams (CSIRTs). Another open source examples of CTI solutions are IntelMQ[8] or Yeti.[9]

The main issue with threat feeds is the abundance of information, i.e., only a small percentage of warnings are important and relevant to the actual system being protected [23]. In the simplest case, the relevancy is determined by which software components are present in the system being protected, which leads to the strategy that warnings about absent software components are not relevant. However, the reality is much more nuanced than this simple case. The net effect of the abundance of information is that the cyber-security specialist has to either regularly spend a large amount of time and effort to filter out and categorize relevant threats or mostly ignore the threat feed, and both situations are undesirable. The standard way of finding potential attacks and threats from logs is writing detection rules by specialists. ML is used a lot less, mostly for anomaly detection in log analysis, aimed at intrusion detection and misuse detection. While automated anomaly detection is widely used in cyber-security, ML and localized categorization of the relevant threats is still a challenge under investigation [24]. Furthermore, current CTI platforms such as MISP, YETI, IntelMQ, do not consider trust models and privacy management in CTI sharing with external sources, which is highly desirable as discussed previously.

Therefore, future secure incident-handling and forensics info-sharing systems should feature practical tools for sharing alerts in a manner which would satisfy the confidentiality and privacy-preserving needs of the domain and participants. Besides, they should include an event/alarm/threat warning filter and categorization component, which would assist a human cyber-security specialist in filtering out and categorizing relevant threats from a large stream/feed of shared alerts. Finally, these platforms should exploit STIX and TAXII as they are well known standards for the communication of threat information used and supported by a large community.

2.4 Cyber Range-Based Training

Cyber ranges (CRs) are often defined as technical environments that simulate ICT infrastructures. CRs have become a common vehicle to deliver hands-on dynamic training using cyber-defence exercises (CDX), to increase the resilience of organizations when facing cyber-incidents or cyber-attacks. CR simulations provide (i) an environment where teams can work together, (ii) an on-the-job experience simulation, and (iii) real-time feedback. Hence, these customized simulations support the enhancement of skills, abilities, processes and, decision making [25]. So far, cyber-

[7] https://www.misp-project.org/.

[8] https://github.com/certtools/intelmq.

[9] https://github.com/casework/case.

security awareness education in the healthcare domain is delivered by lecture-based or online learning courses. However, hands-on training may support experiential learning as it simulates on-the-job experience and real-time feedback. In the long run, it notably contributes to cyber-resilience of organizations [26]. During the last decade, healthcare organizations spent only 1–2% of their annual budget on IT [27]; it is important to deliver a training that supports the specific needs of the healthcare domain. Current CR technologies and applications may use technologies that are also used in the healthcare domain, e.g., IoT, Bluetooth, or Industrial Control Systems (ICS) but so far only few testbeds or CR address specifically healthcare incidents and attacks. Hence, hands-on exercises in the healthcare sector is still at its beginning [28].

After this discussion, it is clear that additional efforts focused on the evolution of current eHealth CRs are needed. This will fuel the development of a standard reference model that will further contribute to the discussion, development, and mitigation of healthcare system design and resilience between public and private organizations, research and industry. This will entail to customize the healthcare CDX scenarios based on vulnerability and breach databases implementing relevant state-of-the-art cyber-attacks in healthcare infrastructures. The aim of these scenarios should be to address not only the most advanced technical incidents for healthcare organizations but also business, social, legal (NIS, GDPR, eIDAS[10]) and other relevant aspects to cyber-security in healthcare.

3 Critical Healthcare Infrastructure: Application Scenarios

As explained previously, the HIT infrastructure expands beyond the limit of medical centers such as hospitals. Thus, it can be identified two broadly differentiated environments: “Smart hospitals”, where infrastructure is homogenized and planned, both at deployment time and upgrades, and “Home/Community care settings”, where the environment is heterogeneous and there is no clear central authority to enforce decisions. Challenges for each infrastructure type differ and therefore the corresponding security assets have to be developed, deployed, and operated differently. The main cyber-security challenges in both scenarios are dissected as follows.

3.1 Challenges in Smart Hospitals Environments

The Smart Hospital concept aims at building a holistic IT environment consisting of automated pieces of both specialized and auxiliary procedures. Thanks to this digitalization process, a huge amount of data, cloud computing services, and AI solutions cooperate to enable advanced eHealth services. Indeed, the provision of healthcare

[10] https://digital-strategy.ec.europa.eu/en/policies/eidas-regulation.

is increasingly using sensors and devices connected to gateways and health services providers' over the Internet. For example, medical devices for the monitoring of patient physiological data such as blood pressure, temperature, heart and respiratory rates, O_2/CO_2 levels, are now routinely operated in a network-connected remote monitored environment [29, 30]. Care devices can be of different nature, such as wearable, implantable, or external units that monitor and transmit personal data and send them to a processing platform, e.g., a handheld controller/monitor, a smart phone, a tablet, or the cloud, for aggregation, analysis, presentation, and storage/archiving; but also for alerts and direct support to care professionals via data or, sometimes automated, commands. These assets, highly dependent on various forms of ICTs, include interconnected clinical information systems, computerized record-keeping and billing systems, laboratory/radiology information services, picture archiving and communication systems, pharmacy information systems, blood bank systems, research labs systems, etc. Other involved systems are mobile or wearable devices, implantable devices, robots and other supportive devices such as alarm and remote-monitoring devices designed to detect falls, measure blood pressure, detect heart arrhythmia, sleeping troubles, respiratory anomalies and various changes in condition.

Therefore, challenges arising in the smart hospital environment related to digital security, privacy and personal data protection are heterogeneous due to the large number of networked devices, hence the large number of potential points of attack, as well as the number of actors, systems and personal data involved. A fundamental one is the communication and data security during data transfers through various networks. Threats of ransomware, network outages, identity theft, insider threats, etc. exist due to the inherent operating systems of health ICT assets. In addition, the smart hospital is an environment where acute safety, continuity of care and trust concerns exist. For example:

1. The loss of medical connected devices, or worse, the manipulation of data coming from these devices, pose a real threat to the life of the patients.
2. When medical devices are integrated into an Electronic Patient Record (EPR)/ Electronic Healthcare Record (EHR) the data are also used by clinical decision support systems to monitor and predict patient status, prompting specific interventions. The loss of such data, or corruption of same, again could have impact on the care and life of a patient.
3. Loss of access to HIS, e.g. critical care EPRs, particularly prescribing and medication administration data, could have a major impact on patient care due to the absence of manually recorded data.

3.2 Challenges in Home/Community Care Settings

The provision of home and community care heavily relies on the use of mIoT devices that include: Panic buttons, door sensors, GPS locators, medication robots, etc. Besides, given the pandemic situation, currently many home care visits take place using the patient's and professionals' tablet, laptop, or mobile devices. The main cyber-security challenges encountered in these environments emanate from mIoT devices and from end-users, namely, patients and family members, informal care givers as well as medical and health professionals, who are targeted by malicious behaviours or actors of industrial espionage. These challenges are related to specific information assets and may have significant negative impacts for the end-user and for the organization (financial, health-related, personal, psychological, trust and safety-related) due to sensitive data loss, lack/loss of monitoring, etc.

Therefore, mIoT devices should be integrated within the security platform in order to evaluate their vulnerability on different use cases attack scenarios while also providing a new set of medical data, which will be available for further analysis and dynamic and active monitoring. The focus should be on challenges related to secure embedded systems as well as stakeholder behaviour for (i) actual physical security; (ii) network security; (iii) stakeholder cooperation for achieving higher levels of protection and dynamic response and recovery; and (iv) social engineering attacks.

All the aspects discussed above are fully considered in the proposed security architecture for the healthcare ecosystem that is described in the following section.

4 Proposed Architecture

In this section, a novel security platform is proposed in order to holistically tackle dynamic security, privacy, and legal risk assessment as well as incident handling in healthcare environments. To this end, it undertakes a series of planes in four areas: (a) cryptographic solutions to provide privacy-preserving and strong provable security guarantees of data sharing of threats, incidents and sensitive data; (b) real-time ML-based analysis of exchanged threats/events from large and diverse streaming sources for filtering and categorization; (c) evidence-based and behavioural-based analysis of cyber-threats and multi-vectors attacks to detect possible unknown cyber-attacks; and (d) CDX with healthcare-specific user domain specificities. As shown in Fig. 2, these planes are distributed in five main components along the platform, namely, (i) Risk Assessment and Management (including a Dynamic Vulnerability Knowledge Base (KB)), (ii) Incident-handling Info Exchange, (iii) Continuous Monitoring, (iv) Response tools, and (v) Cyber Range (CR). These components and their elements are detailed in the following.

4.1 Risk Assessment and Management

This component utilizes the Dynamic Vulnerability KB (described below) to monitor situations and aid the cyber-security expert in decision making. Proactively, it automates attack detection, assists with incident response, and aids reactively with the

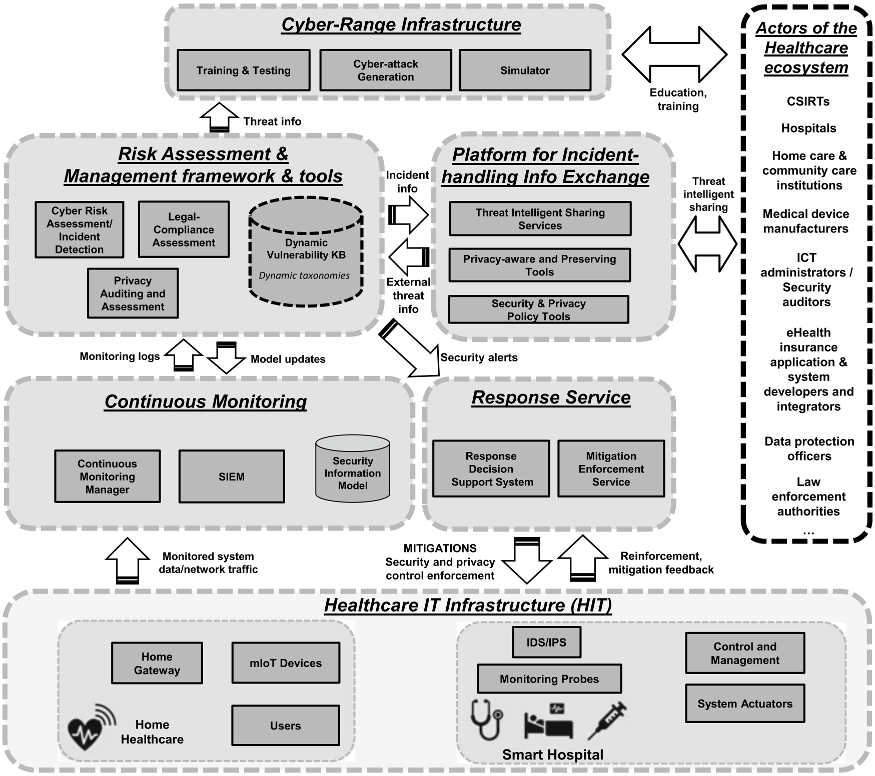

Cyber-Range Infrastructure
Training & Testing
Cyber-attack Generation
Simulator
Education, training
Actors of the Healthcare ecosystem
CSIRTs
Hospitals
Home care & community care institutions
Medical device manufacturers
ICT administrators / Security auditors
eHealth insurance application & system developers and integrators
Data protection officers
Law enforcement authorities
...
Threat info
Risk Assessment & Management framework & tools
Cyber Risk Assessment/ Incident Detection
Legal-Compliance Assessment
Dynamic Vulnerability KB
Dynamic taxonomies
Privacy Auditing and Assessment
Incident info
External threat info
Platform for Incident-handling Info Exchange
Threat Intelligent Sharing Services
Privacy-aware and Preserving Tools
Security & Privacy Policy Tools
Threat intelligent sharing
Monitoring logs
Model updates
Security alerts
Continuous Monitoring
Continuous Monitoring Manager
SIEM
Security Information Model
Response Service
Response Decision Support System
Mitigation Enforcement Service
Monitored system data/network traffic
MITIGATIONS Security and privacy control enforcement
Reinforcement, mitigation feedback
Healthcare IT Infrastructure (HIT)
Home Gateway
mIoT Devices
Home Healthcare
Users
IDS/IPS
Monitoring Probes
Control and Management
System Actuators
Smart Hospital

Fig. 2 Proposed cyber-security platform for the eHealth ecosystem

mitigation steps and future prevention of cyber-attacks. Subcomponents are described as follows:

- Cyber Risk Assessment/Incident Detection subcomponent: It goes beyond intrusion detection or threat data analysis provided by the monitoring services with the aim of recognizing both known and previously unseen attacks. It recognizes the source and target of the attack, as well as the purpose and motivation of the attacker. Big data analysis and ML techniques are used for attack/threat detection. This element correlates behaviour from multiple sources to catch multi-vector attacks, integrating data from diverse sources to fully track attacker activities. In addition, it shows how an entity behaves and how it might behave in the future, predicting attackers' behaviours.
- Legal Compliance Assessment subcomponent: This module provides dynamic, evidence-based, standard-compliant risk assessment for healthcare organizations. It considers cascading effects of threats and propagation of vulnerabilities in interconnected healthcare infrastructures.

- Privacy Auditing and Assessment subcomponent: It provides business process analysis to identify threats and redesign processes using privacy enhancing techniques and technology.
- Dynamic Vulnerability KB: It maintains and manages vulnerabilities specific to ICT-based health and social care ecosystems, technologies, applications, and services. It includes dynamic taxonomies for healthcare-related attacks, and vulnerability collection, upload, maintenance and mechanisms including API development. This component is also in charge of data fusion and harmonization of the data coming from different multi-heterogeneous sources, e.g., external entities (CSIRTs, Law Enforcement Agencies (LEAs), etc.), monitoring alerts, reactions countermeasures, and the current status of the infrastructure. It combines huge security data and log collections from diverse sources to ease the analysis of cyber-security alerts and to find suspicious behaviors in order to have the most adequate response. One key task of this module is to build a dynamic vulnerability connection with the Malware Information Sharing Platform (MISP), which is the reference database used by CSIRTs for storing, sharing, and correlating indicators of compromises of targeted attacks. This strategy permits stakeholders to exchange data about captured/detected malware and its indicators. Thereby, this collaborative knowledge sharing about known malware or threats is highly valuable for users, who benefit from the improved counter-measures used against targeted attacks and the development of preventive actions.

4.2 Platform for Incident-Handling Information Exchange

This module enables healthcare stakeholders to share in a privacy-preserving way, CTI data with external entities (CSIRTs, LEAs, etc.), according to decisions from a trust model. To this aim, a fine-grained, privacy respectful, GDPR-compliant and sustainable trust model governs information exchange that adapts to stakeholders' needs. A multi-dimensional approach to quantify trust among involved stakeholders is devised, combining applicable legal requirements, the peer's reputation, the collaboration maturity, i.e., the degree of contribution of data shared, and the membership to federations. The trust model drives the CTI information exchange. This component includes the following elements:

- Threat Intelligence Sharing Services: This asset enables the exchange of up-to-date cyber-security information and sensitive data with external entities or data providers. This allows platform administrators to receive and share their own system CTI information and sensitive data with other entities, including new inferred cyber-security risk assessments and conclusions. The communication is based on well-known standards such as STIX. Potential external entities for which data sharing will be enabled encompass CSIRTs, third-party security data providers, or Open Source Intelligence (OSINT) data sources, among others.

- Privacy-aware and Preserving Tools: This module integrates a set of novel privacy-preserving mechanisms, crypto-algorithms and tools, including extensions for healthcare stakeholders to access and share information between them, while ensuring the full protection of sensitive personal and business data. These tools will provide a series of crucial features: (i) End-to-end confidentiality and integrity for critical data; (ii) minimum data disclosure; (iii) long-term data security; and (iv) strong authentication and authorization.
- Security and Privacy Policy Tools: These tools encompass user-friendly dashboards for collecting, uploading, maintaining, and disseminating vulnerabilities of ICT-based health and social care ecosystems, technologies, applications, and services. In addition, intuitive tools to configure security, privacy and reaction policies that govern both system and network configurations, such as channel protection, traffic filtering, authentication, authorization, etc. are also integrated. The policies are the input of a specialized interpreter service that drives the monitoring and reaction planes in order to simplify the security management of the infrastructure.

4.3 Continuous Monitoring

This component is in charge of the automatic constant evaluation of the elements composing the HIT infrastructure for checking that the system is running as expected and that security and privacy requirements are met. The SIEM (Security Information and Event Management), by using an automated security evaluation service coordinated by the Continuous Monitoring Manager, performs periodic security scans by employing advanced vulnerability assessment tools. Thus, it is evaluated whether the defined security policies accessible from the Security Information Model are properly applied by performing intelligent data-driven automated and contextual monitoring of the activities in the infrastructure. To this end, different sources of key security information are considered by performing holistic monitoring of traces, status reports, event logs, and other operational information and comparison of the security status with pre-defined security policies, system models, etc. Traditional signature-based attack detection is also analyzed in this component (usually deployed directly in the healthcare infrastructure) and reported to the Risk Assessment and Management Component. Besides, many entities from different heterogeneous parts of the HIT infrastructure, hence, with different particular characteristics, are monitored by this module. For that reason, systemic adapters or bridges, which are installed on the HIT infrastructure per se, should be employed. Each bridge consists of a structured input/output layer and a configuration layer that may be adapted during the bridge instantiation. All these aspects are controlled by the Continuous Monitoring Manager.

4.4 Response Service

The quick and effective mitigation of security breach effects is a crucial aspect as HIT systems and corresponding technologies (peripherals, servers, DBs, etc.) are continually attacked and compromised. Tools to enforce remediation and countermeasures techniques are integrated within the architecture allowing a reaction capability to trigger countermeasures in case of any failure or security issue detection. The countermeasure strategies are linked to the risk level defined during the risk assessment phase and the specific actions are defined by the Response Decision Support System. Once a security issue is detected in the HIT infrastructure, a countermeasure is triggered by the Mitigation Enforcement Service according the root cause(s) identified during the diagnosis phase in the Risk Assessment and Management module. This reaction implementation is aligned with the security policies in order to restore the normal functioning as soon as possible. For this to be done, a set of automatic corrective actions are enforced into the HIT infrastructure that, in addition, allows to build a more resilient infrastructure by applying several strategies like: Network slicing, services isolation, usage of Virtual Private Network (VPN), orchestration of virtualized services, etc. These measures evolve the security capabilities of the infrastructure, which provides a continuous feedback to the Response Service regarding the effectiveness of the taken actions.

4.5 Cyber Range

Aiming at mitigating the lack of operator's experience while facing cyber-security related threats, interactive environments are to be designed and offered in order to train front-line employees/operators and improve their cyber-awareness and cyber-responsiveness through a continuous learning process. Thanks to CRs, the gap between cyber-theory and cyber-practice can be closed. The idea of a CR is tied with the possibility to host cyber-crafts without exposing the real infrastructure, therefore making use of the Simulator component. The simulator provides the controlled environment needed for cyber-security education, training and testing and, in addition, offers support for technical security verification of new devices to be added to the system. On top of the Simulator component, a Cyber-attack Generation component is provided. In order to be able to train operators as well as test new components (or software updates to already tested hardware), models need to be created for each attack vector and intrusion detection techniques that feed the Simulator component to mimic real world situations in a non-deterministic way so that the training and testing is not biased. Consequently, Training and Testing components are needed to provide the meaningful interfaces with the Risk Management plane.

4.6 HIT Infrastructure

As described in Sect. 3, two main application scenarios have been identified: "Smart hospitals" and "Home/Community care settings". Regarding the former, an IDS/IPS component, which detects and/or prevents intrusions in the system, can be deployed. The component depends on the system, network and context information extracted from Monitoring Probes that are dependent on the precise equipment and information to be analyzed. Thanks to the feedback produced by these two components, the security infrastructure can dynamically react and mitigate cyber-security threats. The components in charge of enforcing the decisions are the System Actuators. The Control and Management component is devoted to control and manage these elements, which may consist of different hardware and software elements. It is also in charge of making the decisions on how to perform the mitigations proposed by the Response Service layer as well as the orchestration of the involved elements.

Regarding "Home/Community care settings", the HIT infrastructure is set-up with a Home Gateway, like those present in any residential network, to control and manage medical IoT devices, such as, panic buttons, door sensors, GPS locators, medication robots, i.e., mIoT devices, and the possibility of deploying additional monitoring probes and actuators. The Home Gateway should be easy to deploy, user-friendly even for non-technical users while also certified in terms of cyber-security. The mIoT devices should pass certain security certifications [31] and permit to be easily updated for security purposes.

5 Conclusion

The digitalization process of the eHealth industry will bring advanced and improved services for citizens as well as cyber-security threats that should be carefully addressed. In this chapter, a wide discussion of the eHealth cyber-security landscape has been presented, giving insights on relevant aspects such as cryptography, intrusion detection, cyber-threat information sharing, and cyber range-based training. Two main application scenarios have been identified: Smart hospitals and home/community care settings, each of them with its own cyber-security risks and challenges. Thereafter, aiming at addressing these challenges, a holistic cyber-security platform devoted to protect eHealth infrastructures has been proposed and dissected. This architecture tackles the identified risks and presents innovative advances adopting state-of-the-art cyber-security techniques and procedures. Future research lines include the actual implementation of this holistic platform to make the complete securization of healthcare cyber-infrastructures to be a reality.

Acknowledgements This work has been supported by the European Commission, under the project CyberSec4Europe (Grant No. 830929) and by the Spanish Ministry of Science, Innovation and Universities, under the project ONOFRE 3 (Grant No. PID2020-112675RB-C44).

References

1. Chehri A, Fofana I, Yang X (2021) Security risk modeling in smart grid critical infrastructures in the era of big data and artificial intelligence. Sustainability 13(6):3196
2. Li W, Chai Y, Khan F, Jan SRU, Verma S, Menon VG, Kavita, Li X (2021) A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. Mobile Network Appl 26(1):234–252
3. Muhammad G, Alshehri F, Karray F, El Saddik A, Alsulaiman M, Falk TH (2021) A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Inf Fusion (in press)
4. Coventry L, Branley D (2018) Cybersecurity in healthcare: a narrative review of trends, threats and ways forward. Maturitas 113:48–52
5. Tully J, Selzer J, Phillips JP, O'Connor P, Dameff C (2020) Healthcare challenges in the era of cybersecurity. Health Secur 18(3):228–231
6. Zhang J, Zhao Y, Jie W, Chen B (2020) LVPDA: a lightweight and verifiable privacy-preserving data aggregation scheme for edge-enabled IoT. IEEE Internet Things J 7(5):4016–4027
7. Sanchez-Iborra R, Skarmeta AF (2020) TinyML-enabled Frugal smart objects: challenges and opportunities. IEEE Circ Syst Mag 20(3):4–18
8. Benhamouda F, Joye M, Libert B (2016) A new framework for privacy-preserving aggregation of time-series data. ACM Trans Inf Syst Secur 18(3):1–21
9. Derler D, Ramacher S, Slamanig D (2017) Homomorphic proxy re-authenticators and applications to verifiable multi-user data aggregation. In: Kiayias A (ed) Financial cryptography and data security, pp 124–142
10. Walfish M, Blumberg AJ (2015) Verifying computations without reexecuting them. Commun ACM 58(2):74–84
11. Yu X, Yan Z, Vasilakos AV (2017) A survey of verifiable computation. Mobile Networks Appl 22(3):438–453
12. Thakkar A, Lohiya R (2021) A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. Artif Intell Rev (in press)
13. Tan S, Guerrero JM, Xie P, Han R, Vasquez JC (2020) Brief survey on attack detection methods for cyber-physical systems. IEEE Syst J 14(4):5329–5339
14. Daniya T, Suresh Kumar K, Santhosh Kumar B, Kolli CS (2021) A survey on anomaly based intrusion detection system. Mater Today Proc (in press, Apr 2021)
15. Sun J, Wang X, Xiong N, Shao J (2018) Learning sparse representation with variational autoencoder for anomaly detection. IEEE Access 6:33353–33361
16. Lee J-H, Kim YS, Kim JH, Kim IK (2017) Toward the SIEM architecture for cloud-based security services. In: IEEE Conference on communications and network security (CNS), Oct 2017. IEEE, pp 398–399
17. Rashid F, Miri A (2021) User and event behavior analytics on differentially private data for anomaly detection. In: 7th IEEE 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), May 2021. IEEE, pp 81–86
18. Moukafih N, Orhanou G, El Hajji S (2020) Neural network-based voting system with high capacity and low computation for intrusion detection in SIEM/IDS systems. Secur Commun Networks 2020:1–15
19. Jin H, Luo Y, Li P, Mathew J (2019) A review of secure and privacy-preserving medical data sharing. IEEE Access 7:61656–61669
20. Zibak A, Simpson A (2019) Cyber threat information sharing: perceived benefits and barriers. In: Proceedings of the 14th international conference on availability, reliability and security, Aug 2019. ACM, New York, NY, USA, pp 1–9
21. Du L, Fan Y, Zhang L, Wang L, Sun T (2020) A summary of the development of cyber security threat intelligence sharing. Int J Digit Crime Forensics 12(4):54–67

22. Bass T (2000) Intrusion detection systems and multisensor data fusion. Commun ACM 43(4):99–105
23. Rajamäki J, Katos V (2020) Information sharing models for early warning systems of cybersecurity intelligence. Inf Secur Int J 46(2):198–214
24. Lorenzen C, Agrawal R, King J (2018) Determining viability of deep learning on cybersecurity log analytics. In: IEEE International conference on big data (big data), Dec 2018. IEEE, pp 4806–4811
25. Hallaq B, Nicholson A, Smith R, Maglaras L, Cook A, Janicke H, Jones K (2021) A novel hybrid cyber range for security exercises on cyber-physical systems. Int J Smart Secur Technol 8(1):16–34
26. Vykopal J, Vizvary M, Oslejsek R, Celeda P, Tovarnak D (2017) Lessons learned from complex hands-on defence exercises in a cyber range. In: IEEE Frontiers in education conference (FIE), Oct 2017. IEEE, pp 1–8
27. Martin G, Martin P, Hankin C, Darzi A, Kinross J (2017) Cybersecurity and healthcare: how safe are we? BMJ 358(j3179)
28. Karjalainen M, Kokkonen T (2020) Comprehensive cyber arena; The next generation cyber range. In: IEEE European symposium on security and privacy workshops (EuroS&PW), Sept 2020. IEEE, pp 11–16
29. Drăgulinescu AMC, Manea AF, Fratu O, Drăgulinescu A (2020) LoRa-Based medical IoT system architecture and testbed. Wirel Personal Commun
30. Shamayleh A, Awad M, Farhat J (2020) IoT Based predictive maintenance management of medical equipment. J Med Syst 44(4):72
31. Matheu SN, Hernández-Ramos JL, Skarmeta AF, Baldini G (2021) A survey of cybersecurity certification for the internet of things. ACM Comput Surv 53(6):1–36

Multi-class Classification for the Identification of COVID-19 in X-Ray Images Using Customized Efficient Neural Network

Adnan Hussain, Muhammad Imad, Asma Khan, and Burhan Ullah

Abstract During the global urgency, experts from all over the world searching for a new technology that supports the COVID19 pandemic. The deep learning and artificial intelligence application used the researchers on the previous epidemic, which encouraged a new angle to fight against the COVID19 outbreak. The limited number of COVID19 kits available in hospitals is due to the increasingly high number of cases. Therefore, it is necessary to implement an alternative system that detects and diagnoses the COVID19 and stops spreading among people. This chapter aims to detect and classify COVID19 infected, normal, and pneumonia patients from X-ray images using deep learning techniques (proposed CNN, AlexNet, and VGG16 models). The experiment was performed by combining two datasets, which are available on the Kaggle repository. The result analysis shows that the proposed CNN model achieved the highest accuracy of 95% from other deep learning models (AlexNet 90% of accuracy, and VGG16 94% of accuracy).

Keywords Classification · Deep learning · COVID19 · Pneumonia · X-ray images · CNN AlexNet · VGG16

1 Introduction

In December 2019, the COVID19 pandemic began in Wuhan, China, and spread worldwide. The cause of the COVID19 pandemic disease infection was the acute severe respiratory syndrome coronavirus (SARS-CoV-2) and Middle East Respiratory Syndrome (MERS-CoV). The COVID19 pandemic is spreading throughout the world at an unprecedented rate for any infectious illness. One of the effective approaches offered by the World Health Organization (WHO) to control the spread of viral disease is social distance and contact tracing [1, 2].

A. Hussain (✉) · A. Khan · B. Ullah
Islamia College University, Peshawar, Pakistan

M. Imad
Abasyn University, Peshawar, Pakistan

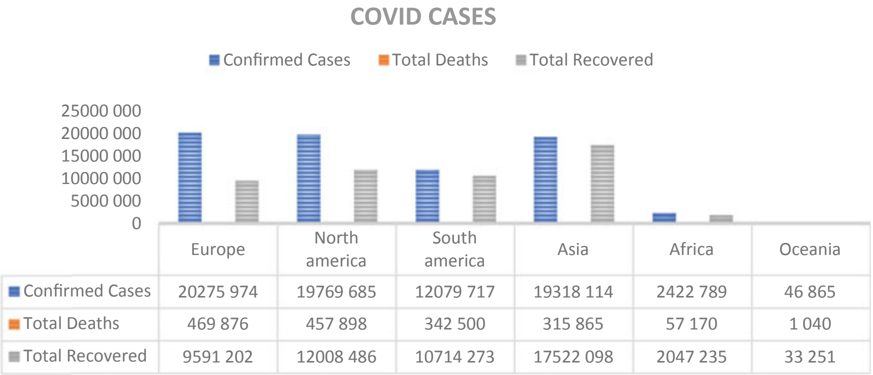
Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_23

According to the World Health Organization (WHO), 9,919,725 cases were reported, while 1,623,064 have died from the COVID19 diseases. In many developed countries, the health framework falls due to the synchronous flare-up of COVID19 and the expanding interest for serious consideration units loaded up with the infected patients [3]. The sign of infection, including respiratory symptoms of the COVID19, are fever, cough, and sore throat. The disease can also be caused by severe acute respiratory syndrome, pneumonia, multi-organ failure, septic shock, and death in some serious cases [4].

It has been determined that women are less affected than men, and children between the ages of 0 and 9 are not affected. In Respiratory rates, the COVID19 have been observed to be infected faster than healthy people. COVID19 is associated with a highly Intensive Care Unit (ICU) due to a rapid transition rate which requiring an urgent quest for fast and precise diagnosis treatments [5]. WHO reported the distribution of COVID19 cases worldwide as of 16 December 2020. Europe, North America, and Asia are the most highly infected countries. Figure 1 shows the region-wise confirmed, total-death, and recovered cases ("Coronavirus", 2021).

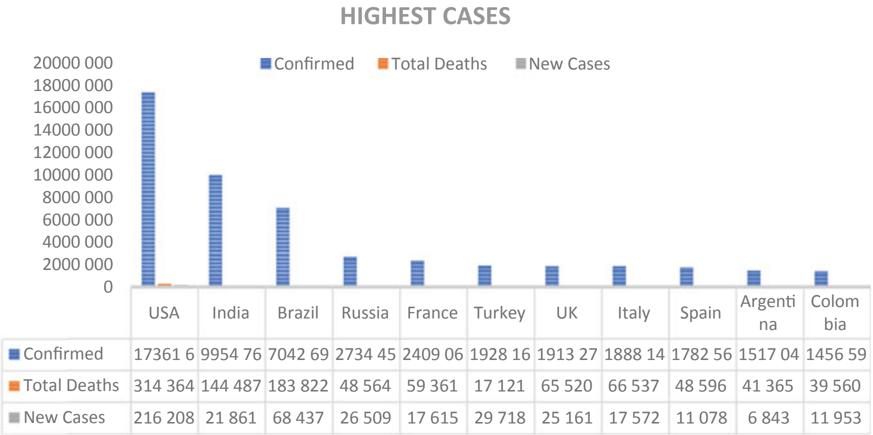
The United States, Spain, Italy, United Kingdom, India, France, Turkey, Russia, Brazil, and Colombia are the highly infected countries with many registered cases of COVID19 patients. According to the WHO, there are currently 17,361 confirmed cases in the United States, rising sharply, and the loss of life has ascended to 314.36. In India, Spain, Italy, Turkey, Argentina, and different nations, by observing serious lockdown and full consideration, mortality and new cases are declining, as presented in Fig. 2 ("Coronavirus", 2021).

Radiological imaging, such as CT scans and chest X-rays, can help people with scars to cope with the epidemic in a timely manner [6]. These methods may, without limitation, differ in the radiological properties of COVID19. The best option for a radiologist is a chest X-ray, as most emergency clinics are equipped with X-ray. As it may be, chest X-ray images obtained from X-ray machines cannot be clearly separate delicate tissues [7]. CT scan of the chest is used to eliminate this problem



	Europe	North america	South america	Asia	Africa	Oceania
Confirmed Cases	20275 974	19769 685	12079 717	19318 114	2422 789	46 865
Total Deaths	469 876	457 898	342 500	315 865	57 170	1 040
Total Recovered	9591 202	12008 486	10714 273	17522 098	2047 235	33 251

Fig. 1 Region-wise COVID19 cases



	USA	India	Brazil	Russia	France	Turkey	UK	Italy	Spain	Argentina	Colombia
Confirmed	17361 6	9954 76	7042 69	2734 45	2409 06	1928 16	1913 27	1888 14	1782 56	1517 04	1456 59
Total Deaths	314 364	144 487	183 822	48 564	59 361	17 121	65 520	66 537	48 596	41 365	39 560
New Cases	216 208	21 861	68 437	26 509	17 615	29 718	25 161	17 572	11 078	6 843	11 953

Fig. 2 Country-wise COVID19 highest cases

and effectively identify the delicate tissues. CT images are needed to examine the chest by Radiologists [8]. In this epidemic, COVID19 requires numerous radiologists to diagnose. Even so, owning one is still beyond the reach of the average person.

As the core technologies of the advancement of artificial intelligence (AI) in recent years, deep learning and machine learning have been reported with significantly improved diagnostic accuracy in medical imaging to detect and classify COVID19 [9]. It surpassed human-level- performance to classify the imageNet dataset, which consists of millions of images in 2015 [10]; also, detection and classification of a skin lesion with machine learning deep learning and obtained impressive results with the help of X-ray images [11].

As a result, COVID19 needs automatic identification. For that machine learning techniques such as, deep learning is widely used for automatic examination of chest radiological images. AI analysts and computer researchers played a key role when COVID19 spread around the world [12].

The rest of the chapter is organized as follows: Sect. 2 reviews and discusses the COVID19 detection and classification. Section 3 discussed the data collection and methodology of COVID19. Section 4 presents the implementation and results of the classification. Finally, Sect. 5 concludes the work.

2 Related Work

Sarker et al. [13] proposed a method based on deep learning. They used Desnet-121 and transfer learning to classify COVID19 patients. They used the exchange learning strategy to eliminate the problem of angles and train deep learning networks. The accuracy of this strategy was 92%.

Öztürk et al. [14] presented a method for covid classification. They use four distinctive feature extraction techniques to extract features from chest CT images. The SVM technique has been used further for classification. For classification purposes, these separated pictures are applied to support vector machines (SVM). They utilized 10-overlap cross approval during the characterization cycle, which is obtained 99% of precision.

Zhang et al. [15] Presented a method for COVID19 classification with greater efficiency based on deep learning models using chest X-ray images leading to faster and more reliable scanning.

Adhikari [16] proposed a network "Auto Indicative Medical Analysis" attempting to discover an irresistible region to help the specialist distinguish the ailing part better. CT and X-rays images were utilized during investigation. The DenseNet technique has been used to eliminate and check the infected area of the lungs.

Narin [17] Proposed a method based on different deep learning models such as InceptionResNet-V2, InceptionV3, and ResNet50 for classification on COVID19 using X-ray images.

Wang [18] presented a technique, COVID19Net, which is implemented on the COVIDx dataset. The dataset consists of 266 COVID19 infected patients with CXR images. The framework was first developed on ImageNet and then achieved the best score of 0.9480 in three-class order.

Shan et al. [3] proposed a method based on deep learning for automatic categorization of infected areas in the lung. The data is consisting of 300 COVID19 positive cases images. This model achieved of 91% accuracy.

Chen et al. [19] Design UNetþþ to differentiate COVID19 and pneumonia. The samples for this model consist of 106 COVID19 and pneumonia patients.

Li et al. [20] created the ConVNet Detector Neural Network (CONVNet) to recognize Covid infection in patients by extracting the features from chest CT images. The COVNet technique was trained on 3322 patients of CT images. The model has been achieved of 95% precision rate.

Xu et al. [5] segmented the infected regions from CT scan and utilized a three-dimensional profound learning model. The model is used to classify COVID19 from normal and Influenza-A viral pneumonia.

Sethy and Behera [21] uses chest X-rays presented a deep learning model with SVM. The SVM and Resnet 50 perform better than other pre-trained models.

Horry et al. [22] presented a technique for analysis of COVID19 using X-ray images. Two methods MobileNet, AOCTNet and shuffleNet with CNN has been used and highlight the extraction feature in the images. The KNN, SVM, RF, and SoftMax classifier has been used further for classification.

Khan [23] presented a pre-trained model Xception to identify the COVID19 infected patients. The dataset consists of COVID19, viral pneumonia, and bacterial cases.

Shi et al. [24] examined the patients who have been harmed from coronavirus pneumonia and have been admitted to a clinical center in Wuhan, China. The CT images of patients with their behavioral characteristics were also analyzed and considered for recognition of Covid 19 diseases.

Table 1 Dataset distribution in classes

Labels	Number of images
COVID19	1266
Normal	1266
Pneumonia	1266
Total	3798

Yadav and Jadhav [25] used CNN-based pre-trained models such as inspirationV3 and VGG16 with SVM using chest X-ray images to perform better. Thejeshwar [26] proposed a KE Sieve Neural Network structure that uses chest X-ray images to find estimates of Covid 19.

Ullah et al. [27] presents a machine learning technique to detect and classify COVID19 from X-ray images. The feature is extracted from a Histogram of oriented gradients and then classify it using a support vector machine and logistic regression. The result shows that the SVM provide better performance than logistic regression.

3 Material and Methods

This section presents a detailed description of the dataset and evaluation of the proposed method, such as pre-processing, augmentation, and experimental setup for classification.

3.1 Dataset

Chest X-ray images from the two datasets combined, which contain a total of 3798 images for training and testing. The dataset is obtained from the open-source Kaggle repositories. The dataset contains a mix of chest X-ray images (COVID19, normal, and pneumonia). In addition, 1266 images of COVID19 infected patients, 1266 for normal, and 1266 for pneumonia patients are presented in Table 1. 70% of the dataset has been used for training purposes and 30% for testing [30].

X-ray images are different in gray surface, features, and dimensions. Figure 3 presents a sample of COVID19 infected, normal, and pneumonia images.

3.2 Image Pre-processing

The pre-processing technique is used to resize the X-ray images from the input data with a fixed size of 224 × 224 × 3, which shows the height, width, and channel. The

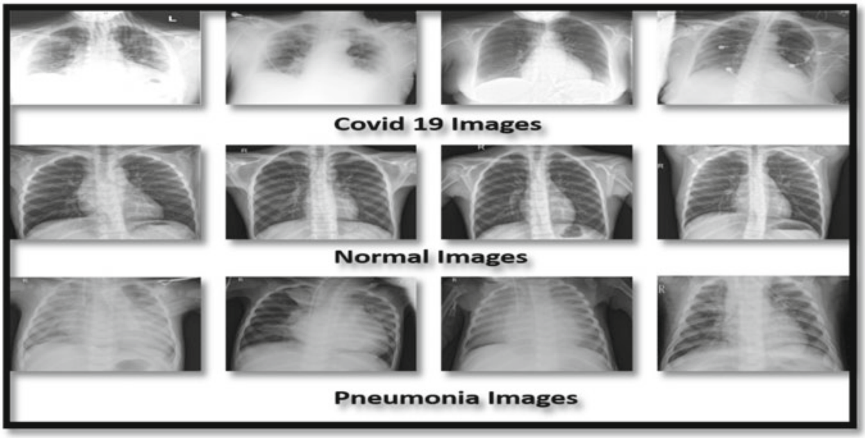


Fig. 3 Sample of COVID19, normal, and pneumonia using X-ray images

performance of the deep learning framework enhances by speeding and anticipating time from the pre-processing.

3.3 Image Augmentation

The data augmentation is applied on an inadequate date during the training process. All images consist of different styles like rotation, horizontal, vertical flip, and zoom-out. The aims are to provide an adequate amount of data into the CNN model to develop the performance and efficiency of the model. Figure 4 presenting the data augmentation with different rotations.

3.4 Classification

In classification, two different convolutional neural networks Alexnet and VGG16 (pre-trained architecture) are used and compared with the proposed model.

3.4.1 AlexNet Model

The AlexNet model is CNNs most illustrative model, which consists of three main points: superior, low training parameters, and solid robustness.

AlexNet is a deep virtual neural organization model comprising hidden layers, including one input and output layer, five polling, and three fully connected layers.

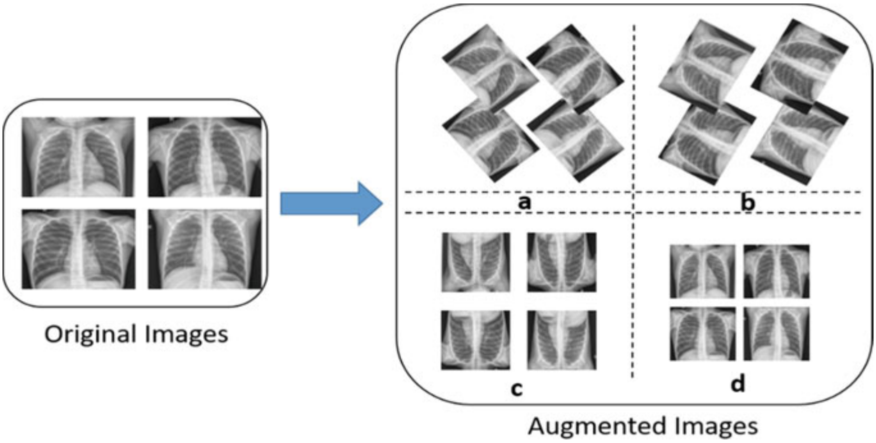


Fig. 4 Data augmentation

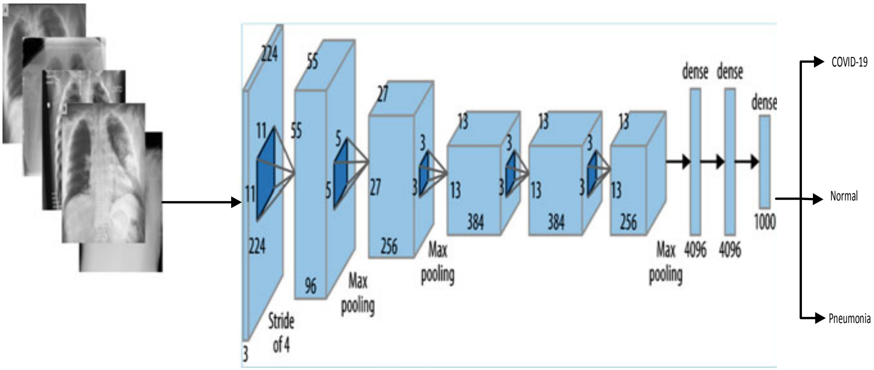


Fig. 5 AlexNet architecture

The learning feature is completed in the approved layer with two channels and passed to the third feature's extraction layer. The fully connected layer is cross-blending for the properties of the two groups, correspondingly. Figure 5 shows the design of the AlexNet architecture [28].

3.4.2 VGG-16 Model

VGG16 is a convolution neural net (CNN) architecture used to win ILSVR(ImageNet) competition in 2014. The VGG16 has a large number of hyperparameters that focused on having convolution layers of 3 × 3 filter and 16 connection layers. The max-pooling layer consists of five layers, and the size of the layers is 2 × 2. The arrangement of convolution and max pool layers consistently throughout

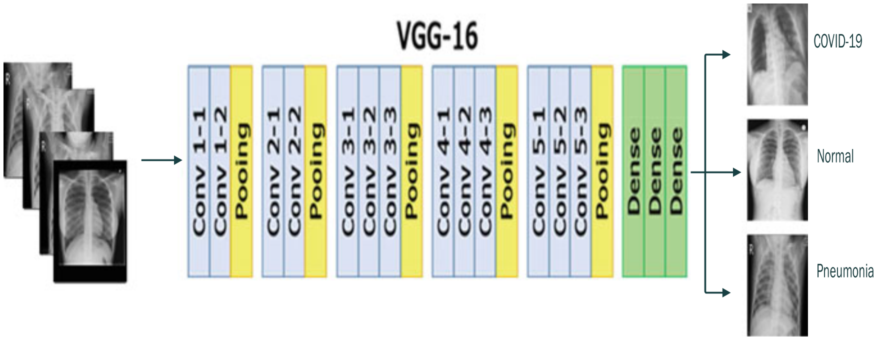


Fig. 6 VGG16 model

the whole architecture and considered one of the excellent vision model architectures. In the end, 2 FC (fully connected layers) followed by a SoftMax for output. A schematic graph of the VGG-16 architecture is presenting in Fig. 6 [29].

3.4.3 Proposed Architecture

The CNN network architecture consists of multiple conventional layers (CONV), subsampling layer (polling layer), and fully connected layers. The proposed CNN model takes image is an input with (128, 128, 3) size. The proposed network consists of four Blocks; each block contains convolutional and pooling layers. In the first block, the convolutional layer having 32 filters with (3, 3) size and the same padding, while the Maxpooling size is (2, 2). The second block contains one convolution layer having 64 filters of size (3, 3) and the same padding, while the Maxpooling size is the same. The third and Fourth blocks contain one convolution layer having 128 filters of size (3, 3) with the same padding and max-pooling size (2, 2). 66,128 features are extracted from the convolution, and max-pooling layers, which is converted into a 1D array called feature vector with 1,4608 sizes: the two fully connected layers (FC) and one out layer. Two layers have (1512) features and a ReLU activation function. In contrast, the output layer has (1,3) classed and SoftMax activation functions which are presented in Fig. 7.

4 Results and Discussion

The performance of each deep learning model has been evaluated using a confusion matrix which is illustrated in Eq. (1) to Eq. (3). Different performance metrics such as accuracy, sensitivity, specificity, precision, and F1-score. have been applied to measure the misclassification of COVID19 during diagnosing from X-ray images. The four term used to describe confusion matrix as follows; True Positive (TP), True

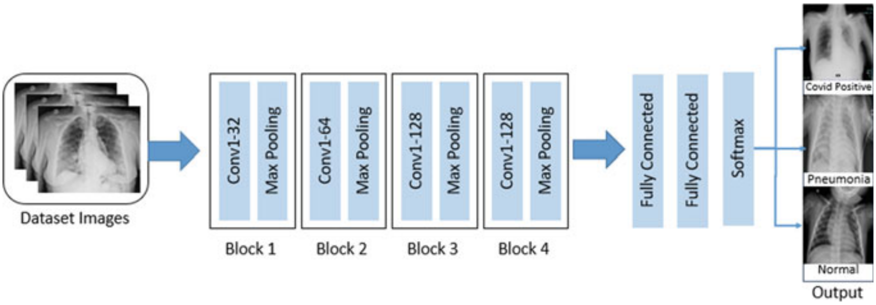


Fig. 7 Proposed CNN model

Negative (TN), False Positive (FP), False Negative (FN). True Positive (TP) refers that the images are correctly predicted and diagnosed. True Negative (TN) is number predicted negative class correctly. False Positive is the number wrongly classify COVID19 images. False Negative (FN) is number of non-detected occurrence of COVID19 ("Evaluating a machine learning model", 2021).

$$\text{Precision} = \frac{\text{Truepositive}}{(\text{Truepositive} + \text{Falsepositive})} \tag{1}$$

$$\text{Recall} = \frac{\text{Truepositive}}{(\text{Truepositive} + \text{Falsenegative})} \tag{2}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FN}) + (\text{TN} + \text{FP})} \tag{3}$$

In this study, chest X-ray images have been used to predict COVID19, normal, and pneumonia. The study examined the performance of three different deep learning models to identify Normal, COVID19, and pneumonia based on CNN, VGG 16, and AlexNet. The performance of different pre-trained models is illustrated in Tables 2, 3 and 4. The proposed CNN model has been achieved 95% of accuracy among other models such as VGG 16 (94%) of accuracy and AlexNet (90%) of accuracy.

The training loss, training accuracy, validation loss, and validation accuracy of the proposed CNN model are illustrated in Fig. 8. The training has been carried out up to 100 epochs to avoid the overfitting of the proposed CNN model. However, it is seen that the Proposed CNN Model shows a fast-training process with low training loss and validation loss.

Table 2 Performance Accuracy of Deep Learning Models

Model	Accuracy
AlexNet	90%
VGG-16	94%
Proposed Model	**95%**

Table 3 Performance of deep learning models using precision and recall

Precision			
Classification	AlexNet (%)	VGG-16 (%)	Proposed model (%)
COVID19	97	99	99
Normal	93	94	95
Pneumonia	82	90	92
Recall			
Classification	AlexNet (%)	VGG-16 (%)	Proposed model (%)
COVID19	96	95	98
Normal	83	93	93
Pneumonia	93	95	95

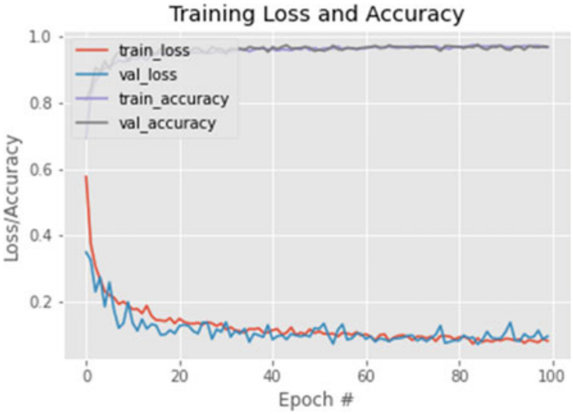


Fig. 8 The performance accuracy of the proposed CNN model

In another detailed performance, comparisons of three models using the test data are shown in Table 3.

The proposed model has achieved the highest precision 99%, 95%, and 92%, and recall 98%, 93%, 95% for COVID19, normal and pneumonia respectively than other which is illustrated in Table 3. As a result, the proposed CNN model provides superiority over the training and testing stage of the other two models.

Moreover, Fig. 9 and Table 2 depicts the graphical representation of three deep learning classifiers with accuracy. The proposed model achieved the best performance accuracy (95%), while the lowest accuracy is 90% achieved by the AlexNet model.

Further, Figs. 10, 11 and 12 represent the confusion matrix for each model. The confusion matrix illustrates the exact number of COVID19, normal, and pneumonia samples.

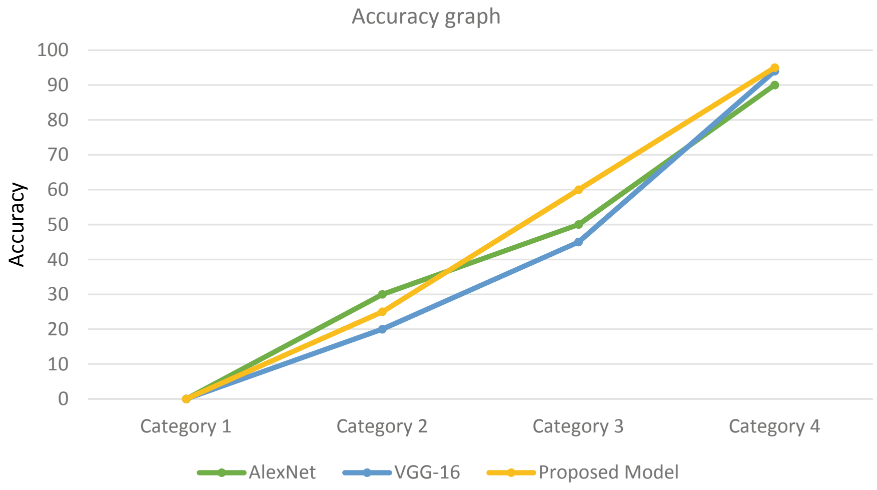
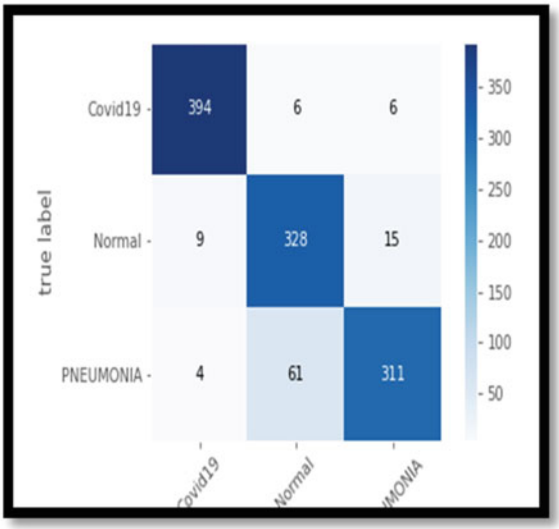


Fig. 9 Accuracy graph of all three models

Fig. 10 Confusion matrix of AlexNet model



5 Conclusion and Future Work

X-ray is the imaging technique that plays a vital role in diagnosing COVID19 and preventing disease among people from the spread. In this study, we used deep learning-based pre-trained models (AlexNet, VGG16) and compared them with fine-tuning of the CNN model. The X-ray dataset consists of COVID19, normal, and pneumonia which differentiate by pre-trained models. The experimental results shows that

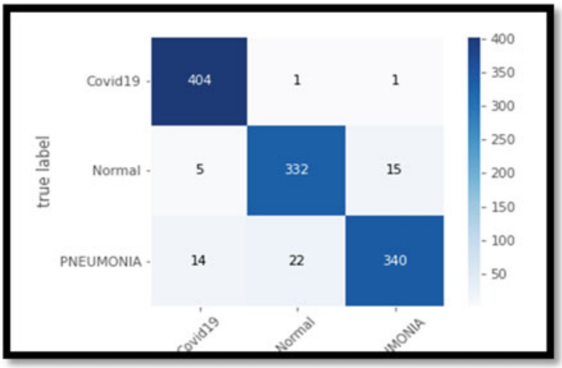


Fig. 11 Confusion matrix of VGG-16 model

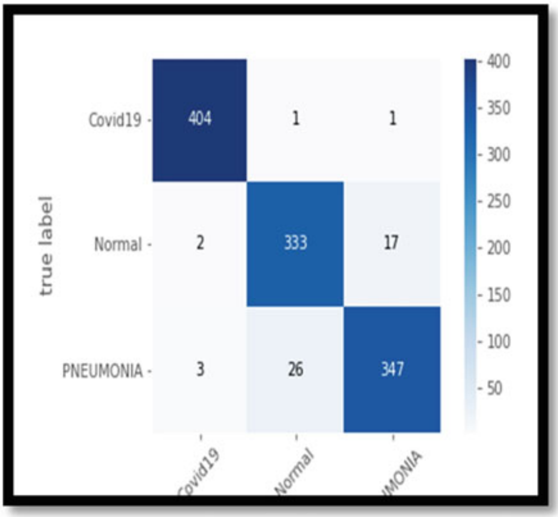


Fig. 12 Confusion matrix of proposed CNN model

our proposed CNN model achieved highest accuracy of 95% among the other two models (AlexNet 90% of accuracy, VGG16 94% of accuracy). In the future, we aim to extend the experimental work using a large dataset of CT and X-ray images. We also aim to use other pre-trained models to enhance performance accuracy and increase efficiency.

References

1. Bai Y et al (2020) Presumed asymptomatic carrier transmission of COVID-19. JAMA 323(14):1406–1407
2. Basavegowda HS, Dagnew G (2020) Deep learning approach for microarray cancer data

classification. CAAI Trans. Intell. Technol. 5(1):22–33

3. Shan F et al (2021) Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction. Med Phys 48(4):1633–1645
4. Singhal T (2020) Uma revisão da doença de Coronavírus-2019 (COVID-19). Indian J Pediatr 87:281–286
5. Xu X et al (2020) A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering 6(10):1122–1129
6. Singh D, Kumar V, Kaur M (2020) Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks. Eur J Clin Microbiol Infect Dis 39(7):1379–1389
7. Tingting Y, Junqian W, Lintai W, Yong X (2019) Three-stage network for age estimation. CAAI Transactions on Intelligence Technology 4(2):122–126
8. Kaur M, Gianey HK, Singh D, Sabharwal M (2019) Multi-objective differential evolution based random forest for e-health applications. Mod Phys Lett B 33(05):1950022
9. Khan N, Ullah F, Hassan MA, Hussain A (2020) COVID-19 classification based on Chest X-Ray images using machine learning techniques. Journal of Computer Science and Technology Studies 2(2):01–11
10. He K, Zhang X, Ren S, Sun, J (2015) IEEE Int. Conf. Computer Vision (ICCV)
11. Salam, A, Ullah, F, Imad M, Hassan MA (2020) Diagnosing of Dermoscopic Images using Machine Learning approaches for Melanoma Detection," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 2020: IEEE, pp 1–5
12. Razzak MI, Naz S, Zaib A (2018) Deep learning for medical image processing: Overview, challenges and the future, *Classification in BioApps,* 323–350
13. Sarker L, Islam MM, Hannan T, Ahmed Z (2020) COVID-DenseNet: a deep learning architecture to detect COVID-19 from chest radiology images. Preprints 2020, 2020050151
14. Öztürk Ş, Özkaya U, Barstuğan M (2021) Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features. Int J Imaging Syst Technol 31(1):5–15
15. J. Zhang *et al.* (2020) Viral pneumonia screening on chest X-ray images using confidence-aware anomaly detection, arXiv preprint arXiv:2003.12338
16. Adhikari NCD (2020) Infection severity detection of CoVID19 from X-Rays and CT scans using artificial intelligence,". International Journal of Computer (IJC) 38(1):73–92
17. Narin A, Kaya C, Pamuk Z (2021) Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Anal Appl 24(3):1207–1220. https://doi.org/10.1007/s10044-021-00984-y
18. Wang L, Lin ZQ, Wong A (2020) Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci Rep 10(1):1–12
19. Chen J et al (2020) Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. Sci Rep 10(1):1–11
20. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K, Liu D, Wang G, Xu Q, Fang X, Zhang S, Xia J, Xia J (2020) Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology 296(2):E65–E71. https://doi.org/10.1148/radiol.2020200905
21. Sethy P, Behera S (2020) Detection of coronavirus disease (COVID-19) based on deep features. Preprints.org; 2020. https://doi.org/10.20944/preprints202003.0300.v1
22. Horry MJ et al (2020) COVID-19 detection through transfer learning using multimodal imaging data. IEEE Access 8:149808–149824
23. Khan AI, Shah JL, Bhat MM (2020) CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, *Computer Methods and Programs in Biomedicine,* 196, 105581
24. Shi H et al (2020) Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis 20(4):425–434
25. Yadav SS, Jadhav SM (2019) Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data 6(1):1–18

26. Thejeshwar C, Chokkareddy, Eswaran K (2020) Precise prediction of COVID-19 in chest X-Ray images using KE sieve algorithm, *medRxiv*
27. Ullah SI, Salam A, Ullah W, Imad M (2021) "COVID-19 Lung Image Classification Based on Logistic Regression and Support Vector Machine," in *European, Asian, Middle Eastern, North African Conference on Management & Information Systems*, Springer, pp. 13–23.
28. Umer M, Ashraf I, Ullah S, Mehmood A, Choi GS (2021) COVINet: a convolutional neural network approach for predicting COVID-19 from chest X-ray images, *Journal of Ambient Intelligence and Humanized Computing,* pp. 1–13.
29. Taresh MM, Zhu N, Ali TAA, Hameed AS, Mutar ML (2020) Transfer learning to detect COVID-19 automatically from X-ray images using convolutional neural networks. Int J Biomed Imaging, 2021
30. Find Open Datasets and Machine Learning Projects | Kaggle. (2021). Retrieved 19 August 2021, from: https://www.kaggle.com/datasets?datasetsOnly=true

Design of an Efficient Rectenna for RF Energy Harvesting for IoT Medical Implants

Arslan Dawood Butt, Faisal Abrar, Muhammad Awais Qasim, Sarosh Ahmad, Muhammad Sajawal, and Muhammad Anas Attiq

Abstract The invention of implantable medical devices is important as it has a direct effect on the lives and safety of humanity. Energy harvesting is a technique to provide enough amount of energy to low power implanted devices such as Pacemaker, Cardio defibrillators, drug pump, Neurostimulator, etc. An approach to developing a prototype rectenna to harvest radio-frequency (RF) energy at Industrial, Scientific and Medical (ISM) band at 900 and 2450 MHz. Different antennas have been designed and compared efficiencies at different bands using Computer Simulation Technology (CST) software. A full-wave bridge rectifier circuit is designed in Advanced Design System (ADS) to rectify the harvested RF energy coming from the antenna to implanted device in the human body. Power Schottky diodes are used in the rectifier circuit as it has high switching frequency and low threshold voltage i.e., HSMS-2860. The rectenna is much efficient and produces a sufficient amount of power which is enough for low-power implantable medical devices (IMDs).

Keywords Energy harvesting · RF power harvesting · RF rectenna · Rectenna for IoT implants · Rectenna for medical implants

1 Introduction

Energy harvesting is a way of scavenging energy from the ambient energy, present in the environment, into usable electrical energy to power an application. Many procedures are adopted to harvest waste energy in the form of solar energy, infrared energy, kinetic energy, thermal energy, and at the end radio-frequency (RF) energy. A tree diagram of some methods for Energy Harvesting is shown in Fig. 1. There are a lot of ways to harvest energy and then convert it into usable electrical power for powering devices. If we talk about RF energy harvesting then the first thing that comes into mind is how to harvest RF energy and then the most suitable way from, we are going to do is through rectenna. Rectenna is the integration of antenna and

A. Dawood Butt · F. Abrar · M. A. Qasim (✉) · S. Ahmad · M. Sajawal · M. A. Attiq
Department of Electrical Engineering and Technology, Government College University Faisalabad (GCUF), Faisalabad, Pakistan

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*, Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_24

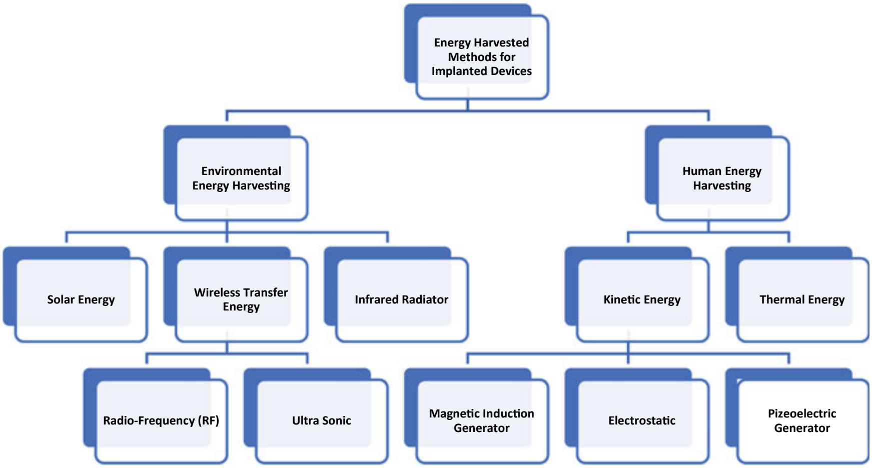


Fig. 1 Tree diagram of energy harvesting methods

rectifier circuits. The first antenna receives the RF waves and then produces AC power and with the help of a rectification circuit, the AC power is converted into DC power.

RF energy harvesting is one of the most suitable and human-friendly ways of scavenging energy from the atmosphere. In this process, RF waves in the environment are harvested by the antenna and then converted into DC with the help of a rectification circuit. The most commonly used antenna for this purpose is a patch antenna, we can get the maximum results by using this type of antenna.

2 Literature Review

In [1] author explained the RF energy harvesting basics and also its background in the submitted paper as energy harvesting technology. It is one of the best ways of powering low-power electronics. We can use ambient or residual energy to power up the devices. A design of RF energy harvesting is given, also the design of the antenna and rectifier circuit. Which can produce sufficient energy to power up low-power electronics. In [2] author designed a dipole antenna with a rectifier circuit having good efficiency and a matching circuit is also introduced. This prototype is tested in a model with an artificial provided source, the artificial source which is used is a type of horn antenna. However, the power which is harvested can be useful for implantable devices. In [3] author designed a wearable antenna for RF harvesting which can be used for biomedical implants on the human body. Rectennas consist of a patch antenna and rectifier circuit. A full-wave rectifier is used with RFC to cause an increment in the output voltage. In [4] author gave us an overview of the chip rectifier circuit and the usage of the circuit for different purposes. This rectifier is designed for energy harvesting whose power can be used for the internet of things, RFID, and wearable gadgets. Different types of RF to DC conversion circuits are used and compared to get the maximum efficiency. In [5] author shown us the 2.45 GHz monopole antenna for RF energy applications. The design rectenna in which monopole antenna is connected with a rectifier circuit and gave us the required efficiency. The monopole antenna is fabricated with the rectifier circuit to give us more efficiency which enables the rectenna for wearable applications. In [6] author wrote an article explaining the brief history that how energy is harvested for low power electronics, different energy harvesting techniques, how power is converted, and how it is managed. The explanation of different ways of energy harvesting like piezoelectric material, heat, and RF is done. How energy harvesting and its ways are developing day by day. In [7] author designed a patch antenna and full-wave rectifier circuit for health care application. It is a wearable rectenna that can power up low-power electronics on the human body. The rectenna is designed at a 2.45 GHz frequency.

3 Methodology

The methodology includes the proposed research procedure that will be carried out to achieve the objectives of the proposed research. This will explain every step taken to meet the goals by using the proposed model, and the choice of components that will be used in this research for making a rectenna prototype that will be used to harvest energy for low-powered IoT medical implanted devices. The design simulation and optimization will be carried out by using CST Studio Suite 2019 for the designing of antennas and Keysight ADS Studio 2020 for designing of rectifier circuit and conclusion. The target is to design three different antennas for this prototype which could operate at different frequencies like 405 MHz (MedRadio band) and 900 MHz and 2.45 GHz (ISM/GSM band). For the other portion of the rectenna, a bridge rectifier circuit will be used to convert the AC into DC which will be filtered out by using a capacitor at the output. This output will be used by the IMDs to operate inside the human body. A pacemaker will be used as a reference implantable medical device in this prototype.

3.1 Research Procedure

The procedure adopted for this proposed methodology is listed below:

(a) First of all, a detailed background of energy harvesting, existing work, and literature review of previous research is required for stepped towards the objectives. This helps in understanding the basics to advance knowledge about energy harvesting.

(b) Based on this, a prototype model is to be made. The model differentiates the proposed research into two major portions; antenna and rectifier.

(c) The selection of frequency band for the antenna is also important. Different bands have a different prospectus. Thus, we have designed different antennas on two bands (ISM and GSM bands). CST Microwave Studio 2019 is used for designing these required antennas.

(d) Like an antenna, a rectifier has a vital role in a rectenna. It rectifies the power coming from the antenna. The choice of diodes for this purpose is very significant. We have to look at every parameter in the selection.

(e) The model, then, is to be implemented in Keysight ADS 2020. This software simulated it and we can get the desired results and conclusion of our proposed prototype.

3.2 Proposed Model

Antennas are to be designed on an ISM band at 900 and 2450 MHz frequencies. A rectifier circuit will be used to rectify the power coming from the antenna. The model of the proposed research is shown in Fig. 2.

3.3 Choice of Diode

The diode is the main source of losses and its characteristics calculate the predominant execution of the circuit. It must have:

- High-speed switching characteristics
- A low cut-off voltage
- A junction voltage and a breakdown voltage adapted to the input voltage

Thus, a zero-biased Schottky diode is required which has high switching capacity, low voltage drops, low junction capacitance, and low voltage threshold. The most used commercial Schottky diodes are HSMS 285x and HSMS 286x, which operate able at ISM and GSM bands.

3.4 Schottky Diode

Schottky diode is a kind of semiconductor that has very low forward bias voltage loss and has very fast switched. It has a forward voltage of 150–450 mV which required fast switching high efficiency. The symbol of the Schottky diode is shown in Fig. 3.

When we apply a sufficient amount of forwarding voltage in the forward direction current starts flowing.

Fig. 2 Model of proposed research

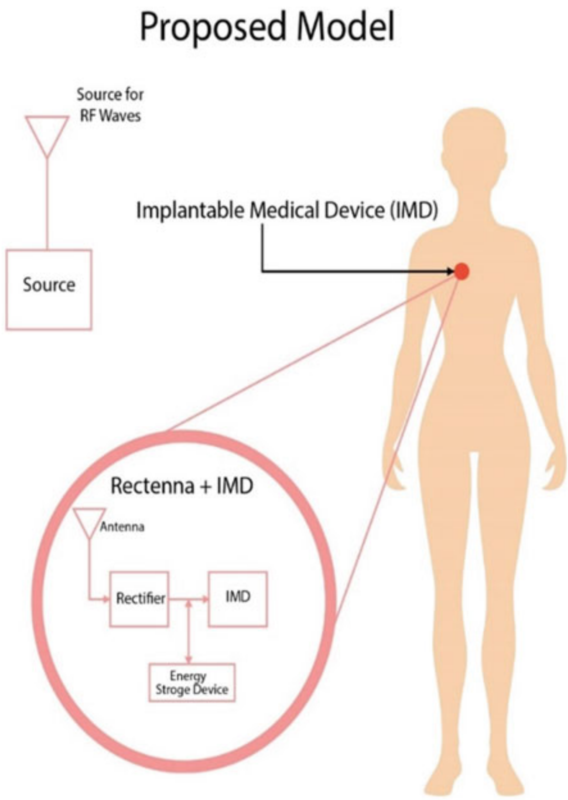
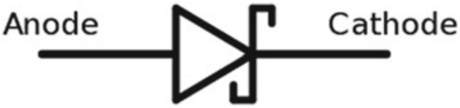


Fig. 3 Symbol of Schottky diode



I. **Advantages**:

Schottky diodes can be used in various applications. It has followed many advantages such as follows:

- Low turn-on voltage
- Fast recovery time
- Low junction capacitance

II. **Features:**

Schottky diodes have the following features:

- High-efficiency rate
- Have low forward voltage loss
- Very small capacitance

- Very small surface-mount package
- For any kind of stress protection guard ring is present

3.5 Bridge Rectifier

The full-bridge rectifier is a circuit that is used to convert AC to DC. Its symbol is shown in Fig. 4. It is one of the most used designs of circuits for this purpose. Because of its high efficiency and capability, the rectifier is used for energy harvesting through rectenna and in many other applications where an individual has to do rectification.

A complete full-wave bridge rectifier circuit along with labels is shown in Fig. 5.

Fig. 4 Bridge rectifier symbol

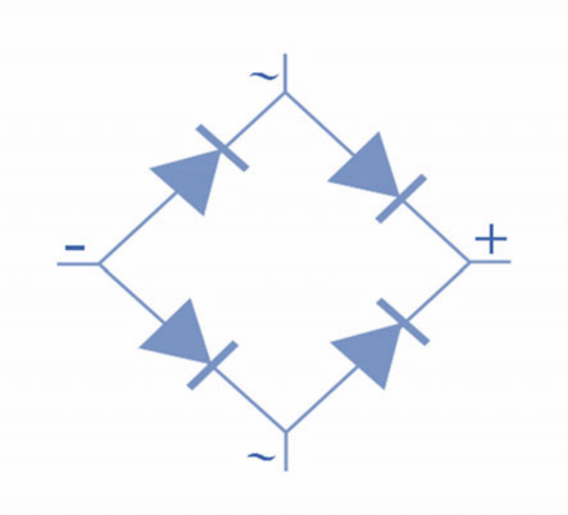
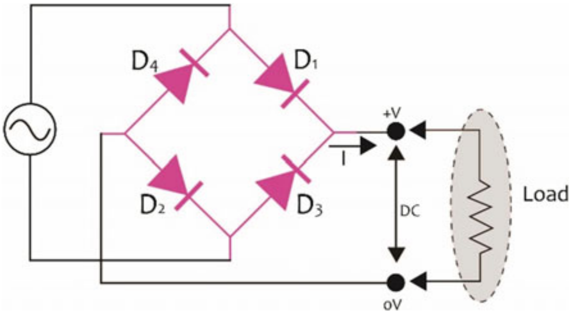


Fig. 5 Bridge rectifier circuit with labels



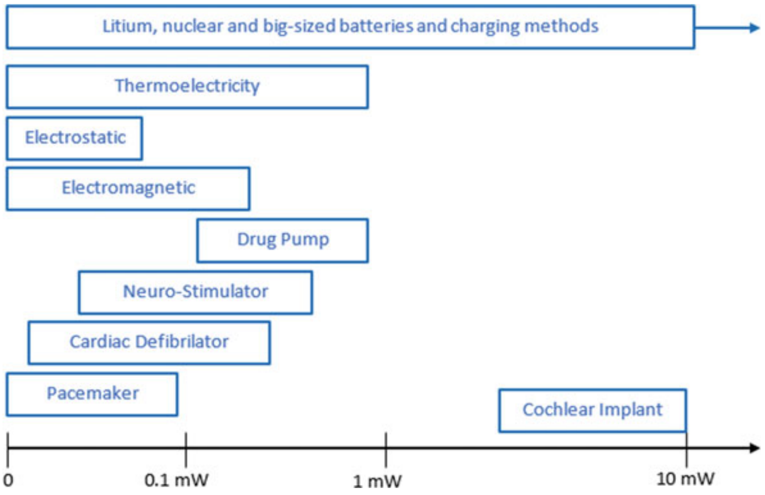


Fig. 6 Power requirement for IoT medical implants

3.6 Power Requirement for IoT IMDs

The requirement of power for IoT-based medical implantable devices, shown in Fig. 6, depends on its usage like how many watts of power is required to operate an IMD. It depends on its usage and at which position, an IMD is to be placed in the body. For example, a pacemaker required 0–100 uW and a drug pump required 100 uW–1 mW power, etc.

3.7 Software

There are mainly two software used in designing of entire prototype, which are:

- CST Studio Suite 2019
- Keysight ADS 2020

4 Design and Simulations

This section will cover the design and simulation of a prototype rectenna for energy harvesting. It includes detailed information about the software used in this research along with antenna designing at different frequencies and rectifier designing, etc. The design simulation and optimization will be carried out by using CST Studio Suite 2019 for the designing of antennas and Keysight ADS Studio 2020 for designing of rectifier circuit and conclusion. The target is to design different antennas for this prototype that could operate at different frequencies like 900 and 2450 MHz (ISM band).

4.1 Antenna Designing

As discussed above, antennas are designed in CST Studio Suite (version 2019). For this proposed researched project, it is decided to design different antennas of different frequencies. Antenna specifications and parameters with their results are going to be explained in the coming topics.

4.2 Receiver Antenna A (@2450 MHz).

This is the first microstrip patch antenna, shown in Fig. 7, designed at 2450 MHz using *ROGER 3010* (lossy) as a substrate. The material of the antenna is copper. The

Fig. 7 Microstrip patch antenna operating at 2450 MHz frequency

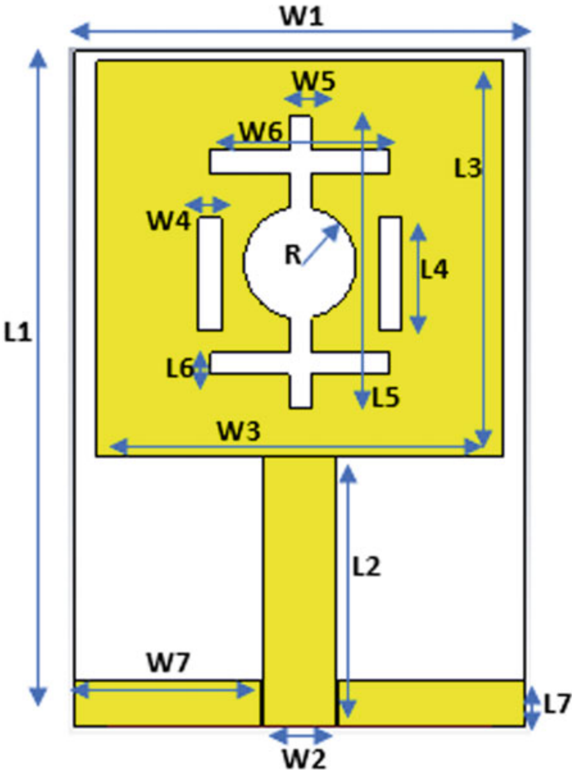


Table 1 Dimensions of antenna A @2450 MHz

Dimension	Values (mm)	Dimension	Values (mm)
L1	30	**W1**	20
L2	12	**W2**	3.2
L3	17.6	**W3**	18
L4	5.0	**W4**	1.0
L5	13	**W5**	1.0
L6	1.0	**W6**	8.0
L7	2.0	**W7**	8.3
R	2.5		

Table 2 Specifications of Antenna A

Receiver Antenna A	
Antenna	Microstrip patch
R. Frequency	2.45 GHZ
Directivity	2.21 dBi
Gain	1.44 dBi
Realized Gain	0.498 dBi
S11	-27.8 dB
VSWR	1.08

benefit of *ROGER 3010* (lossy) is that this material exhibits dimensional stability with an expansion coefficient matched to copper. It has a dielectric constant of 11.2.

Tables 1 and 2 represent the dimensions of the respective antenna operating at 2450 MHz frequency, respectively.

This antenna has a return loss of -27.98 dB, as shown in Fig. 8.

Voltage Standing Wave Ratio (VSWR) is the measure of the power transmitting efficiently from source to load via a transmission line/medium. This is considered good when it is less than 1.2. In this case, this is 1.08, as shown in Fig. 9.

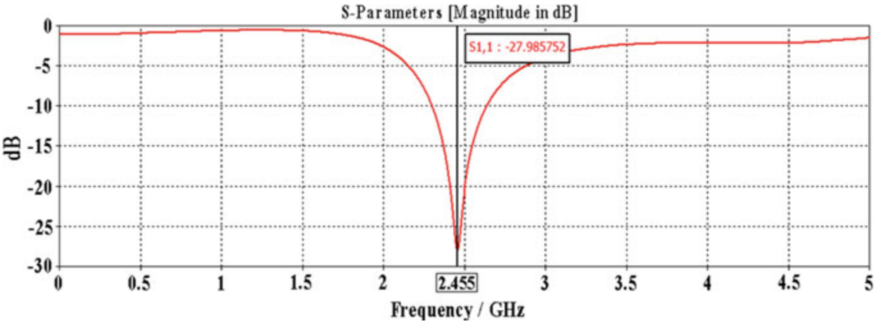


Fig. 8 Return loss of Antenna A

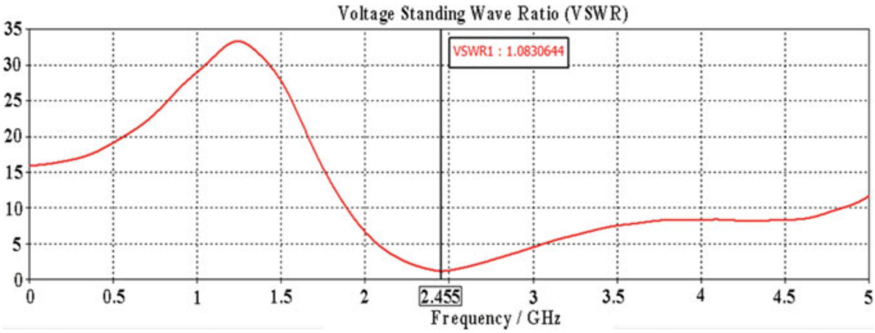


Fig. 9 VSWR of Antenna A

4.3 *Receiver Antenna B (@900 MHz).*

The second monopole antenna operating at ISM band i.e., 900 MHz frequency, is shown in Fig. 10. The radiating patch is printed on a flexible *ROGER RT 5880* (lossy) substrate. The relative permittivity of *ROGER* is 2.2 and loss tangent 0.0009.

Tables 3 and 4 represent the dimensions of the respective antenna operating at 900 MHz frequency, respectively.

This antenna has a return loss of -17.88 dB, as shown in Fig. 11.

This antenna has a voltage standing wave ratio of 1.29, as shown in Fig. 12.

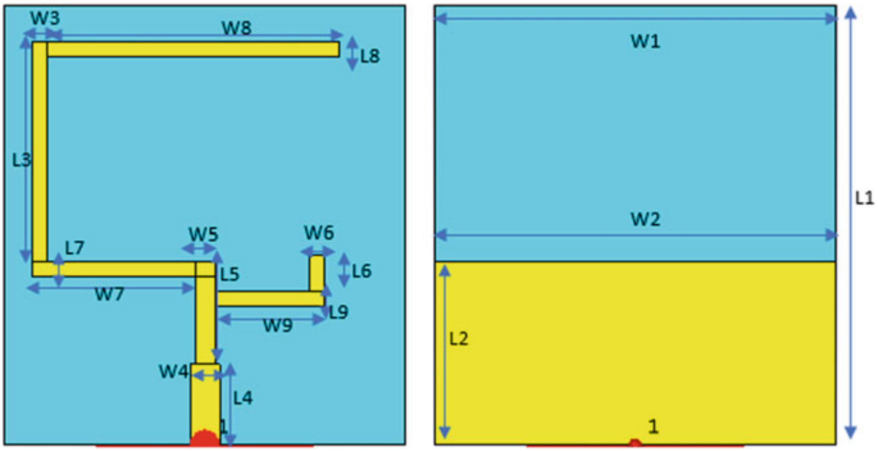


Fig. 10 Monopole antenna operating at 900 MHz frequency

Table 3 Dimensions of antenna B @900 MHz

Dimensions	Values (mm)	Dimensions	Values (mm)
L1	68.81	**W1**	63.07
L2	28.67	**W2**	63.07
L3	34.4	**W3**	2.29
L4	12.61	**W4**	4.59
L5	16.05	**W5**	3.06
L6	5.73	**W6**	2.29
L7	2.29	**W7**	25.61
L8	2.29	**W8**	45.87
L9	2.29	**W9**	17.2

Table 4 Specifications of Antenna B

Receiver antenna B	
Antenna	Monopole patch
R. Frequency	900 MHz
Directivity	1.55 dBi
Gain	1.44 dBi
Realized gain	1.36 dBi
S11	-17.88 dB
VSWR	1.29

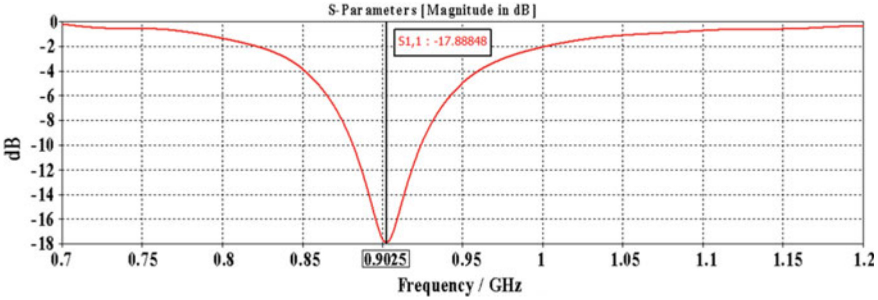


Fig. 11 Return loss of Antenna B

4.4 Rectifier Designing

A rectifier is a very essential part of during design of a rectenna. Different terms should be considered during designing like voltage drop, heat dissipation, peak inverse voltage, etc. We preferred the *HSMS-2860* power Schottky diode for our prototype. It is operatable on GSM and ISM band frequencies. We took four *HSMS2860* Schottky diodes and made a bridge of them.

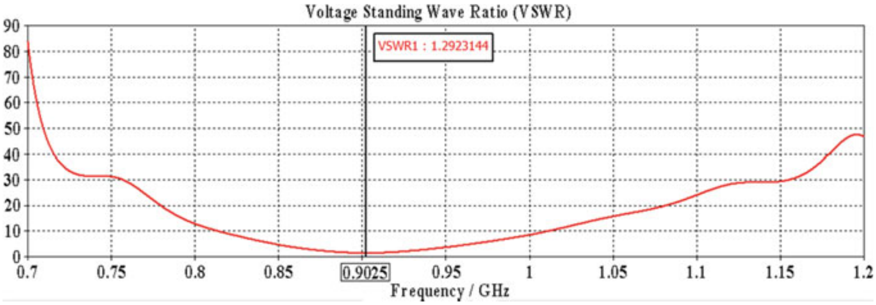


Fig. 12 VSWR of Antenna B

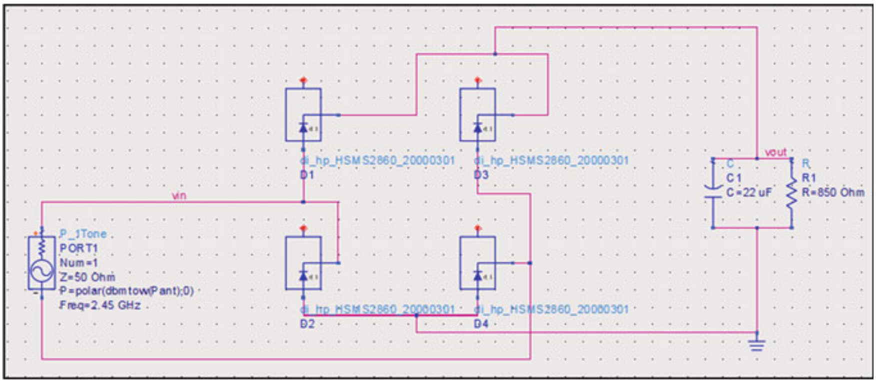


Fig. 13 Proposed rectifier circuit along with power source in the place of designed Antenna

Path Wave ADS 2020 is used to design the rectifier circuit for the proposed rectenna. In Fig. 13, there is a bridge of Schottky diodes following a capacitor before load. The capacitor filtered the DC power coming from the rectifier and then it will forward the filtered DC power to load. The load is of about 850-Ω resistance. The IMD for which we are designing this prototype is a pacemaker as a reference. The pacemaker has internal resistance in the range of 800–900 Ω. So, we choose a number between this range like 850 Ω.

An RF power source is attached at the input of the bridge rectifier. This will provide power coming from the antenna to the rectifier. We designed antennas in CST Studios Suite and Rectifier in Keysight ADS, therefore, the RF power source, used in Keysight ADS, will provide the same power to the rectifier which is received by an antenna. It is to be noted that the power, which is going towards the rectifier, is the power after deducting all the losses which occurred in the antenna.

5 Results and Discussion

In this section, the overall results of antennas and rectifiers have been discussed along with necessary safety measurements and considerations. It includes received power by antennas of different frequencies and then the calculation of output power at load after rectification.

5.1 Received Power by Antennas

This portion includes the reception of power by antennas attached with rectifier of rectenna. In simple words, it is the power coming from a source received by the antenna and forward this to rectifier after removing metallic losses occurs in the antenna.

5.1.1 Friis Transmission Equation (FTE)

The Friis transmission equation is used in telecommunications engineering which gives the power received by one antenna given another antenna at some distance away transmitting a known amount of power. A model for the application of FTE is shown in Fig. 14.

For the operating two antennas, it is the ratio of available power at i/p of receiving antenna "Pr" to/p power of transmitter antenna "Pt" which is mentioned in Eq. (1).

$$P_r = \frac{p_t G_t G_r \lambda_0^2}{(4\pi)^2 r^2} \tag{1}$$

where "Pr" is the power received by the receiver antenna, **"**Pt" is the power transmitted by transmitter antenna, **"**Gt" is the gain of transmitter antenna, **"**Gr" is the gain of receiver antenna, **"**λ" is the wavelength and "r" is the distance between transmitter and receiver antenna. Power harvested by antenna A and B is shown in Tables 5 and 6, respectively.

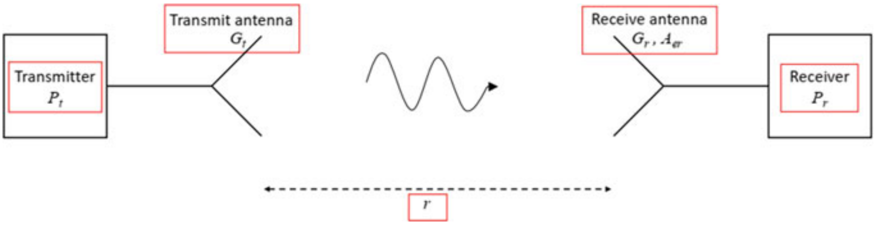


Fig. 14 General model for application of Friis transmission equation

Table 5 Power harvested by Antenna A

r	Pr (@2450 MHz)	
cm	mW	dBm
100	0.703	-1.5304
50	2.815	4.49478
20	17.601	12.45537

Table 6 Power harvested by Antenna B

r	Pr (@900 MHz)	
cm	mW	dBm
100	14.325	11.5609
50	57.3	17.5815
20	358.12	25.5402

After removing metallic losses, antenna A yields 12.45 dBm power at 20 cm and -1.53 dBm at 100 cm distance while antenna B yields 25.54 dBm power at 20 cm and 11.56 dBm at 100 cm distance.

5.1.2 Output power

The Rectenna A is the rectenna with a receiver antenna is operating at 2450 MHz. Rectenna B is the rectenna with receiver antenna is operating at 900 MHz These are simulated and all the results are calculated and mentioned in Tables 7 and 8, respectively.

Table 7 Overall output power from Rectenna A

Output power Rectenna A (@2450 MHz)			
R (cm)	Pin (dBm)	Vout (V)	Pout (W)
100	-1.5304	95.32 m	10.689u
50	4.49478	504.8 m	0.299 m
20	12.45537	1.923	4.35 m

Table 8 Overall output power from Rectenna B

Output Power Rectenna B (@900 MHz)			
R (cm)	Pin (dBm)	Vout (V)	Pout (W)
100	11.5609	1.687	0.003
50	17.5815	3.88	0.018
20	25.5402	6.732	0.053

Table 9 Efficiency of rectifier

I/p Power (mW)	I/p Power (dbm)	O/p Voltages (V)	Efficiency (%)
1	0	0.163	3.12
3.16	5	0.559	11.6
7.94	9	1.137	19.15
15.84	12	1.800	24.06
31.62	15	2.748	28.09
63.09	18	4.098	31.29
125.89	21	6.014	33.80
199.52	23	6.723	26.65

5.1.3 Calculation of Rectifier Efficiency

The efficiency of the rectifier converting RF to DC can be calculated by the ratio of DC o/p power to RF i/p power.

$$\eta = \frac{P_0}{P_i} \tag{2}$$

Where,

$$P_o = \frac{V_{out(DC)}^2}{R_L} \tag{3}$$

The efficiency of the proposed rectifier is calculated at various inputs and is shown in Table 9.

The rectifier is highly efficient with max. the efficiency of approximately 34% for the 21 dBm i/p power. This research-based rectenna is capable to yield enough power which can be utilized in operating low-power IoT medical implanted devices. The input power vs output voltages at load and efficiency graphs are shown in Figs. 15 and 16, respectively.

Figure 16 represents the efficiency of the full-wave bridge rectification circuit. It shows that the rectifier is approximately 34% efficient and can yield enough power for low-powered IoT implantable medical devices by harvesting RF energy over the given frequency band.

6 Conclusion

An efficient rectenna for RF energy harvesting applications in IoT medical implants is designed in the ADS-2020 simulator. It is based on an antenna and rectifier. Two antennas are designed in CST Studio Suite 2019 at 900 and 2450 MHz. *HSMS-2860*

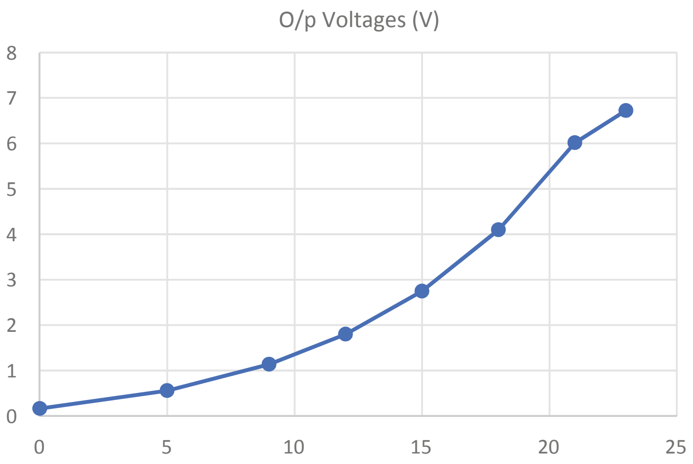


Fig. 15 Input power versus output voltages

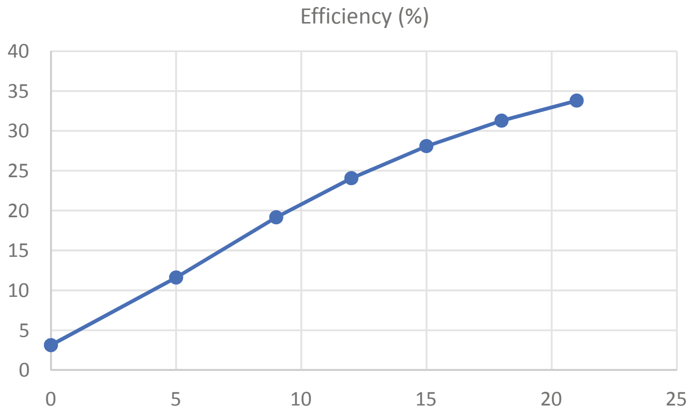


Fig. 16 Input power versus efficiency

power Schottky diode is used for rectifier circuits that have a low voltage drop and fast switching capacity. Rectenna operated in ADS-2020 and works efficiently yielding maximum output voltages of 6.723 V. From Rectenna A, operating at 2450 MHz frequency, we concluded that it can yield the output 24% efficiently, while from Rectenna B which is operating at 900 MHz frequency, can yield 33.8% efficient output. Thus, both the proposed rectenna produce enough power required for low-powered IoT-based implantable medical devices (IMDs).

References

1. Asif SM, Braaten BD (2016) Design of a compact implantable rectenna for wireless pacing applications. IEEE International Symposium on Antennas and Propagation (APSURSI) 2016:167–168
2. Parna Kundu (datta), Juin Acharjee and Kaushik Mandal," Design of an Efficient Rectifier Circuit for RF Energy Harvesting System", International Journal of Advanced Engineering and Management, Technical and Scientific Publisher, 2017, 2(4), pp. 94–97.
3. A. Eid, J. Hester, A. Nauroze, T.H. Lin, J. Costantine, Y. Tawk, A.H. Ramadan and M. Tentzeris, A Flexible Compact Rectenna for 2.4GHz ISM Energy Harvesting Applications", 2018 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, 2018, pp.1887–1888.
4. Wang L (2016) Menglong He, and Zhao Wang", Radio Frequency Energy Harvesting Technology". International SoC Design Conference (ISOCC) 2016:219–220
5. H. Dai, Y. Lu, M. Law, Sai-Weng Sin, U. Seng-Pan and R.P. Martins," A review and design of the on-chip rectifiers for RF energy harvesting", 2015 IEEE International Wireless Symposium (IWS 2015), 2015, pp.1–4.
6. Harb A (2011) Energy Harvesting: State-of-the-art. Renewable Energy 36:2641–2654
7. S. Ding, S. Koulouridis and L. Pichon," Design and characterization of a dual-band miniaturized circular antenna for deep in body biomedical wireless applications", International Journal of Microwave and Wireless Technologies, Volume 12, Special Issue 6: EuCAP 2019 Special Issue, July 2020, pp. 461–468.
8. C. Lin, C.Chiu and J. Gong," A Wearable Rectenna to Harvest Low-Power RF Energy for Wireless Healthcare Applications", 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018, pp.1–5.
9. Sofia Bakogianni, and Stavros Koulouridis, "A Dual-Band Implantable Rectenna for Wireless Data and Power Support at Sub-GHz Region", IEEE Transactions on Antennas and Propagation
10. Kush Agarwal, Rangarajan Jegadeesan, Yong Xin Guo, and Nitish V. Thokar, "Wireless Power Transfer Strategies for Implantable Bioelectronics: Methodological Review", IEEE Reviews in Biomedical Engineering, Vol.10, 2017
11. Tran LG, Cha HK, Park WT (2017) RF power harvesting: a review on designing methodologies and applications. Micro and Nano Syst Lett 5:14

A Review: Recent Automatic Algorithms for the Segmentation of Brain Tumor MRI

Asra Rafi, Zia Khan, Faiza Aslam, Soyeba Jawed, Ayesha Shafique, and Haider Ali

Abstract Medical imaging techniques are a vital tool in disease diagnosis. The images are being developed to satisfy the growing need for important information from medical image scans by anticipating constitutional tissues for clinical analysis. The application of deep learning techniques is increasing with the demand for automatic diagnosis of medical imaging. Different layers are used in deep learning models to represent data abstraction and construct computational models. Imaging techniques allow medical experts such as radiologists to correctly recognize a patient's condition, making medical procedures more accessible and automated. The review's primary goal is to present a study on recent brain tumor detection segmentation and classification approaches. Brain tumors are reviewed because of their importance compared to other tumors and their high illness rate. Many brain tumor segmentation models have been described to grasp these methodologies well, along with their limits and benefits. The study focuses primarily on contemporary deep learning-based brain tumor detection technologies, such as deep generative and deep learning networks. The more advanced and recent techniques available in the literature are also reviewed to describe the methods for performing image segmentation and to emphasize the importance of segmentation models that are not used in real-time due to little or no interaction between clinicians and developers. Most research does not consider the data augmentation element of brain tumor segmentation, which is critical for improving performance. The most challenging feature, or limitation, is the fluctuation in the morphology of tumors or the intensity degree of tumors, both of which still require study in this arena.

A. Rafi · F. Aslam
Department of Computer Science, COMSATS University, Islamabad, Pakistan

Z. Khan (✉) · S. Jawed
Department of Electrical and Electronic Engineering, Universiti Teknologi Petronas, Seri Iskandar,, Malaysia

A. Shafique
Department of Computer Science, Nanjing University of science and technology, Nanjing, China

H. Ali
IT Artificer Peshawar, Khyber Pakhtunkhwa, Pakistan

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_25

Keywords Brain tumor segmentation · Deep learning networks · Deep generative networks

1 Introduction

The human brain is a porous stack of soft tissue, and a brain tumor is an abnormal growth of undesirable cells in the interior section or over the human brain. This anomaly may cause pressure within the skull and swelling of the brain in the brain. The brain tumor is divided into two sub-categories [1].

Primary Brain Tumor

This type of tumor can start in the brain region. It can be 'Benign' or 'Malignant' [2]. The structure of benign tumor type is uniform, and it has no cancerous cells. However, malignant is a worse tumor type which is non-uniform in a structure containing cancer cells.

Secondary Brain Tumor

The secondary or metastatic tumor may start in any body part and spread to the human brain. As it is metastasis as knowledge-based that could be in the body of patients as one tumor or multiple tumors like breast, lungs etc.

Brain Tumor is detected by imaging tests which are performed on Magnetic Resonance Imaging (MRI) [3] and Computed Tomography (CT) scan [4]. MRI is considered more appropriate for brain tumor detection as it can provide better contrast resolution in comparison with CT [5, 6]. In the MRI technique, ionizing radiation is not used, which protects healthy brain tissues from harm [7]. The analysis of brain tumor segmentation is a step to find the information in the clinical aspect and extract diagnostic characteristics of patients from images. To extract this information, multidimensional image data is provided through which brain tumor detection and localization is performed through different techniques available in the literature [8]. These techniques are used for the processing data, extracting useful patterns from data, labeling the examples and representing the information from image data. In medical imaging, precise detection of brain tumors is playing a vital role in 2019 [9]. There are many technical tools available for brain tumor detection like MRI, CT and Positron Tomography. These tools are useful to examine the region of the tumor in various shapes, sizes and positions. Brain tumor segmentation is still a demanding task in medical imaging.

Recently, Convolutional Neural Network (CNN) [10] based technique used for tumor detection, specifically brain tumor. The CNN uses convolutional layers to convolve an image or signal with filters to get feature maps in return and links with previous layers [11]. The whole process of segmentation is done most of the time by following these steps: MR scans pre-processing, patch extraction as MRI data is volumetric data and has background pixels as well which are not helpful so patches from brain part are extracted, patch pre-processing in the next step to normalize

Table 1 Summary of image based segmentation work on the brain tumor detection

Platform	Dataset	Proposed system	Year	Simulation ENV	Results
ML based	BraTs 2013, 2015	Potential field (PF) clustering, Local binary pattern (LBP), Gabor wavelet transform (GWT) [43]	2019	Matlab 2015	DSC = 0.96, 0.98
ML based	212 MRI images	Gray level Co-occurence Matrix (GLCM), Multi-Layer perceptron (MLP), Naive Bayes [44]	2014	Matlab 2012	Acc = 0.98
ML based	944 MRI images	Orthogonal Gamma distribution machine learning appraoch (OGMLA) [45]	2017	Not Given	Acc = 0.98
Region based	BraTs 2012	Localized active contour model with background intensity (LACM-BIC) [46]	2017	Not Given	DSC = 0.93
Super pixel based	BraTs 2012	Random majority down-sampling-synthetic minority sampling technique [47]	2018	N4050 system core2duo 2.10 GHZ	DSC = 0.91
Super pixel based	BraTs 2012	Simple linear iterative clustering [48]	2016	Not Given	DSC = 0.88
Region based	Brain MRI dataset	Fuzzy C mean [49]	2014	Not Given	DSC = 0.94

patches so that they could be able to pass through the network, in the next step model is used for the training of images and post-processing in the next step to refine the results further.

Structure: First, in Sect. 2, we have reviewed the recent work in the literature on brain tumor segmentation using deep learning and machine learning models, and we have provided a Table 1 containing relevant articles details. Second, in Sect. 2 Brain Tumor Segmentation [12] along with the detailed review of Image segmentation techniques, deep generative models and deep learning models applications in brain tumors. The challenges phased by researchers in brain tumor detection are discussed in Sect. 3 along with the future directions. The survey is summarized in Sect. 4.

1.1 Search Engine

To identify related contributions, different digital databases such as Scopus, PubMed, scienceDirect, IEEExplore are utilized to collect important literature. The query for papers containing is "machine learning" "convolutional", "deep learning", "brain", and "image segmentation" in title or abstract. Additionally, conference proceedings for BraTs 2014 , BraTs 2015, BraTs 2017 and BraTs 2018, are searched based on each paper's title. The most recent paper reviewed in this article is the one published on April 2021 [13] (Fig. 1).

2 Related Work of Brain Tumor Segmentation

2.1 Image Segmentation Techniques

Image segmentation is one of the tasks in image processing in which input image scan is partitioned into different areas having similar pixels features. The work on medical imaging has been increased with time, and rapid progress can be observed in this field as various techniques have been proposed recently [14, 15]. Image segmentation aims to modify the image representation in a more meaningful and understandable way. Its techniques are divided based on significant features assisting the segmentation [16].

2.1.1 Edge-Based Segmentation

Edge-based segmentation [17] is a type of segmentation technique that is based on the intensities differences instead of defining the similarities of pixels to discover the closer borders analogous to the image scan object [18]. This technique can be used to get control of the size modification outcome of the segmented tumor caused because

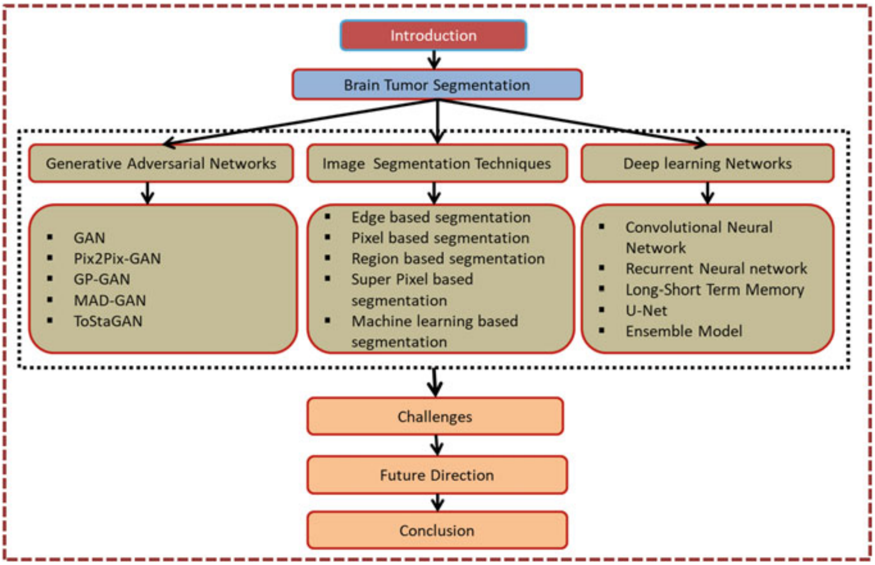


Fig. 1 Organization of the survey paper

of inappropriate thresh-holding process, which has been used for segmentation [19]. This approach is designed to be variation sensitive, and it checks either pixel is on edges [20]. The drawback of this approach is that the edges identified by this technique do not encircle the segmented object entirely. Various post-processing steps have been taken to combine the linked edges for the improvement of image edges representation as these are sensitive to noise in the image [21–23].

2.1.2 Pixel-Based Segmentation

Pixel-based segmentation is a threshold-based technique. It is the simplest method of segmentation which is in practice for two-dimensional image scans. In this method, neighboring pixels values are discarded, and the current pixel value is considered only. The threshold is measured from the image histogram in pixel-based approaches [24]. If the object segmentation is performed by one threshold, then it is called global thresh-holding. On the other hand, local thresh-holding is done when we have two or more objects to be segmented [25, 26]. The limitation of this method is that relation between pixels is ignored, and only pixel information or intensity value is being considered. As a result, certain pixels are unable to join the required or the image background portion. It can be noted that threshold-based segmentation techniques have been used to differentiate between background portion and brain area of images [15].

2.1.3 Region-Based Segmentation

The input image is partitioned into sub-regions based on similarity criteria defined in the start. The start is taken with one or multiple pixels known as seeds. The neighboring group of pixels is evaluated, and the groups which fulfill the similarity criteria are added [27]. The similarity of the pixels is defined based on intensity data or edges in the image. This process of adding pixels is repeated till the addition of the last pixel in the formation of interest. The region growing approach is based on the segmentation of similar regions and creates associated regions [21]. The Partial Volume (PV) effect is the major disadvantage of this region-growing approach, limiting the MRI segmentation accuracy. The borders between multiple tissues in MRI are blurred by PV, as voxel might contain a different type of tissues [28]. Region growing approaches may create holes in the extracted regions as they are noise sensitive [29]. Another point to be noticed in this approach is that if the seed point is not selected appropriately, then the extracted region might grow outside of the section of interest or combines with another irrelevant region [18].

2.1.4 Superpixel Based Segmentation

Superpixel is a technique of grouping similar looking pixels in an image. A superpixel based approach is used for region extraction in an image. Several graph-based superpixel segmentation and gradient-based superpixel segmentation techniques are available in the literature [30–35]. Mainly superpixel segmentation is divided into: (1) Graph-based techniques and (2) Gradient ascent based technique. Graph-based techniques treat each pixel of an image as a nodes in a graph. The weight of the edges is calculated between two adjacent nodes of a graph, the weight of these nodes are proportional to similar looking neighbouring pixels. Superpixels are based on the cost function defined over the graph. It can be done using the following techniques which include normalised cut [33], efficient graph based image segmentation [31] and entropy based segmentation [34]. The normalised cuts technique [33] is recursive in nature and divides the graph of pixels based on texture and contour obtaining the global minimum of a cost function defined on the edges at the partition boundaries. This technique produced regular superpixels with less varying size and shapes. The drawback is boundaries of the superpixels are not very well defined and it is quite slow, the segmentation process requires more than two minutes to complete. Another approach [31] where clustered pixel in an image are defined as nodes on a graph, where each superpixel is based on the minimum spanning tree of pixels. These superpixels have well defined boundaries of regions but are irregular (widely varying size and shape) in shape.

In [34], a graph-based approach is proposed for superpixel segmentation, an objective function is proposed based on: (1) entropy rate of a random walk on a graph; (2) a balancing term based on cluster division. Compact and homogeneous clusters are defined based on their entropy rate (entropy measures the randomness of the elements), which further divides the image on perceptual boundaries where the balancing term is used to maintain the clusters of equal size. The objective function is optimised using an algorithm based on submodular function. Solving this submodular function is NP hard problem for which a global optimum is difficult to achieve. Therefore, the algorithm utilises the greedy approach thus the formulation obtains a 1/2 bound for the optimal solution. Results shows that it performs better than the state of the art [30–32, 34] in terms of computational complexity and accurate boundary's. Moreover, it also produces regular shape superpixels with accurate boundary adherence [34].

The gradient ascent based technique makes an initial assumption about the clustering of the pixels and refines it with iterations until a set criterion is met for the superpixel. The gradient ascent based techniques include mean shift [30] and watershed [32].

2.1.5 Machine-Learning-Based Segmentation

The most advanced procedure used for the automatic diagnosis and analysis of medical images is machine learning. Through machine learning algorithms, complex patterns from the empirical data are learned for accurate decision making [25]. There are three categories of machine learning methods used for image segmentation named unsupervised [36], semi-supervised [37] and supervised [38] segmentation approach. Automatic labeling of train data by similar grouping pixels numerically, the segmentation is called unsupervised segmentation [39]. The texture features and pixel intensities are used to segment MRI scans into notable sections. When manually labeled training data is considered, then the segmentation approach is supervised. Brain tumor segmentation algorithms are based on Artificial Neural Network (ANN) [40], Fuzzy C-Means (FCM) [41] and Support Vector Machine (SVM) [42].

2.2 Deep Learning Networks

Deep Learning Networks are depending on searching for the best architecture to solve a particular problem. As deep learning solves complex and challenging tasks, the network architectures become more complex to design. The learning networks discussed in this article are CNN, Deep Neural Network (DNN) [50], Recurrent Neural Network (RNN) [51], Long Short Term Memory (LSTM) [52].

2.2.1 CNN

The CNN has been reviewed in this section then we move toward the other platforms. In this article, to recognize the glioma, deep learning approaches are used to cover the region or finding the area where the glioma is present. MRI is playing a vital role still. The manual effort is time taking and tedious. At the last of the year, 2019 Simulated Benchmark dataset (SBDL) perform acute for the tumor segmentation, which is validated by Dice similarity [53]. Later Dominant Rotated Local Binary Patterns (DRLBP)[54] feature for deep learning-based fusion which are optimized by Particle Swarm Optimization (PSO) [55]. For this purpose inception, V3 CNN is used to extract features by using the softmax classifier. The results are collected over the BRATS dataset with 10 fold cross-validation, and the comparison conducted between BRATS 2013, BRATS 2014, BRATS 2017, BRATS 2018 dataset [56]. The average results on this dataset are collected, and it is outperformed in segmentation and classification. For further improvement, we can use it on capsule net.

Around the middle of the 2019 year, Neutro-sophic Set Expert Maximum Fuzzy-Sure Entropy (NS-EMFSE) [57] is presented to automate the edge detection of the tumor using MRI. Whereas for detecting brain tumors, the Hybrid technique Neutro-sophic with CNN is used with 5 fold cross-validation. The ensemble model NS-EMSE + CNN + SVM outperforms the benchmark scheme and shows 95.62 accuracies with the softmax classifier. The experiment is performed over The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) [58] dataset. To measure the success of the system youden index [59], precision and accuracy are used as evaluation parameters [60]. In the future, the system could be enhanced by using the different neutrosophic techniques with CNN.

2.2.2 DNN

DNN is the neural network with a definite complexity, having models following sophisticated modeling based on mathematics. An efficient methodology DNN and Discrete Wavelet Transform (DWT) [61] by[62] is presented, which is an ensemble for the brain tumor detection in MRI on 66 real scans in which 22 represent normal and the remaining 44 are categorized into the types of tumor. Classification is done with 7 fold cross-validation technique with 7 hidden layers of DNN. The model is evaluated with average precision, average F-measure, average Receiver Operating Characteristic (ROC) curve [63] into different classes like normal image glioblastoma, sarcoma and metastatic bronchogenic carcinoma tumor.

2.2.3 RNN

To discriminate the various type of brain tumors on MRI scans is a still difficult task in medical imaging [64]. A novel approach to holistic 3D brain tumor detection is presented. The proposed architecture used the holistic label instead of pixel or slice-wise labeling. To detect brain tumor types like glioma, meningioma, and metastasis tumor, a dataset of 422 MRIs is conducted. While doing experiments, both public and proprietary datasets are used for tumor screening and classification with two proposed schemes, Dense Net-NSTM and Dense Net-Dense Net. It shows that the approach proposed by the authors is very operative because of public and proprietary. The proposed approach would become more effective if the model could be extended to solve the other type of tumor.

Automated segmentation of High Graded Glioma HGG is necessary to save the life of people. A novel approach RNN is introduced at first 2019 [65]. LSTM approach helps have global context at once instead of using conventional scheme CNN which gained the global context by combining local features. HGG with 3D-Hilbert Filter curve is used over the BraTs-17 dataset. Standard evaluation metric Dice Similarity Coefficient (DSC) [66] is used for to the evaluation of the respective model. The validation score of the dice over the model is 0.62, 0.77, 0.64, which further improves the quality of the whole and core tumors.

2.2.4 LSTM

LSTM is an artificial neural network having feedback connections. It can process images as well as sequences of large data [52]. LSTM model is used by [67] for automated brain tumor detection by using MRI. To achieve the high-quality N4ITK and Gaussian 5 X 5 matrix filter are used in multi-sequence MRI. Experiments are conducted over Brats dataset and Sub-acute Ischemic Stroke Lesion Segmentation-Ischemic Stroke Lesion Segmentation (SISS-ISLES) dataset [68] included five MRI sequences T1, T1C, T2, Fluid Attenuated Inversion Recovery (FLAIR)[69], Diffusion-Weighted Imaging (DWI) [70]. To evaluate the model's performance, Dice Similarity Coefficient is used with 1.00 on 2012 synthetic, 0.95 on 2013, 0.99 on 2014, 0.98 on SISS-ISLES 2015, 0.99 2018 on 2015, 0.97 DSC precision is better for the real-world patient. Our proposed model attained 0.95 on BRATS.

2.2.5 U-Net

U-Net [71] is a deep learning model which was first introduced in 2015 and has been used for the segmentation tasks in medical imaging. This architecture is based on CNN, consisting of an encoder/contracting path to extract the context followed by an asymmetric path known as decoder, expanded for compressed localization of object within an image. Most recently, CNN-based methods perform brain tumor segmentation, which generally adds limited layers and small receptive fields, affecting contextual information quality. In this context, [72] has used the U-Net in combination with loss function, which is class sensitive and data-augmentation, which makes a sophisticated model able to train on few examples and achieved consistent results. Recently, U-Net-based models are being used for automatic brain tumor segmentation as [73] has applied this architecture based on fully convolutional networks on the BraTS 2015 dataset and performed cross-validation to evaluate their model and have achieved comparable results.

2.2.6 Ensemble Model

In 2019, to diagnose the brain tumor, a new approach by the integration of small kernels two-path convolutional neural network (SK-TPCNN) [74] with Random Forest (RF) algorithm is presented, which has improved the feature extraction quality. This architecture combines small and large convolution kernels, which increased the non-linear mapping to overcome the over-fitting. By the integration of RF, the feature is reduced. Experiments are conducted over the Brats 2015 instead of lesion dataset using cross-entropy function. It shows that training or learning ability is well defined as we increased the number of paths. The results are validated using cross-validation techniques through the DSC 0.92 performance-enhancing region, and the whole region contains 0.89. A new framework is introduced in 2018 by the combination of cascaded CNN with the support of LSTM is proposed [75]. This model has remarkable results over the conventional scheme. LSTM clearly distinguishes between low and high-grade glioma. It is specially designed for 3D brain tumor classification. The results show that VGG-16 outperforms for extracting the high-level feature as compared to AlexNet and ResNet. It would be more exciting pixel-by-pixel labeling with less no of instances in the future. In MRI segmentation RNN model was used in 2018 [76] to scan the sequence of data in a prostate. Interaslice information is

Table 2 Summary of deep learning based work on the brain tumor detection

Platform	Dataset	Proposed system	Year	Simulation ENV	Results
CNN based	BraTs 2017, 2018, 2019	BrainSeg-Net [78]	2021	Not given	DSC= 0.90
CNN based	BraTs 2018	DCU-Net [79]	2020	Keras Framwork, Nvidia GeForce GTX 1060 GPU	DSC = 0.90
CNN based	BRaTs 2017–2018	CNN+DRLBP, PSO [53]	2019	Matlab 2018	Acc = 0.98
CNN based	TCGA-GBM dataset	NS-EMSE [57]	2019	Matlab 2017	Acc = 0.85
DNN based	122 MRI images	DNN+DWT [62]	2017	Matlab2015	Acc = 0.87
RNN based	422 MRI	DenseNet-LSTM+ DenseNet-DenseNet [64]	2019	Nvidia Titan Xp GPU [80]	
RNN based	BraTs 2017	LSTM [65]	2019	Not given	DSC = 0.62,0.17,0.64
LSTM	MICCAI and SISS-ISLES dataset	N4ITK and Gaussian Filter [67]	2019	Matlab2015	Acc = 0.98
Fusion based	BraTs dataset	Non-subsampled contourlet transform [81], Adaptive Neuro Fuzzy Inference System (ANFIS) [82]	2018	MatlabR 2014b	Acc = 0.95
Fusion based	MICCAI dataset	Efficient multi-sequence [83]	2018	MatlabR	DSC = 0.70, 0.63
Deep Learning model	56 of 136 of MRI dataset	Multi-parametric DLM [84]	2019	Not given	DSC= 0.81, 0.78, 0.27
Ensemble model	BraTs 2015 dataset	SK-TPCNN with RF [74]	2019	Matlab	DSC = 0.92, 0.89
Ensemble model	BraTs 2015 dataset	LSTM with cascaded CNN [75]	2018	MATLAB 2018b	Acc = 0.84
Ensemble model	BraTs dataset	RNN with BDC-LSTM [76]	2018	Not given	Acc= 0.86

also helpful in prostate segmentation. The proposed scheme used the neighboring slice into the network, and the network extracted the following interslice context based on previous segmentation results. To check the performance of RCNN over CNN, we replace the Fully Convolutional Neural Networks (FCNs) with Bidirectional Convolutional LSTMs (BDC-LSTMs) [77]. The results are matched with the benchmark scheme on U-Net, V-Net. FCN and U-Net better prostate performance using the inter-slice scheme to reduce the information loss and monitor the feature extraction. In this way, segmentation results are improved (Table 2).

Table 3 Summary of GAN based work on the brain tumor detection

Platform	Dataset	Proposed system	Year	Simulation ENV	Results
GAN based	ANDI dataset, BRaTs 2015	Pix2pix- GAN	2018	Nvidia GDX	DSC = 0.81
GAN based	BraTs 2014	GP-GAN	2020	Pytorch, GPU Nvidia Titan XP	DSC = 0.88
GAN based	3064 MR images	GAN+CNN	2020	Python 3.6, GTX 1060 GPU	DSC =0.95
GAN based	BraTs 2015- 2017	RescueNet	2019	Nvidia Tesla V1004	DSC = 0.94
GAN based	AD dataset: OASIS-3	MADGAN	2019	Nvidia Quadro GV100 GPU	AUC = 0.92
GAN based	BraTs 2018	Vox2Vox: 3DGAN	2021	GeForce RTX 2080 Ti	DSC= 0.90
GAN based	BraTs 2015	ToStaGAN	2021	Nvidia GTX 2080Ti GPU	DSC = 0.85

2.3 Deep Generative Networks

Deep generative networks [85] are used to generate data samples that follows the exactly same probabilistic distribution of a specific training dataset. Deep generative models have used the idea of deep learning to create generative models and learning algorithms. In medical imaging, these networks are used for data augmentation. As in medical imaging, we have fewer data provided to perform operations, and deep learning models require more data to train and learn the image patterns, so in this regard, deep generative networks are highly preferred. Recently, work was done on the data augmentation for one-shot segmentation of brain MRI scans in [86] by the spatial transformation of images. They claimed boosted performance of the model by the augmentation of realistic synthesis of images. Generative Adversarial Network (GAN) [87] is an artificial neural network proposed for the generation of images effectively.

2.3.1 GAN

GAN is a generative model based on a machine learning system introduced by [87] in 2014. In [88, 89] authors have presented a method for generating abnormal MRI scans affected by brain tumors using the GAN model. By the augmentation of MRI scans, they claim that their model has achieved better brain tumor segmentation performance. In [90] implemented for the first time a technique called GP-GAN for identification of brain tumor with noval objective function.The noval objective function enhanced the prediction of brain tumor galioma.Navid Ghassemi et al. [91] performed the data augmentation of brain MRI using GAN. The augmented are then fed into CNN network for classification of brain MRI tumor. Shubhangi Nema et al.[92] introduced residual cyclic unpaired encoder-decoder network (RescueNet) for segmentation of brain tumor MRI images. The RescueNet utilizing the unpair adversarial training to segmentation the tumor of brain MRI images.Changhee Han et al. [13] introduced the unsupervised medical anomaly detection generative adversarial network (MADGAN), an unique two-step technique for detecting brain anomalies at different stages on multisequence structural MRI utilising GAN-based multiple adjacent brain MRI slice reconstruction. Furthermore, Marco Domenico Cirillo et al. [93] introduced a 3D volume-to-volume Generative Adversarial Network (GAN) for segmentation of brain tumours. The proposed approach, called Vox2Vox, performed the segmentation of multi-channel 3D brain MRI images. Yi Ding et al. [94] introduced a noval approach called two-stage generative adversarial neural network (ToStaGAN) to improve the brain tumor segmentation accuracy. In first stage the coarse segmentation is performed, which is then forward to the second stage to produce better segmented output of brain tumor MRI (Table 3).

3 Challenges and Future Work

In the medical imaging domain, the main challenges are relevant to benchmark machine learning challenges. Some of the major challenges are overfitting, class imbalance, variation in the shape and intensities of the tumor. When we do not have enough labeled data in medical image diagnosis, then the issue of **overfitting** occurs. To overcome the problem, data augmentation is recommended in the future. Instead of memorizing the pattern, the model gets variance in training data and could learn the usage patterns. The issue of **class imbalance** is a major hurdle time in the divergence of models, especially in the domain of medical imaging. This problem affects brain tumor segmentation as we have a huge amount of healthy tumor cells compared to abnormal cells, which causes the problem of class imbalance. People have used multiple sampling techniques to decrease the majority class or increase the minority class to deal with this problem. On the other side, many loss functions are proposed in the literature, used to overcome the class imbalance. In the future, generalize dice loss function [95] can be used to deal with this problem. It would be best practice in the future to use the 3D Convolution for the better performance of the model to look over the sampling loss plan, and layer loss [96].

The multi-phase technique is used to segment brain tumors based on a similarity region-based approach with a random walk algorithm. The architecture does not ensure expressive segmentation. DNN and DWT is an ensemble used for the brain tumor diagnosis in MRI on 66 real images in which 22 represent normal, and the remaining 44 are categorized into the other types of tumor. CNN takes less time in training than DNN, so better results are achieved by making an ensemble of CNN + DWT in the future. A novel approach to end-to-end segmentation is useful by introducing the CNN replacing U-net with effective cascade training strategy followed beside it pixel-by-pixel segmentation [97]. Dynamically sampling trains complex architecture U-net with data augmentation and class sensitive loss on few data. LSTM outperforms for the detection of benign and malignant tumors [2]. The model performed well over the high-dimensional 3D data. It would still be challenging to predict the glioma accurately, and pixel label tumor along with the fewer amount of training data [98]. With the acute development in medical image processing and its research venues being in the limelight to detect brain tumors, the SbDL technique is recommended for tumor detection. For more advancement, CNN could be implemented with the capsule net. In image processing, the neutrosophic technique plays a vital role in the ensemble of CNN. However, its performance can be boosted by applying other variants of CNN. Brain tumor has round about 201 types, so it is challenging to discriminate. MRI-based techniques are used nowadays to find the location of the brain tumor region appropriately. With the advancement in MRI techniques, different modalities of MRI would be incorporated effectively in future practice. The modalities could be Magnetic Resonance Spectroscopy (MRS), Diffusion Tensor Imaging (DTI) [99], and Perfusion Imaging (PI) [100] Holistic 3D brain tumor detection is presented in [64] for three types of brain tumors. However, it would be more challenging to detect another type of tumor by providing model understandability through weekly supervised pathology localization. This domain has gaps that could become a working area in future research.

4 Conclusion

It is concluded that the demand for automatic diagnosis has been increased as time passes. So by taking motivation from the recent models proposed for the brain tumor diagnosis, which are deep learning-based, a brief review has been taken place in the form of an article. In this article, different deep learning-based algorithms are discussed in detail. This article has discussed the challenges and provides future directions that could be followed to deal with the challenges. The findings of this article and future directions are summarized below:

- It is observed that although deep learning models are available with usage patterns, the deep learning interpret-ability needs to be examined in the future.
- As in healthcare specifically for brain tumor segmentation, we have a small amount of labeled data for the deep learning model training. For one-shot learning, recent works have been done.
- Most of the work in deep learning is supervised learning-based. However, we need to give attention to unsupervised learning [36] and semi-supervised learning [37] approaches for handling examples without labels.
- To overcome the issue of class imbalance discussed in the review, there is still room to propose a new loss function or provide sampling techniques.
- The problem of overfitting is discussed in the article is a challenging and attention seeking task which needs to be addressed in the future either by data augmentation or by parameter tuning.

References

1. El-Melegy M (2014) Tumor segmentation in brain MRI using a fuzzy approach with class center priors. EURASIP J Image Video Process 2014:21
2. Walker AE, Robins M, Weinfeld FD (1985) Epidemiology of brain tumors: the national survey of intracranial neoplasms. Neurology 35(2):219 (1985)
3. Huettel SA, Song AW, McCarthy G (2004) Functional magnetic resonance imaging, vol 1. Sinauer Associates, Sunderland
4. Naidich DP et al (1982) Computed tomography of bronchiectasis. J Comput Assist Tomogr 6(3):437–444
5. Anitha V, Murugavalli S (2016) Brain tumor classification using two-tier classifier with adaptive segmentation technique. IET Comput Vis 10(1):9–17
6. Huang M, Yang W, Wu Y, Jiang J, Chen W (2014) Brain tumor segmentation based on local independent projection-based classification. IEEE Trans Biomed Eng 61(10)
7. Bahadure NB, Ray AK, Thethi HP (2017) Image analysis for MRI based brain tumor detection and feature extraction using biologically inspired BWT and SVM. Hindawi Int J Biomed Imaging 12
8. Wong K (2005) Handb Biomed Image Anal Segment Models Part B 2:111–182
9. Tiwari A, Srivastava S, Pant M (2019) Brain tumor segmentation and classification from magnetic resonance images: review of selected methods from 2014 to 2019. Pattern Recogni Lett
10. Chua Leon O, Roska Tamas (1993) The CNN paradigm. IEEE Trans Circuits Syst I Fund Theory Appl 40(3):147–156
11. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imag 35(5) (2016)
12. Havaei M et al (2017) Brain tumor segmentation with deep neural networks. Med Image Anal 35:18–31
13. Han C, Rundo L, Murao K, Noguchi T, Shimahara Y, Milacski ZÁ, Satoh SI et al (2021) MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. BMC Bioinf 22(2):1–20
14. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Lanczi L et al (2014) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34(10):1993–2024
15. Gordillo N, Montseny E, Sobrevilla P (2013) State of the art survey on MRI brain tumor segmentation. Magn Reson Imaging 31(8):1426–38
16. Poonam J, Dass R, Duhan M (2019) A comparative analysis of various image segmentation techniques. In: Proceedings of 2nd international conference on communication, computing and networking. Springer, Singapore
17. Liu Cheng, Liu Weibin, Xing Weiwei (2019) A weighted edge-based level set method based on multi-local statistical information for noisy image segmentation. J Vis Commun Image Rep 59:89–107
18. Dougherty G (2009) Digital image processing for medical applications. Cambridge University Press
19. Jähne B (2005) Digital image processing, 60th edn. Springer, Berlin
20. Birry RAK (2013) Automated classification in digital images of osteogenic differentiated stem cells. University of Salford
21. Rogowska J (2009) Overview and fundamentals of medical image segmentation (Chap. 5). In: Bankman IN (ed) Handbook of medical image processing and analysis. Elsevier/Academic Press
22. Sonka M, Hlavac V, Boyle R (2014) Image processing, analysis, and machine vision. Cengage Learning
23. William K (2001) Digital image processing, 3rd edn. Wiley-Interscience, Canada, p 656
24. Petrou M (2010) Texture in biomedical images. In: Deserno MT (ed) Biomedical image processing. Springer, Berlin, pp 157–76

25. Jin L, Min L, Jianxin W, Fangxiang W, Tianming L, Yi P (2014) A survey of MRI-based brain tumor segmentation methods. Tsinghua Sci Technol 19(6):578–95
26. Naji SA, Zainuddin R, Jalab HA (2012) Skin segmentation based on multi pixel color clustering models. Digit Signal Process 22(6):933–40
27. Solomon C, Breckon T (2011) Fundamentals of digital image processing: a practical approach with examples in matlab. Wiley
28. Mie S, Lakare S, Ming W, Kaufman A, Nakajima M (2000) A gradient magnitude based region growing algorithm for accurate segmentation. In: Proceedings 2000 International conference on image processing . Vancouver, BC. IEEE, pp 448–451
29. Pham D, Xu C, Prince JL (1999) A survey of current methods in medical image segmentation. Technical report, Johns Hopkins University, Baltimore
30. Comaniciu D, Meer P (2002) Pattern analysis and machine intelligence. IEEE Trans Mean Shift Robust Appr Toward Feat Space Anal 5:603–619
31. Felzenszwalb PF, Huttenlocher PD (2004) Segmentation efficient graph based image. In J Comput Vis 59(2):167–181
32. Vincent L, Soille P (1991) Pattern analysis and machine intelligence. IEEE Trans Watersheds Digit spaces Effic Algorithm Based Immers Simul 13(6):583–598
33. Levinshtein A, Kutulakos ASKN, Fleet DJ, Dickinson SJ, Siddiqi K (2009) Pattern analysis and machine intelligence. IEEE Trans TurboPixels Fast Superpixels Using Geometr Flows 12:2290–2297
34. MingYu L, Tuzel O, Ramalingam S, Chellappa R (2011) Proceedings of IEEE Computer Society conference on computer vision and pattern recognition. In: Entropy rate Superpixel segmentation, 2011, June, pp 2097-2104
35. Achantaand R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2012) Pattern analysis and machine intelligence. IEEE Trans SLIC Superpixels Compar State Art Superpixel Methods 11:2274–2282
36. Barlow HB (1989) Unsupervised learning. Neural comput 1(3):295–311
37. Chapelle O, Scholkopf B, Zien A (2006) In: Chapelle O et al (eds) Semi-supervised learning
38. Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning. ACM
39. Larose DT (2005) Discovering knowledge in data: an introduction to data mining. Wiley, USA
40. Zhang Z (2018) Artificial neural network. In: Multivariate time series analysis in climate and environmental research. Springer, Cham, pp 1–35
41. Arora J, Khatter K, Tushir M (2019) Fuzzy c-means clustering strategies: a review of distance measures. Software engineering. Springer, Singapore, pp 153–162
42. Tavara S (2019) Parallel computing of support vector machines: a survey. ACM Comput Sur (CSUR) 51(6):123
43. Amin J, Sharif M, Raza M, Saba T, Anjum MA (2019) Brain tumor detection using statistical and machine learning method. Comput Methods Progr Biomed 177:69–79
44. Sharma K, Kaur A, Gujral S (2014) Brain tumor detection based on machine learning algorithms. In J Comput Appl 103(1):7–11
45. Inguva MSC, Goud VM, Srikanth N, Manjula Y (2018) Machine-learning approach based Gamma distribution for brain abnormalities detection and data sample imbalance analysis
46. Ilunga-Mbuyamba E, Avina-Cervantes JG, Cepeda-Negrete J, Ibarra-Manzano MA, Chalopin C (2017) Automatic selection of localized region-based active contour models using image content analysis applied to brain tumor segmentation. Comput Biol Med 91:69–79
47. Rehman ZU, Naqvi SS, Khan TM, Khan MA, Bashir T (2019) Fully automated multiparametric brain tumour segmentation using superpixel based classification. Exp Syst Appl 118:598–613
48. Soltaninejad M, Yang G, Lambrou T, Allinson N, Jones TL, Barrick TR, Ye X (2017) Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI. Int J Comput Assist Radiol Surg 12(2):183–203

49. Preetha R, Suresh GR (2014). Performance analysis of fuzzy c means algorithm in automated detection of brain tumor. In: 2014 World Congress on computing and communication technologies. IEEE Feb 2014, pp 30–33
50. Miikkulainen R et al (2019) Evolving deep neural networks. In: Artificial intelligence in the age of neural networks and brain computing. Academic Press, pp 293–312
51. Zhang Lei, Wang Shuai, Liu Bing (2018) Deep learning for sentiment analysis: a survey. Wiley Interdiscipl Rev Data Mining Knowl Discov 8(4):e1253
52. Zhao Z et al (2017) LSTM network: a deep learning approach for short-term traffic forecast. IET Intell Transp Syst 11(2):68–75
53. Sharif MI, Li JP, Khan MA, Saleem MA (2019) Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images. Pattern Recognit Lett
54. Szűcs J, Péter B (2019) Binary tomography using variants of local binary patterns as texture priors. In: International conference on computer analysis of images and patterns. Springer, Cham
55. Pan X et al (2019) Hybrid particle swarm optimization with simulated annealing. Multimed Tools Appl 78(21):29921–29936
56. Simpson AL et al (2019) A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063
57. Özyurt F, Sert E, Avci E, Dogantekin E (2019) Brain tumor detection based on convolutional neural network with neutrosophic expert maximum fuzzy sure entropy. Measurement 147:106830
58. https://wiki.cancerimagingarchive.net/display/Public/TCGA-GBM . 27 Dec 2019
59. Martínez-Camblor P, Pardo-Fernández JC (2019) The Youden index in the generalized receiver operating characteristic curve context. Int J Biostat 15(1)
60. Campana SE (2001) Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. J Fish Biol 59(2):197–242
61. Zhang D (2019) Wavelet transform. Cham, Fundamentals of Image Data Mining. Springer, pp 35–44
62. Mohsen H, El-Dahshan ESA, El-Horbaty ESM, Salem ABM (2018) Classification using deep learning neural networks for brain tumors. Fut Comput Inf J 3(1):68–71
63. Hajian-Tilaki K (2013) Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Internal Med 4(2):627
64. Zhou Y, Li Z, Zhu H, Chen C, Gao M, Xu K, Xu J (2018) Holistic brain tumor screening and classification based on DenseNet and recurrent neural network. In: International MICCAI Brainlesion Workshop. Springer, Cham, Sept 2018, pp 208–217
65. Grivalsky S, Tamajka M, Benesova W (2019). Segmentation of gliomas in magnetic resonance images using recurrent neural networks. In: 2019 42nd International conference on telecommunications and signal processing (TSP). IEEE, pp 539-542, July 2019
66. Tustison NJ, Gee JC (2009) Introducing Dice, Jaccard, and other label overlap measures to ITK. Insight J 2
67. Amin J, Sharif M, Raza M, Saba T, Sial R, Shad SA (2019) Brain tumor detection: a long short-term memory (LSTM)-based learning model. Neural Comput Appl 1–9
68. Maier O, Wilms M, Handels H (2015) Image features for brain lesion segmentation using random forests. In: BrainLes 2015. Springer, Cham
69. Eltayeb EN, Salem NM, Al-Atabany W (2019) Automated brain tumor segmentation from multi-slices FLAIR MRI images. Bio-medical Materi Eng. Preprint, pp 1–13
70. Tufton N et al (2019) Diffusion-weighted imaging (DWI) highlights SDHB-related tumours: a pilot study. Clin Endocrinol
71. Ronneberger O, Fischer P, Thomas B (2015) U-Net: convolutional networks for biomedical image segmentation. Medi Image Comput Comput Assist Intervent (MICCAI) 9351:234–241
72. Fabian I (2017) Brain tumor segmentation using large receptive field deep convolutional neural networks. In: Bildverarbeitung für die Medizin. Springer, Berlin, pp 86–91
73. Dong H et al (2017) Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. Annual conference on medical image understanding and analysis. Springer, Cham

74. Yang T, Song J, Li L (2019 A deep learning model integrating SK-TPCNN and random forests for brain tumor segmentation in MRI. Biocybern Biomed Eng
75. Shahzadi I et al (2018) CNN-LSTM: cascaded framework for brain tumour classification. In: 2018 IEEE-EMBS conference on biomedical engineering and sciences (IECBES). IEEE
76. Zhu Q, Du B, Turkbey B, Choyke P, Yan P (2018) Exploiting interslice correlation for MRI prostate image segmentation, from recursive neural networks aspect. Complexity
77. Chen J et al (2016) Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: Advances in neural information processing systems
78. Rehman MU, Cho S, Kim J, Chong KT (2021) BrainSeg-Net: brain tumor MR image segmentation via enhanced encoder-decoder network. Diagnostics 11(2):169
79. Yang T, Zhou Y, Li L, Zhu C (2020) DCU-Net: multi-scale U-Net for brain tumor segmentation. J X-ray Sci Technol 28(4):709–726
80. Nickolls John, Dally William J (2010) The GPU computing era. IEEE micro 30(2):56–69
81. Do MN, Vetterli M (2005) The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans Image Process 14(12):2091–2106
82. Polat Kemal, Güneş Salih (2007) Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Appl Math Comput 187(2):1017–1026
83. Lim KY, Mandava R (2018) A multi-phase semi-automatic approach for multisequence brain tumor image segmentation. Expert systems with applications 112:288–300
84. Laukamp KR, Thiele F, Shakirin G, Zopfs D, Faymonville A, Timmer M, Borggrefe J (2019) Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI. Eur Radiol 29(1):124–132
85. Nguyen A, Yosinski J, Clune J (2019) Understanding neural networks via feature visualization: a survey. arXiv preprint arXiv:1904.08939
86. Zhao A et al (2019) Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
87. Goodfellow I et al (2014) Generative adversarial nets. In: Advances in neural information processing systems
88. Goodfellow I (2016) NIPS 2016 tutorial: generative adversarial networks. arXiv preprint arXiv:1701.00160
89. Shin H-C et al (2018) Medical image synthesis for data augmentation and anonymization using generative adversarial networks. International Workshop on Simulation and Synthesis in Medical Imaging. Springer, Cham
90. Elazab A, Wang C, Gardezi SJS, Bai H, Hu Q, Wang T, Lei B (2020) GP-GAN: brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images. Neural Netw 132:321–332
91. Ghassemi N, Shoeibi A, Rouhani M (2020) Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images. Biomed Signal Process Control 57:101678
92. Nema S, Dudhane A, Murala S, Naidu S (2020) RescueNet: an unpaired GAN for brain tumor segmentation. Biomed Signal Process Control 55:101641
93. Cirillo MD, Abramian D, Eklund A (2020) Vox2Vox: 3D-GAN for brain tumour segmentation. arXiv preprint. arXiv:2003.13653
94. Ding Y, Zhang C, Cao M, Wang Y, Chen D, Zhang N, Qin Z (2021) ToStaGAN: an end-to-end two-stage generative adversarial network for brain tumor segmentation. Neurocomputing 462:141–153
95. Sudre CH et al (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 240–248
96. Isensee F, Kickingereder P, Bonekamp D, Bendszus M, Wick W, Schlemmer HP, Maier-Hein K (2017) Brain tumor segmentation using large receptive field deep convolutional neural networks. In: Bildverarbeitung für die Medizin 2017. Springer, Berlin, pp 86-91

97. Li H, Li A, Wang M (2019) A novel end-to-end brain tumor segmentation method using improved fully convolutional networks. Comput Biol Medicine 108:150–160
98. Shahzadi I, Tang TB, Meriadeau F, Quyyum A (2018). CNN-LSTM: cascaded framework for brain tumour classification. In: 2018 IEEE-EMBS conference on biomedical engineering and sciences (IECBES). IEEE, Dec 2018, pp 633–637
99. Law M (2009) Advanced imaging techniques in brain tumors. Cancer Imaging 9, Special issue A:S4
100. Liu J, Li M, Wang J, Wu F, Liu T, Pan Y (2014) A survey of MRI-based brain tumor segmentation methods. Tsinghua Sci Technol 19(6):578–595

IoT Based Machine Learning and Deep Learning Platform for COVID-19 Prevention and Control: A Systematic Review

Muhammad Imad, Adnan Hussain, Muhammad Abul Hassan, Zainab Butt, and Najm Ul Sahar

Abstract The world has been facing a challenging issue of COVID-19, which has infected 223 countries and 223 billion deaths worldwide. It is believed that the world will need to take preventive measures against the COVID-19 pandemic until a vaccine is developed. Early detection of COVID19 infections is a major challenge for healthcare professionals, governments, and organizations to combat against the virus. Therefore, its need an intelligent monitoring system that detects COVID19 and tracks the infected person may improve clinical decision-making and stop spreading the virus among people. The Internet of things (IoT), machine learning, and the Deep learning approach changed our lives in the healthcare sector. This survey presents how IoT, machine learning, and deep learning are incorporated into the pandemic prevention and control system by detection, diagnosis, monitoring, tracing, and social distance finding. We examine and review the most recent literature and present the role of IoT, machine learning, and deep learning in combating the current COVID-19 pandemic. Further, we have also identified a few issues and research directions while using IoT during the COVID-19 pandemic.

Keywords COVID-19 · Internet of Things (IoT) application · Smart healthcare · Machine learning · Deep learning · Coronavirus · Pandemic

1 Introduction

The Internet of Things (IoT) refers to a wireless-based internet-connected object that collects, transfers, and stores data over a defined network without human intervention. All the devices are connected with a particular unique number and code. IoT combines

M. Imad (✉) · M. A. Hassan · N. U. Sahar
Abasyn University, Peshawar, Pakistan

A. Hussain
Islamia College University, Peshawar, Pakistan

Z. Butt
Allama Iqbal Open University, Islamabad, Pakistan

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_26

the network sensors that collect data locally and remotely for Electronic Health management (E-Health) [1]. The E-Health management system manages the data such as heart rate, blood pressure, temperature of the body to monitor the patients when they need a medical consultancy and medication. The E-Heath management system can easily access the location of patients and provide distant medical facilities for emergency purposes [2].

IoT-based healthcare provides an essential role in the medical field to monitor patients' health-related issues during an emergency. The IoT in Medical covered remote health services, telecare, and telemedicine [3].

The global pandemic of COVID-19, that effect worldwide and spread among several countries. The IoT-based services provided monitoring of patients in quarantine and collected the data to stop spreading in this pandemic. The COVID-19 infected patients are isolated in a separate room for better treatment. The IoT, ML/DL based framework systems have improved the performance and diagnosis of COVID-19 infected patients for providing treatment during the emergency stage [4].

According to WHO, COVID-19 is a global pandemic issue with a massive outbreak and infected 223 countries worldwide. People who already have medical issues are affected mostly by COVID-19. The COVID-19 spread through saliva from the nose when the infected patient cough or sneezes. The severe symptoms that need medical attention are difficulty in breathing, chest pain or pressure, and loss of speech. The procedures required to prevent and slow down the spreading of COVID-19 are as follows: Washing your hands, keeping at least a one-meter distance, wearing a facemask, and stop smoking that makes your lungs week [5].

Currently, COVID-19 is having a tremendous influence on the healthcare system and global economy. The spreading of COVID-19 has required an urgent need for technologies to assist governments and healthcare systems in combating the pandemic, minimizing its transmission the infection rate, and developing a method to detect it [6].

According to the World Health Organization (WHO), the total confirms cases, newly reported cases, and the death rate from the top countries are registered. As of 26 June 2021, the USA is the most affected country, with 33,451,965 are confirmed cases, 108,004 are new cases, and a death rate is 601,231 are reported. The other countries are listed in Figs. 1, 2, and 3 [5].

The IoT is generally used in manufacturing, transportation, utility organizations, healthcare industries, agriculture, infrastructure, and home automation industries. Furthermore, IoT technology can access information anywhere using any device [7].

The major contribution of the paper is as follows:

- The role of different IoT-based applications that have been used during the COVID-19 pandemic.
- The recently published paper on machine learning and deep learning to detect, classify, and diagnose COVID-19.
- It also highlights the COVID-19 challenges and future research directions with the help of IoT, ML/DL.

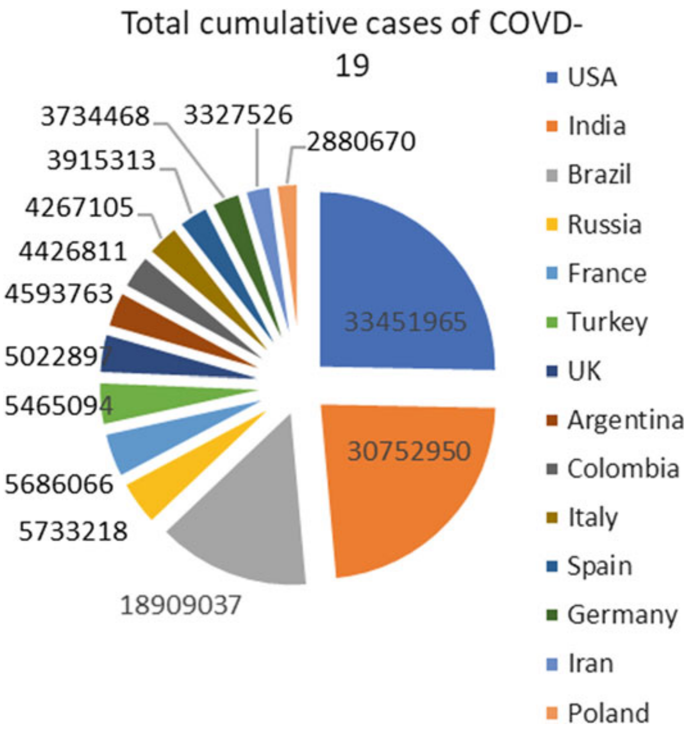


Fig. 1 Total cumulative case of COVID-19

The rest of this paper is structured as follows. Section 2 describes the systematic overview of IoT application, machine learning, and deep learning. Section 3 explains the challenges and research direction related to COVID-19. Finally, the conclusion is discussed in Sect. 4.

2 Literature Review

The previous study related to the COVID-19 pandemic using IoT, ML/DL has been discussed in this chapter.

2.1 IoT in COVID-19

IoT is a new technology that can connect small objects through the network without human interactions. More simply, the small objects can be combined with the internet to monitor and transfer data through IoT devices [8].

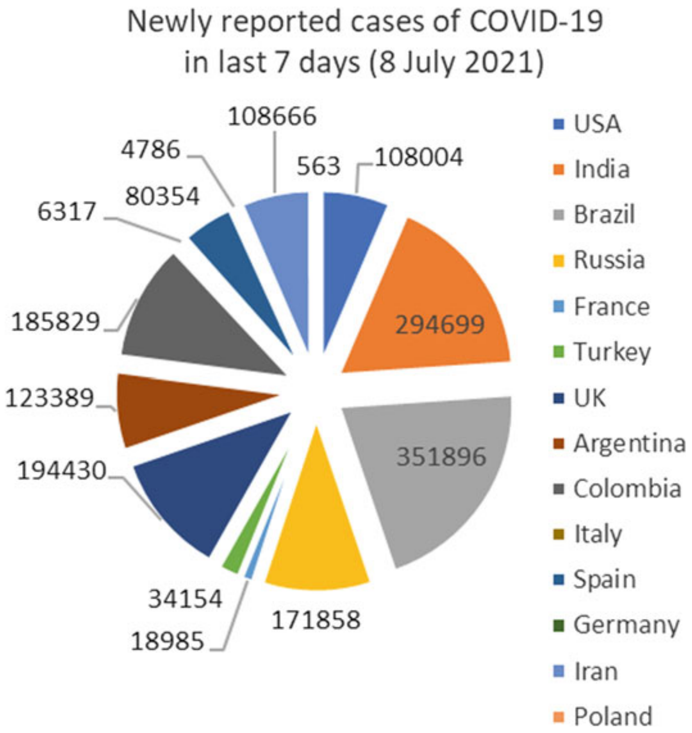


Fig.2 Newly reported cases of COVID-19

In recent years, IoT technology plays a key role in the healthcare industry to identify various infectious diseases. As the contingency of COVID-19 is high in the present pandemic, patients must be linked with and followed by their doctors proactively at different stages of COVID-19 [9]. Figure 2 shows various applications which play a key role in different sectors.

The first phase of COVID-19 is an urgent need for rapid identification due to the high contagiousness of COVID-19, where infected patients can easily transfer the virus to others. The sooner the patient is diagnosed, the better the virus's spread may be halted, and the patient can receive appropriate treatment. The IoT devices help speed up the detection process by collecting data from patients. This can be accomplished by capturing body temperatures with various sensors and collecting samples from the body. The second phase of the disease occurs when the patient has been identified as infected and isolated for treatment. IoT devices can remotely monitor the patients' treatments (Fig. 4) [10].

Furthermore, the numerous IoT devices and apps, such as smart wearables devices, drones, robots, smart glasses, smart helmets, thermometers, and smartphone applications, are used to combat COVID-19. The specifications of these technologies about the pandemic are listed in Table 1.

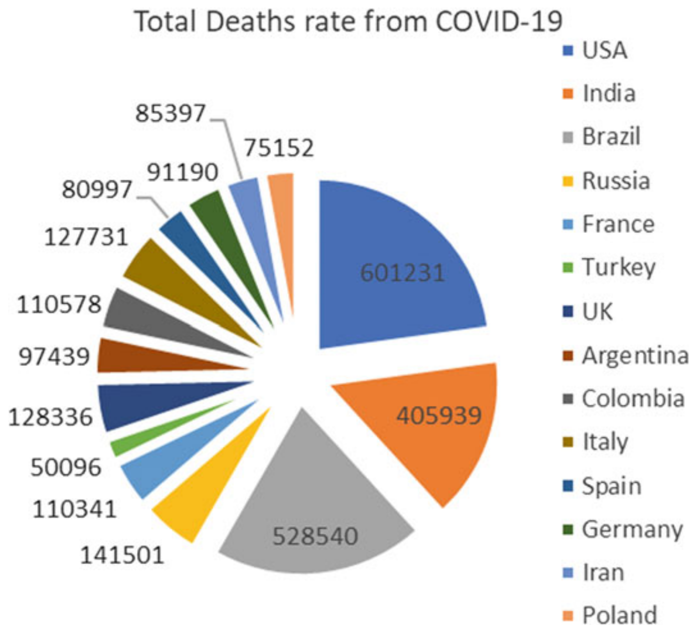


Fig. 3 The total death rates of COVID-19

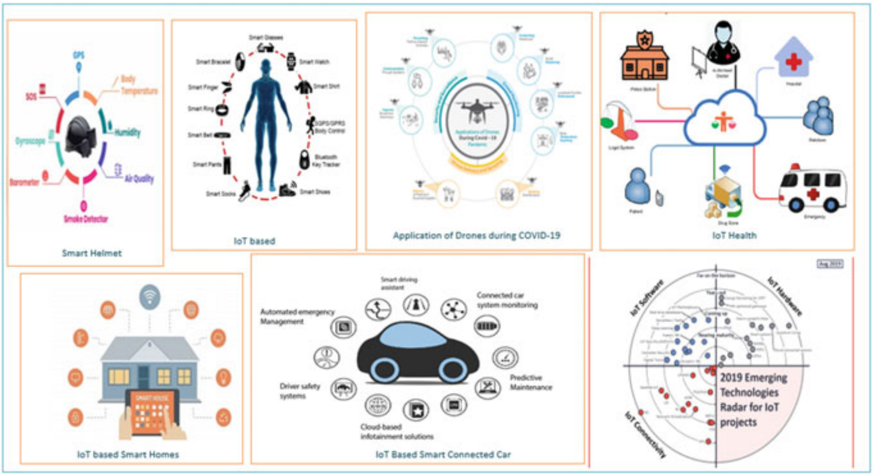


Fig. 4 IoT for healthcare application

Table 1 IoT based smart applications for COVID-19

Reference	Applications	Functions
[11]	IoT based smart helmet	Take capturing of location and face images, temperature monitoring
[12]	IoT based smart glasses	Monitoring the temperature and fewer human
[13]	IoT based smart thermometer	Capturing the temperature in a crowd
[14]	IoT based thermal imaging drones for combat	Capturing the temperature in a crowd
[15]	Surveillance drone	Monitoring and support the police
[16]	Multipurpose drone	Monitoring the crowd, capturing the temperature
[17]	IoT based ambulance, forecasting	Monitoring health and gadgets also find social distancing and forecasting using AI
[18]	Real-time tracking	Prevention, Monitoring, and diagnosing of a patient
[19]	Designed an IoT based wearable quarantine band	Detection, track the person
[20]	Smart sensors base device	Measure body temperature also identified the infected patients
[21]	IoT-based application for COVID-19 diagnosing	Prediction and diagnosing of COVID-19 cases, monitoring and treatment response of COVID-19 infected patients
[22]	Autonomous robots	Detection of the symptoms and monitoring social distancing
[23]	Application for stopping corona	Monitoring health reports and tracking the symptoms
[24]	IoT base easy band	Monitoring and alert the social distancing of people

2.2 *Deep Learning for COVID-19*

To identify the global pandemic, WHO, scientists, and medical doctors search for new technology to identify the various stages, control and diagnose the virus, as well as trace and contact the infected patients. In the last decades, machine learning, deep learning, and AI provide promising results in speed up processing power, reliability and outperform various healthcare tasks [25]. The rapid diagnostic and screening procedure aids in preventing pandemic illnesses such as SARS-CoV-2 is cost-effective and speeds up the associated diagnosis. As a routine technique to supplement traditional diagnosis and screening, the medical expert can use radiological imaging techniques such as X-rays, CT scans, and MRI [26].

In COVID-19, various deep learning and machine learning methods were applied to detect, diagnose, and classify diseases. The CNN (convolutional neural network, LSTM (long short-term memory), and GAN (adversarial network) are used for the detection of COVID-19. The pre-trained models of deep learning such as Visual Geometry Group Network, Residual Neural Network, DenseNet201, InceptionV3, Inception ResNetV2, Xception, and MobileNetV2 [27] deep learning [28, 29] ResNet18, ResNet50, ResNet101, VGG16, VGG19 [30] has been used to solve various problems of COVID-19 which is presented in Table 2.

2.3 Machine Learning for COVID-19

Furthermore, researchers use ML to examine the virus, diagnose patients, assess public health implications, test new remedies, and much more. Machine learning and deep learning have made rapid progress in clinical research, biomedical detection, precision medicine, and medical diagnostics. Such tools can provide new probabilities for patients, clinicians, and researchers, allowing them to make more informed decisions and achieve better outcomes. Currently, as the pandemic spreads due to the spread of the COVID-19, institutions, business developers, and medical providers are considering Machine learning and Deep Learning ways to combat the risk of this current disaster [37].

Many solutions in the medical area have been built using ML and DL approaches, as mention in Table 3 and, they are currently being used to combat COVID-19. The AI community is working hard to distribute apps that can help contain the effects of the pandemic using DL methods. While Table 3 presents the various diagnostic technique which used during COVID-19 pandemic for the detection and classification (Fig. 5).

3 Discussion and Future Work

In this review, the author thoroughly analyzed recent research literature and examined reports supporting the key role of machine learning, deep learning and IoT applications during COVID-19. However, numerous organizations, industrial communities, and research labs use various technologies to detect and trace the virus to mitigate its effects. IoT, ML, and DL have shown promising results during quarantine time and recovery from COVID-19.

Different diagnosing and detection technique has been implemented to control and stop the spreading of COVID-19 pandemic. However, in IoT, there will be more research and development need for privacy and security concerns. Also, more application needs to identify the COVID-19 virus and track the patients. Collecting patients' information and providing the prescription during an emergency case to minimize the lifetime risk of the patient's life. Also, provide the location to the patients and find the nearest hospital.

The X-ray, CT, and MRI images give optimal results in distinguishing COVID-19 from other types of virus diseases. However, it must be more efficient in order

Table 2 Covid-19 detection and classification using deep learning technique

Reference	Methodology used	Results	Classes	Dataset	Data type
[27]	Visual geometry group network, residual neural network, DenseNet201, InceptionV3, Inception ResNetV2, Xception and MobileNetV2	VGG19 = 90% ResNetV2 = 70% DenseNet201 = 90% InceptionV3 = 50% Inception Res NetV2 = 80% Xception = 80% MobileNetV2 = 60%	2-Classes COVID-19, normal	GitHub repository	X-ray images
[28]	InceptionV3, ResNet50 and Inception-ResNetV2	InceptionV3 = 97% ResNet50 = 98% Inception-ResNetV2 = 87%	4-Classes COVID-19, normal, viral-pneumonia, bacterial-pneumonia	GitHub repository	X-ray images
[30]	ResNet18, ResNet50, SqueezNet and DenseNet-121	ResNet18 = 98% ResNet50 = 96% SqueezNet = 98% DenseNet-121 = 96%	2-Classes COVID-19, normal	ChexPertdatase dataset	X-ray images
[31]	Alex Net, Squeez Net, Google Net, VGG, Mobile Net, ResNet18, ResNet50, ResNet101 and Dense Net	AlexNet = 96% Squeez Net = 98% GoogleNet = 98% VGG = 98% Mobile Net = 96% ResNet18 = 98% ResNet50 = 98% ResNet101 = 98% Dense Net = 98%	4-Classes COVID-19, viral-pneumonia, bacterial-pneumonia, normal	GitHub repository	CT-images
[32]	DeCoVNet	DeCoVNet = 95%	2-Classes Covid-19, normal	Dataset is collected from HUST china	CT-images

(continued)

Table 2 (continued)

Reference	Methodology used	Results	Classes	Dataset	Data type
[33]	COVID-NET and COVID ResNet	COVID-NET = 83% COVID ResNet = 96%	4-Classes COVID-19, normal, bacterial-pneumonia, viral-pneumonia	GitHub repository	chest X-rays
[34]	COVID-CAPS without pre-training and pre-trained COVID-CAPS	COVID-CAPS without pre-training = 95.7% Pre-trained COVID-CAPS = 98.3%	4-classes normal, bacterial-pneumonia Viral-pneumonia, COVID-19	GitHub repository	X-ray images
[35]	VGG19-CNN, ResNet152V2	VGG19 + CNN model = 98.05% of accuracy	4-Classes normal, COVID-19, pneumonia, and lung cancer	GitHub repository	X-ray images
[36]	Neural network that is a concatenation of the Xception and ResNet50V2 networks	The average accuracy COVID-19 cases are 99.50%, and for all classes is 91.4%	3-Classes Normal, pneumonia, and COVID-19	GitHub repository	X-ray images

Table 3 Machine learning for COVID-19 detection, classification

Author	Methodology	Result	Classes	Dataset	Data type
[38]	SVR, simple LR and PR	SVM provides a better result	Active cases, recovery, and death cases region and country wise	Data is collected from different country detail	Country and region-wise data
[39]	LR, multinomial NB, SVM and DT	Logistic regression and multinomial Naïve Bayes = 92% accuracy	4-Classes COVID, SARS, ARDS COVID, ARDS	GitHub repository	Clinical reports
[40]	SVR, DNN, LSTM and PR	PR yielded a minimum root mean square error	Conform, death and recovery cases	Dataset has been downloaded from the Johns Hopkins dashboard	Country-wise data from hospital
[29]	GLCM, LDP, GLRLM, GLSZM, DWT, and SVM	GLSZM 99.68% accuracy with tenfold cross-validation	2-Classes Covid Normal	Dataset download from Italian Society of Radiology Center	CT images
[41]	Machine learning hybrid approach	Accuracy = 0.84	2-Classes COVID positive COVID negative	Hospitalized patients with COVID-19	Hospitalized patients' data with COVID-19, pneumonia
[42]	RIDGE, RF, XGBoost and LASSO, RM	XGBoost model = 92.5% sensitivity and 97.9% specificity	3-Classes COVID-19, influenza patients, normal	GitHub repository	Patient based clinical data
[43]	XGBoost, GBM, SVM, Random Forest, and Decision Tree	XGBoost algorithm performed with the highest accuracy (>85%)	2-Classes Covid Normal	GitHub repository	History of patients including travel and clinical details
[44]	LR, multilayer perceptron, and vector autoregression	95% accuracy with LR	Confirmed, death and recovered cases in India	Kaggle repository	Hospital data

(continued)

Table 3 (continued)

Author	Methodology	Result	Classes	Dataset	Data type
[45]	SVM, DT, NB, KNN and RF	SVM = 96% DT = 82%, Naive Bayes = 90%, k-Nearest Neighbors and Random Forest = 92%	2-Classes Covid Normal	Kaggle repository	X-ray images
[46]	Support vector machine and logistic regression	SVM = 96% LG = 92%	2-Classes Covid Normal	Kaggle repository	X-ray images
[47]	MRFO, MRFODE, HHO HGSO, WOA, SCA and GWO	98.09% Accuracy	2-Classes Covid Normal	Kaggle repository	X-ray images
[48]	ANN, SVM, RBF, k-NN, RF, DT and CNN and Deep Learning	SVM = 95% RBF = 94%	2-Classes Covid Normal	GitHub and Kaggle's repository	X-ray images
[49]	Logistic, linear, logarithmic, quadratic, cubic, compound, power, exponential, GA, PSO, and GWO	The logistic model provides a better result	Total cases, cumulative statistics	Data downloaded from worldmeter site	Italy, Germany, Iran, USA, and China

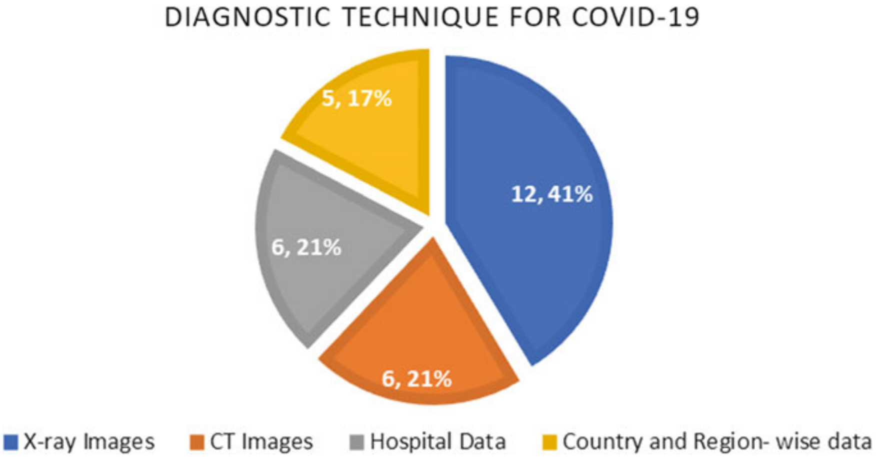


Fig. 5 Various diagnostic technique for COVID-19

to properly detect the feature from noisy, ambiguous, or incorrect datasets, which might lead to misclassification. Moreover, the radiology images of ultrasound and MRI are limited to detect the COVID-19 patients. There is still a need to focus on categorizing COVID-19 with various disease symptoms such as fever, cough, sore throat, diarrhea, and breathing to identify the chronic patient's issue by utilizing modern machine learning and deep learning approaches. Finally, the article provides information about the importance of IoT, ML and DL during COVID-19.

4 Conclusion

The global outbreak of COVID-19 has affected millions of lives around the world. The current era of advanced technology with IoT, DL and ML has improved numerous medical aspects of human existence and enhanced the detection of chronic and contagious diseases. The healthcare sector has been enhanced, and clinical decisions have been made more effectively due to the Internet of Things (IoT). Although IoT-based technology can revolutionize the way we live beyond COVID-19, as discussed in the paper, it requires further research and validation before widespread adoption and deployment. This review discusses the IoT, machine learning, and deep learning techniques to predict, classify COVID-19. Several researchers have used X-ray, CT, MRI, and clinical data to diagnosed COVID-19. Based on systematic reviews, it shows that the deep learning model achieved the best performance results in every domain, along with medical research and radiology. Finally, it can be concluded that the IoT applications, Machine learning, and Deep Learning approaches played a vital role in predicting, categorizing, screening, and limiting the spread of the COVID-19 pandemic.

References

1. What is IoT (Internet of Things) and How Does it Work? IoT Agenda, 2021. Available: https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT
2. Ndiaye M, Oyewobi SS, Abu-Mahfouz AM, Hancke GP, Kurien AM, Djouani K (2020) IoT in the wake of COVID-19: a survey on contributions, challenges and evolution. IEEE Access 8:186821–186839
3. Sharma N et al (2021) A smart ontology-based IoT framework for remote patient monitoring. Biomed Sig Process Control 68:102717
4. Kallel A, Rekik M, Khemakhem M (2021) Hybrid-based Framework for COVID-19 prediction via federated machine learning models
5. Coronavirus, Who.int, 2021. Available: https://www.who.int/health-topics/coronavirus#tab=tab_2
6. Alqahtani M (2021) IOT within the Saudi Healthcare Industry during Covid-19. EasyChair 2516–2314
7. What is the Internet of Things (IoT)?. (2021). Retrieved 5 November 2021, from https://www.oracle.com/internet-ofthings/what-is-iot/

8. Haddad Pajouh H, Dehghan Tanha A, Parizi RM, Aledhari M, Karimipour H (2019) A survey on internet of things security: requirements, challenges, and solutions. Internet of Things 100129
9. Phelan AL, Katz R, Gostin LO (2020) The novel coronavirus originating in Wuhan, China: challenges for global health governance. JAMA 323(8):709–710
10. Rahman MS, Peeri NC, Shrestha N, Zaki R, Haque U, Ab Hamid SH (2020) Defending against the Novel Coronavirus (COVID-19) outbreak: how can the Internet of Things (IoT) help to save the world? Health Policy Technol 9(2):136
11. S. Kurkute, N. Ahirao, R. Ankad, and V. Khatal, "IOT based smart system for the Helmet detection. In: Proceedings of international conference on sustainable computing in science, technology and management (SUSCOM), Amity University Rajasthan, Jaipur-India
12. Mohammed M et al (2019) Novel coronavirus disease (Covid-19): detection and diagnosis system using IoT based smart glasses. Pesquisa.bvsalud.org. Available: https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/covidwho-828827
13. Chamberlain SD, Singh I, Ariza CA, Daitch AL, Philips PB, Dalziel BD (2020) Real-time detection of COVID-19 epicenters within the United States using a network of smart thermometers. MedRxiv
14. How drones are being used to combat COVID-19. Geospatial World, 2021. Available: https://www.geospatialworld.net/blogs/how-drones-are-being-used-to-combat-covid-19
15. Sharma K (2021) Cyient provides drone-based surveillance technology to support Telangana state police in implementing COVID-19 lockdown. Cyient.com. (2021). Available: https://www.cyient.com/prlisting/corporate/cyient-provides-drone-basedsurveillance-technology-to-support-telangana-state-police-in-implementing-covid-19-lockdown
16. Delhi Civic Body Begins Thermal Screening People on Balconies With Drones. NDTV.com, 2021. Available: https://www.ndtv.com/delhi-news/coronavirus-delhi-civic-body-using-drones-to-check-temperature-of-people-on-balconies-2209832
17. Kamal M, Aljohani A, Alanazi E (2020) IoT meets COVID-19: status, challenges, and opportunities. arXiv preprint arXiv:2007.12268, 2020
18. Swayamsiddha S, Mohanty C (2020) Application of cognitive Internet of Medical Things for COVID-19 pandemic. Diab Metabol Syndr Clin Res Rev 14(5):911–915
19. Singh V, Chandna H, Kumar A, Kumar S, Upadhyay N, Utkarsh K (2020) IoT-Q-band: a low cost internet of things based wearable band to detect and track absconding COVID-19 quarantine subjects. EAI Endorsed Trans Internet of Things 6(21)
20. Kumar K, Kumar N, Shah R (2020) Role of IoT to avoid spreading of COVID-19. Int J Intell Netw 1:32–35
21. Otoom M, Otoum N, Alzubaidi MA, Etoom Y, Banihani R (2020) An IoT-based framework for early identification and monitoring of COVID-19 cases. Biomed Signal Process Control 62:102149
22. Autonomous robots are helping kill coronavirus in hospitals. IEEE Spectrum. Available: https://spectrum.ieee.org/autonomous-robots-are-helping-kill-coronavirus-inhospitals#toggle-gdpr
23. GIS cloud offers technology and support for free to projects fighting to contain coronavirus|GIS cloud. GIS Cloud. Available: https://www.giscloud.com/blog/gis-cloud-offers-technology-and-support-for-free-to-projects-fighting-to-contain-coronavirus/
24. Foresman B (2021) This wristband vibrates if you break social distancing rules|EdScoop. EdScoop. Available: https://edscoop.com/university-florida-social-distancing-wristband/
25. Davenport T, Kalakota R (2019) The potential for artificial intelligence in healthcare. Fut Healthcare J 6(2):94
26. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A (2020) Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. Comput Biol Med 121:103795
27. Hemdan EE-D, Shouman MA, Karar ME (2020) Covidx-net: a framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055
28. Narin A, Kaya C, Pamuk Z (2021) Automatic detection of coronavirus disease (covid-19) using X-ray images and deep convolutional neural networks. Pattern Anal Appl 1–14

29. Barstugan M, Ozkaya U, Ozturk S (2020) Coronavirus (covid-19) classification using CT images by machine learning methods. arXiv preprint arXiv:2003.09424
30. Minaee S, Kafieh R, Sonka M, Yazdani S, Soufi GJ (2020) Deep-covid: predicting covid-19 from chest x-ray images using deep transfer learning. Med Image Anal 65:101794
31. Rehman A, Naz S, Khan A, Zaib A, Razzak I (2020) Improving coronavirus (COVID-19) diagnosis using deep transfer learning. MedRxiv
32. Zheng C et al (2020) Deep learning-based detection for COVID-19 from chest CT using weak label. MedRxiv
33. Farooq M, Hafeez A (2020) Covid-resnet: a deep learning framework for screening of covid19 from radiographs. arXiv preprint arXiv:2003.14395
34. Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A (2020) Covid-caps: a capsule network-based framework for identification of covid-19 cases from x-ray images. Pattern Recogn Lett 138:638–643
35. Ibrahim DM, Elshennawy NM, Sarhan AM (2021) Deep-chest: multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. Comput Biol Med 132:104348
36. Rahimzadeh M, Attar A (2020) A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. Inf Med Unlocked 19:100360
37. Xu X et al (2020) A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering 6(10):1122–1129
38. Yadav M, Perumal M, Srinivas M (2020) Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos Solit Fract 139:110050
39. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Din MMU (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12(3):731–739
40. Punn NS, Sonbhadra SK, Agarwal S (2020) COVID-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv
41. Ferrari D et al (2020) Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia—challenges, strengths, and opportunities in a global health emergency. PloS One 15(11):e0239172
42. Li WT et al (2020) Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med Inform Decis Mak 20(1):1–13
43. Ahamad MM et al (2020) A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Exp Syst Appl 160:113661
44. Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Env Res Risk Assess 34:959–972
45. Khan MIN, Ullah F, Hassan MA, Hussain A (2020) COVID-19 classification based on Chest X-Ray images using machine learning techniques. J Comput Sci Technol Stud 2(2):1–11
46. Ullah SI, Salam A, Ullah W, Imad M (2021) COVID-19 lung image classification based on logistic regression and support vector machine. In: European, Asian, Middle Eastern, North African conference on management & information systems. Springer, Berlin, pp 13–23
47. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT (2020) New machine learning method for image-based diagnosis of COVID-19. Plos One 15(6):e0235187
48. Abed M et al (2021) A comprehensive investigation of machine learning feature extraction and classification methods for automated diagnosis of Covid-19 based on X-ray images. Comput Mater Continua 3289–3310
49. Ardabili SF et al (2020) Covid-19 outbreak prediction with machine learning. Algorithms 13(10):249

Oncology with Artificial Intelligence: Classification of Cancer Using Deep Learning Techniques

S. Mala, B. Nagarajan, G. Sangeetha, and J. Suganthi

Abstract Diagnosing cancer in the earliest stage has the chance of remedy since Cancer is a genetic disease and it is caused by changes to genes. We have numerous technologies for detecting cancer. Artificial Intelligence is one of the technologies to observe cancer cells and it is highly energized by its structure and function of brain named Artificial Neural Networks (ANN). Artificial Intelligence is a clone of Human intelligence processes by machines particularly computers. Multilayer Perceptrons, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are the famous Neural Networks (NN) in Deep learning. National Cancer Institute has listed all kinds of cancer. Some of them can be discovered using Artificial Neural Network mechanism. This type of cancer and the mechanism of deep Artificial Neural Network to perceive it are discussed below in this article.

Keywords Breast cancer · Brain cancer · Skin cancer · Artificial neural network · Recurrent neural network · Convolutional neural network

1 Introduction

Artificial Intelligence (AI) is a prolific technology for analyzing complex medical images and draw out the relationship among the huge data set. Computer aided

S. Mala (✉)
Department of Computer Applications, Madurai Kamaraj University College, Madurai, Tamil Nadu, India

B. Nagarajan
Arunai Engineering College, Tiruvannamalai, Tamil Nadu 606603, India

G. Sangeetha
Department of Computer Science and Information Technology, S.S. Duraisamy Nadar Mariammal College, Kovilpatti, Tamil Nadu, India

J. Suganthi
Department of Information Science and Engineering, T. John Institute of Technology, Bengaluru, India

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_27

cancer diagnosis methods are highly accurate. Cancer is one of the primary reasons of death over the past decades. Cancer is referred as a condition in which there is abnormal mutational growth and spread of the cells towards other parts of the body. The cancerous cells are classified as malignant cells and benign cells as depicted in Fig. 1. There are various kinds; breast cancer, melanoma, lung cancer, colon cancer, leukemia, liver cancer, prostate cancer, colorectal cancer, Non-Hodgkin's lymphoma, etc. [1].

There has been a continuous evolution in the cancer research over the previous decades [2]. Researchers applied various screening techniques to facilitate the early diagnosis of the cancer that reduces the mortality rate. Earlier, manual examination of the patient images is used for the cancer diagnosis. But, it required more time consumption and highly prone to the diagnosis errors. CAD systems and AI techniques are used to assist the healthcare practitioners in the interpretation of the medical images [3].

With the rise of the most recent computerization advances, gigantic measure of malignancy information could be gathered and made accessible to the clinical examination local area. However, accurate classification and prediction of the cancer is a challenging task for the diagnostic experts. Machine Learning (ML) Techniques play a vital role in the accurate classification and prognosis of the cancer. ML is an application of AI that enables automatic learning of the input data and cancer classification without requiring detailed programming. This learning process starts with the observation of the input patient data to establish a certain pattern in the data and to make reliable decisions in the future. The main aim of the ML techniques is to allow the cancer diagnosis systems to learn automatically and provide better results without requiring human intervention. ML carries a major role in the effective analysis of massive amounts of data. Use of these ML techniques resulted in the generation of quicker detection result and eradication of the human errors. It also facilitated in the classification of the cancer patients into high or low risk groups.

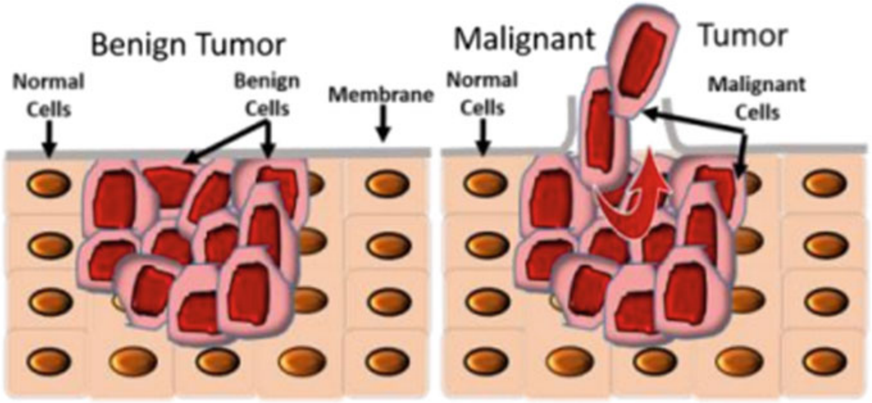


Fig. 1 Benign and malignant tumor

Deep learning [4] is a part of the ML that could learn the input data and make intelligent decisions on its own. Deep learning techniques formulate top quality representation from the input descriptions directly [5].

2 Artificial Neural Network (ANN)

Neural Networks (NNs) represent network of neurons that mimic the operations of a human brain for recognizing the relationship between huge amounts of data. ANN is an interconnected group of artificial neurons as depicted in Fig. 2, with the self-learning capability that produces better classification results [5]. ANN is addressed as a weighted coordinated chart where the fake neurons structure the hubs. The relationship between the info and yield of the neurons is seen as the coordinated edges with loads. ANN gets the information signal as vector. Then, at that point, each information is increased by its comparing loads that signify the interconnection strength between the neurons. The weighted information sources are summed up. In the event that the weighted aggregate is equivalent to nothing, inclination is added to work on the yield reaction. The total of the weighted input falls in the range of 0 to $+\infty$. Back propagation algorithm is an important way to train the multilayer ANNs. In this method, the error gradient corresponding to the weights for a given input is calculated through the propagation of error from the output layer to the hidden layer and input layer.

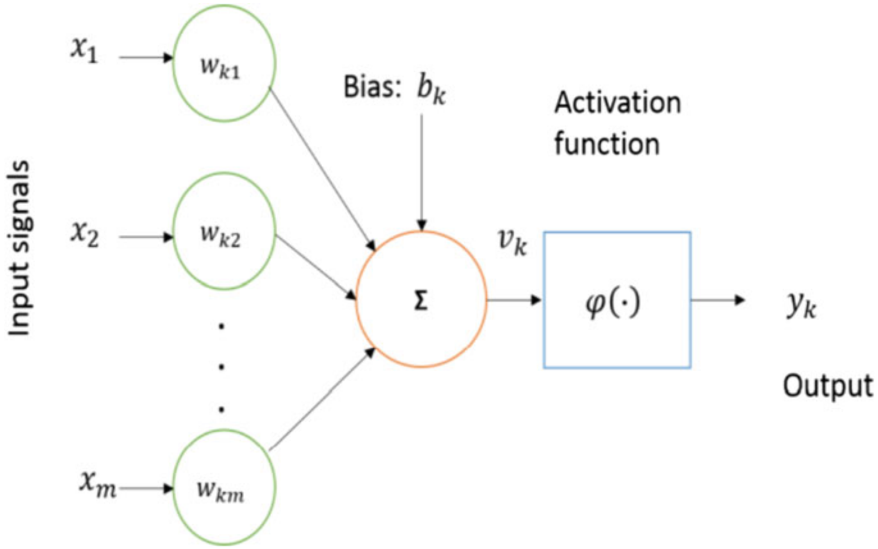


Fig. 2 Basic model of neuron

Deep NNs are trained by using Gradient Descent (GD), Stochastic GD (SGD) and Back propagation [6].

2.1 *Types of ANN*

The ANNs are classified as various types based on the neuron of human brain [7] and network functions. Most of the ANNs share similarities and effective at the classification and segmentation tasks [5].

2.1.1 Feedback ANN

Feedback ANN is also referred as recurrent network. In this type of network, the output is applied back to the network for achieving best results. Being a dynamic network, it is best suited for solving the optimization problems. There is bidirectional flow of information in the feedback ANN. This network is in transient state until an equilibrium point is reached. Until there is a new input, it will remain in the equilibrium point. When a new input is received, there is a need to find a new equilibrium point. Because of the feedback paths, the inputs to each neuron are modified frequently, which leads the network to enter a new state every time when a new input pattern is applied.

2.1.2 Feed Forward ANN

A feed-forward ANN comprises an input layer, an output layer and either a single or multiple layers of neurons. By evaluating the output through the review of the input, the network power could be identified based on the collective behavior of the connected neurons and the network output could be decided. In this network, the information flow is unidirectional i.e. from the input layer to the output layer. Hence, there is no feedback or loops and the inputs and outputs are fixed. It is a straight forward network that associates the inputs with the outputs. Feed forward ANNs are used in the pattern generation, recognition and classification. The Feed-forward network is static, such that it produces a single set of output values rather than a sequence of output values from a given input. The feed-forward networks are devoid of memory so that the response to an input is independent of the previous network state.

2.1.3 Convolutional Neural Network (CNN)

Fukushima [8] initially proposed a hierarchical network structure for visual object recognition. In any case, it was not utilized because of the enormous equipment necessities required for preparing the organization. LeCun et al. [9] applied slope based figuring out how to the CNNs and got fruitful grouping of transcribed digit. CNN yields better advantages than the DNNs and produce highly optimized weights. Two primary sections of the CNN engineering are including extractors and classifier. Each element extraction layer gets the yield of the past layer and passes its yield as contribution to the ensuing layer.

CNN is a feed-forward NN wherein the information signal is prepared straightforwardly without requiring any circles or cycles. The CNN includes three sorts of layers: convolutional layers, pooling layers and completely associated layers as displayed in Fig. 3. The yield hubs of the convolutional layer and max pooling layers are joined together into highlight planning. Blend of a solitary plane or various planes of the first layers frames each plane of a layer. The hubs of a plane are associated with a little locale in each associated plane of the past layer. Through the convolution procedure on the information hubs, every hub of the convolutional layer performs extraction of components from the info pictures.

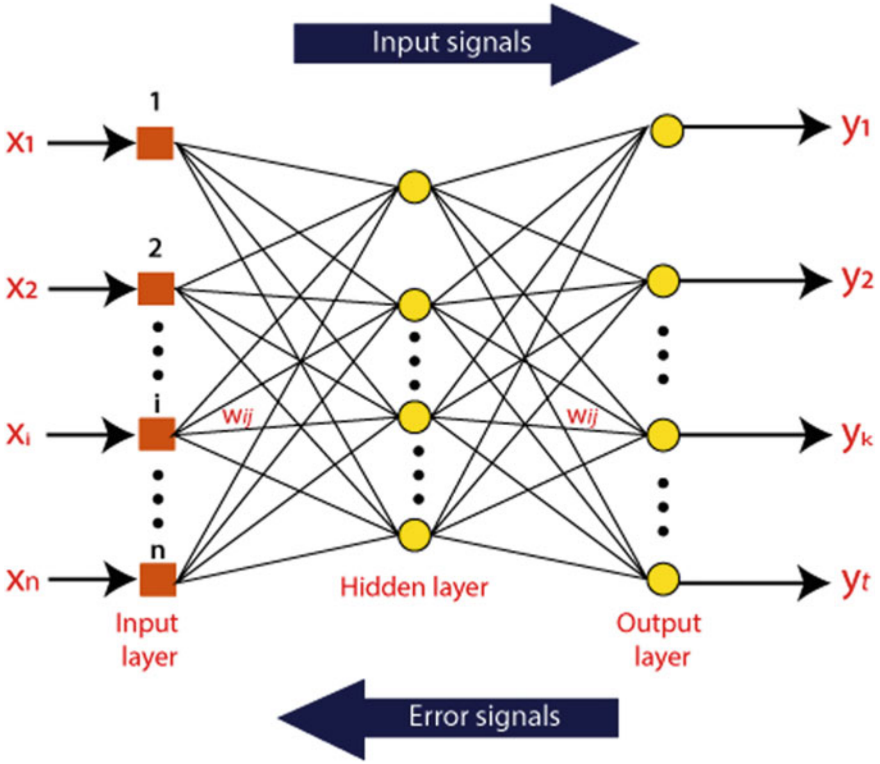


Fig. 3 Artificial neural network

More elevated level provisions are gotten from the components moved from the lower-level layers. There is event of dimensionality decrease because of the component proliferation to the higher layer, contingent upon the size of part utilized for the convolutional and max-pooling activities.

The information layer holds the pixel upsides of picture. The convolutional layers decide the yield of the fake neurons of which are associated with the neighborhood areas of the contribution by ascertaining the scalar item between their loads and the district associated with the info volume. The amended direct unit applies sigmoid enactment capacity to the yield of the actuation delivered by the past layer. The pooling layer performs down examining alongside the spatial dimensionality of the given contribution, through the further decrease in the quantity of boundaries inside that initiation. The completely associated layers produce class scores from the initiations to be utilized for classification.

A convolutional organization could effectively catch the Spatial and Temporal conditions in a picture by utilizing the significant channels. It performs better fitting to the picture dataset because of the critical decrease in the quantity of boundaries and reusability of loads. All in all, the organization could be prepared for better comprehension of the picture complexity (Fig. 4).

Convolutional layer

Kernel is concerned in convolutional operation. In a 3X3X1 framework, the piece shifts multiple times, while playing out the lattice duplication between the part and a piece of the picture on which the bit drifts around. The channel moves to the right side with a particular step esteem, till the channel parses the total width. Further, the channel jumps down to the left half of the picture with a similar step esteem. This interaction is rehashed constantly until the whole picture is navigated. For the pictures with different channels the profundity of the Kernel is same as the information picture. Lattice augmentation is performed and all the increase results are summarized along

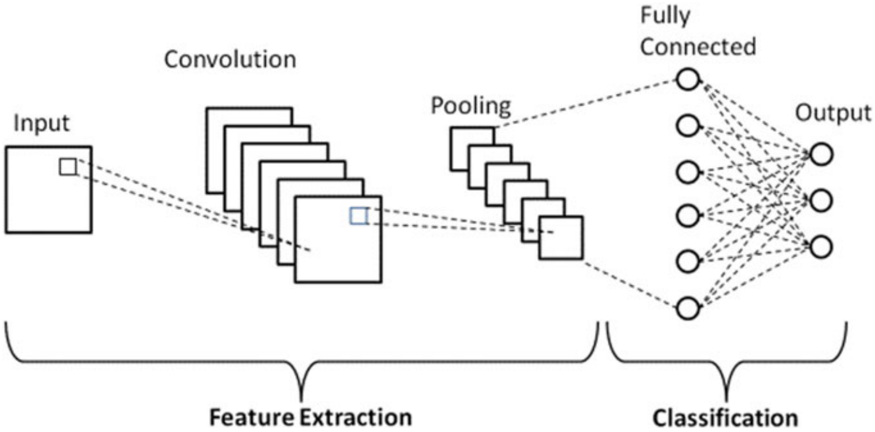


Fig. 4 Convolutional neural netowork

with the inclination for giving the tangled component yield. The primary goal of the convolution activity is to separate the significant level provisions. The consequences of this activity are either decrease or expansion in the dimensionality of the convolved highlight.

Pooling layer

The pooling layer is utilized to lessen the spatial size of the convolved highlight. Thus, the computational power needed for data processing through dimensionality reduction is reduced. It is highly useful to extract the dominant features that are invariant to the rotation and position. Thus, the effective training of the model is achieved. Pooling is classified as Max Pooling and Average Pooling. Max pooling returns the maximum value from the portion of the image covered by the Kernel. Average Pooling returns the average of all the values from the portion of the image covered by the Kernel.

Average pooling simply reduces the dimensionality as a noise suppression process, while the Max Pooling completely discards the noisy activations and removes the noise along with the dimensionality reduction. Hence, Max Pooling performs better than the Average Pooling.

Fully-connected (FC) layer

It is the economical way of learning the non-linear combinations of the high-level features represented by the output of the convolutional layer. The main purpose of the fully-connected layer is to combine the features into several attributes that facilitate better classification. The fully connected layer has the ability to mix signals as every neuron in a layer has connectivity with every neuron in the subsequent layer. Hence there is an information flow among each input and each output attributes.

Softmax layer

The output of the fully-connected layer is applied to the softmax layer. It consists of same number of nodes as the output layer. When summed up, the decimal probabilities add up to 1.0. Softmax has two variants: Full Softmax calculates the probability for every possible class and candidate sampling which calculates the probability for all the positive labels. It is otherwise called the multi-class logistic regression or soft-argmax function, because it is generalization of logistic regression which is applied for multi-class classification, and used for logistic regression as its formula is similar to the sigmoid function. The softmax architecture comprises of four units: (1) The input data are mapped in the pre-processing unit, (2) Calculation part calculates the exponential function, (3) Probability output part focuses on calculating logarithmic function and (4) Control part coordinates all the three parts of the architecture [10].

2.1.4 AlexNet

Alex et al. [9] proposed a more profound CNN model when contrasted with LeNet, for the visual acknowledgment of the article. Alexnet-a CNN has an enormous effect in the uses of profound learning. Its engineering has eight layers with five convolutional layers and three completely associated layers. It permits multi-GPU preparing that utilizes half of the neurons on one GPU and the lay on the other GPU. Alexnet is actually an amazing model with high exactness on different testing datasets yet evacuation of any traditional layer will radically corrupt the presentation. AlexNet is a prevailing engineering incase of item identification assignments and discovers applications in AI issues. With the SoftMax layer toward the end, Alexnet's engineering has two completely associated layers with dropout. Neighborhood Response Normalization (LRN) and dropout, two new ideas, are expressed as such. LRN can be done twoly: first applying on single channel browsed same component map and the Second LRN can be applied across the channels.

In the AlexNet engineering, there are five convolutional layers, two FC layers and delicate max layer. The info information is applied to the first convolutional layer that performs convolution activity alongside max pooling and Local Response Normalization (LRN). 3 × 3 channels are taken on for the maximum pooling tasks with a step size of 2. 5 × 5 channels are used for the maximum pooling activity in the second convolutional layer. For the maximum pooling tasks in the third, fourth and fifth convolutional layers, 3 × 3 channels are utilized alongside 384, 384 and 296 component maps correspondingly. The Two FC layers are utilized followed by the softmax at the last stage. LRN is applied on the component guides or single channel or across the channels.

2.1.5 VGGNet

VGG (Visual Geometry Group) Net depicts that the network's depth is a vital component for classification accuracy. The design comprises of a few convolutional layers that utilizes the ReLU enactment work followed by a solitary max pooling layer and different completely associated layers that utilizes a similar initiation work. The last layer of the model is a SoftMax layer for order. Characterized by its simplicity, the depth is increased by 3 × 3 convolutional layers piled up on the top of each other. Volume size is reduced by max pooling. Two fully-connected layers, with 4096 nodes each are then followed by a SoftMax classifier. Despite of using VGGNet in various image classification problems in deep learning, it has two major inconveniences that it is quite slow and tiresome to train the network and as the weight of the network architecture is quite large regarding the bandwidth.

2.1.6 CapsuleNet/CapsNet

CapsNet is made of capsules which are a group of neurons with the activity vector representing the parameters of a particular type of entity like an object or an object part [11]. Active capsules at one level decides predictions, for the representation parameters of higher-level capsules. A higher-level capsule gets activated, when multiple predictions are agreed. The architecture is very simple with only two convolutional layers and a fully connected layer. Capsules can deal with numerous different transformations of various objects or object parts simultaneously.

2.1.7 GoogleNet

The target of GoogleNet is to decrease the intricacy of calculation when contrasted and the customary CNN. The commencement layers that had fluctuating responsive fields, made by portions of various sizes. These open fields made tasks that caught inadequate relationship designs in the new component map stack. GoogleNet utilizes a heap of beginning layers to ad lib the cutting edge acknowledgment exactness. It thoroughly comprises of 22 layers, which was significantly more noteworthy than some other organization. All things considered, Google Net uses lesser number of organization boundaries and computations as compared to AlexNet or VGG. The GoogleNet Architecture is deeper with 22 layers, 27 pooling layers and linearly stacked 9 inception modules. The ends of the inception modules are connected to the pooling layer. GoogleNet is trained with the help of distributed machine learning systems with a moderate amount of model and data parallelism.

A CNN based on GoogleNet using Median Intensity Projections (MIP) extracts accurate information from CT scans hence finds application in detecting lung cancers. For faster convergence and higher accuracy, it employs transfer learning method.

2.1.8 ViPCNN

In Visual Phrase directed Convolutional Neural Network (ViPCNN), each visual relationship is accepted as an expression of three parts. Visual expression discovery is the task of limiting an expression (subject-predicate-object), where predicate clarifies the connection between the subject and item, it includes recognizing and confining sets of associated objects in a picture and furthermore classify the association between them. For distinguishing the visual relationship, two normal methodologies are utilized. The base up plan distinguishes items and afterward recognize the potential associations/cooperations between every one of them, the hierarchical plan all the while distinguishes the subject-predicate-object state by thinking about the relationship as an incorporated entirety. VGG Net is utilized as the essential structure square of ViPCNN that partitions the plan method into two sections: trio proposition and stage acknowledgment. Celebrity CNN utilizes the whole picture as information that is taken care of into a few convolutional and max-pooling layers to make the component map. The organization is then parceled into four parts of which one is allotted for trio proposition and the remainder of the three were allocated for express location [12].

2.1.9 LCNN

A Lookup-based CNN (LCNN) encodes convolutions by couple of queries to a word reference that is prepared to encase the space of loads in CNNs. It offers considerable speed at inference and efficient training. LCNN dictionaries are architecture independent so they can be transferred across the layers which let us to train a dictionary using a shallow network and then transfer it to a deeper one [13].

SqueezeNet

SqueezeNet is a convolutional neural organization that is 18 layers profound. Fire module is the structure square of squeezeNet. Three significant methodologies are followed here: 1. 3 × 3 channels are supplanted by 1 × 1 channels, 2. Number of contributions to 3 × 3 channels are decreased by crushing layers, 3. Postponed down testing to have enormous enactment maps [11]. The initial two methodologies safeguard precision and lessen the quantity of boundaries in a CNN while the last methodology boosts the exactness. A Fire module comprises of a crush convolution layer having just 1 × 1 channels and a grow layer that has a mix of 1 × 1 and 3 × 3 convolution channels.

ZFNet

ZFNet (Zeiler and Fergus network) is a standard convolutional neural organization planned by envisioning transitional element layers and the activity of the classifier. Contrasted with other CNN organizations, it has diminished channel size and the convolution steps are decreased. The info, a RGB picture to ZFNet was quantized to 224 × 224 pixels which were handled by the organization that outcomes in the likelihood that the picture has a place with each class. The organization comprises of five shareable convolutional layers, max-pooling layers, dropout layers, and three completely associated layers. It utilizes a 7 × 7 size channel and a decreased step esteem in the main layer. The last layer is the softmax layer that is utilized to change over the score into the likelihood that the picture has a place with every classification. Here, an altered direct unit (ReLU) is paced behind each secret layer as well as pooling layer for nonlinear change, to speed up combination of the organization.

Densely Connected Network (Densenet)

Dense Convolution Network (DenseNet) is one among the latest neural network used for visual object recognition. When compared to other CNN network designs, higher accuracy and fewer parameters can be achieved using dense connection. Here, each layer receives additional inputs from all the previous layers and outputs its own feature-maps to all the successive layers, this result in a thinner and compact/dense network with fewer numbers of channels and it has higher memory and computational efficiency.

1 ×1 convolution layer which is followed by 2 ×2 average pooling layer is used as the transition layers between two adjacent dense blocks. The size of the feature map is the same within the dense block such that they can be easily concatenated together. Average pooling is performed at the end of the last dense block, and finally SoftMax classifier is attached. DenseNets are good feature extractors for variety of computer vision tasks owing to their decreased feature redundancy and compact internal representations.

Highway Networks

Highway networks permits data move through many layers on data parkways, enlivened by Long Short-Term Memory repetitive organizations and utilizes versatile gating units to smooth out the progression of data. Though there may be hundreds of layers, simple gradient descent can be used to directly train the highway networks which enable the study of extremely efficient and deep architectures. Deep networks with restricted computational budget can also be trained directly in a single stage when it is converted to highway networks. Highway networks do not suffer if we increase the depth of the network. Here too fully connected plain layers form the first layer of the network which if then followed by fully connected highway layers and at last the output is produced by a SoftMax layer [14].

One advantage of the highway architecture over other network connections is that the network itself learns to adjust the information routing dynamically depending on the present input. Highway networks have smaller errors when compared to plain networks and better accuracy with lesser number of parameters.

Network in Network (NiN)

In NIN, the GLM summed up direct model is supplanted by a miniature organization structure, a general non straight approximator, in which multi-facet perceptron is chosen as the portrayal of the miniature organization, that is a widespread capacity approximator and back-proliferation is utilized to prepare the neural organization. The NiN has two key components: convolutional MLP layer and the global average pooling layer. Multi-layer perceptron MLP replaces the GLM for input convolution and ReLU (Rectified linear unit) is used as the activation function for MLP. The cross channel. The cross channel pooling layer is used here which is similar to a convolution layer made of 1 ×1 convolution kernel. Global average pooling method replaces the conventional fully connected layers in CNN that generates one feature map for each category of the classification task in the final MLP convolutional layer. Despite of adding fully connected layers which were on the top of the feature maps, the average of each feature map is taken into consideration, and the resulting vector is provided directly into the SoftMax layer. The advantage of global average pooling are the feature maps can be interpreted easily and overfitting is avoided. Additionally, as it aggregates out the spatial data, it is more powerful to enter spatial interpretations. As an underlying regularizer, worldwide normal pooling straightforwardly force the component maps as certainty guides of ideas. This is refined by the convolutional layers, as better guess to the certainty maps is made than that by GLMs.

NIN is a heap of MLP convolutional layers, with the worldwide normal pooling and the target cost layer on the top. In the middle of the MLP convolutional layers we can add sub-examining layers as in CNN organizations. A three-layer perceptron is available inside each MLP convolutional layer. The quantity of layers in both NIN and the miniature organizations is adaptable and these layers can be tuned for explicit errands [15].

FractalNet

FractalNet, is the foremost simple substitute for ResNet which shows that to build ultra-deep neural networks, explicit residual learning is not required. During training, it has the ability of transition from effectively shallow to deep. As a fractal network widens it also gets deepens. The number of convolution layers present in the longest path between the input and output defines the depth of the Fractalnet. The result of a convolution layer is termed as blob and the join merges two blobs into one. The size of the filter set in the previous convolution layer is called as the channel count. As we expand the fractal, we combine the adjacent joins into a single join layer that spreads multiple columns. The join layer combines all of its input feature blobs into a single output blob. In FractalNet, no signal is powerful when compared to others. Each input to a join layer is the output of previous convolution layer [16].

Inception Recurrent Residual CNN (IRRCNN)

The objective of IRRCNN is to improve the performance recognition using the same number or lesser computational parameters as compared other deep learning techniques. The engineering incorporates Repetitive Convolutional blocks (RCL), trailed by initiation units lastly remaining units. The sources of info are taken care of into the information layer, then, at that point the sources of info are sent through commencement units where RCLs are applied, lastly the yields of the beginning units were added to the contributions of the IRRCNN-block. The intermittent convolution activity is acted in the initiation unit. In view of this repetitive construction inside the convolution layer, the yields at the current time step are added with that of the past time step.

The outputs at the current time step are then used as inputs for the upcoming time step. This ensures better recognition accuracy with the same count of network parameters [17].

U-Net FCNN

U-Net for biomedical picture division created by O. Ronneberger's engineering comprises of two ways. Encoder—the principal constriction way catches the setting in the picture and is a heap of convolution and pooling layers. The subsequent way—decoder is the symmetric extending way, which uses translated convolutions and empowers the exact limitation. It is a start to finish FCN network with just convolutional layers and no thick layer. Thus, it can acknowledge a picture of any size.

Generative Adversarial Networks (GANs)

GAN is a two-player min–max game with a generator as the primary player and discriminator as the subsequent player. The change from the earlier circulation of the arbitrary clamor to sensibly looking pictures is finished by generative organization G. Discriminator D organization groups the phony example that is produced, from the genuine preparing information appropriation. Generator G boundaries are changed utilizing the input data from the discriminator D so discriminator in the arrangement assignment can be tricked by the generator tests. While G learns and creates the genuine examples, D delivers better and more reasonable phony examples. GAN can plan the irregular to a practical dissemination and is utilized for applications including reproduction, division, space transformation and recognition. GANs can likewise be utilized for the manufactured information age where the generator maps the commotion to the engineered picture vector.

Long Short-Term Memory (LTSM)

Long transient memory (LTSM) is a type of intermittent neural organization intended to stay away from the drawn out reliance issue with the default conduct—Remembrance of data. It has input associations; henceforth it is likewise named as broadly useful PC with the capacity to handle groupings of information.

Limited Boltzmann Machine (RBM)

RBM is depicted by an amazingly clear plan contained an observable layer or the data layer, and a mysterious layer, planned as a bipartite graph as there is intra-layer correspondence in RBM, which is the huge limit in this designing. These machines are ready to grow the consequence of probabilities designated to every model in a given getting ready set, by a contrastive distinction estimation performing Gibbs testing.

Autoencoders (AEs)

AEs are a kind of counterfeit neural organizations used to learn proficient information codings in an unaided way that figures out how to duplicate its contribution to its yield. It has an interior/stowed away layer that depicts a code to address the info, and it is has two principle parts: an encoder that maps the contribution to the code, and a decoder that maps the code to a reproduction of the information.

Stacked Autoencoders

The crucial plan of SAEs is stacking of n autoencoders into n stowed away layers using solo layer-wise learning followed by the changing using an oversaw strategy, including three phases: Firstly, using input data, the first autoencoder is ready and the component vector is outlined. Additionally, this component vector is the commitment of the accompanying layer and the cycle is reiterated until the completion of the readiness of stowed away layers. Third, a retrogressive spread/BackPropagation (BP) plot is used for minimization of the cost work after the readiness of the mysterious layers and the heaps are invigorated with planning set to gain the aligning.

Inadequate Autoencoders SAE

Here sparsity is introduced in the mysterious units by making the amount of centers in a mysterious layer greater than the data layer. Heap of SAE (SSAE) is ready in greedy plan anyway they are related with the encoding part figuratively speaking. In any case, the mysterious layer is autonomously ready as SAE, and the yield of this layer transforms into the commitment of the accompanying layer getting ready. Arrangements are removed by using low-level SAE, later unique SAEs are stacked together where these eliminated parts are dealt with to the commitment of obvious level SAE for the extraction of more significant components. A totally independent inadequate convolutional autoencoder (CAE) network is made out of six convolutional layers and two ordinary pooling layers which is isolated into three branches: the center ID branch, the nearer see feature branch, and the establishment branch. The reproduced pictures of nearer view and establishment are made by translating the bleeding edge and establishment feature maps. The last picture is created by adding the two momentary pictures.

Convolutional Autoencoders CAE

Convolutional autoencoders, a class of solo learning estimations, take in features from the unlabelled pictures by using beginning to end learning plan. The spatial association between the image pixels improves it than the stacked autoencoders. Components can be removed from them once the channels have been learned and are viably be used to reproduce the information. In CAEs, the amount of limits expected to make an inception map is reliably the extremely, that makes it proper for the scaled high dimensional pictures. If we displace the show layer as opposed to the totally related layers of a direct autoencoder, it transforms into the convolutional autoencoder. The degrees of the data layer and the yield layer proceed as before as in the clear autoencoder, besides the unravelling part, which changes to the convolutional network.

Profound Belief Networks (DBN)

DBN is a probabilistic generative framed by stacking a few Restricted Boltzmann Machines (RBMs). DBN was intended to show an apparent conveyance among the info and secret layers' space to such an extent that there are immediate associations among the lower layer hubs and backhanded associations among the upper layer hubs. The preparation interaction is done layer-wise and simultaneously by changing the weight boundaries utilizing contrastive union (CD) to build up a reasonable gauge of the learning likelihood.

Versatile Fuzzy Inference Neural Network (AFINN)

The versatile fluffy surmising neural organization (AFINN) joins the derivation capacity of fluffy, human information aptitude and versatile learning of neural organization, making it a more remarkable methodology than those dependent on neural organizations or fluffy rationale alone. At learning, stage enrolment work is naturally tuned thusly loads are changed in by back spread.

Genome Profound Learning Philosophy

GDL model for malignancy ID comprises of element choice, highlight quantization, information channels and profound neural organizations including various secret layers among information and yield layers. GDL comprises of information handling and model preparing.

Information handling includes three stages: Initial advance is the sequencing information are contrasted with a reference with acquire a point change record, which is then changed over into a configuration of the model information. Last advance is to channel the information after change particularly in explicit model, as there is a need of just chose restricted variety destinations. GDL model involves four completely associated layers and a softmax relapse layer and uses Rectified Linear Unit (ReLU) as non-liner enactment work and to streamline the model, L2 regularization was utilized. Toward the finish of the preparation, a characterization model is gotten.

3 Conclusion

Artificial Intelligence is a boon creation of human which reproduce human brain in terms of intelligence and learning. Different methodologies and mechanisms for cancer classification with Deep learning are bluntly explained in this paper. Simulated intelligence has the further developed preparing procedures to learn complex examples in huge measure of information. Computer based intelligence is an extraordinary component for discovering malignancy qualities as well as in the fields of Virtual Assistant or Chat bots, Agriculture and Farming, Autonomous Flying, Retail, Shopping and Fashion, Security and Surveillance, Sports Analytics and Activities, Manufacturing and Production and Live Stock and Inventory Management.

References

1. Agarwal A, Saxena A (2018) Malignant tumor detection using machine learning through Scikit-learn. Int J Pure Appl Math 119(15):2863–2874
2. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674
3. Goyal M, Knackstedt T, Yan S, Hassanpour S (2020) Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. Comput Biol Med 127:104065
4. Sadoughi F, Kazemy Z, Hamedan F, Owji L, Rahmanikatigari M, Azadboni TT (2018) Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer Targets Therapy 10:219–230
5. Munir K et al (2019) Cancer diagnosis using deep learning: a bibliographic review. Cancers 11(9):1235
6. Alom MZ et al (2018) The history began from alexnet: a comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164
7. Segato A, Marzullo A, Calimeri F, De Momi E (2020) Artificial intelligence for brain diseases: a systematic review. APL Bioeng. https://doi.org/10.1063/5.0011697
8. Fukushima K (1988) Neocognitron: a hierarchical neural network capable of visual pattern recognition. Neural Netw 1(2):119–130
9. LeCun Y et al (1998) Gradient-based learning applied to document recognition. In: Proceedings of the IEEE 86.11, pp 2278–2324
10. Du G et al (2019) Efficient softmax hardware architecture for deep neural networks. In: Proceedings of the 2019 on Great Lakes Symposium on VLSI 2019
11. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. arXiv preprint arXiv:1710.09829
12. Li Y et al (2017) Vip-cnn: visual phrase guided convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017
13. Bagherinezhad H, Rastegari M, Farhadi A (2017) Lcnn: lookup-based convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017
14. Srivastava RK, Greff K, Schmidhuber J (2015) Training very deep networks. arXiv preprint arXiv:1507.06228
15. Lin M, Chen Q, Yan S (2013) Network in network. arXiv preprint arXiv:1312.4400

16. Larsson G, Maire M, Shakhnarovich G (2016) Fractalnet: ultra-deep neural networks without residuals. arXiv preprint arXiv:1605.07648
17. Alom MZ et al (2020) Improved inception-residual convolutional neural network for object recognition. Neural Comput Appl 32(1):279–293

AI, Social Media, and Big Data Analytics

A k-Mean Classification Study of Eight Community Detection Algorithms: Application to Synthetic Social Network Datasets

Mohamed El-Moussaoui, Mohamed Hanine, Ali Kartit, and Tarik Agouti

Abstract An incredible importance has been devoted to the community detection algorithms applied to complex networks. In fact, a large variety of approaches and proposals have enriched both academic and commercial purposes to deal with the identification of communities and structures in many different fields such as sociology, biology, transportation, statistical physics, computer science and so on. In this paper, we propose a k-mean classification study of eight algorithms (Fast Greedy, Walktrap, Spinglass, Leading Eigen, Label propagation, Infomap, Optimal and Louvain), applied to two social network synthetic datasets. The aim of this paper is to highlight both convergences and divergences between them, taking into consideration the twofold: Modularity measure (Q) and the number of detected communities (CN). The experimentation has been fulfilled using the R language without focus on time and space complexity measurements. This contribution builds an experimental state of the art, designed to reach beginners audiences on the topic of community detection identification.

Keywords Community detection · Clustering · Modularity · Complex network · Social network analysis

1 Introduction

Complex Network Analysis has been imposed as an important field of research, widely involved in social networks, biological networks, food networks, web and metabolic networks and so on. Nowadays, with all accorded attention, complex network analysis is considered as one of the basics of the discrete mathematics, while originally representing a form of mathematical graph theory [10, 17, 22, 41]. With this rapid expansion of social network platforms, in particular Facebook, Instagram,

M. El-Moussaoui (✉) · M. Hanine · A. Kartit
Chouaib Doukkali University—ENSAJ, Route AZEMMOUR, El Jadida, Morocco

T. Agouti
Cadi Ayyad University—FSSM, Route SAFI, Marrakesh, Morocco

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_28

Twitter, YouTube and so on [18, 43], the data generated by these various platforms used in our daily-lives, has increased significantly and has begun to enhance the scientific community for analysis purposes. In fact, community detection creates valuable opportunities through statistical analysis [20, 21], and brings answers to the future of our society (i.e. Identification of communities in social networks; profiling goods and services and even users in social networks, …, etc.), in order to understand behaviors of systems and individuals based on their interactions or activities [8].

The interest in community detection approaches was expressed earlier by Newman and Girvan in their valuable contributions [11, 22], where specific characteristics were introduced to help understanding community structures, and to discover common properties of users representing similar characteristics.

In this sense, a complex network by default, exhibits a set of properties which help to define community structures composing the network. For instance, if a user in a social network shares the same properties as another user or other users of the same social network, then we can say, this network displays at least a community structure [42]. Therefore, a community in social networks represents a group of users sharing a maximum of common properties between them, and a minimum of similarities with other users of the same social network [37] (Fig. 1).

Identifying clusters in social networks is mainly about discovering similarities which allow to construct cohesive groups, which represent important opportunities in various applications for marketing, statistical and commercial purposes (e.g. Friends and videos recommendations in online social networks platforms, Product recommendations in e-commerce platforms [3], …, etc.).

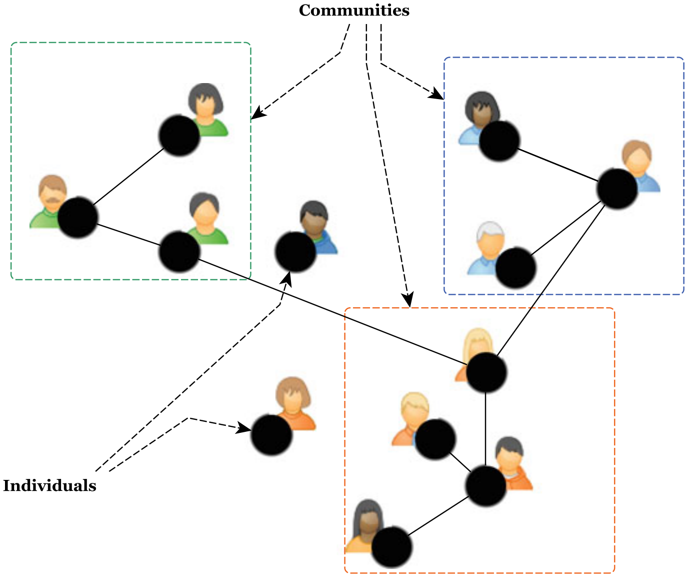


Fig. 1 A sample presentation of a social network with three communities

The interest behind the community detection field of study applied to social networks, is to understand the behaviors of users, groups. Then, identify and classify characteristics of the network structures. Therefore, a voluminous number of approaches has been published to deal with community detection in social networks, many of them have been applied differently through various synthetic datasets.

In this paper, we provide a comparative overview of different approaches dealing with the problem of community identification. Moreover, we focus our study on two synthetic datasets in order to highlight important conclusions on the obtained results. We based the current comparative study on papers written in English, with valuable content for practitioners interested in community detection, and which provide the subject matter of the problem based on free datasets. This paper is structured as follows: In Sect. 2, we provide a review of the primer concepts and related definitions. Where in Sect. 3, we give a literature review of some existing community detection approaches. Then, we describe in Sect. 4 an overview and brief definitions of the used approaches and algorithms. We present details of the comparison study in Sect. 5, and we discuss the obtained results in Sect. 6. Finally, we conclude with observations and a number of perspectives for future works.

2 Primer Concepts

2.1 Complex Network

A complex network is a real life representation of complex systems around everyday life. A complex network represents the intersection of two fields, the graph theory and the statistical mechanisms. The characteristics of such systems are described as follows:

- The emergence of behaviors
- Autonomy of management and self-organization
- A Non-linearity of interactions
- System evolution.

Online social networks are depicted by social network normal representation, namely called a graph. A graph definition is based on nodes denoting users or individuals, and edges denoting relationships between them. Groups, clusters or communities represent cohesive groups of users sharing similarities between them more than with other users of the network. In addition, properties of users can either be topological related to the structure of the network, or to topical properties related to semantic characteristics.

2.2 Network Properties

Topological-based properties: Network measures represent the topological properties for which valuable metrics can be assigned to each node. For instance, a degree measure is a fundamental metric based on the number of edges linked to a node. Moreover, degree distribution of all nodes defines characteristics of the network structure. In fact, the out-degree and in-degree measures are designed to represent, respectively outcoming and incoming edges of two interconnected nodes.

The following are the most used network measurements without regard to the directness of the network:

- Degree
- Free networks
- Shortest path
- Transitivity
- Centralities.

Topic-based properties: The topic properties are not necessarily real connections between network components. Textual information networks represent a kind of network, which uses the topic-based properties to identify similarity matrices of the network component. Thus, depending on the network type, topic properties can be extracted from nodes and edge contents.

2.3 Clustering

Clusters means groups of users in case of social networks, obtained by processing the similarity selection of user properties. Clustering techniques (Hierarchical and partitioning methods) were used frequently to deal with similarities of node properties at a high level of computation. To deal with the clustering problem, it may be imperative to improve similarities of the network properties since clusters share at least a level of similarity, which is computed through different measures such as clustering coefficient, betweenness, node degree, centrality and modularity.

In this work, the computation of the modularity measures will be applied through several community detection algorithms on two different networks. The formula of the modularity measure is expressed as follows in Eq. 1 and according to the definition of Girvan et al. [11]:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \tag{1}$$

where C_i is a set of partitions (subgroups), m is the number of nodes in the network, k_i is the node degree. A_{ij} is the set of elements of the adjacency matrix of the network, and the δ-function yields one if nodes i and j belong to the same community $C_i = C_j$, and yields zero when they belong to different communities.

3 Literature Review

A valuable number of research in complex network analysis has been interested in community detection during the last few years [10, 14, 17, 39, 45]. Furthermore, platforms using community detection algorithms have been increased. For instance, recommendation systems in social networks such as friend's recommendations on Facebook, video's recommendations on YouTube, product recommendations and identification of customers shopping habits on Amazon, eBay or Wish and so on. In fact, community detection provides a valuable theoretical instrument to understand network structures and to enhance the user's real-life expectations.

The literature has been enriched with valuable approaches published by Girvan, Newman, Fortunato and so on [2, 4, 10, 24, 25, 30]. Where recent approaches gained attraction with their innovative aspects in dealing with specific characteristics or environments [35, 38]. However, no one of these approaches can provide a global and efficient resolution to a particular scenario or a final objective.

Agrawal et al. [1] based his approach on spectral clustering and minimum cut problem, where Fortunato [10] based his proposed approach on statistical inference perspectives, Schaeffer [36], proposed his approach for clustering problem as an unsupervised learning task based on similarity measure over the data of the network. Girvan and Newman have based their proposal for the community detection on betweenness calculation focusing on the weights of nodes to assess groups boundaries in the graph, where the modularity measurement is the overall consistency [11, 19, 22].

The weight employed by [22, 23] is intended as the betweenness metric [9], reflecting the number of shortest paths connecting all moving pairs of nodes. Nevertheless, the question of community identification has largely been explored for undirected graphs. Various approaches have been proposed in the empirical comparison performed by Leskovec et al. [19], which inspire many disciplines to address the issue.

Fortunato [10] noted interestingly, that there are few ways to expand methods from undirected to directed case studies. Moreover, they mentioned that the directness of edges is one of the multiple difficulties to deal with in graph clustering issues. Additionally, various real-world social networks datasets are directed, noticing the importance of the directionality of edges in storing information behind. Therefore, many community detection approaches and methods can not be applied directly to weighted and directed graphs, where there are specific granularity scales in the number of communities not always known before.

As there has been more attention devoted to the community identification issue, in particular for social networks, multiple researchers have focused on topological metrics [6, 9–11, 13, 22, 23, 45]. Recent studies therefore begin to concentrate on both topological and topical measurements [6, 29] to bypass limited performances of topological metrics in identifying communities.

4 Approaches Overview

In this section, we describe the definition of each approach and highlight the used principle. In addition, we provide the author of the approach and the related reference of the main contribution.

Fast Greedy Proposed by [4], the Fast greedy algorithm takes advantage of modularity optimization parameters, where every node in the network is considered as a single community, then merged to other communities according to its modularity. The process is applied to all nodes of the network and stops when modularity is not moving up.

Walktrap Introduced by [27], the Walktrap algorithm is classified as a hierarchical clustering algorithm, and based on the random walk principle. In fact, the short distance random walks allow a node to belong to a community or not.

Spinglass Initially proposed by [31], this algorithm is a representation of a model of statistical mechanics. Each node in the network will be updated according to an optimal spin configuration, in order to be assigned to a spin state or a community.

Leading Eigen Proposed initially by [26], this algorithm belongs to hierarchical approaches category, and based on a top-down representation optimizing the modularity measures of the network, this configuration splits the network initially to two partitions according to the evaluation of the leading eigenvector matrix.

Label Propagation Introduced by [32], this approach identifies clusters through assigned labels of node's neighbors. Where initially a random node is assigned to a unique label, and moved iteratively to the label of the most frequent labels of its neighborhood.

InfoMap Introduced by [33, 34], this algorithm takes principal from random walk approach. The process of the Infomap algorithm is based on the minimum description length of the walker trajectory. Communities are formed together iteratively while keeping optimization of information flow until no optimization is possible.

Optimal Used under the constraint of the network size limitation, the goal of this algorithm is to optimize the modularity function itself, in order to find the corresponding clusters to the maximum modularity. The proposed implementation in R igraph library [5] is based on modularity optimization as a linear integer problem, initially introduced by Girvan and Newman [11].

Louvain Introduced by Blondel et al. [2], this algorithm belongs to the multilevel algorithms family, represented as an efficient approach based on local dynamical optimization of the modularity measure. The Louvain approach process is made in two steps: the first step aims to assign every single node to a community, taking into account the maximization of the network modularity. Then, the second step aims to form network structures based on clusters previously identified in the first step. Further, the process is repeated until improvement of the network modularity.

5 Experimentation

In this section, we perform the classification study based on the selected algorithm, where we present the obtained results of the performed experimentation, taking into consideration the available resources. In fact, the experimentation contribution promotes the CN and Modularity measure computation than the time and space complexity measures. All selected algorithms have been implemented and executed separately using RStudio version 1.2.1335, on MacOS operating system, with an Intel core i5 processor of 2.5 GHz and 8 GB of memory.

We used for this experimentation the igraph library [5], provided publicly in the R distribution (version 3.3.3 ©2017). The following sections describe first the detailed technical description of the selected synthetic datasets. Then, we describe the comparative study based on the computation results of the proposed community detection algorithms.

5.1 Datasets

The igraph Library [5] provides a bunch of synthetic datasets for different purposes. In fact, we have been interested in the social network datasets category, for which we choose two datasets, representing synthetic social networks, in order to apply different community detection algorithms and compute both the modularity measures (Q) and identify the community number (CN) related to each algorithm. The following are the graph representations of the selected datasets (Fig. 2).

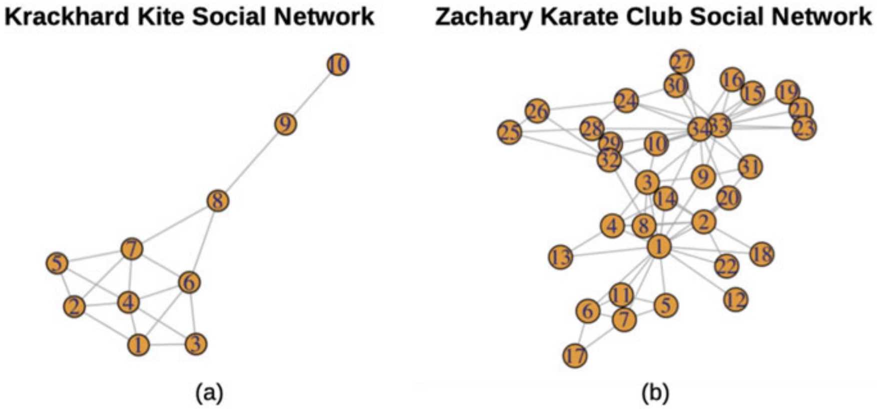

Krackhard Kite Social Network
Zachary Karate Club Social Network
1
2
3
4
5
6
7
8
9
10
(a)
(b)

Fig. 2 **a** The graph representation of the Zachary Karate Club [44] with 34 members and 78 edges. **b** The graph representation of the Krackhardt Kite social network [16] with 10 nodes and 18 edges

5.2 Comparative Overview

In this section, we compute each one of the eight community detection algorithms separately to two synthetic social network datasets. All the obtained results are compared based on two important parameters: (i) the number of identified communities; and (ii) the modularity computed through the formula expressed in Eq. 1.

We visualized in Figs. 3 and 4, the obtained results of applying the community detection algorithms respectively to Zachary Karate Club social network [16] and Krackhardt Kite social network [44]. In addition, we computed the modularity measures (Q) and determined the number of identified communities (CN).

The results of the above visualizations are summarized respectively in Tables 1 and 2.

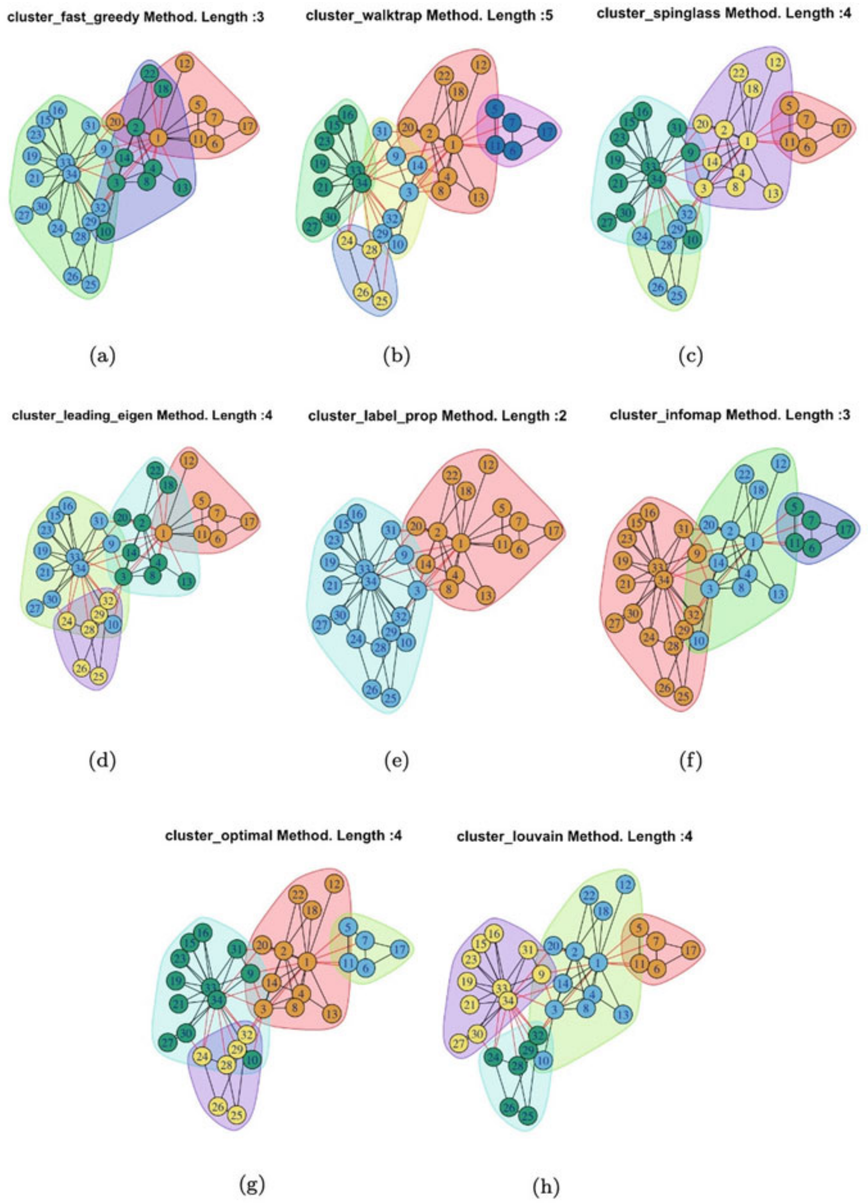
We summarized the eight algorithm visualizations in Figs. 3 and 4, where we highlight the obtained results respectively for Zachary Karate Club social network [44] and Krackhardt Kite social network [16] datasets. The obtained measures for every twofold (Q and CN) are presented in Figs. 3 and 4 for comparative overviews (Figs. 5 and 6).

Table 1 Modularity (Q) and community number (CN) measures for each community detection algorithm, applied to Zachary Karate Club social network

Community detection algorithms	Modularity (Q)	Community number (CN)
Fast greedy (a)	0.38	3
Walktrap (b)	0.35	5
Spinglass (c)	0.42	4
Leading eigen (d)	0.40	4
Label propagation (e)	0.37	3
Infomap (f)	0.40	3
Optimal (g)	0.42	4
Louvain (h)	0.42	4

Table 2 Modularity (Q) and community number (CN) measures for each community detection algorithm, applied to Krackhardt Kite social network [16]

Community detection algorithms	Modularity (Q)	Community number (CN)
Fast greedy (a)	0.22	3
Walktrap (b)	0.12	3
Spinglass (c)	0.22	3
Leading eigen (d)	0.16	3
Label propagation (e)	0.10	2
Infomap (f)	0.10	2
Optimal (g)	0.22	3
Louvain (h)	0.22	3

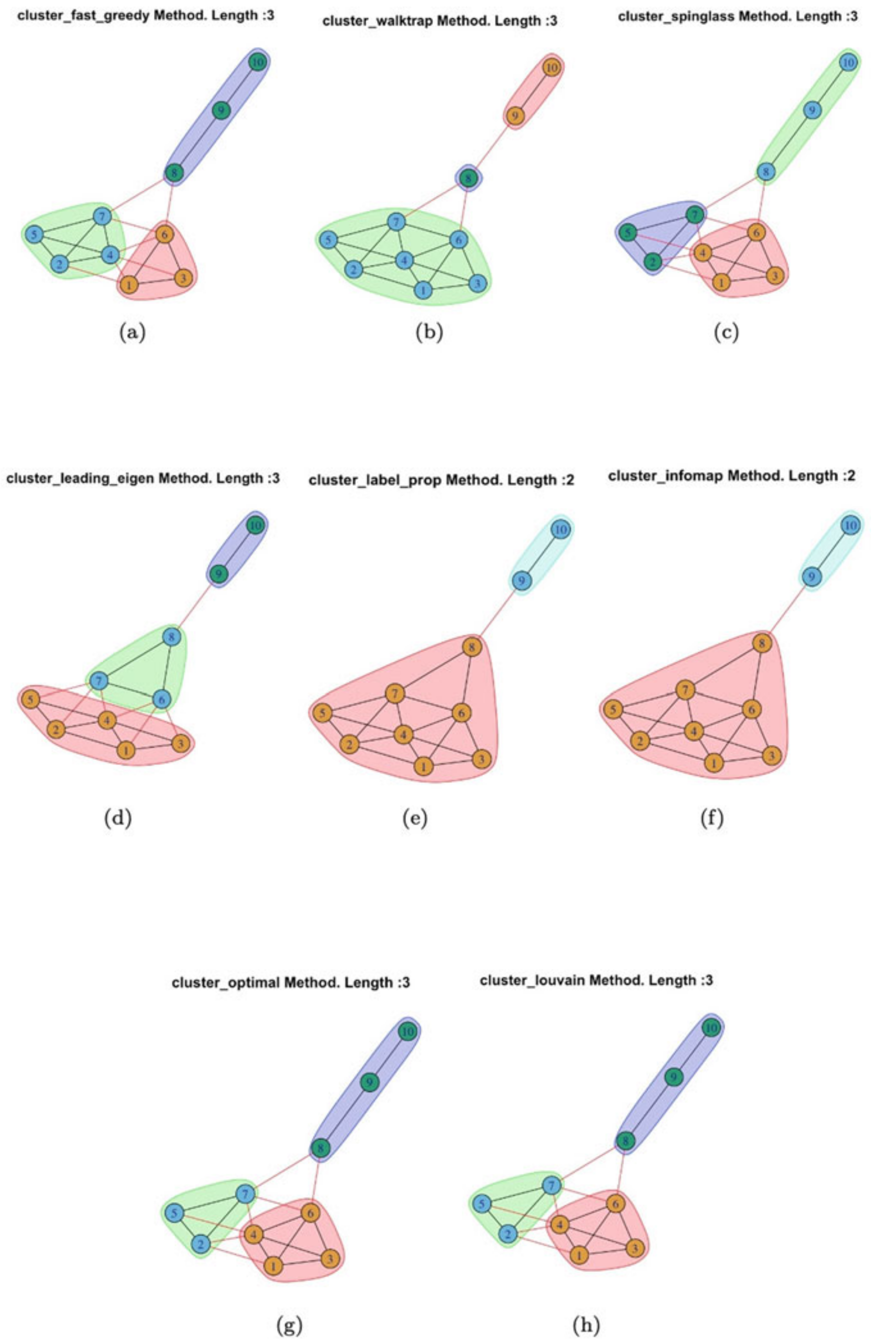

cluster_fast_greedy Method. Length :3
cluster_walktrap Method. Length :5
cluster_spinglass Method. Length :4
cluster_leading_eigen Method. Length :4
cluster_label_prop Method. Length :2
cluster_infomap Method. Length :3
cluster_optimal Method. Length :4
cluster_louvain Method. Length :4

(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 3 Comparison of eight community detection algorithms on Zachary Karate Club social network. A colored circle represents a community in the network. Algorithms: **a** Fast greedy, **b** Walktrap, **c** Spinglass, **d** Leading eigen, **e** Label propagation, **f** Infomap, **g** Optimal and **h** Louvain



cluster_fast_greedy Method. Length :3

cluster_walktrap Method. Length :3

cluster_spinglass Method. Length :3

(a)

(b)

(c)

cluster_leading_eigen Method. Length :3

cluster_label_prop Method. Length :2

cluster_infomap Method. Length :2

(d)

(e)

(f)

cluster_optimal Method. Length :3

cluster_louvain Method. Length :3

(g)

(h)

Fig. 4 Comparison of eight community detection algorithms on Krackhardt Kite social network [16]. A colored circle represents a community in the network. Algorithms: **a** Fast greedy, **b** Walktrap, **c** Spinglass, **d** Leading eigen, **e** Label propagation, **f** Infomap, **g** Optimal and **h** Louvain

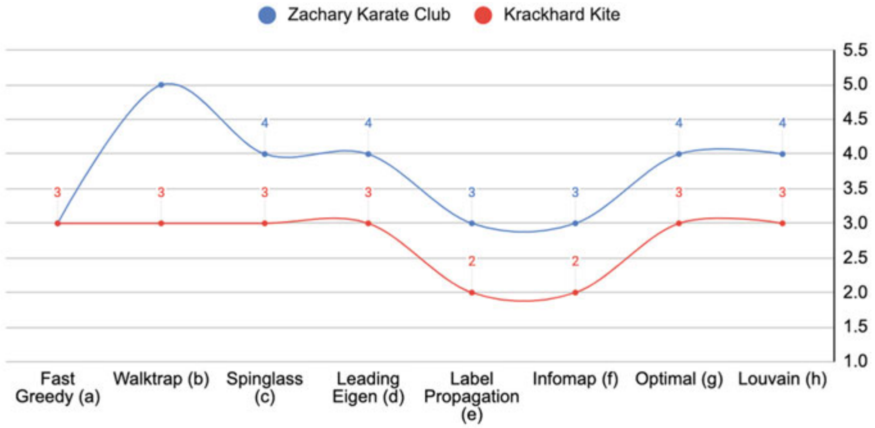


Fig. 5 Comparative of the obtained measures CN for both Zachary Karate Club and Krackhardt Kite datasets

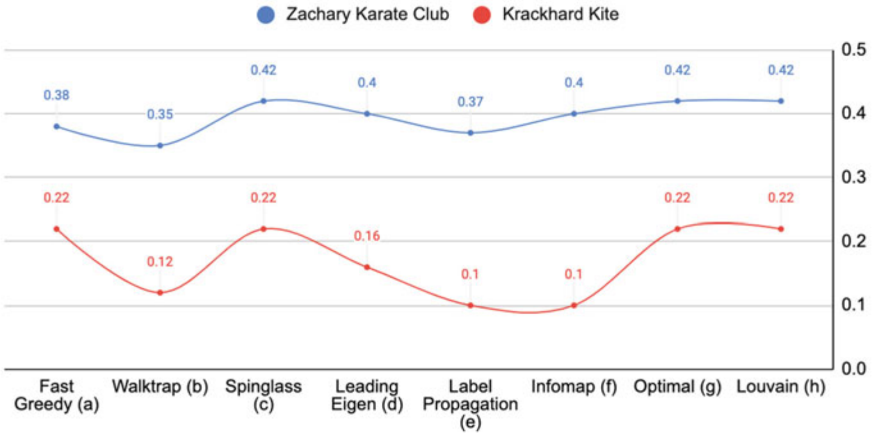


Fig. 6 Comparative of the obtained measures Q for both Zachary Karate Club and Krackhardt Kite datasets

5.3 k-Mean Classification

In this section, we process the k-mean classification for the obtained results on separate datasets, to print out the shared similarities that gather the experimented algorithms.

We processed the k-mean classification in two steps: (i) The identification of the optimal k by employing the Elbow method which uses the Within-cluster Sum of Square (noted WSS), and (ii) the k-mean classification for both social network datasets. The following figure displays the results of the optimal k computation for both synthetic datasets (Fig. 7).

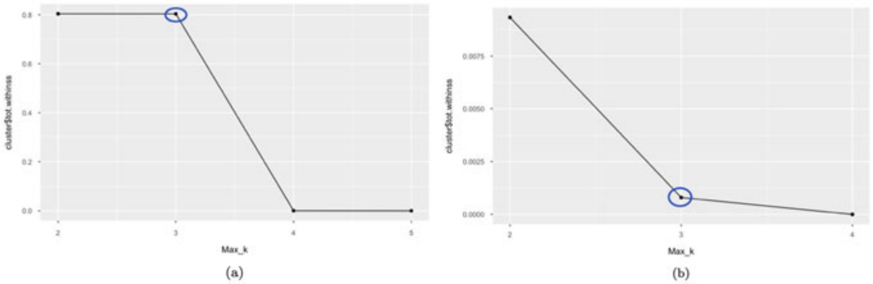


Fig. 7 Identification of $k = 3$ for the optimal k-mean values for both datasets **a** Zachary Karate Club and **b** Krackhardt Kite datasets

Table 3 k-mean classification results applied to Zachary Karate Club dataset

k-mean classification	Average (Q)	Community number (CN)
Class 1: (b)	0.35	5
Class 2: (c), (d), (g) and (h)	0.41	4
Class 3: (a), (e) and (f)	0.38	3

Table 4 k-mean classification results applied to Krackhardt Kite dataset

k-mean classification	Average (Q)	Community number (CN)
Class 1: (b) and (d)	0.14	3
Class 2: (a), (c), (g) and (h)	0.22	3
Class 3: (e) and (f)	0.10	2

Once the optimal k is determined, we compute the k-mean classification for both datasets with optimal k equal to two ($k = 3$), the obtained result is as detailed in Tables 3 and 4.

As shown in Table 3, the k-mean classification reveals three families of algorithms. (i) The first class includes only one (Walktrap algorithm) which creates the maximum of communities ($CN = 5$). (ii) The second class includes four algorithms with an average modularity of $Q = 0.41$ (Spinglass; Leading Eigen; Optimal and Louvain), and creates four communities ($CN = 4$). Where (iii) the third class includes three algorithms with an average modularity of $Q = 0.38$ (Fastgreedy; Label Propagation and InfoMap), and creates three communities ($CN = 3$).

As shown in Table 4, the k-mean classification reveals three classes of algorithms. (i) The first class which creates three communities ($CN = 3$) with a low average of modularity ($Q = 0.14$), and includes two algorithms (Walktrap and Leading eigen), where the (ii) second class creates the same number of communities ($CN = 3$ communities) and with high average of modularity ($Q = 0.22$), this class includes Fast Greedy, Spinglass, Optimal and Louvain algorithms. Where (iii) the third class includes just two communities and concerns the Label propagation and Infomap algorithms. This class is characterized with the lowest average of modularity ($Q = 0.1$).

6 Discussion

Various proposals have been published in the last few years, providing literature guides to several community detection approaches in social networks. For instance, Dao et al. [7] mentioned in their precious contribution the most popular approaches with a large variety of experimental datasets. Consequently, this literature comparative facilitates for readers to understand the field of community detection and assimilate the variety of approaches and methods.

Contrariwise, other contributions [15, 28, 40] described a detailed comparison approach regardless of the overlapping aspect of the community detection methods. In fact, this study highlights the characteristics of real-world networks to a multiple community membership configuration, where a node can belong to more than one community. Moreover, the contribution of Ghasemian et al. [12] mentioned an important comparison based on the number of detected communities versus the maximum capacity of communities to be detected, deduced from exercising various theoretical models.

From the above list of selected approaches in Sect. 4, the choice of a given approach stands for the two major advantages: (i) the optimization perspectives of the algorithm, and (ii) the gained performances displayed by the algorithm in solving a community detection problem. Therefore, the adoption of the appropriate community detection approach to apply in the context of social networks, still an ambiguous phase and belongs to the network configuration and the implementation factors.

In this paper, we presented a classification study that may facilitate the understanding of the variety of approaches for researchers and readers of the community detection topic. We based the contribution on two parameters: (i) the number of detected communities by applying each one of the proposed algorithms, and (ii) the modularity measure of each algorithm.

Therefore, we presented a summary of the observed parameters for the eight selected algorithms, applied on two synthetic datasets. Furthermore, we provided plots of the created communities for every algorithm separately on both datasets, to conclude on both the number of detected communities and the computed modularity. Moreover, given the obtained results, we proposed a k-mean classification by employing the Elbow Method to compare the behaviors of all proposed algorithms, where we classified the experimented algorithms into three classes for each dataset as described in Tables 3 and 4.

7 Conclusion

The conducted classification study leads to a better understanding of both the differences and the similarities shared between the studied approaches. Particularly, when using multitudes of datasets with various characteristics, helping to reveal hidden configurations and bring better proposals for a given analysis. Thus, there is no reason to ignore the network contexts and factors in adopting the appropriate community detection approach in handling a complex problem. For these reasons, the current contribution can only provide overviews of existing methods and approaches, but not all exclusive arguments to decide on the efficiency of an approach against others.

Practically, we have been interested in two famous synthetic social network datasets, the Zachary Karate Club dataset with 34 nodes and 78 edges, and the Krackhardt Kite dataset containing 10 nodes with 18 edges. Therefore, this contribution study reveals many observations related to the chosen parameters (Q and CN). The analysis of the behavior of every algorithm of the ones mentioned above, displays three classes of algorithms for both datasets. The premium classes of algorithms have been distinguished by creating the highest number of communities regardless of the average of modularity measures. Where the second classes maintained the most moderated results, with approximately the same behavior. The third classes revealed the lowest number of communities.

In addition, the Walktrap algorithm created the highest communities for both datasets, contrariwise to the Label propagation algorithm, which created the lowest number of communities in both datasets, where the Spinglass algorithm displayed the highest modularity measure unlike the Infomap algorithm, with the lowest modularity measure. The Spinglass, Optimal and Louvain algorithms represent the most stable results with respect to both dataset volumes and configuration.

Rather than the identification, the quality of the detected communities taking into account all possible parameters such as time/space complexity and overlapping communities are not included in this contribution, and will be the subject of future work. In addition, we hope the future comparative contributions refine the obtained results according to other voluminous real-world datasets, and integrate more community detection approaches to help improve shared results.

Acknowledgements The authors would like to thank the editors stuff and anonymous reviewers for providing insightful suggestions and comments to enhance the quality of the paper.

References

1. Aggarwal CC, Wang H (2010) A survey of clustering algorithms for graph data. Adv Database Syst 40:275–301. https://doi.org/10.1007/978-1-4419-6045-0_9
2. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2010) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 10. https://doi.org/10.1088/1742-5468/2008/10/P10008

3. Bedi P, Sharma C (2016) Community detection in social networks. WIREs Data Mining Knowl Discov 6:115–135. https://doi.org/10.1002/widm.1178
4. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70. https://doi.org/10.1103/PhysRevE.70.066111
5. Clauset A, Newman MEJ, Moore C (2006) The igraph software package for complex network research. InterJ Complex Syst 70. https://igraph.org
6. Ding Y (2011) Community detection: topological vs. topical. J Informetr 498–514. https://doi.org/10.1016/j.joi.2011.02.006
7. Dao VL, Bothorel C, Lenca P (2020) Community structure: a comparative evaluation of community detection methods. Netw Sci 1–42. https://doi.org/10.1017/nws.2019.59
8. El-Moussaoui M, Agouti T, Tikniouine A, Eladnani M (2019) A comprehensive literature review on community detection: approaches and applications. Procedia Comput Sci 151:295–302. https://doi.org/10.1016/j.procs.2019.04.042
9. Fortunato S, Latora V, Marchiori M (2004) Method to find community structures based on information centrality. Phys Rev E 70. https://doi.org/10.1103/PhysRevE.70.056104
10. Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174. https://doi.org/10.1016/j.physrep.2009.11.002
11. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99:7821–7826. https://doi.org/10.1073/pnas.122653799
12. Ghasemian A, Hosseinmardi H, Clauset A (2018) Evaluating overfit and underfit in models of network community structure. IEEE Trans Knowl Data Eng 32:1722–1735. https://doi.org/10.1109/TKDE.2019.2911585
13. Gui Q, Deng R, Xue P, Cheng X (2018) A community discovery algorithm based on boundary nodes and label propagation. Pattern Recogn Lett 109:103–109. https://doi.org/10.1016/j.patrec.2017.12.018
14. Jin S, Yu PS, Li S, Yang S (2015) A parallel community structure mining method in big social networks. Math Probl Eng 109:1–13. https://doi.org/10.1155/2015/934301
15. Javed MA, Younis MS, Latif S, Qadir J, Baig A (2018) Community detection in networks: a multidisciplinary review. J Netw Comput Appl 108:87–111. https://doi.org/10.1016/j.jnca.2018.02.011
16. Krackhardt D (1990) Assessing the political landscape: structure, cognition, and power in organizations. Adm Sci Q 35:342–369. https://doi.org/10.2307/2393394
17. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78. https://doi.org/10.1103/PhysRevE.78.046110
18. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. Phys Rev 80:342–369. https://doi.org/10.1103/PhysRevE.80.056117
19. Leskovec J, Lang KJ, Mahoney MW (2010) Empirical comparison of algorithms for network community detection. In: WWW'10: proceedings of the 19th international conference on World wide web, pp 631–640. https://doi.org/10.1145/1772690.1772755
20. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS ONE 6. https://doi.org/10.1371/journal.pone.0018961
21. Meo PD, Ferrara E, Provetti A (2014) Mixing local and global information for community detection in large networks. J Comput Syst Sci 80(1):72–87. https://doi.org/10.1016/j.jcss.2013.03.012
22. Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256. https://doi.org/10.1137/S003614450342480
23. Newman MEJ, Girvan M (2003) Mixing patterns and community structure in networks. Lect Notes Phys 66–87. https://doi.org/10.1007/978-3-540-44943-0_5
24. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69. https://doi.org/10.1103/PhysRevE.69.026113
25. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74. https://doi.org/10.1103/PhysRevE.74.036104
26. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci U S A 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103

27. Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: Computer and information sciences—ISCIS 2005. Lecture notes in computer science, vol 3733, pp 284–293. https://doi.org/10.1007/11569596_31
28. Palla G, Deenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818. https://doi.org/10.1038/nature03607
29. Qin M, Jin D, Lei K, Gabrys B, Musial K (2018) Adaptive community detection incorporating topology and content in social networks. Knowl-Based Syst 161:342–356. https://doi.org/10.1016/j.knosys.2018.07.037
30. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci U S A 101(9):2658–2663. https://doi.org/10.1073/pnas.0400054101
31. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74. https://doi.org/10.1103/PhysRevE.74.016110
32. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76. https://doi.org/10.1103/PhysRevE.76.036106
33. Rosvall M, Bergstrom CT (2007) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A 115(4):1118–1123. https://doi.org/10.1073/pnas.0706851105
34. Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. Eur Phys J Spec Top 178:13–23. https://doi.org/10.1140/epjst/e2010-01179-1
35. Riolo MA, Cantwell GT, Reinert G, Newman MEJ (2017) Efficient method for estimating the number of communities in a network. Phys Rev E 96. https://doi.org/10.1103/PhysRevE.96.032310
36. Schaeffer SE (2007) Graph clustering. Comput Sci Rev 1(1):27–64. https://doi.org/10.1016/j.cosrev.2007.05.001
37. Saoud B, Moussaoui A (2018) Node similarity and modularity for finding communities in networks. Phys A Stat Mech Appl 492:1958–1966. https://doi.org/10.1016/j.physa.2017.11.110
38. Tasgin M, Bingol HO (2018) Community detection using preference networks. Phys A Stat Mech Appl 495:126–138. https://doi.org/10.1016/j.physa.2017.12.060
39. Xia Z, Bu Z (2012) Community detection based on a semantic network. Knowl-Based Syst 26:30–39. https://doi.org/10.1016/j.knosys.2011.06.014
40. Xie J, Szymanski BK (2012) Towards linear time overlapping community detection in social networks. In: 16th Pacific-Asia conference on advances in knowledge discovery and data mining, vol 7302, pp 25–36. https://doi.org/10.1007/978-3-642-30220-6_3
41. Xiang J, Wang ZZ, Li HJ, Zhang Y, Chen S et al (2017) Comparing local modularity optimization for detecting communities in networks. Int J Mod Phys C 28:6. https://doi.org/10.1142/S012918311750084X
42. Yang B, Liu D, Liu J (2010) Discovering communities from social networks: methodologies and applications. In: Handbook of social network technologies and applications. Springer, pp 331–346. https://doi.org/10.1007/978-1-4419-7142-5_16
43. Yang Y, McAuley J, Leskovec J (2013) Community detection in networks with node attributes. https://doi.org/10.1109/ICDM.2013.167
44. Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:4. https://doi.org/10.1086/jar.33.4.3629752
45. Zhang XS, Wang RS, Wang Y, Wang J, Qiu Y, Wang L, Chen L (2009) Modularity optimization in community detection of complex networks. EPL (Europhys Lett) 87:3. https://doi.org/10.1209/0295-5075/87/38002

Topic Modeling for Short Texts: A Novel Modeling Method

Badr Hirchoua, Brahim Ouhbi, and Bouchra Frikh

Abstract Topic modeling is one of the major concerns in the short texts area, and mining these texts could uncover meaningful insights. However, the extreme short texts' sparsity and imbalance bring new challenges to conventional topic models. In this paper, we combine a new ranking method with hierarchical representation for short text. Words ranking proves to be inexorable in generating value from disorganized short texts; thus, a novel ranking paradigm is developed and is referred to as the ordered biterm topic model (OBTM). OBTM models the semantic connection between every two words, regardless of whether they show up in similar short content, reinforcing the capacity to reveal the genuine semantic examples behind the corpus. The intense contextual information maintained in the space of word-to-word assures the word sensation in recognizing relevant topics with reliable quality. Then, the paradigm is associated with a hierarchical representation that captures the relations connecting the created topics. OBTM learns topics at the corpus level. This makes the inference more effective and robust, referring to hidden semantics patterns. Experiments on real-world collection reveal that OBTM can discover more relevant and coherent topics. It achieves high performance in various tasks and outperforming the state-of-the-art baselines.

Keywords Topic modeling · Short text · Biterm · Content analysis · Information systems · Sparseness

B. Hirchoua (✉) · B. Ouhbi
Industrial Engineering and Productivity Department, National Higher School of Arts and Crafts (ENSAM), Moulay Ismal University (UMI), Meknes, Morocco

B. Frikh
LIASSE Laboratory, National School of Applied Sciences (ENSA), Sidi Mohamed Ben Abdellah University (USMBA), Fez, Morocco

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_29

1 Introduction

Short texts have grown as influential data sources, including social media and web page snippets, among others. Short texts topic modeling has attracted increasing consideration in the last years [23, 32, 37]. The short texts are short, low signal, noisy, high volume and velocity, topic drift, and redundant data. Notwithstanding, enormous signals produced by the short texts raise it as a reliable topic modeling source, and mining these texts could uncover meaningful insights. Topic modeling [2, 14, 15] uses text mining mechanisms for hidden themes extraction among a collection of documents. Their assumptions on the training corpus are varied where it can minimize the amount of time to understand, synthesize, and exploit the short texts. Hence, the good models applied under various training conditions for long texts, such as probabilistic models including Probabilistic Latent Semantic Analysis (PLSA) [17] and Latent Dirichlet Allocation (LDA) [2] are not extensible for short texts. The before-mentioned models run on the hypothesis that every document contains a distribution over topics, comprising a subset of words. The proportions of different topics and their distribution across the vocabulary are modeled from the corpus. Precisely, this is done by employing statistical methods such as variational inference [2], and Gibbs sampling [12]. On the one hand, the efficiency and strength of such models depend on the word co-occurrences patterns in the input dataset, which is sufficiently and accurately defined in an extensive dataset containing long texts. On the other hand, the short texts' sparsity is confirmed to be faulty in displaying the connection linking words, which persists as a difficulty to the conventional methods.

As a solution for short texts, traditional topic models are extended [20, 30] by exploiting external knowledge, pre-trained topics, or metadata to inject extra valuable word-to-word co-occurrences. Phan et al. [30] used the discovered latent topics from Wikipedia to enhance and infer topics for short texts. Lim et al. [22] proposed the Twitter-Network topic model, which exploited the additional Twitter information such as authorship, hashtags, and the user-follower network. Zeng et al. [42] proposed a topic memory method to encode hidden topics characteristic of class labels. Moreover, Sridhar et al. [39] represented this space by the dense distributed words representations. Injecting strong assumptions for short text, such as unigrams model [27] or bag of words [13] can push the modelization process further. Nevertheless, the efficiency of the before-mentioned mechanism is heavily data-dependent, which is not always available. Additionally, certain assumptions lose the model's extensibility to capture several topic factors inside a document.

Since the short texts suffer from the short length, there is an expanding demand over words representation. the biterm topic model (BTM) [5] is one of the most significant solutions. The biterm is an unordered word-pair co-occurring in a short context. The BTM handles the sparsity by transferring the learning from the document level to the whole corpus, which earns the global corpus distribution. BTM assumes that a biterm shared the same mixture of topics which is represented as words distribution. The major difference between the BTM and bigram boils down to a correlations modeling level. The bigram modelizes ordinal correlations between words in document level, while BTM attempts to model the semantic conditions in the whole corpus. Contrary to the existing models, which learn the underlying topic elements by modeling the generation of documents, BTM models the biterms generation.

Compared to BTM topic models, the significant differences and advantages of the proposed approach lie in two main aspects: Firstly, the BTM models the word co-occurrence in the entire corpus to enhance topic modeling. Still, it breaks the patterns of the short texts by modeling the ordered and unordered biterm as the same. Secondly, the BTM ignores the biterm context, where a given biterm represents a context to other biterms in the corpus. The ordered biterms are semantically dependent more than the unordered biterms; thus, they should be ranked differently. Contrary to the BTM, the OBTM enables the ordered word pair co-occurring and penalizes the unordered biterms. The OBTM hypothesizes that the biterm words share the related topic selected from a mixture of topics across the entire corpus. Still, it differentiates the ordered biterms, which represent a high weight, from the unordered biterms.

To this aim, the OBTM ranks each biterm differently, where this process is stimulated in two steps. The first one contains the ordered biterms, where the second step includes the rest of the unordered biterms. The proposed methodology can illustrate semantically linked words that are not connected in the original short content. OBTM shapes the semantic connection between biterms, whether they are in similar short content or not, which reinforces its capacity to reveal the genuine semantic examples behind the corpus [21]. With the corpus-level patterns, each instance is enhanced, which consequently assists in relieving the sparsity issues. To summarize, this paper performs the following main contributions:

- This work demonstrates a novel biterm ranking method for jointly learning the topic distribution at the corpus level rather than document/sentence level.
- The OBTM catches the relationship between topics level and illustrates that the topic modeling for short texts is improved through the novel ranking method.
- The OBTM pushes the learning process further and efficiently maintains higher-order semantics among biterms.
- The OBTM improves semantic connections among words, whether or not they show up together in the input sentence, and simultaneously reduces sparsity.

The remainder of the paper starts with a summary of related work (Sect. 2). Section 3 presents the Chinese restaurant processes. The proposed system is explained in Sect. 4. Section 5 presents the probabilistic inference used in this work. Section 6 is devoted to the experiments and results. Section 7 summarizes the paper with the main conclusions and proposals for future work.

2 Related Work

In this section, some of the short texts topic modeling methods will be discussed. Firstly, the BTM extensions are discussed. Secondly, the short texts topic modeling improvement using deep neural networks is demonstrated. Last, the short texts classification models are also introduced.

Short texts topic modeling presents interesting relevant patterns retrieved from different corpus including social media and news headlines. Cheng et al. [5] introduced the BTM approach to model the short text topics by analysing word co-occurrence patterns in the entire dataset. Jiang et al. [19] built the biterm pseudo-documents (BPDTM) to extend the BTM. This model adopted the word co-occurrence network that encouraged the word representations with their semantic nearest biterms. Pang et al. [28] applied the BTM to sentiment classification [16]. First, the words-documents similarity is computed using the singular value decomposition. Then, each short document is extended by the most similar words. Heng-Yang et al. [24] exploited word embeddings to obtain semantically related word couples from the entire dataset, ignored by BTM. Ruan et al. [35] combined the traditional BTM with semantic dependency connections for the micro-blog topic detection. Zuo et al. [44] handled the short texts topic modeling by proposing a word co-occurrence network-based model for topic modeling (WNTM). Contrary to the BTM, which assumed two-word biterms, the WNTM acted in the word level, increasing the semantic density.

In addition to the BTM based methods, deep neural networks are widely used to improve short text topic modeling. Shen et al. [36] performed short text classification by adding word-cluster embedding using the deep neural network. Gao et al. [10] utilized the embedding-based minimum average distance to formulate a regular-sized pseudo-documents instead of the short texts. Lim et al. [22] referred to the Twitter metadata to propose the Twitter-Network (TN) for topic modeling. These metadata included hashtags, authorship, and the user-follower network, among others. The TN topic model combined the hierarchical Poisson-Dirichlet processes (PDP), a random function model based on a Gaussian process for text modeling, and social network modeling. Moreover, the TN enabled the automatic topic labeling and the general inference framework which handled other topic models with embedded PDP nodes.

In [41], authors have proposed a topic modeling technique to learn the latent topic patterns, which involves regularized non-negative matrix factorization for short texts. The model exploits global word-to-word co-occurrence obtained from a big dataset to mitigate the constraint of data sparseness. It additionally incorporates a clustering procedure to further enhance the thematic inference quality. To select the most relevant words, the suggested approach relies on word distribution properties. This results in reducing the sensitivity to length or topic distribution of documents and makes it more adapted to browsing short text topics. A variational auto-encoder for topic modelling has been introduced in [43]. The authors have embedded pre-trained word vectors and entities representations in the knowledge graph to enhance the performance of the neural network model, which promotes the learning process and allows the short text's sparsity avoidance. Moreover, the work has incorporated a supervised model to monitor the inherited latent mapping of both topic distributions and generation, given the labelled data in the training set. Similarly, a work, presented by [29], aims at an online detection topics to deal with streaming data related to sentiment analysis. The work has used a long short-term memory network to build a framework that consists of an online latent semantic indexation, respecting some regularization constraints. A topic-level attention scheme was embedded in the long-term memory network for retrieving the feelings from the identified topic. Authors have hypothesized having only one topic in each sentence as most of the commented texts are short.

Short texts topic modeling for documents classification are essentially generated from the following two perspectives. On the first one, documents containing short texts are extended utilising topics learned from big external corpora. Phan et al. [30] built a classifier on a set of latent topics discovered from a large-scale data collection. Chen et al. [4] provided discriminative characteristics for short text classification through integrating multi-granularity hidden topics. Duc-Thuan et al. [40] investigated big external collections including Wikipedia, LNCS, and DBLP to uncover underlying topics. The topic models capture low dimensional representations that output a dense and lower-dimensional vector from the second perspective. These representations usually denote a particular semantic sense known as a topic [6, 33]. Zeng et al. [42] proposed topic memory networks, which encoded latent topic representations using a novel topic memory mechanism. This approach did not refer to any external resources such as features extending or pre-trained topics, but instead, it used the memory networks to jointly surveys topic inference and text classification. Vskrlj et al. [38] introduced the tax2vec algorithm, which investigated word taxonomies in order to establish novel semantic characteristics and to enhance the robustness and performance of the learned classifiers. Sridhar et al. [39] introduced an unsupervised short texts topic model, which accomplished soft clustering based distributed representations of words. Therefore, the low-dimensional semantic vector space is modeled using the Gaussian mixture of distributed word representations, which captured the latent topics models, which overtook the word co-occurrence patterns sparsity issue.

Notwithstanding the rich researches discussed above, topic modeling for short texts persists as an open problem. On the one hand, the BTM ignored the semantically related and infrequently co-occurrent word pairs. In contrast to the BTM, which assumed two words biterms, the WNTM acted in the word level, where a word is semantically linked to solely nearby words. Specifically, WNTM created a pseudo-document for every word using its neighbouring words. Nevertheless, this model extended the sparsity level from short texts to pseudo-documents sparsity. Moreover, WNTM considered the first level words correlation, where real-world performance assumed the second level, as well as the words' context. On the other hand, the BPDTM employed triangle connections in the network of co-occurrence. This idea constructed a considerable amount of unnecessary edges in the network, influencing the results performance.

3 Chinese Restaurant Processes

The Chinese restaurant process (CRP) is a distribution over partitions of integers obtained by assuming unlimited tables in a Chinese restaurant. Thus, the first customer occupies the initial table, and then the m_{th} following client occupies a table picked as follows:

$$\begin{array}{l} p(occupied\ table\ i|previous\ customers) = \frac{m_i}{\gamma+m-1} \\ p(next\ unoccupied\ table\ i|previous\ customers) = \frac{m_i}{\gamma+m-1} \end{array} \quad (1)$$

The number of preceding clients at table i is m_i, and γ controls how often a new client takes a different table versus choosing an occupied table, dependent on the number of clients in the restaurant.

After M clients choose their tables, the arrangement plan provides a distribution of M items. The partition structure is equivalent to the Dirichlet process distribution [8]. However, the basic rule variations in Eq. (1) are considerably allowed; specifically, the data-dependent choice of γ, as well as the general functional dependence on the current partition [31]. Every client is linked to the vector β at its table. Notice that each table (i.e., customers table) shares the same parameter vector.

The CRP handles the uncertainty across several segments in a mixture model. Each table, in a sampling mixture [18], is linked to a draw from $p(\beta|\eta)$. The β is a parameter of the mixture component, and η is the symmetric Dirichlet hyperparameter. Given a data set, where each data point is generated using Eq. (1), the posterior model reveals two components. On the one hand, the seating plans distribution, where the number of occupied tables defines the mixture components number. On the other hand, each table that uses one of the determined setting plans activates a partition on the linked β parameter. The Markov chain Monte Carlo [26] can approximate the posterior.

Associating to the CRP, a nested CRP (nCRP) is defined as follows: The city contains an infinite number of Chinese restaurants, where a restaurant is a root, and its tables are links to other restaurants. The restaurant's tables are links to other restaurants. Each restaurant is assigned once; therefore, the city's restaurants are arranged as a hierarchical tree, where each restaurant is linked to the tree's level.

Assuming an nCRP process, a tourist on his first visit attends the root and chooses a table based on Eq. (1). He visits the restaurant indicated on the last table and selects a different table on his next visit. After L day, the tourist has occupied exactly L restaurants. The process represents a pathway from the root restaurants to the L_{th} tree's level. Repeating this process for T tourists taking L-day holidays, the gathering paths describe the infinite tree's particular L-level subtree. Using the same analogy, the biterms are seen as two friends who share the same behavior and choices.

This work uses this topic hierarchy before model a hierarchy mechanism, which postulates a relationship between levels. The nCRP can express uncertainty over the possible L-level trees, identical to the standard CRP.

4 Topic Modeling Over Short Texts

In this section, the proposed approach is further presented, and the detailed process is also discussed.

Topic modeling is a text processing technique based on statistical analysis, which extracts themes within a collection of documents. These techniques produce a similar results form, where topics are represented by the correlated words, enabling documents to be drawn from topics. However, topic modeling methods such as PLSA, LDA, and hierarchical LDA (HLDA) [11] are probabilistic; they allow the weighted words to topics assignments. Nevertheless, the word co-occurrence patterns' sparsity and unreliability make them sensitive to text shortness. The aggregation of all patterns in the corpus suffers from the instability of frequencies, which does not explicitly exhibit the correlation of the unordered words. Given this, we introduce a novel biterm topic approach. It assumes a new style to expose the underlying topic components by directly modeling word co-occurrence patterns. The OBTM does not consider the predefinition of topics, the biterms nested partitioning, or the allocation of topics-to-levels.

The system starts with the biterms extraction based on two levels technique, where the first level extracts the ordered words, and the second level formulates the combination of the unordered words. For example, a document including four different words will produce the following two levels:

$$(w_1, w_2, w_3, w_4) \Rightarrow \begin{cases} Level\ 1: \{(w_1, w_2), (w_2, w_3), (w_3, w_4)\}. \\ Level\ 2: \{(w_1, w_3), (w_1, w_4), (w_2, w_4)\}. \end{cases}$$

If two words show up together frequently, they are expected to refer to the same topic or at least to be adjacent in the hierarchical model. Hence, the OBTM considers two words of a biterm are drawn separately from a topic, which is, in turn, sampled from a mixture across the entire corpus. Besides, every biterm is ranked based on its frequencies and behavior inside the corpus, specifically being a part of the biterm or being a context for other biterms.

The general process for biterms (w_j, w_i) ranking is expressed as follows:

1. **Step 1: Corpus transformation**
   - Transform the whole corpus into biterms levels as described above.
2. **Step 2: First order biterms ranking (biterm co-occurrence)**
   - $R_1 = x_1 + \vartheta x_2$: where x_1 and x_2 are the number of times the two words in the biterm (w_j, w_i) appear together in the first and second level respectively, and the parameter ϑ penalizes the words in the second level.
3. **Step 3: Second order biterms ranking (exchangeable biterm context)**
   - $R_2 = 2 * R_1 + \frac{1}{\epsilon + y_1} + \frac{\vartheta}{\epsilon + y_2}$
     where y_1 and y_2 are the number of times the w_j is in the context of w_i in the

first, and second level respectively. The first rank R_1 is multiplied by 2 because it represents the exchangeable rank for each word in the biterm.

4. **Step 4: Third order biterms ranking (non interchangeable biterm context)**
    - $R_b = R_2 + \frac{1}{\epsilon+z_1} + \frac{\vartheta}{\epsilon+z_2}$
      where z_1 and z_2 are the number of times the word w_j or w_i serves as a context for other words in the first, and second level respectively.

We have tuned the parameters ϑ via a grid search on a small dataset to obtain the most stable value, resulting from the exact value of $\vartheta = 0.001$. The final biterms ranking mechanism uses the following proposed formula:

$$R_b = 2(x_1 + \vartheta x_2) + \frac{1}{\epsilon + y_1} + \frac{\vartheta}{\epsilon + y_2} + \frac{1}{\epsilon + z_1} + \frac{\vartheta}{\epsilon + z_2} \tag{2}$$

Once the ranking steps are done, the system defines a hierarchical model to distinguish biterms beside the paths created by the nCRP. In order to obtain the generative model for the ranked biterms, firstly, the system associates the topic with a node in the tree; then, a path elects an infinite set of topics. Secondly, a probability distribution on the topics is defined for a given path using the GEM distribution [31]. Given a GEM distribution draw and its probabilities, the system generates a biterm couple by frequently selecting topics, then representing each word in the biterm using the probability distribution determined by its chosen topic.

Formally, suppose the nCRP determines the infinite tree and assigns c_b to the path in that tree for the b_{th} biterm (i.e., customer). In the proposed approach, the biterms are considered formed from the generative method presented in Algorithm 1, where $z \sim Discrete(\theta)$ denotes the discrete distribution setting $z = i$ with probability θ_i. This generative process assigns a probability distribution over potential corpora.

Algorithm 1 The OBTM's generative process

```
for e doach table $k \in T$ in the infinite tree
    Draw a topic $\beta_k \sim$ Dirichlet($\eta$).
end for
for b doiterm, $b \in \{1, 2, \ldots, B\}$
    Draw $c_b \sim$ nCRP($\eta$).
    Draw a distribution over levels in the tree, $\theta_b \,|\{m, \pi\} \sim GEM\,(m, \pi)$.
    for e doach word $(w_i, w_j)$
        Choose level $z_{b,n} \,|\theta_b \sim Discrete(\theta_b)$.
        Choose word $w_{b,n} \,|\{z_{b,n}, c_b, \beta\} \sim Discrete(\beta_{c_b}\,[z_{b,n}\,])$.
    end for
end for
```

The biterms ranking mechanism allows an extensive weight to different words. The topic at the root level has a high probability over words in the corpus instead of a short sentence (document). In other words, assuming that each biterm is attached to a particular path, the direct level under the root produces a poor distribution on the biterms (general topics), where that level's topics will assign a significant probability on the valuable terms. Moving down, the biterms nested partitions get admirable. Therefore, the related topics will be more specific comparing to the particular biterms in these paths.

Diving into different levels of the constructed tree, if the system visits m nodes at most m topics and words will be generated. Thus, the $(m+1)_{th}$ word can belong to the previously generated topics, or generate a new one. In contrast, suppose that the biterm b_t has previously been generated, the next biterm can select an old topic using an earlier path, or it can generate a new topic along a new branch at any point in the tree.

The Dirichlet hyperparameter η controls the topic's sparsity. The smaller values will turn up at topics of probability mass on a small words' arrangement. Furthermore, the hyperparameter η stick-breaking parameters for the topic proportions $\{m, \pi\}$. In other words, it controls how many biterms evolve from topics of diversifying abstractions. Precisely, if m is large ($m = 0.5$), then the posterior will assign extra biterms to higher levels of abstraction. If π is large, the allocations of the word will not expect to diverge from such a setting if π is large ($\pi = 100$).

The proposed approach is a hybrid model, where every biterm displays various routes within the tree. However, OBTM uses a three levels distribution for word generation: first, it ranks the word in the whole context where words with similar context should occupy close levels positions, then it chooses a path through the tree, and finally, it picks the word's level.

Section 5 provides the probabilistic inference details and mechanisms. The used inference methods are also discussed to demonstrate the usability and robustness of our approach over the state-of-the-art methods.

5 Probabilistic Inference

This section aims at providing the probabilistic inference details and mechanisms. The adapted inference methods are also discussed to demonstrate the usability and robustness of our approach over the state-of-the-art methods.

The inference attempts to define the hidden variables that reveal the topic model. In other words, to generate the corpus, the system uses a collection of latent variables that completely describe the transformation from topics to allocations, then to assignments. However, modifying a topic assignment for the word w_i preserving the other assignments will thoroughly illustrate the dataset excluding the particulate word w_i. The main goal in this work is to perform posterior inference by inverting the generative biterms process for estimating the covered topic structure. Given a set of short texts, the system attempts to discover the posterior distribution on the infinite sets of hierarchies, path assignments, and level allocations of biterms.

Markov chain Monte Carlo (MCMC) [3] algorithm can approximate the posterior distribution. The MCMC constructs a Markov chain via sampling from a stationary distribution. Moreover, it uses this chain to sufficiently approach the objective, gathers the sampled states after that, and collects states to estimate the target. The general model starts with the values set, which the hidden variables can have in the Markov chain state space. The targeted distribution is conditional over the underlying variables. Specifically, we choose the Gibbs sampling algorithm. Thus, every underlying variable is sampled and conditioned on the rest of the variables. The collapsed Gibbs sampling is employed, by marginalizing out some latent variables, so the convergence time of the chain is minimized.

Notations: let c_b be the biterm paths, $z_{b,n}$ is the word level allocations, β_i is the topic marginalization parameters, and θ_b refers to the per-biterm topic proportions. The hyperparameter γ reflects the tendency of the biterm to share topics, η reflects the expected variance of the underlying, and m and π indicate the expectation regarding the words allocation over levels inside a biterm. Finally, $z_{-(b,n)}$, and $w_{-(b,n)}$ represent the level allocations vector representations and remarked words ignoring respectively the $z_{b,n}$ and $w_{b,n}$.

Intuitively, the parameters γ and topic prior η are the CRP parameters, which examine the inferred tree size. The choice of γ and η is crucial to the model, where a large γ and small η construct a tree with more topics. The small η favors fewer words with a high probability for a topic. Therefore, the posterior requires extra topics to describe the input corpus. On the other hand, a large η improves the likelihood, which leads to new ways while crossing the nested CRP. Overall, a larger π controls the specificity and generality that points to more interpretable hierarchies. Therefore, the target is to approximate the posterior:

$$p(c_{1:B}, z_{1:B} \mid \gamma, \eta, m, \pi, w_{1:B}) \tag{3}$$

The GEM parameter m indicates the balance of general words corresponding to specific ones, while π returns how surely the system expects the short texts to adhere to the proportions. Based on the actual path assignments and all other variables, the model samples the variable of level allocation $z_{b,n}$ from its distribution:

$$p(z_{b,n} \mid z_{-(b,n)}, c, w, m, \pi, \eta) \propto p(z_{b,n} \mid z_{b,-n}, m, \pi) p(w_{b,n} \mid z, c, w_{-(b,n)}, \eta) \tag{4}$$

The sampling is performed for the first term in Eq. (4) over the current levels' space (biterm), which is the $\max(z_{b,-n})$. The first distribution components for $k \leq \max(z_{b,-n})$ are:

$$\begin{aligned} p(z_{b,n} = k | z_{b,-n}, m, \pi) &= E\left[V_k \prod_{j=1}^{k-1} (1 - V_j) | z_{b,-n}, m, \pi \right] \\ &= E\big[V_k | z_{b,-n}, m, \pi \big] \prod_{j=1}^{k-1} E\big[(1 - V_j) | z_{b,-n}, m, \pi \big] \\ &= \frac{m\pi + \#\big[z_{b,-n} = k\big]}{\pi + \#\big[z_{b,-n} \geq k\big]} \prod_{j=1}^{k-1} \frac{(1-m)\pi + \#\big[z_{b,-n} > j\big]}{\pi + \#\big[z_{b,-n} \geq j\big]} \end{aligned} \tag{5}$$

where #[...] represents the number of elements satisfying the given condition.

Based on the possible assignment, the word's probability is represented by the second term in Eq. (4). The β_i parameters are produced from a symmetric Dirichlet distribution with hyperparameter η. Hence, the system gets the smoothed frequency of observing word $w_{b,n}$ that contributes to the topic on level $z_{b,n}$ with the path c_d:

$$p(w_{b,n}\ |z, c, w_{-(b,n)}, \eta) \propto \#\big[z_{-(b,n)} = z_{b,n}, c_{z_{b,n}} = c_{b,z_{b,n}}, w_{-(b,n)} = w_{b,n}\big] + \eta \tag{6}$$

The last distribution component is:

$$p(z_{b,n} > max(z_{b,-n})|z_{b,-n}, w, m, \pi, \eta) = 1 - \sum_{j=1}^{max(z_{b,-n})} p(z_{b,n} = j|z_{b,-n}, w, m, \pi, \eta) \tag{7}$$

Each biterm path adapted over all paths and all observed words, is examined after the level allocations variables sampling:

$$p(c_b|w, c_{-b}, z, \eta, \gamma) \propto p(c_b|c_{-b}, \gamma)p(w_b|c, w_{-b}, z, \eta). \tag{8}$$

where $p(w_b|c, w_{-b}, z, \eta)$ is the data probability given the selection of a remarkable path, and the nested CRP indicates the prior on paths as $p(c_b|c_{-b}, \gamma)$.

Given the corpus, the conditional distribution of the latent variables acts as the Markov chain stationary distribution. The overall procedure of Gibbs sampling is summarized in Algorithm 2. Given the current status of the sampler $\{c_{1:B}^{(t)}, z_{1:B}^{(t)}\}$, each feature adapted on the remainder is iteratively sampled.

Algorithm 2 The OBTM's Gibbs sampling procedure

```
for each biterm $b \in \{1, \ldots, B\}$ do
    Randomly draw $c_b^{(t+1)}$ from Eq. (8).
    for each word $i \in b$ / $i \in \{w_1, w_2\}$ do
        Randomly draw $z_{b_i,n}^{(t+1)}$ from Eq. (4).
    end for
end for
```

The values of hyperparameters are included in the inference method by providing them with prior distributions, which also contain parameters and hyperparameters. However, fixing the original hyperparameters, these parameters do not affect the inference.

6 Experiments

In this section, the effectiveness of the proposed approach is examined and verified. Moreover, the proposal's usefulness is also confirmed via complete analyses.

6.1 Datasets

To demonstrate the effectiveness and efficiency of the proposed approach, three different short texts collections are used for evaluation:

- Questions dataset: StackOverflow repository contains ~20,000 questions from 20 different categories.
- Arabic dataset: Arabic tweets collection is a standard short texts collection particularly used in sentiment analysis research. This collection provides approximately ~10,000 Arabic tweets.
- Reviews collection: This dataset contains the surveys of fine foods from Amazon, containing ~500,000 inspections for more than 10 years.

Table 1 highlights the resulted preprocessed data statistics.

6.2 Experimental Results

6.2.1 Evaluating the Quality of Topics

The performance of different models is compared. Topic coherence and Umass-Scores are used to assess the quality of topic-word distribution. The Umass-Scores achievements are shown in Table 2 on the questions and Arabic tweets collections with a variation of the external collection size ranging from 10 to 300. The nearest Umass-Scores to zero indicates a more generative topic model. The proposed model outperforms the BTM method and produces better topics showing a relevant result. Additionally, Table 3 shows the coherence results on the reviews collection with topic size ranging from 20 to 300. The OBTM presents a high coherence overall sizes compared to the BTM. The results demonstrate that the discovered topics using the OBTM system are more coherent than other methods.

The topics content is investigated for qualitative analysis. For each topic, Tables 4 and 5 list the N most probable and representative words for a topic where $N \in \{5, 10, 20\}$. Besides, to examine the topic coherence more comprehensively, a multilingual human judgment is used. On the one hand, on the Arabic topics, the OBTM identifies highly correlated words comparing to BTM, which includes more than two irrelevant words as well as the extracted topics become less coherent when the topic

Table 1 Summary of the three short texts collections

	Questions	Reviews	Arabic
#*Sentences*	19,281	568,455	10,006
#Words	8067	95,889	33,393
AvgDocLen	5.20	39.67	10.72

Table 2 Umass-Scores of the proposed algorithm versus the BTM (the nearest Umass-Scores to zero indicates more generative topic model)

		External collection size			
Collection	Method	10	100	200	300
Question	BTM	−6.85	−17.69	−18.51	−17.84
	OBTM	−0.84	−5.28	−7.65	−9.36
Arabic	BTM	−7.67	−15.05	−17.06	−17.72
	OBTM	−0.44	−2.04	−3.24	−4.08

Table 3 Comparison of topics coherence performance of our model versus the BTM

Topic size	OBTM	BTM
20	0.82	0.73
40	0.79	0.66
50	0.8	0.62
60	0.81	0.61
80	0.78	0.57
100	0.78	0.51
150	0.76	0.55
200	0.81	0.52
250	0.8	0.52
300	0.82	0.53

Contrary to the BTM, the topic size in the OBTM model does not affect the topic coherence performance

Table 4 The *N* most probable words for the Arabic collection

Topic size	OBTM	BTM
5	Datatable + instance + firefox + share + unexpectedly	instance + apache + local + upgrade + install
10	Datatable + instance + firefox + share + unexpectedly + click + simple + project + commit + fade	change + studio + visual + tab + ctrl + move + behavior + document + local + upgrade
20	Multiple ++ database ++ jquery ++ big ++ webpage + really + ignore + oracle + need + find + software + sign + unexpectedly + clipboard + select + state ++ session + autom + water + setup	world + height + nstextview + apache + instal + instance + local + backend + upgrade + bash + foo + liner + txt + system + really + reset + proper + tip + nhibernate + ssis

size gets bigger. On the other hand, using the questions dataset, the OBTM returns more prominent and relevant words contrary to BTM, which includes two separate groups to words (two different subtopics). The proposed approach includes more relevant words that means more coherent topics than BTM.

The proposed approach outputs a hierarchical topics tree, which reveals the collation between topics as well. Figure 1 shows the results of a processed sentence *[Fill*

Table 5 The *N* most probable words for the question collection

Topic Size	OBTM	BTM
5	سيد + المعارضة + حلب + بلد + كاملة Master + Opposition + Aleppo + Country + Complete	جمهورية + انتخابات + اطالب + عمل + استبعاد Republic + elections + demand + action + exclusion
10	الهولوكوست + تشغيل + الوجوه + موقف + فتح الصباع + وفاء + أهلية + جامعة المنصورة الثورة العراقية Holocaust + Run + Face + Position + Open AL Mansoura University + Fulfillment + eligibility The Iraqi Revolution	ضبط + مناديل + أسفر + بخاخات مركز + حوزته + جنسية + مخدرات سوهاج + طبية Tuning + Tissues + Resulting + Sprays Center + possession + nationality + drugs Sohag + Medical
20	الدولة + رعاية + أعد + دستور + منشآت محافظات + خوف + قانوني + المقابلة + السلفيين حنفي + قصد + الإزدواجية + د.عمرو + يمكن الطايفي + عزل + الجهود + أحوال + رفض State + sponsorship + prepared + constitution + facilities Provinces + fear + legal + interview + salafis Hanafi + intention + duality + Dr. Amr + can Taif + isolation + efforts + conditions + rejection	مصر + إفريقيا + الوضوء + المغرب داخل + رووس + فرائض + إلي + الثلاثة + الإعلام + نيوز + اما + انظر المتخصصة + طريق الله + وزيرة + ماسبيرو مصطفى + حركة + ستجدها Egypt + Africa + Ablution + Morocco Inside + Rouss + Statutes + To + Three Media + News + + See specialized + + Way of God + Minister + Mustafa + movement + You will find it

dataset data link query resultset] from the question collection. The model returns three topics levels, where the two first levels are already constructed, and the third level (orange elips) is inferred using the new short sentence. The system accords the two words [*fill—data*] to the first leave, [*dataset—link—query*] to the second leave, and [*resultset*] to the last leave. These results are relevant, where the first two words [*fill—data*] are co-occurred always together in the collection. Moreover, the three words [*dataset—link—query*] are clearly must refer to the same topic because they are a database main concepts. The last word [*resultset*] is pertinent to the third leave since it represents an output of the programming languages such as scala, spring, etc.

6.2.2 Word Similarity Tasks

This section shows the performance comparison between OBTM and other methods, including BTM, LDA, and WNTM, on the word similarity task. The results demonstrate the model's efficiency in building a dense semantic word representation on short texts.

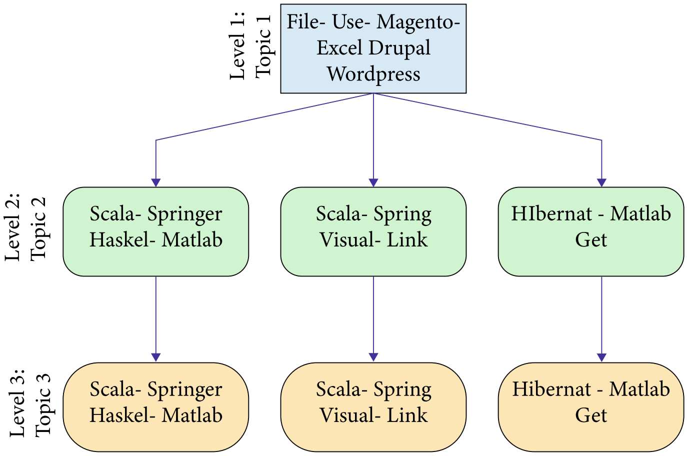


Fig. 1 Inference a topic hierarchy for the sentence [Fill dataset data link query resultset]

Table 6 Samples of biterms with their human similarity score [0,10]

First word	Second word	Similarity
Stock	Jaguar	0.92
Money	Cash	9.15
Development	Issue	3.97
Lad	Brother	4.46

Word similarity tasks evaluate distributional semantic spaces by measuring the notion of word similarity using the word vector representations according to humans' judgment. Table 6 shows an example of biterms along with their similarity from [9]. The accurate learned topics can assume related words such as "bank" and "money" to have similar semantic representations.

The words similarity is evaluated by the JS (Eq. 17) and Cosine (Eq. 18). Higher similarity expresses valuable word semantic forming. We fixed the size of the topics to 150 for all models. The ranked similarity results on the Arabic dataset are highlighted in Fig. 2. The results reveal that the OBTM outperforms the WNTM, BTM, and LDA significantly in both measures JS and Cosine. The OBTM surpasses the WNTM, which operates similarly to BTM on JS and exceeds BTM remarkably on Cosine. The OBTM straight models the word in the corpus level instead of the document/sentence level. Therefore, OBTM can learn more precise word representations compared to the baseline methods.

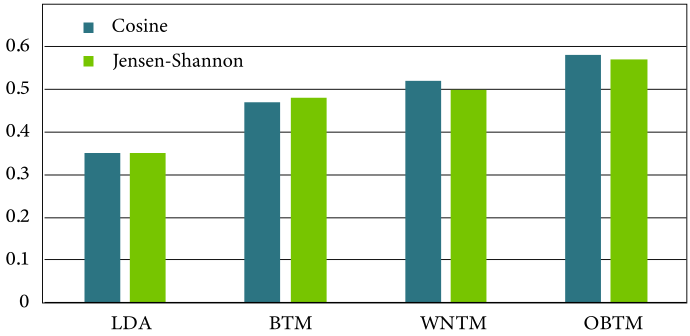


Fig. 2 The correlation results of the word similarity task using the Arabic dataset. The results prove that OBTM outperforms NWTN, LDA, and BTM to discover better semantic word representations over short texts

Notwithstanding, WNTM utilizes the LDA to shape the word's context extracted from the term network. Hence, compared to BTM and WNTM, the OBTM model's assumption is more powerful. Thus although all models can learn good word semantic representations differently, OBTM is more reliable and stable. Despite the coherence achievement in short texts, the OBTM creates dense and high semantic representation for words.

6.2.3 Classification Task

Another critical use of short text topic modeling is the text classification task. To examine the models' strength in learning semantic representation on short texts, this section shows the classification performance evaluation and explores the effectiveness of the novel OBTM's ranking mechanism on BBC news articles. Firstly, the preprocessing step is conducted to clean the input data, which outputs 1780 entries for the training task, and 445 for the test. Table 7 lists different label and their count over the BBC dataset.

Table 7 Label counts for BBC news articles used in document classification tasks

Label	Count
Sport	511
Business	510
Politics	417
Tech	401
Entertainment	386

The accuracy measure represents OBTM's influence on short texts representation examined to different models, including BTM, WNTM, BPDTM, and LDA. The accuracy is the relationship of true positives and true negatives with the total number of cases examined. Formally,

$$Accuracy = \frac{(true\ pos + true\ neg)}{(true\ pos + true\ neg + false\ pos + false\ neg)} \tag{9}$$

Topic modeling has different promising utilization, including dimensionality reduction, where the corpus can be reduced into a fixed collection of topics, featuring for text classification. The input entries are first ranked using the novel ranking method. Then the classification system uses these ranks as features values. The weighted averages accuracy for the used data set is shown in Fig. 3. Overall, the document classification task over short texts suffers from data sparsity, making the LDA accuracy lower than the other models. OBTM accuracy represents less variation than BPDTM, WNTM, and BTM and works powerfully related to the before examined task.

The overall results confirm the OBTM superiority over the state-of-the-art approaches. In particular, LDA works poorly compared to other methods, meaning that it cannot handle the short text topic classification. BPDTM is better than BTM and WNTM, demonstrating that representing words via adjacent biterms is beneficial for obtaining topic structure. Last, the OBTM surpasses the BPDTM by a significant accuracy margin, illustrating that word ranking via the novel mechanism is beneficial for building topic structure and short text classification.

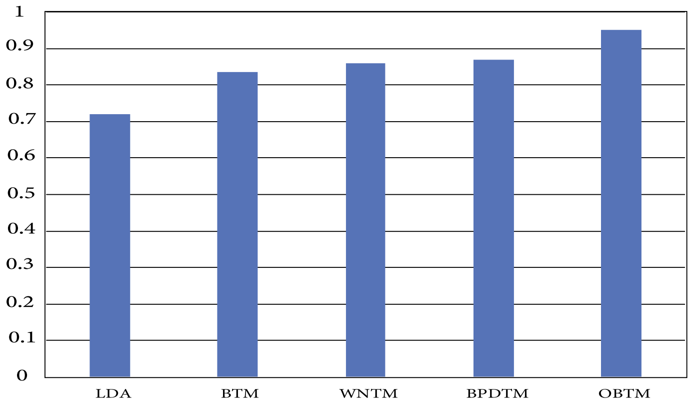


Fig. 3 Classification performance comparison between OBTM, BPDTM, WNTM, BTM, and LDA on short texts on BBC text data. The results prove that the ranking mechanism of OBTM outperforms BPDTM, NWTN, LDA, and BTM on text classification tasks

7 Conclusion

In this paper, a new biterm ranking method for short text topic modeling has been introduced. The proposed method relies on a novel biterm ranking paradigm for topic modeling, which can discover higher topics quality and accurately infer the documents' topic proportions. In contrast to existing methods, this system explicitly models the word co-occurrence patterns in the corpus level rather than document/sentence level. The proposed approach can simultaneously improve semantic relations linking words and overcome the short text's sparsity. Moreover, the learning process based on the biterm ranking paradigm has been extended to efficiently preserve the higher-order semantics between words. Comparisons with other similar systems proved the solidity and reliability of our method. The results show that the discovered topics are more outstanding and coherent than the other systems ad algorithms. Meanwhile, the classification accuracy and word semantic representation are highly caught using the novel OBTM system.

In future work, our model will be extended to resolve the continuous topic modeling in the big data context, which typically involves several issues from different backgrounds.

Appendix

Parameters Tuning

The manual tuning of the α, β hyper-parameters, is avoided, since the model figures out the exact values based on the statistical data distribution. In the preceding experiments, the model always achieves the best performances when $\alpha = 0.01$ and $\beta = 0.001$. The α is used as the prior hyper-parameters in the Dirichlet process to generate topics, while β is used to generate words. The proposed approach determines the number of the hidden topic using a Dirichlet process, where the finite distribution of topics is sampled from a selected common base distribution, which considers the countably infinite set of possible topics. For all baseline models, Table 8 illustrates the experimental setting used in the original paper.

Table 8 The main common parameters setting for OBTM and the baselines methods

Parameter	LDA	BTM	WNTM	BPDTM	OBTM
α	0.05	$\frac{50}{K}$	0.1	$\frac{50}{K}$	0.01
β	0.01	0.01	0.1	0.1	0.001

Evaluation Metrics

Short text topic models evaluation is an open problem, where a lot of metrics have been proposed for measuring the quality of the topics. To provide a good evaluation, this section highlights the evaluation metrics used in this paper.

Topic Coherence

Human topic ranking is the highest standard, and consequently a topic interpretability measure. In recent years, a new automatic evaluation methods are developed to evaluate the topic model quality. The topic coherence reflects the homogeneity for words, which contribute to the topic formulation. The proposed approach adopts the C_V coherence measure proposed by Roder et al. [34]. Notably, the C_V consists of four major parts. The first step is the data segmentation pairs, more formally, let $W = \{w_1, \dots, w_N\}$ be the set of top-N words that describes a topic, then $S_i = \{(W', W^*)|W' = w_i; w_i \in W; W^* = W\}$, is the set of all pairs. For example, if $W = \{w_1, w_2, w_3\}$, then the pair $S_1 = \{(W' = w_1), (W^* = w_1, w_2, w_3)\}$. Douven et al. [7] assume that the segmentation measures the extent to which the subset W^* supports or conversely undermines the subset W'.

The second step retrieves the probability of a single word $p(w_i)$, or the joint probability of two words $p(w_i, w_j)$, which can be guessed using their frequency over the corpus. The coherence measure C_V creates a new virtual document using a frequency sliding window calculation. The window size creates a slid over the document by one-word token per step. The final probabilities $p(w_i)$ and $p(w_i, w_j)$ are calculated from the total number of virtual documents.

Given a pair $S_i = (W', W^*)$, the third step calculates a confirmation measure ϕ, which reflects the strength of W^* supports W'. Similarly to Aletras et al. [1], W' and W^* are represented as a means context vectors, that captures the semantic support of words in W using Eq. (10). Thus, the agreement between individual words w_i and w_j is calculated using Eq. (11) along normalized pointwise mutual information. Furthermore, the log operator is smoothed by adding ϵ to $\log p(w_i, w_j)$, and the γ parameter controls the weight on higher NPMI values. ϕ is the confirmation measure for a given pair S_i which is obtained by calculating the cosine vector similarity of all context vectors $\phi_{S_i}(\vec{u}, \vec{w})$ (Eq. 12).

$$\mathbf{v}(W') = \left\{ \sum_{w_i \in W'} NMPI(w_i, w_j)^{\gamma} \right\}_{j=1,\dots,|W|} \tag{10}$$

$$NMPI(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{p(w_i,w_j)+\epsilon}{p(w_i).p(w_j)}}{-\log p(w_i, w_j) + \epsilon} \right)^{\gamma} \tag{11}$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{|| \vec{u} ||_2 \cdot || \vec{w} ||_2} \tag{12}$$

The final step returns the arithmetic mean of all confirmation measures ϕ as the final coherence score.

Pointwise Mutual Information

Since the C_V metric evaluates the topic models quality internally, the proposed approach evaluated using the PMI-Score [25], which measures the topic coherence based on pointwise mutual information using external sources. Besides, these external data are model-independent, which makes the PMI-Score fair for all topic models. Given a topic k and its n probable words $(w_1, \ldots, w_n)$, the PMI-Score measures the pairwise association between them is:

$$PMI_{Score}(k) = \frac{1}{n(1-n)} \sum_{1<i<j<n} PMI(w_i, w_j) \tag{13}$$

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)} \tag{14}$$

This is an empirical conditional log-probability $\log p(w_j|w_i) = \log \frac{p(w_i, w_j)}{p(w_j)}$ smoothed by adding ϵ to $p(w_i, w_j)$.

Word Similarity

The conditional topic distribution for the word w can be defined as its semantic representation Eq. (15), where the $p(z_k|w)$ value is obtained after completing the Gibbs sampling step (Eq. 16).

$$s_w = [p(z_1|w), p(z_2|w), \ldots, p(z_k|w)], \tag{15}$$

$$p(z_k|w) = \frac{n_{w|z_k}}{n_w} \tag{16}$$

where $n_{w|z_k}$ stands for how many times w has be assigned with topic k during the sampling, and n_w is the total occurrence of w in the given corpus. The distance between two words w_i and w_j represented by their semantic representations s_i and s_j is calculated using the Jensen–Shannon divergence:

$$JS(s_i, s_j) = \frac{1}{2} D_{KL}(s_j||m) + \frac{1}{2} D_{KL}(s_i||m) \tag{17}$$

where $m = \frac{1}{2}(s_i + s_j)$ and $D_{KL}(p||q) = \sum_i p_i \ln \frac{p_i}{q_i}$ is the Kullback–Leibler divergence. The cosine similarity is also adopted to measure the distance between two words' vectors, which is defined as:

$$Cosine(s_i, s_j) = \frac{s_i.s_j}{||s_i|| \quad ||s_j||} \tag{18}$$

References

1. Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th international conference on computational semantics (IWCS 2013)—long papers, pp 13–22
2. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
3. Carlo CM (2004) Markov chain Monte Carlo and Gibbs sampling. Notes
4. Chen M, Jin X, Shen D (2011) Short text classification improved by learning multi-granularity topics. In: Twenty-second international joint conference on artificial intelligence
5. Cheng X, Yan X, Lan Y, Guo J (2014) BTM: topic modeling over short texts. IEEE Trans Knowl Data Eng 26(12):2928–2941
6. Dai Z, Sun A, Liu XY (2013) Crest: cluster-based representation enrichment for short text classification. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 256–267
7. Douven I, Meijs W (2007) Measuring coherence. Synthese 156(3):405–425
8. Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 209–230
9. Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, Wolfman G, Ruppin E (2002) Placing search in context: the concept revisited. ACM Trans Inf Syst 20(1):116–131
10. Gao W, Peng M, Wang H, Zhang Y, Xie Q, Tian G (2018) Incorporating word embeddings into topic modeling of short text. Knowl Inf Syst 1–23
11. Griffiths TL, Jordan MI, Tenenbaum JB, Blei DM (2004) Hierarchical topic models and the nested Chinese restaurant process. In: Advances in neural information processing systems, pp 17–24
12. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235
13. Gruber A, Weiss Y, Rosen-Zvi M (2007) Hidden topic Markov models. In: Artificial intelligence and statistics, pp 163–170
14. Hirchoua B, Ouhbi B, Frikh B (2017) A new knowledge capitalization framework in big data context. In: Proceedings of the 19th international conference on information integration and web-based applications & services, iiWAS'17. Association for Computing Machinery, New York, NY, pp 40–48. https://doi.org/10.1145/3151759.3151780
15. Hirchoua B, Ouhbi B, Frikh B (2019) Topic hierarchies for knowledge capitalization using hierarchical Dirichlet processes in big data context. In: Ezziyyani M (ed) Advanced intelligent systems for sustainable development (AI2SD'2018). Springer International Publishing, Cham, pp 592–608
16. Hoang T, Le H, Quan T (2019) Towards autoencoding variational inference for aspect-based opinion summary. Appl Artif Intell 1–21
17. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 289–296
18. Ishwaran H, James LF (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models. Stat Sin 1211–1235

19. Jiang L, Lu H, Xu M, Wang C (2016) Biterm pseudo document topic model for short text. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 865–872
20. Jin O, Liu NN, Zhao K, Yu Y, Yang Q (2011) Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM international conference on information and knowledge management. ACM, pp 775–784
21. Kou F, Du J, Yang C, Shi Y, Liang M, Xue Z, Li H (2019) A multi-feature probabilistic graphical model for social network semantic search. Neurocomputing 336:67–78
22. Lim KW, Chen C, Buntine W (2016) Twitter-network topic model: a full Bayesian treatment for social network and text modeling. arXiv preprint arXiv:1609.06791
23. Lin T, Tian W, Mei Q, Cheng H (2014) The dual-sparse topic model: mining focused topics and focused terms in short text. In: Proceedings of the 23rd international conference on world wide web. ACM, pp 539–550
24. Lu H, Ge G, Li Y, Wang C, Xie J (2018) Exploiting global semantic similarity biterms for short-text topic discovery. In: 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI), pp 975–982. https://doi.org/10.1109/ICTAI.2018.00151
25. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 262–272
26. Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9(2):249–265
27. Nigam K, McCallum AK, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39(2–3):103–134
28. Pang J, Li X, Xie H, Rao Y (2016) SBTM: topic modeling over short texts. In: International conference on database systems for advanced applications. Springer, pp 43–56
29. Pathak AR, Pandey M, Rautaray S (2021) Topic-level sentiment analysis of social media data using deep learning. Appl Soft Comput 108:107440. https://doi.org/10.1016/j.asoc.2021.107440
30. Phan XH, Nguyen LM, Horiguchi S (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on world wide web. ACM, pp 91–100
31. Pitman J et al (2002) Combinatorial stochastic processes. Technical report 621. Department of Statistics, UC Berkeley. Lecture notes for St. Flour course
32. Qiang J, Chen P, Ding W, Wang T, Xie F, Wu X (2016) Topic discovery from heterogeneous texts. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 196–203
33. Razavi AH, Inkpen D (2014) Text representation using multi-level latent Dirichlet allocation. In: Canadian conference on artificial intelligence. Springer, pp 215–226
34. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining. ACM, pp 399–408
35. Ruan D, Han J, Dang Y, Zhang S, Gao K (2017) Modeling on micro-blog topic detection based on semantic dependency. In: 2017 9th international conference on modelling, identification and control (ICMIC), pp 839–844. https://doi.org/10.1109/ICMIC.2017.8321571
36. Shen Y, Zhang Q, Zhang J, Huang J, Lu Y, Lei K (2018) Improving medical short text classification with semantic expansion using word-cluster embedding. In: International conference on information science and applications. Springer, pp 401–411
37. Shi T, Kang K, Choo J, Reddy CK (2018) Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: Proceedings of the 2018 world wide web conference on world wide web. International World Wide Web Conferences Steering Committee, pp 1105–1114
38. Škrlj B, Martinc M, Kralj J, Lavrač N, Pollak S (2019) tax2vec: constructing interpretable features from taxonomies for short text classification. arXiv preprint arXiv:1902.00438

39. Sridhar VKR (2015) Unsupervised topic modeling for short texts using distributed representations of words. In: Proceedings of the 1st workshop on vector space modeling for natural language processing, pp 192–200
40. Vo DT, Ock CY (2015) Learning to classify short text from scientific documents using topic models with various types of knowledge. Expert Syst Appl 42(3):1684–1698
41. Yi F, Jiang B, Wu J (2020) Topic modeling for short texts via word embedding and document correlation. IEEE Access 8:30692–30705. https://doi.org/10.1109/ACCESS.2020.2973207
42. Zeng J, Li J, Song Y, Gao C, Lyu MR, King I (2018) Topic memory networks for short text classification. arXiv preprint arXiv:1809.03664
43. Zhao X, Wang D, Zhao Z, Liu W, Lu C, Zhuang F (2021) A neural topic model with word vectors and entity vectors for short texts. Inf Process Manag 58(2):102455. https://doi.org/10.1016/j.ipm.2020.102455
44. Zuo Y, Zhao J, Xu K (2016) Word network topic model: a simple but general solution for short and imbalanced texts. Knowl Inf Syst 48(2):379–398

Prediction and Analysis of Moroccan Elections Using Sentiment Analysis

Ahmed Oussous, Zakaria Boulouard, and Benjelloun Fatima Zahra

Abstract Sentiment analysis (SA) or opinion mining constitutes an important scientific field that uses advanced methods to mine population's views and determine their feelings. In fact, SA is exploited by many fields. In politics, SA can mine the citizens' opinions posted online and use this data as a means to predict the results of an ongoing election. Morocco will witness its legislative elections, early on September 8, 2021, and many Moroccans use web sources like social media and electronic journals to express their opinions in favor or against a certain party's candidates. The objective of this paper is to mine and analyze citizens' sentiments towards Morocco's upcoming elections in general, as well as their favorite parties and candidates. The first step will be performing a sentiment analysis on posts or comments related to the elections from the Moroccan Hespress website using supervised machine learning (ML) techniques. We will apply these algorithms on 3503 comments being mined during two months; June and July 2021 in order to get the sentiment polarity of the commentators. The second step will be to implement some ML-based classifiers in order to determine the polarity of each comment. The third step a thematic analysis on both positive and negative comments that will reveal the public favorite parties and candidates, as well as the motives behind their choices. Finally, we discuss our analytical findings. The results of this study offer a precise view of Moroccans' opinion about the Moroccan elections. Such extracted insight can be exploited by electoral candidates in their political campaigns to improve their campaign strategy and to attract the largest possible number of people to vote on their electoral program.

Keywords Politics · Election · Presidential candidate · Sentiment analysis · Natural language processing · Supervised machine learning

A. Oussous (✉) · Z. Boulouard
LIM, Hassan II University, Casablanca, Morocco
e-mail: ahmed.oussous@fstm.ac.ma

Z. Boulouard
e-mail: zakaria.boulouard@fstm.ac.ma

B. F. Zahra
LGS, Ibn Tofail University, Kenitra, Morocco

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_30

1 Introduction

Sentiment analysis (SA) is a valuable tool for exploring big data and extracting valuable information. This is very helpful for many sectors. Indeed, directors and managers can get a precious insight to improve their strategies or better understand in nearly real-time what is going within their entity and surrounding environment. That is why an increasing number of scientists have taken interest in this research area.

Indeed, sentiment analysis is a discipline that combines the best of Natural Language Processing (NLP), Information Retrieval (IR), and linguistics. The main objective of SA is to mine texts for opinions and to classify them according to their semantic orientations that can be either positive or negative [1].

Research is active in the field of sentiment analysis in the sense that we are witnessing an abundance of textual data sources comprising opinions on the web (search engines, forums, social networks, etc.). The online analysis of the posted opinions is necessary in order to extract at an overview of citizens' feeling regarding a given subject.

These opinions are of great importance to companies as they allow them to adjust their strategies and improve their products or brands according to the results revealed by the advanced sentiment analysis of the available data.

In general, Sentiment Analysis is of particular interest to those interested in discovering or analyzing public opinion. This could be for personal, business or political reasons [2].

For many years, researchers used SA to explore voters' opinions over popular social media networks such as Facebook or Twitter. Several studies were able to provide thorough predictions and analyses of the elections in several countries. Examples of such elections include the US Senate and Presidential Elections [3, 4], the Dutch Senate Elections [5], Pakistani General Elections [6], Presidential Election in Indonesia [7], and General Election in the United Kingdom [8].

The literature has presented some work that went deeper enough to provide explanations behind their poll result predictions. For instance, [9] found that the voters are more likely to distance themselves from a failing candidate and be more inclined to vote for a more promising counterpart, while others tend to remain loyal to their favorite party or candidate.

Arab countries are no exception to this rule, they have a large number of social media users. The Arabic language is used in many countries. In fact, The Arab population represents around 5.6% of the world's population and around 4.8% of internet users.[1]

In our study, we explored the possibility to apply SA on the Arabic language as a means to classify Arabic texts at sentence level, as the data on social media can be either a blog, a post, a comment, an online review, a tweet, etc.

The choice of Arabic is motivated by different reasons: it is the native language of different countries, as well as millions of people around the world [10]. Moreover,

[1] http://www.internetworldstats.com/stats19.html (last viewed 02/09/2021).

it holds rich historical and cultural backgrounds that have a great influence on every social aspect of the Arab speaking community. Furthermore, its complex morphology and structure make Arabic a very challenging language when it comes to SA related tasks.

The next general election will be held in Morocco on September 8, 2021. Consequently, discussion and prediction about which party will win the election in Morocco has become a hot and interesting conversation among Moroccan citizens over the past few months.

Throughout this period, many Moroccans have been using social media, blogs and electronic journals to express their opinions in favor or against the parties candidates.

Therefore, this paper aims to identify and analyze public sentiments towards the Moroccan general election in general, and towards candidates in particular. The goal is to determine their chances of being elected based on reviews and comments of Moroccan peoples.

We organized the rest of this chapter as follows: Sect. 2 provides an overview of related works. Section 3 describes the solution we proposed, while Sect. 4 presents the results of our study as well as their interpretations. The last section concludes this work and opens up to ideas for future works.

2 Related Works

During last years, sentiment analysis was used in many fields for various purposes that aims to extract insight based on opinions. Many researches tackled social media in order to mine data for election purposes. For instance, they used SA for analyzing public sentiments regarding different candidates and for predicting election results.

Bansal et al. [11] introduced a method called "Hybrid Topic Based Sentiment Analysis" (or HTBSA) and applied it to predict election results based on short tweets. This method bases its analysis on capturing word relations and co-occurrences in the abovementioned tweets.

In [12], DiGrazia et al. demonstrated the strong correlation between the presence of a US candidate on Twitter and his electoral votes' share.

Safiullah et al. [13] tried to assess the influence of Twitter on the outcome of the 2014 general elections in India. They collected 8,877,275 tweets between January 01, 2014 and April 09, 2014 concerning 12 Indian political parties. They found that social media buzz had a great impact on the 2014 general election poll results.

Marozzo et al. [14] analyzed tweets related to the Italian constitutional referendum of December 4, 2016. Their study displayed the voters' intentions weeks before the voting day.

Vepsäläinen et al. [15] demonstrated that Facebook "Likes" were a weak indicator of the voters' tendency in the 2015 Finnish parliament elections.

Razzaq et al. [6] provided a thorough analysis on the impact of social media in predicting the results of the general election in Pakistan. Their analysis covered different aspects such as the individual's political behavior, sentiment detection, as well as tweet classification.

Oyebode et al. [16] applied lexicon based solutions as well as supervised machine learning to analyze public sentiments towards two favorite candidates for the Nigerian presidential elections, and to determine their chances of winning.

Cram et al. [17] mined and analyzed over 35 million tweets to provide an overview on the tendencies related to the candidates for the British general elections of 2017.

Budiharto et al. [7] proposed a new approach to predict the polarity of sentiments based on counting important data and top words. They applied their approach on tweets that focused on the Indonesian elections of 2019 and were able to predict the favorite candidates.

Wang et al. [18] analyzed French tweets to predict the outcome of the presidential election of 2017 in France and to scout the popularity of the favorite candidates.

A more thorough overview covering a large scope of approaches and studies related to the prediction of poll results in different countries worldwide is available on this survey by Chauhan et al. [19].

Our paper focuses on the Moroccan election; it proposes a new framework for sentiment analysis to predict the election result. It is based on posted opinions and comments related to the Moroccan Election in 2021. Our framework combines tags counting and sentiment analysis as the pre-processing work.

3 Methodology

3.1 Research Overview

The research process mainly includes four steps: the first step is to collect opinions related to the Moroccan general election 2021. The second step is text processing of the collected data using our Framework "ASA Arabic" [20]. The third step is to construct the features model. The fourth and last step is to assess the public opinions based on words frequency. In the following subsections, we describe each step in detail.

3.2 Data Collection

As explained in the previous section, numerous research works have focused on Sentiment Analysis (SA) during the last years. A part of this work was concentrated on Arabic. Building a new corpus for Arabic language can be a great contribution to the SA field since there is a growing audience generating each day millions of Arabic opinions, tweets and comments.

We created a publicly available SA dataset by collecting reviewers' opinions from Hespress website (An Arabic-language Moroccan online news website) against various published articles dealing with Moroccan general election. We used Hespress[2] as a data source because it is the Morocco's first, largest, and most popular news website in Morocco.

To do that, we implemented a web scraping solution based on Python. It helped us collect 3503 comments from news articles published on Hespress between June 1 and July 30, 2021 and were directly related to the Moroccan elections.

The goal is to understand the behavior and feeling of the Moroccan people regarding various topics including Moroccan election, governmental decisions and trust in political parties.

3.3 Preprocessing

Comments and reviews are usually challenging when it comes to natural language processing. The words are usually misspelled, or may include duplicated characters. Another challenging situation is the use of slang and "urban" words that may not exist on common dictionaries. Applying direct SA on such texts may yield poor results. It is important to note that Arabic is a challenging language due to its complex morphology and structure that require careful pre-processing [21].

Before classifying a certain text, a data pre-processing stage is needed.

For that, we have used our framework ASA (Arabic Sentiment Analysis) [20], which include many pre-processing task such as:

- Cleaning: in order to improve the detection of polarity and to guarantee data quality, we have cleaned our data by removing irrelevant items such as tags and non-textual contents. Replications of usernames, hashtags, links, and unwanted white spaces are also removed.
- Tokenization: it consists of splitting the text into words (tokens) separated by whitespaces or punctuation characters. The result of this operation is a set of words.
- Normalization: Arabic is a language where a word can be represented in different forms for different reasons (tenses, gender, plurality, etc. …). Reducing this word to its basic form ensures the consistency needed for Sentiment Analysis.
- Stop Words Removal: Words that do not carry a meaning themselves, and yet have their role to play in sentences are called stop words. Conjunctions, articles, or prepositions are among the most common stop words in a text. For text mining or sentiment analysis, these words are not useful, so we remove them in the pre-processing step.

[2] https://www.hespress.com/ (last viewed 02/09/2021).

3.4 Feature Selection

It is a technique of data cleaning that removes noisy data and only leaves the most relevant ones. It reduces dimensionality and processing time [22].

During this step, the text is transformed to a vector representation, where the weight of the word (feature) is calculated according to the document containing that word.

The literature present several types of weighting. The most popular are: Term Frequency (TF), Boolean, Inverse Documents Frequency (IDF), and Term Frequency Inverse Document Frequency (TF-IDF).

Boolean (or binary) weighting is the simplest one since it returns 1 in case the word is present on the document, or 0 otherwise.

TF computes the number of times a word appears in a document, while IDF inspects whether this word is common across the documents.

TF-IDF is a combination of both the abovementioned techniques.

For our study, TF is the feature selection technique we adopted to produce our feature vectors.

An n-gram is one of many text features considered for Sentiment analysis; it represents a contiguous sequence of n terms from a given sequence of text. For a better performance, we adopted both unigram and bigram models in our experiments.

3.5 Classification

To predict and determine the sentiment of our data collected, we used the trained model of [23] which is a Sentiment analysis dataset of Arabic Facebook comments about the Moroccan elections of 2016, with a size of 10,254 comments divided into positive and negative. Our framework offers various classification algorithms. We developed five ML models to predict sentiment polarity using SVM, Naive Bayes (NB), Adaboost, Logistic regression (LR).

4 Experimental Results and Discussions

Since people are spending more time on social media and less outside home, political parties need to invest more on their online presence so they can increase their potential reach. This can either be by posts on their own social media pages, or through web news regarding their activities. Both options provide comment sections where the common people (potential voters) can express their opinion in favor or against parties or candidates.

These comments constitute valuable data sources for sentiment analysis purposes as they can predict which party or candidate is more likely to win an election for example.

In our study, we calculated the frequency of words, the scores of negative emotional indicators (negative sentiments) and positive emotional indicators (positive sentiments) of the various posts and comments collected from data.

The goal is to analyze the attitudes, interests and behavior of Moroccans regarding the 2021 Moroccan general election. Another goal is to identify also all discussed subjects during this period in Morocco and the sentiments of the people towards the parties and their electoral program.

The data mining of the collected data was performed using Python language and a word cloud was generated which describes the sentiments expressed in the tweets.

4.1 Frequency and Distribution of Search Keywords Related to the Moroccan General Election

Figure 1 shows the tag cloud (Word cloud) which represents the keywords most used by Moroccans during two months, June and July 2021.

According to Fig. 1, the most dominant tags are «السياسة، المغاربة،الاحزاب، البرلمان، الانتخابات، »(Elections, parties, Moroccans, politics, Parliament). These tags show that Moroccans have been following the elections since the beginning of election campaigns in Morocco.

In fact, the public has started to seek information on the new candidates to know about their electoral programs.

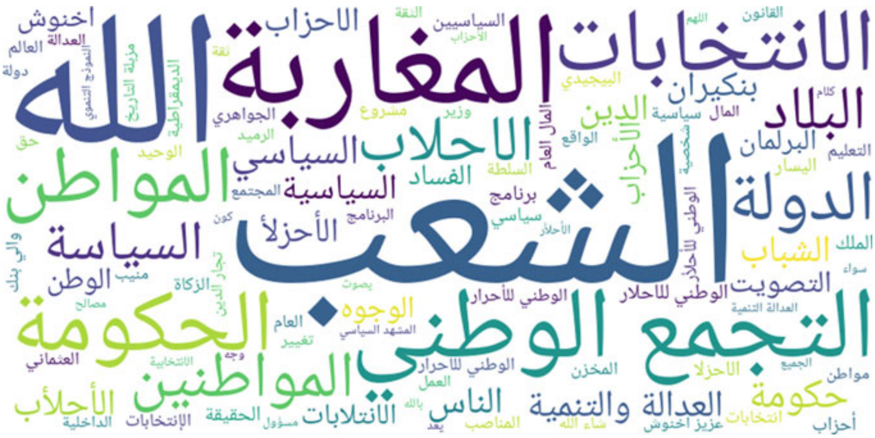


Fig. 1 Word Cloud representing the most used keywords by Moroccans in June and July 2021

Upon the Election campaigns start in Morocco (where the tag الانتخابات(Elections), and البرنامج(the program) comes from), all users began to share information and discuss the subject on the web. This reflects their interest about the political scene in Morocco in general and about Moroccan elections in particular.

Indeed, Moroccan citizens are starting to express their clear and frequent opinions on what they expect from political parties in general, and their voters in particular, in order to support them before the next elections, which explain the appearance of tags التعليم، الفساد الثقة، المواطنين، الحكومة، الشعب،(People, government, citizens, trust, education, corruption).

With the launch of the electoral campaigns in Morocco, which are being waged by political parties, the analysis of Fig. 1 shows that the most important competing parties will be the Justice and Development Party and the National Rally of Independents. These parties are participating in the current government coalition. As shown in Fig. 1, the main party related keywords appearing are التجمع الوطني(National Rally of Independents party), العدالة و التنمية(Justice and Development Party). Figure 1 also highlights, Albeit to a lesser extent and the Federation of the Democratic Left (اليسار).

The figure also shows some of the candidates and representatives of these political parties in Morocco, like the tags اخنوش and عزيزاخنوش(Aziz Akhannouch) which is a Moroccan businessperson, Secretary-General of the National Rally of Independents party, and current Minister of Agriculture since 2007. Tags also refer to بنكيران(Abdelilah Benkirane), a former Prime Minister of Morocco (from November 2011 to March 2017), as well as منيب(Nabila Mounib) a Moroccan female politician who currently serves as the Secretary-General of the Unified Socialist Party (PSU).

In the next section, we will identify and analyze public sentiments towards the two most dominant political parties (as appeared in the word cloud of Fig. 1), which are the Justice and Development Party (JDP), and the National Rally of Independents (NRI). Based on comments on the Moroccan web news platform “Hespress,” we will try to determine which of NRI and JDP parties will have greater chances of winning the Moroccan elections. We will also analyze the popular candidates of each of these parties.

4.2 Sentiments’ Analysis Related to the Election Based on the Moroccans Comments

In order to study in more detail the reactions, opinions, and sentiments expressed about the Moroccan election, we used our corpus of collected data. Our corpus is then imported into our ASA framework. The ASA framework was used to classify the sentiments regarding the Moroccan election as either positive or negative. Figure 2 illustrates the overall sentiment classification (i.e. Positive versus negative) based on the application of five supervised ML algorithms (NB, Adaboost, SVM, LR).

The advanced analysis confirms that the hot subjects related to the Moroccan election tended to be negative.

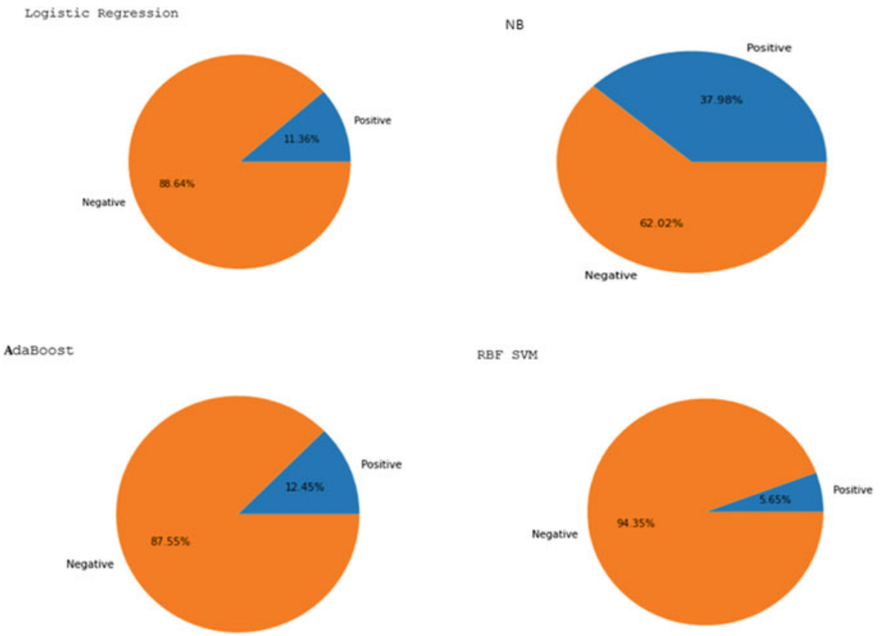


Fig. 2 Sentiment classification results

In fact, the prediction of all the classifiers used gives the negative feeling. In fact, negative sentiments represented 94.35%, 62.02%, 87.55%, and 88.64% for SVM, NB, Adaboost, and Logistic Regression respectively. While, the percentage of positive sentiments was 5.65%, 37.98%, 12.45% and 11.36% respectively.

Moroccan citizens have expressed clear and frequent opinions about what they want from political parties in general and from their electors in particular, in order to support them before the upcoming elections.

The results of this study indicate a high level of lack of confidence among Moroccans in political parties, especially on the part of young people, who constitute the large group of Moroccans.

Based on Fig. 2, it becomes clear that Moroccan youth do not wait for hope of change, as the gap widens day after day between youth and political parties. This is due to the loss of their confidence over the past years in the party institutions and in the electoral process as a whole. The latters are considered by most young people as a political game that is far from their preoccupations and does not respond to their expectations regarding many areas including employment, health and education. This is why the tags التغيير الواقع, التعليم, الثقة, الشباب appeared in Fig. 1.

This is confirmed by the statements of the Governor of the Central Bank of Morocco, Abdellatif Jouahri, during a press conference. He affirmed that people no longer trust parties, and that the reluctance to vote is increasing, considering that the main problem is a problem of trust not only in politicians, but even in those who belong to the public sector, which justifies the appearance of the tags (Jouahri) "الجواهري" ،"والي البنك"(Bank Governor) in Fig. 1.

Most see that they only want realistic, achievable promises that can be handled. Moreover, participants relate their intention to participate in the upcoming local elections with the extent to which political parties will respond to citizens' needs. Most of the participants expressed their desire to see the actions of the parties on the ground, the commitments with ordinary citizens and the start of launching concrete projects. A large number of participants also believe that the parties should give the youth the opportunity to participate in decision-making within the party system. A number of them also expressed their desire for the parties to open more of their branches and to have a presence throughout the year.

Among the reasons for the lack of confidence in political parties, is the lack of renewal of the elites and the absence of democracy, which makes partisan youth reluctant to vote and party affiliation. The political discourse of most parties is still classic and does not take into account the social transformations that the country is experiencing. The digital reputation and presence have become an important factor for political parties to ensure effective communication with Moroccans (The tags السياسيين(politicians), البرنامج(program), الديموقراطية(democracy)).

Consequently, political parties are forced to improve their work and create communicative and organizational mechanisms that restore confidence in them.

Some Moroccans accuse the parties of corruption, as they adopt practices such as buying votes during elections and adopting nepotism in their internal policies, while others sell their votes for money in order to vote on a particular candidate.

Some people also see that the only goal of the candidates is to reach positions of responsibility and authority and to achieve their personal goals, not to reform and develop the country. They also call for holding public funds looters accountable and abiding by the responsibility-accountability nexus (Tags: الفساد(corruption), العام المال(public money), السلطة ,الدين تجار(power) in Fig. 1).

As we have shown, the two parties that appear mostly in Fig. 1 are the National Rally of Independents party (NRI) and the Justice and Development Party (JDP). So, we tried to analyze the trends and patterns about people's sentiments towards these popular political parties.

Figure 3 shows the result of sentiment analysis from popular candidate parties. We notice that the NRI is a more popular party than the JDP. This is because the number of positive opinions for the NRI party far exceeds those in favor of the JDP. This means that among web users who comment political commentary pertaining to the 2021 Moroccan Elections, there are more mentions (positive) of the NRI party.

Therefore, it seems that NRI has a more positive response from Moroccan citizens, while JDP has a more negative sentiment. This comes from the fact that JDP has some negative issues either within the party itself, or among its supporters.

Another reason for this result is that NRI is conducting a good electoral campaign compared to the JDP Party. For example, it is sufficient to compare the National Rally Party's campaign on Facebook with its counterpart the JDP Party.

As we can see in Figs. 4 and 5 that NRI Party is more popular on Facebook (1,455,761 likes) compared to JDP (1,191,030 likes).

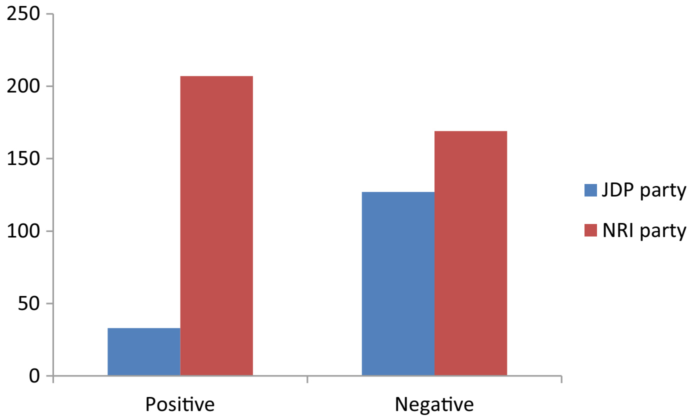
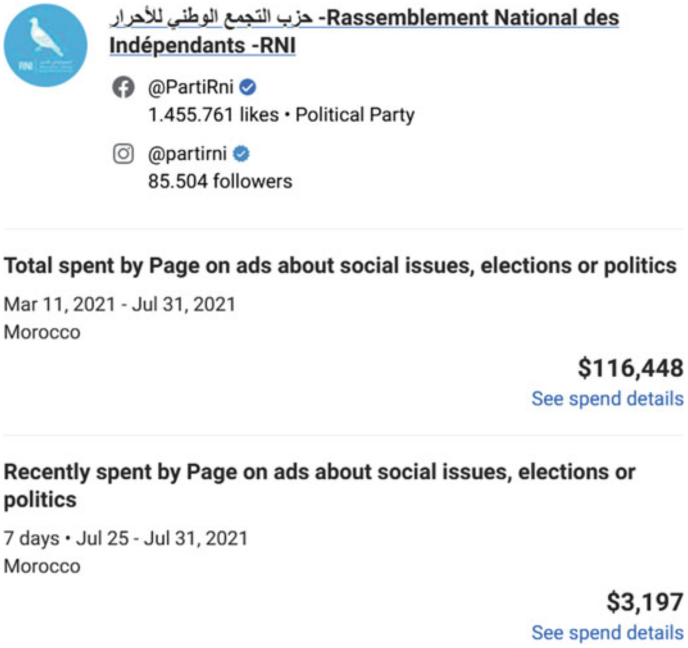


Fig. 3 Sentiment analysis of candidate parties based on comments from popular tags



حزب التجمع الوطني للأحرار -Rassemblement National des Indépendants -RNI

@PartiRni
1.455.761 likes • Political Party

@partirni
85.504 followers

Total spent by Page on ads about social issues, elections or politics

Mar 11, 2021 - Jul 31, 2021
Morocco

$116,448
See spend details

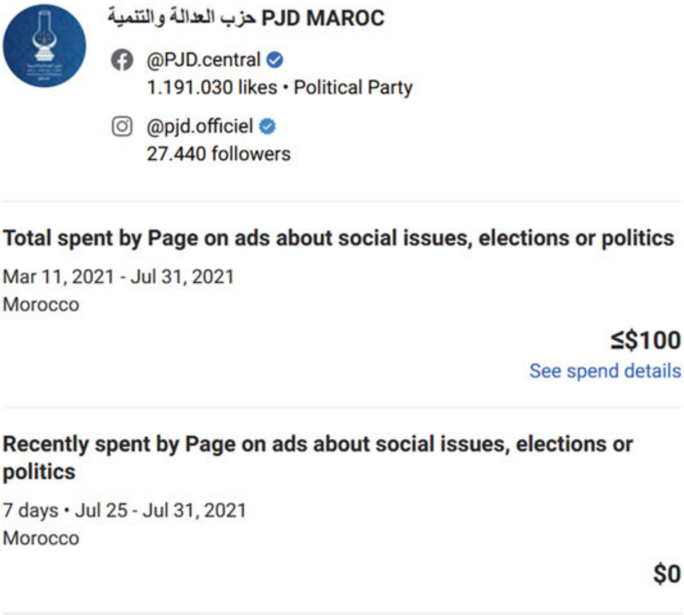
Recently spent by Page on ads about social issues, elections or politics

7 days • Jul 25 - Jul 31, 2021
Morocco

$3,197
See spend details

Fig. 4 Data behind the Ad for RNI page

As with the traditional electoral campaign, the virtual campaign is allocated various tools, funds, and personnel to monitor it, in order to persuade the electorate to vote on the party's electoral list. This is what the NRI has done, as it realized the importance of Social media in its campaigns, in parallel with its work on the ground.



حزب العدالة والتنمية PJD MAROC

@PJD.central
1.191.030 likes • Political Party

@pjd.officiel
27.440 followers

Total spent by Page on ads about social issues, elections or politics

Mar 11, 2021 - Jul 31, 2021
Morocco

≤$100
See spend details

Recently spent by Page on ads about social issues, elections or politics

7 days • Jul 25 - Jul 31, 2021
Morocco

$0

Fig. 5 Data behind the Ad for JDP page

This is what Figs. 4 and 5 embody; we can see that the total spent by RNI Page on ads about social issues, elections or politics is greater than what the JDP page spent.

The above figures and results mean that based on the political views of users about these political parties and the analysis of Facebook pages, the NRI party has a great potential to gain a comfortable victory over the JDP.

5 Conclusion

Often overlooked, electronic journals have shown a great potential as a reliable source of data for sentiment analysis purposes. Especially if the study concerns general opinion subjects, such as elections or other political events.

In this research, we focused on data related to 2021 Moroccan election, in order to understand trends and patterns regarding the Moroccan elections, we have collected data from Hespress from the beginning of June 2021 until the end of July 2021 (a month before the elections day).

Using these data, we were able to provide a solution that can predict the election results. Our method focuses on word cloud and sentiment analysis based on the advanced machine learning approach.

The experimental results predicted that the favorite party would be RNI as it emerges as the most popular among the comments from the analyzed data.

References

1. Oussous A, Lahce AA, Belfkih S (2018) Improving sentiment analysis of moroccan tweets using ensemble learning. In: International conference on big data, cloud and applications. Springer, Cham, pp 91–104
2. Duwairi RM, Marji R, Sha'ban N, Rushaidat S (2014) Sentiment analysis in arabic tweets. In: 2014 5th international conference on information and communication systems (ICICS), pp 1–6. IEEE
3. Shi L, Agarwal N, Agrawal A, Garg R, Spoelstra J (2012) Predicting US primary elections with Twitter. http://snap.stanford.edu/social2012/papers/shi.pdf
4. Macafee T, McLaughlin B, Rodriguez NS (2019) Winning on social media: candidate social-mediated communication and voting during the 2016 US presidential election. Social Media + Soc 5(1):2056305119826130
5. Sang ETK, Bos J (2012) Predicting the 2011 Dutch senate election results with twitter. In: Proceedings of the workshop on semantic analysis in social media, pp 53–60
6. Razzaq MA, Qamar AM, Bilal HSM (2014) Prediction and analysis of Pakistan election 2013 based on sentiment analysis. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014). IEEE, , pp 700–703
7. Budiharto W, Meiliana M (2018) Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. J Big data 5(1):1–10
8. Burnap P, Gibson R, Sloan L, Southern R, Williams M (2016) 140 characters to victory?: Using Twitter to predict the UK 2015 general election. Elect Stud 41:230–233
9. Boatwright B, Mazer JP, Beach S (2019) The 2016 US presidential election and transition events: a social media volume and sentiment analysis. South Commun J 84(3):196–209
10. Oussous A, Lahcen AA, Belfkih S (2019) Impact of text pre-processing and ensemble learning on Arabic sentiment analysis. In: Proceedings of the 2nd International conference on networking, information systems & security, pp 1–9
11. Bansal B, Srivastava S (2018) On predicting elections with hybrid topic based sentiment analysis of tweets. Procedia Comput Sci 135:346–353
12. DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More tweets, more votes: social media as a quantitative indicator of political behavior. PloS One 8(11):e79449
13. Safiullah M, Pathak P, Singh S, Anshul A (2017) Social media as an upcoming tool for political marketing effectiveness. Asia Pac Manag Rev 22(1):10–15
14. Marozzo F, Bessi A (2018) Analyzing polarization of social media users and news sites during political campaigns. Soc Netw Anal Min 8(1):1–13
15. Vepsäläinen T, Li H, Suomi R (2017) Facebook likes and public opinion: predicting the 2015 Finnish parliamentary elections. Gov Inf Q 34(3):524–532
16. Oyebode O, Orji R (2019) Social media and sentiment analysis: the Nigeria presidential election 2019. In: 2019 IEEE 10th annual information technology, electronics and mobile communication conference (IEMCON), pp 0140–0146. IEEE, Oct 2019
17. Cram L, Llewellyn C, Hill R, Magdy W (2017) UK general election 2017: a Twitter analysis. arXiv preprint arXiv:1706.02271
18. Wang L, Gan JQ (2017) Prediction of the 2017 French election based on Twitter data analysis. In: 2017 9th computer science and electronic engineering (CEEC). IEEE, Sept 2017, pp 89–93
19. Chauhan P, Sharma N, Sikka G (2021) The emergence of social media data and sentiment analysis in election prediction. J Ambient Intell Humaniz Comput 12(2):2601–2627
20. Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S (2020) ASA: a framework for Arabic sentiment analysis. J Inf Sci 46(4):544–559
21. Boudad N, Faizi R, Thami ROH, Chiheb R (2018) Sentiment analysis in Arabic: a review of the literature. Ain Shams Eng J 9(4):2479–2490
22. Liu B, Zhang L (2012) A survey of opinion mining and sentiment analysis. In: Mining text data. Springer, Boston, pp 415–463
23. lecMorocco (2017) Dataset of facebook comments about moroccan elections 2016. https://github.com/sentiprojects/elecmorocco2016

Analysis of COVID-19 Trends in Bangladesh: A Machine Learning Analysis

Nishat Ahmed Samrin, Md. Mahmudul Hasan Suzan, Md. Selim Hossain, Mohammad Sarwar Hossain Mollah, and Md. Dulal Haque

Abstract The globe has reached a critical juncture in the last recent years. According to data we collected from an official internet source, Bangladesh recorded 913,258 confirmed cases, 14,646 death cases with a 1.60% mortality rate, and 85% recovery rate as of June 30, 2021. Furthermore, the delta variant currently has a significant impact on improving the current COVID situation in Bangladesh. So, one of the efficient ways to prevent this outbreak is to building multiple Bangladesh outbreak prediction models to analyze historical data and predict with it for making decisions and implementing appropriate COVID-19 control measures. In this study, a machine learning model, an Auto-Regressive Integrated Moving Average (ARIMA) and Prophet, were developed using time series analysis to forecast new cases in Bangladesh in the coming days. This study examined the model outputs, compared their performance, and created predicted values from these models using the Python programming language. The ARIMA model is the best fit model among the algorithms used to predict the new COVID-19 situation in Bangladesh. The primary goals of this paper are to analyze COVID-19 trends and predict the new upcoming cases and assist decision-makers in controlling the Bangladesh outbreak.

N. A. Samrin · Md. M. H. Suzan · Md. S. Hossain (✉) · M. S. H. Mollah
Department of Computing and Information System, Daffodil International University, Dhaka, Bangladesh

N. A. Samrin
e-mail: rnishat16-380@diu.edu.bd

Md. M. H. Suzan
e-mail: mahmudul16-410@diu.edu.bd

M. S. H. Mollah
e-mail: headcis@daffodilvarsity.edu.bd

Md. D. Haque
Department of Electronics and Communication Engineering, Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur 5200, Bangladesh
e-mail: dhaque@hstu.ac.bd

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_31

Keywords COVID-19 · Pandemic · Prediction · Machine learning · Python · ARIMA model · MAD · RMSE · MAPE

1 Introduction

In recent years, the world has seen various changes in human life. The effect of COVID-19 has very severely hit on health, education, economy, jobs, and many more things. This pandemic situation threatens to reverse decades of growth in the fight against poverty, economically and educationally. To restrain this quick spread of the virus, many countries were prosecuted for lockdown. The effect of the coronavirus case has first emerged in China. The virus has broken out in almost 220 countries around the world, including Bangladesh. As of July 31, 2020, there were 17,106,007 confirmed cases and 668,910 deaths worldwide. COVID-19 has become a global public health threat. As of July 31, 2020, there were 17,106,007 confirmed cases and 668,910 deaths worldwide. The pandemic situation has sickened almost 176 million people in the world. In Bangladesh, the first Coronavirus was discovered on March 8, 2020. Due to this pandemic situation, like other countries globally, Bangladesh has been obliged to implement mandatory lockdowns such as house quarantine and isolation, social distancing, and local or international flying bans. At the start of the novel, the coronavirus rate in Bangladesh was slowly increasing. But the infected cases got worse from June 2020. So far, six types of vaccines are approved for use—Pfizer-BioNTech vaccine, Sputnik V vaccine, Johnson & Johnson vaccine, Serum Institute of India Covishield vaccine, Sinopharm vaccine, Sinovac vaccine. It is still difficult to cover the whole country under vaccination because most vaccines still have not arrived. Lack of awareness and consciousness is the main reason for this COVID-19 rising rapidly. Even after observing a significant number of months, not many notable changes were seen in the COVID-19 situation. It is tough to say how much COVID-19 will affect in the future. However, Bangladesh, a lower-middle-income economy with one of the world's most populations, faces numerous obstacles. Without predicting the new cases, it is difficult to control the pandemic situation. This paper presents data analysis, visualization, and prediction of new cases in Bangladesh. Further sections will explain the literature review, methodology, commentary, discussion, comparison, and conclusion. Figure 1 illustrates line plots of the confirmed, death, and recovered cases with lab tests of Bangladesh from March 8, 2020, to June 30, 2021. In Table 1, we show division-wise confirmed cases with case rates.

It is observed that Khulna's confirmed cases rate has the top rate of 19.94% of the eight-division, while Sylhet has the lowest confirmed cases rate of 9.75%. Coronavirus in Bangladesh. We used Prophet and ARIMA to choose a better predictive model for the Covid dataset. We believe our prediction can assist authorities and decision-makers in determining how long the lockdown should last, how substantial meals should be imported, how effectively medical services we need, and so on. Some research tried to predict the COVID-19 situation with the machine learning

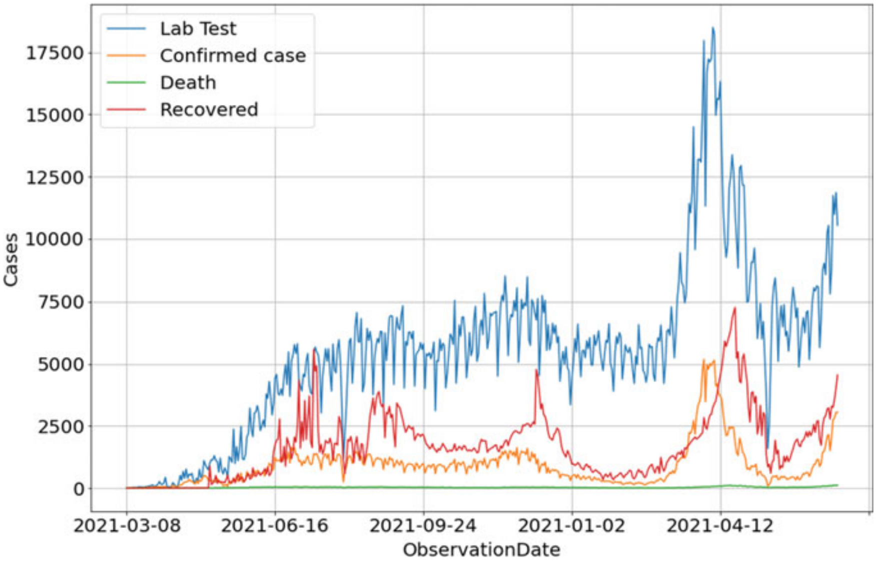


Fig. 1 Bangladesh covid cases trend from March 8, 2021 to June 30, 2021

Table 1 Divisional-wise Covid-19 situation of confirmed cases and case rate in Bangladesh

Division	Lab test	Confirmed	Case rate (%)
Dhaka	2,721,916	464,314	17.06
Chittagong	1,307,372	134,077	10.26
Barisal	154,150	22,002	14.27
Mymensingh	151,946	17,535	11.54
Rangpur	166,327	30,169	18.14
Sylhet	293,016	28,580	9.75
Rajshahi	480,428	63,797	13.28
Khulna	307,638	61,341	19.94

(ML) approach. In our paper, we proposed a machine learning approach to predict new cases of Bangladesh.

Therefore, the significant contributions we have proposed in this paper is:

- We gather time-series data on COVID-19 cases, lab test rates, confirmed cases, death cases, and recovery cases.
- We have applied machine learning models ARIMA and Prophet to predict the new cases of Bangladesh and achieved the highest prediction in the Prophet model.
- Understand the predictive capability of the models; a comparison of the model's performance and prediction values is also shown.

2 Literature Survey and Review

The present situation shows the importance of analyses and prediction of COVID-19. Many research works have been executed to predict the outbreak of COVID-19. Shahriare et al. [1], the researchers analyze the development and assessment of prevention strategies and identify factors that most affect the spread of COVID-19 infection in Bangladesh. They evaluated the time-series COVID-19 dataset and used regression models and the prophet algorithm. Sina et al. [2] presented their paper by comparative analysis of machine learning and soft computing to predict the COVID-19 attack. They proposed two ML models, Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA). The Researchers explored ways to be prepared to advance any rapid surge in an outbreak to apply optimal management of existing resources, construct a predictive model, and facilitate decision-making in Bangladesh in [3, 4]. Aman et al. [5] demonstrate in their study how to plan for any abrupt increase in outbreaks to ensure optimal resource management. They also proposed many ML algorithms like ARIMA, Logistic regression (LR) and found that LR gives the best accuracy of 99.93%. The main aim of research [6–8] is that these researchers have focused on the COVID-19 situation in Bangladesh. In papers [6, 7], researchers conducted ML models and discovered that ML models work extraordinarily well in providing precise information about COVID-19 to authorities for decision-making. Also, ensure early identification of cases among the population. Their research's primary objectives were to analyze how to control the rapid rising of COVID-19 and create awareness among general people in Bangladesh. In [9], the researcher predicted the student adaptability level using machine learning during the COVID-19 online education. Murad et al. [8] proposed in their research predicted that domestic violence in Bangladesh during the COVID-19 epidemic. Their primary goal was to employ ML algorithms to predict domestic violence with information from an online survey. They used various methods, including Random Forest, LR, and Naive Bayes, and discovered that RF is the most effective model for their dataset with 77% accuracy in the model. Dhruv et al. [10] focused on covid-19 disease diversity. They have used algorithms. Random Forest classifier performed the best with an accuracy of 0.80.

3 Research Methodology

3.1 Data Source

In this research, for the dataset, we have used the report of COVID-19 cases in Banladesh, which is and Govt. official source available online [11]. Although there are many online sources available for covid data during this pandemic, we have chosen this source because it records Bangladesh's most authentic covid cases. We

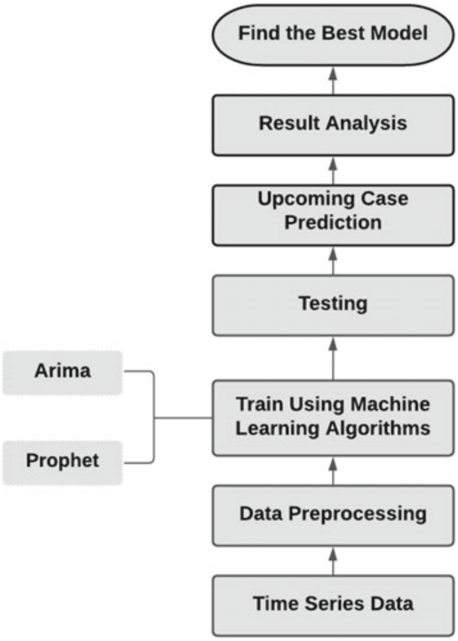


Fig. 2 Methodology for predicting the new case of COVID-19 cases

gathered the time series data from this online source from March 8, 2020, to June 30, 2021 (Fig. 2).

3.2 Data Used

We have processed the data after gathering it from Data Source [11]. In this dataset, we have used observation date, total Cases of Covid affected people, the number of tests performed, the number of total deaths, and recovered patients. We have used only the Observation Date and the total Case of Covid affected people to predict future cases. We organized these data in a time series format for easy machine understanding. All these data were stored in an excel supported CSV format for easy accessibility by ML models.

3.3 Description of Model

Although many Machine Learning models exist for future forecasting of time series data. But we employed two distinct ML models. Prophet and ARIMA, the two most prominent forecasting ML models, were used to anticipate the COVID-19 future confirmed cases, recovered cases, and death cases.

Prophet: Facebook created the open-source forecasting model Prophet. It is a trendy, flexible model for high-quality and automated forecasting. A decomposable time series model is used here. In this research paper, we use trend, seasonality, and holidays as the model components. They were combined as the following equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon t \tag{1}$$

Here, g(t) denotes the modeling of non-periodic changes in time series; the piecewise linear or logistic growth curve is used. s(t): periodic changes (e.g., weekly/yearly seasonality) h(t): effects of holidays (provided by the user) on schedules that are irregular; εt: error term accounts for any unexpected changes that the model does not account for.

ARIMA: ARIMA is also a popular time series forecasting model for Auto-Regressive Integrated Moving Average. We used ARIMA to train our time-series covid data in our research for forecasting COVID-19 future cases. Before applying the ARIMA model, we need to ensure that the data set is not stationary by reducing its degree of differencing [12, 13]. The differencing order equation is equivalent to $z^i = Y_{i-2}y_{i-1} + y_{i-2}$. In the ARIMA (p, d, q) model, the arguments to auto. ARIMA model provides for many variations on the algorithm. The number of autoregressive terms, or "lag observations," is defined by the parameter p. q representing the number of forecast errors in the model, d represents the degree of difference required to make the time series stationary, et represents the amount of white noise. In terms of y, the general forecasting equation is denoted as follows:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi 3 Y_{t-p} + \omega 1 e_{t-1} + \omega 2 e_{t-2} + \cdots + \omega q e_{t-1} + et \tag{2}$$

The weights ($\Phi_1, \Phi_2 \ldots \Phi_3$) and ($\omega_1, \omega_2 \ldots \omega_3$) for the AR and MA, respectively, are calculated based on the correlations between the lagged and current observations [11]. For our dataset, we have chosen ARIMA (2, 0, 0).

4 Result Analysis and Discussion

In this section, we tried to analyze data and predict the COVID-19 outbreak in Bangladesh. The running results of these ML models, ARIMA and Prophet, are described in this section.

Table 2 analyzed the COVID-19 data from March 8, 2020, to June 30, 2021, totaling 480 days of data. Following this analysis, we can see that the highest tested cases, confirmed cases, dead cases, and recovered daily are 18,500, 5178, 119, 816,250, and lowest 10, 3, 0, 0 respectively for Bangladesh.

Figure 3 shows the overall lab test performed, Confirmed Cases, and Case's rate for the observation period. Notably, the covid situation was stable from August 2020

Table 2 Descriptive statistics of daily cases of Covid-19 pandemic

Day wise	Minimum	Maximum	Sum	Mean	Standard deviation
Observation date	08-Mar-20	30-Jun-21	157,689,657	44,363.86	96.75
Lab tested	10	18,500	2,655,891	5533.11	3415.12
Confirmed cases	3	5178	441,678	920.16	872.35
Dead cases	0	119	14,503	30.214	22.4454
Recovered cases	0	16,833	816,250	328,520.12	1504.37
Valid observations (listwise)	480				

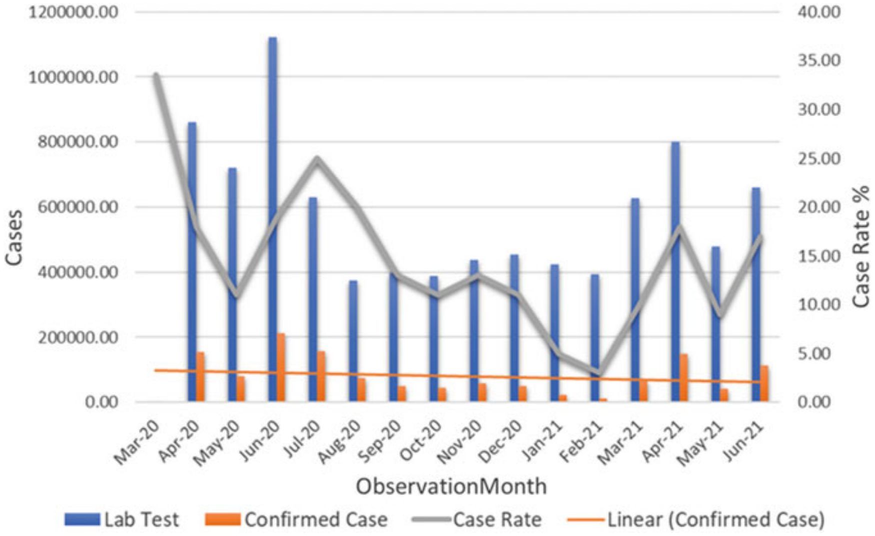


Fig. 3 Month-wise COVID-19 situation of confirmed, case rate, lab test cases in Bangladesh

to February 2021. In contrast, it is continuously rising from March 2021 in terms of both Confirmed and Case rates. This graph demonstrates the ups and down situation of COVID-19.

The graph Fig. 4 shows, the division-wise lab test rate, confirmed case rate, case rate, and linear (confirmed Case) are visible. Lab Test case rate in Dhaka was 23%, Chittagong 12%, Barisal 3%, Mymensingh 4%, Rangpur 3%, Sylhet 3%, Rajshahi 5%, Khulna 4% respectively.

It is also observed that Dhaka has the highest infected case rate of 23%, while Sylhet has the lowest rate of 3%. On the other hand, the confirmed case rate in Dhaka is as high as 22%, and it is also observed that Mymensingh has the lowest rate. The case rate in Khulna is high at almost 18%. In the second-highest is Rangpur at 17%. And the lowest one is in Sylhet, where the rate is 8%.

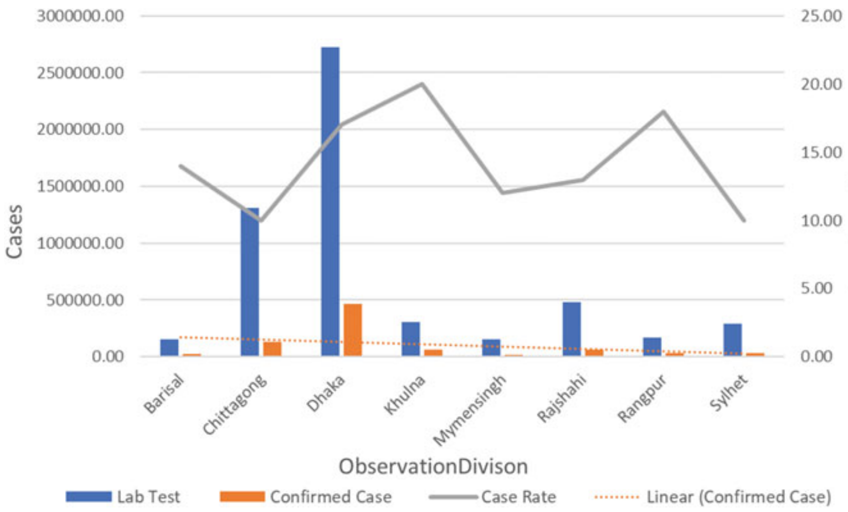


Fig. 4 Division-wise COVID-19 situation of confirmed, case rate, lab test cases in Bangladesh

4.1 Performance Evaluation

Many evaluation pointers can be used to assess the performance of a model. In this research, we have sed Root Mean Square Error (RMSE), Median Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Correlation of Coefficient to measure our model performance. The formulas used to measure the model performance are in Eqs. 3, 4, 5, and 6, respectively.

$$\text{M.A.D} = \frac{\sum ABS(A_i - F_i)}{n} \tag{3}$$

$$\text{RMSE } = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(A_i - F_i)^2} \tag{4}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100 \tag{5}$$

$$\text{Correlation of coefficient} = \frac{n\left(\sum A_i F_i\right) - \left(\sum A\right)\left(\sum F\right)}{\sqrt{\left[n\sum A^2 - \left(\sum A\right)^2\right]\left[n\sum F^2 - \left(\sum F\right)^2\right]}} \tag{6}$$

Here, the Actual value is represented as A_i, Forecasted Value F_i, and n represents the total number of observations.

4.2 Performance Analysis of Model

We have used two distinct machine learning methods, including the Prophet and ARIMA, to forecast the new cases of the COVID-19 virus. Our objective was to find a better predictive model by comparing their performances. We used 100% of the dataset to train and test these models; also predicted the next three months (July 1, 2021, to September 30, 2021) of future forecasting.

Figures 5 and 6 show the predicted situation of COVID-19 for the next three months by the Prophet and ARIMA model, respectively.

Table 3 shows a comparison of performance measuring attributes between two models. It also shows that the overall precision in the ARIMA model is better than the Prophet model. Table 4 also observed that the ARIMA model achieved the lowest RMSE, MAD, MAPE of 5511.08, 11,330.25, and 5.17%, respectively, and the highest correlation coefficient 0.999985325. From the above-aforementioned analysis of the result, it is benignly perceived that the ARIMA model beat the other performance indicators. So, the model ARIMA is a good choice for predicting the new cases of COVID-19.

Figure 7 shows a two-line chart comparing the ARIMA and Prophet model prediction where it is visible that there will be a high spike in covid confirmed cases from July 1 to September 30, 2021.

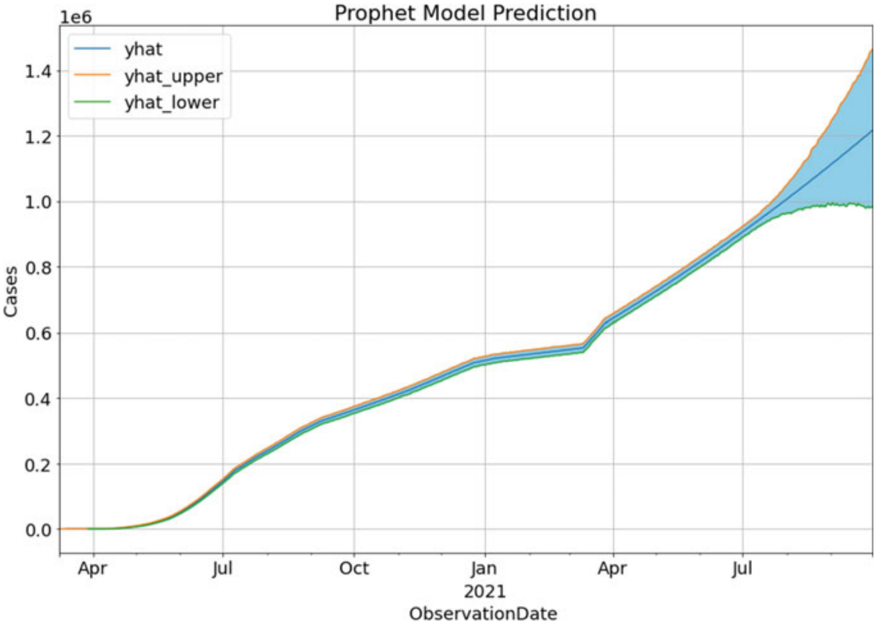


Fig. 5 Prophet model predicted COVID-19 cases

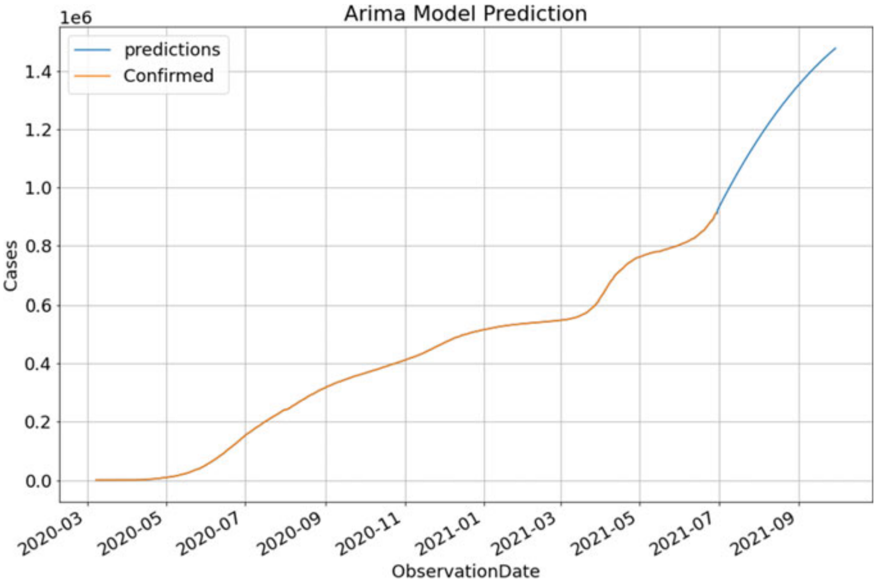


Fig. 6 ARIMA model predicted COVID-19 cases

Table 3 Performance comparison of different models

Pointers	Model	
N = 480	Prophet	ARIMA
MAD	12,004.61283	1875.593859
RMSE	15,794.50401	2390.525123
MAPE	59.5%	8.08%
Correlation of coefficient	0.999054138	0.999985325

5 Discussion and Result Analysis

From March 8, 2020, to June 30, 2021, we studied the distinct instances of the COVID-19 outbreak in Bangladesh using a daily and monthly dataset and compared the results to the confirmed recovery rate and death rate. We used the Prophet and ARIMA model to help in predicting new cases in Bangladesh. We used time-series data which helps us work on modeling supervised learning and predicting from present data. Similarly, it is more critical to predict the upcoming corona situation in Bangladesh than in other countries, but it is most important to maintain the circumstances. When such an outbreak occurs, having readily available data and knowledge is crucial for continuing the evaluation required to identify threats and initiate outbreak containment steps.

With this research, we also visualize the Division-wise covid cases and test cases. We also figured out the descriptive statistics and found the minimum daily test rate on

Table 4 Comparison of predicted confirmed cases between prophet and arima model

S. N.	Date	Predicted (prophet)	Predicted (ARIMA)	S. N.	Date	Predicted (prophet)	Predicted (ARIMA)
1	2021-07-01	923,737	922,016	47	2021-08-16	1,133,960	1,259,299
2	2021-07-02	926,406	930,711	48	2021-08-17	1,139,514	1,265,278
3	2021-07-03	929,978	939,342	49	2021-08-18	1,154,668	1,271,201
4	2021-07-04	933,054	947,910	50	2021-08-19	1,162,538	1,277,071
5	2021-07-05	936,510	956,415	51	2021-08-20	1,166,268	1,282,886
6	2021-07-06	939,619	964,858	52	2021-08-21	1,168,832	1,288,648
7	2021-07-07	943,690	973,238	53	2021-08-22	1,174,025	1,294,356
8	2021-07-08	946,127	981,555	54	2021–08-23	1,187,206	1,300,010
9	2021-07-09	950,092	989,811	55	2021-08-24	1,192,778	1,305,611
10	2021-07-10	953,404	998,004	56	2021-08-25	1,195,534	1,311,158
11	2021-07-11	955,843	1,006,135	57	2021-08-26	1,205,411	1,316,653
12	2021-07-12	959,714	1,014,205	58	2021-08-27	1,208,393	1,322,095
13	2021-07-13	963,038	1,022,214	59	2021-08-28	1,215,461	1,327,484
14	2021-07-14	969,400	1,030,161	60	2021-08-29	1,222,175	1,332,821
15	2021-07-15	973,279	1,038,047	61	2021-08-30	1,227,071	1,338,106
16	2021-07-16	976,822	1,045,872	62	2021-08-31	1,234,581	1,343,339
17	2021-07-17	980,854	1,053,636	63	2021-09-01	1,245,013	1,348,520
18	2021–07-18	986,214	1,061,340	64	2021-09-02	1,251,158	1,353,649
19	2021-07-19	986,401	1,068,984	65	2021-09-03	1,257,644	1,358,727
20	2021-07-20	991,275	1,076,567	66	2021-09-04	1,265,348	1,363,754
21	2021-07-21	996,059	1,084,091	67	2021-09-05	1,268,075	1,368,730
22	2021-07-22	1,001,992	1,091,555	68	2021-09-06	1,281,284	1,373,655
23	2021-07-23	1,006,397	1,098,959	69	2021-09-07	1,289,712	1,378,529
24	2021-07-24	1,010,202	1,106,304	70	2021-09-08	1,299,027	1,383,353
25	2021-07-25	1,015,039	1,113,590	71	2021-09-09	1,302,507	1,388,126
26	2021-07-26	1,016,513	1,120,817	72	2021-09-10	1,308,467	1,392,850
27	2021-07-27	1,026,764	1,127,985	73	2021-09-11	1,315,172	1,397,524
28	2021-07-28	1,028,263	1,135,094	74	2021-09-12	1,321,176	1,402,148
29	2021-07-29	1,033,789	1,142,145	75	2021-09-13	1,332,490	1,406,722
30	2021-07-30	1,041,242	1,149,138	76	2021-09-14	1,336,663	1,411,248
31	2021-07-31	1,045,678	1,156,073	77	2021-09-15	1,344,752	1,415,724
32	2021-08-01	1,049,525	1,162,950	78	2021-09-16	1,356,033	1,420,151
33	2021-08-02	1,055,789	1,169,769	79	2021-09-17	1,362,897	1,424,530
34	2021-08-03	1,062,495	1,176,531	80	2021-09-18	1,365,521	1,428,860
35	2021-08-04	1,066,322	1,183,236	81	2021-09-19	1,372,064	1,433,142

(continued)

Table 4 (continued)

S. N.	Date	Predicted (prophet)	Predicted (ARIMA)	S. N.	Date	Predicted (prophet)	Predicted (ARIMA)
36	2021-08-05	1,062,495	1,176,531	81	2021-09-20	1,383,932	1,437,376
37	2021-08-06	1,066,322	1,183,236	83	2021-09-21	1,382,634	1,441,562
38	2021-08-07	1,062,495	1,176,531	84	2021-09-22	1,397,862	1,445,700
39	2021-08-08	1,066,322	1,183,236	85	2021-09-23	1,404,312	1,449,790
40	2021-08-09	1,066,322	1,183,236	86	2021-09-24	1,413,126	1,453,834
41	2021-08-10	1,099,595	1,222,273	87	2021-09-25	1,423,591	1,457,830
42	2021-08-11	1,105,218	1,228,583	88	2021-09-26	1,431,923	1,461,779
43	2021-08-12	1,113,607	1,234,837	89	2021-09-27	1,436,335	1,465,681
44	2021-08-13	1,120,653	1,241,035	90	2021-09-28	1,453,204	1,469,537
45	2021-08-14	1,124,494	1,247,178	91	2021-09-29	1,458,763	1,473,347
46	2021-08-15	1,126,547	1,253,266	92	2021-09-30	1,463,527	1,477,110

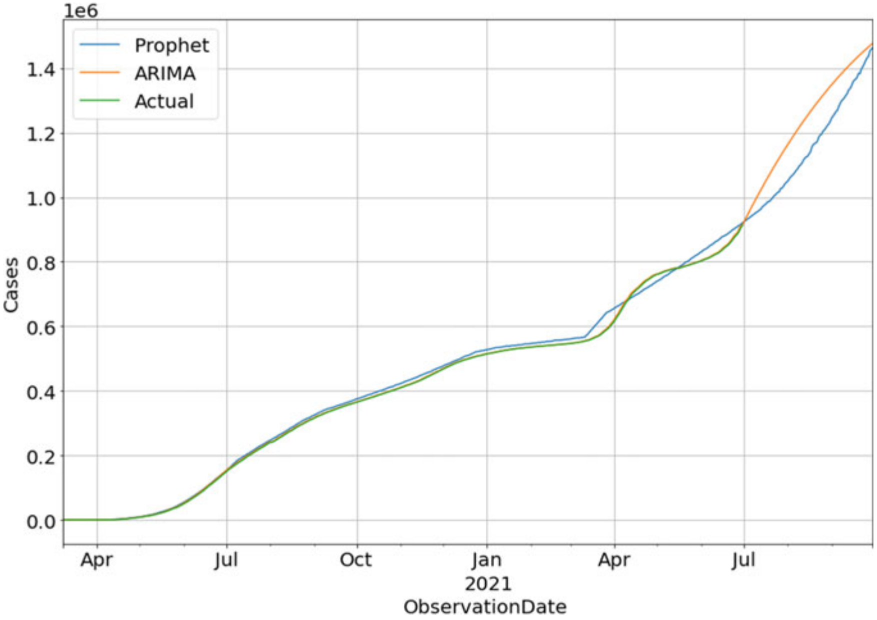


Fig. 7 Performance comparison of different models ARIMA and Prophet

March 8, 2020, was 10, but on June 30, 2021 maximum tested rate was 18,500. The Daily confirmed rate was a minimum of 3 on March 8 and June 30; it is 5178, and the death rate is also excellent. On March 8, it was 0, but on June 30, it was 119. It is also notable that the number of deaths confirmed cases rises fast over time. We also make out of the monthly rate of COVID-19. In March 2020, the rate was around 25%, and in April, it was 23% lab tests, but we saw a massive outbreak in June 2020. It was 38%. Between two months, 13% increase in cases showed fast corona speeded. In July it becomes 22%. August 13%, September 15%, October 14%, November 17%, and end of the year in December 18%. After 2021, in January of 2021, the lab test has become low down to around 17%, and in February, it decreased by 14%. However, In March, it again expanded by 21%. In April 25%, In May 17% and Lastly in June was 23%. In Table 5, we have tried to note other authors' papers' works with our paper to show an overall comparison. From the comparison, we found that several researchers also got a better prediction result with the ARIMA model.

6 Conclusion and Future Work

This global pandemic caused the deaths of millions of lives and heavily affected the world economy. Humans will have to battle the Coronavirus in the future and obey the Government's rules. The use of a lockdown isn't the only sole option; it is also the social and moral responsibility of every citizen to observe the rules. So, in this study, we tried to reports a comparative analysis of the ML model for predicting new COVID-19 cases in Bangladesh. We also examined the status of the first 480 days from March 8, 2020. We have made use of the country's first publicly available covid dataset. The two ML models are applied ARIMA and Prophet for the prediction using machine learning by python programming language. The above trend data analysis and line plots show that the predicted values for the ARIMA model are more fitted than the Prophet model. We can forecast the future COVID-19 scenario in Bangladesh based on the model analysis. This trend analysis also depicts COVID-19's case rate, lab test rate, recovery rate, and mortality rate. To control new coronavirus instances, we must pay attention to existing analyses and future predictions. This paper will help us to take the necessary steps to handle the situation of the COVID-19 pandemic. We'll keep an eye on upcoming Cases and look at new machine learning techniques, such as deep learning, to enhance the findings more consistently and precisely in the future.

Table 5 Comparison table of our proposed model along with previous research work

Authors	Focus	Used algorithms	Best-model
Shahriare et al. [1]	Analyze the development and assessment of prevention strategies	Regression and prophet	Regression
Sina et al. [2]	They focus on comparing ML and soft computing models to forecast the COVID-19 outbreak	MLP and ANFIS	MLP
Samrat et al. [3]	Focused on experimental data analysis and collected them to analyze data on COVID-19	No use	null
Aman et al. [5]	To prepare for any unexpected rise in an outbreak and secure optimal management of accessible staff	ARMA, ARIMA, BRR, SVR, RFR, XGB, HW, LRP, LR	ARIMA
Anjir et al. [14]	Development for facilitating the decision-making process	LSTM, ANFIS	LSTM,
Vikas et al. [4]	Predicting the new cases in India in the next coming days	ARIMA, AR	ARIMA
Mazharul et al. [6]	They work on ML models to help authorities to make decisions easy	PR, HWA, ARIMA, Prophet	FB Prophet Model
Abdullah et al. [7]	To ensure early awareness among general people to identify infection. It also focuses on forecasting the future situation of COVID-19 in Bangladesh	Linear Regression model	Linear Regression model
Murad et al. [8]	Analyze domestic violence in Bangladesh during the COVID-19 situation	random forest, Logistic Regression, and Naive Bayes	Random Forest
Our proposed	Analysis and prediction of a new case of Coronavirus situation using machine learning approach in Bangladesh	ARIMA, Prophet	ARIMA

References

1. Satu MS, Howlader KC, Islam SMS (2020) Machine learning-based approaches for forecasting COVID-19 cases in Bangladesh. SSRN Electron. J. https://doi.org/10.2139/ssrn.3614675
2. Ardabili SF et al (2020) COVID-19 outbreak prediction with machine learning. Algorithms 13(10). https://doi.org/10.3390/a13100249
3. Dey A, Kumar S, Rahman MM, Siddiqi UR, Howlader (2020) Exploring epidemiological behavior of novel coronavirus (COVID-19) outbreak in Bangladesh. SN Compr Clin Med 2:1724–1732. https://doi.org/10.1007/s42399-020-00477-9
4. Kulshreshtha N, Vikas, Garg (2021) Predicting the new cases of coronavirus [COVID-19] in India by using time series analysis as machine learning model in Python. J Inst Eng Ser B:1—7. https://doi.org/10.1007/s40031-021-00546-0
5. N. Khakharia, Aman and Shah, Vruddhi and Jain, Sankalp and Shah, Jash and Tiwari, Amanshu and Daphal, Prathamesh and Warang, Mahesh and Mehendale, "Outbreak prediction of COVID-19 for dense and populated countries using machine learning. Ann Data Sci 8:1—19. https://doi.org/10.1007/s40745-020-00314-9
6. Leon KA, Islam M, Iqbal MI, Azim SM, Al Mamun KA (2021) Predicting COVID-19 infections and deaths in Bangladesh using machine learning algorithms. In: 2021 International conference on information and communication technology for sustainable development, pp 70–75. ICICT4SD50815.2021.9396820
7. Al Shuaeb M, Abdullah SM, Kamruzaman M, Al-Amin S (2020) COVID-19 outbreak prediction and forecasting in Bangladesh using machine learning algorithm. Int J Trend Sci Res Dev 5(1)
8. Sifat RI (2020) Impact of the COVID-19 pandemic on domestic violence in Bangladesh. Asian J Psychiat 53:102393. https://doi.org/10.1016/j.ajp.2020.102393
9. Mahmudul Hasan Suzan MAPM, Samrin NA, Biswas AA (2021) Students' adaptability level prediction in online education using machine learning approaches. In: 12th International conference on computing, communication and networking technologies (ICCCNT 2021). https://doi.org/10.13140/RG.2.2.27516.46726
10. Patel D, Kher V, Desai B, Lei X, Cen S, Nanda N, Gholamrezanezhad A, Duddalwar V, Varghese B, Oberai AA (2021) Machine learning based predictors for COVID-19 disease severity. Sci Rep 11(1)
11. D. C. and R. Institute of Epidemiology (2020) Institute of Epidemiology, Disease Control and Research. https://iedcr.gov.bd/COVID-19/
12. Pramanik MA, Rahman MM, Anam ASMI, Ali AA, Amin MA, Rahman AKMM (2021) Modeling traffic congestion in developing countries using google maps data. Adv Intell Syst Comput AISC 1363:513–531. https://doi.org/10.1007/978-3-030-73100-7_36
13. Raguib H, Sayem DA, Kumar MJ, Hassan KT, Abdullah Al N, Salehin SM, Moinul H, Sultana JN (2020) Prediction of epidemics trend of COVID-19 in Bangladesh. Front Public Health 8:631. https://doi.org/10.3389/fpubh.2020.559437
14. Chowdhury KKS, Ahmed A, Hasan KT, Hoque KKS (2021) Analysis and prediction of COVID-19 pandemic in Bangladesh by using ANFIS and LSTM network. Cognit Comput 13:761–770. s12559-021-09859-0

Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions

Lamiae Demraoui, Siham Eddamiri, and Lamiae Hachad

Abstract A digital future is inevitable in today's world. Digitalization is a necessity due to the increasing expectations for productivity and competitiveness. User experience is streamlined, and the contact between firms and their customers is accelerated, thanks to digitization. Better products and more profits are also the results of this process. Recently, the concept of digital transformation has developed as a focus for enterprises in emerging countries. Thus, human and corporate processes, as well as technology elements, are all being transformed to give better services. In this work, we present our system that predicts loan repayment or default based on customer's digital data using classification models to save time and energy. Our system achieved promising results in terms of accuracy, AUC, and ROC which demonstrate the effectiveness of our approach.

Keywords Digital transformation · User experience · Business · Machine learning · Banking

1 Introduction

Digital transformation is a disruptive process that is driven by digital technologies. Its effects on organizational value generation, strategy, and structure processes are immense [1]. Digital Transformation is the adoption of disruptive technology in order to boost production, value creation, and social welfare. Many national govern-

L. Demraoui (✉)
Systems Engineering Laboratory, Sultan Moulay Slimane University, Beni Mellal, Morocco
e-mail: l.demraoui@usms.ma

S. Eddamiri
Department of Mathematics and Computer Science, University Moulay Ismail, ENSAM, Meknes, Morocco

L. Hachad
Moroccan School of Engineering Sciences, EMSI, Casablanca, Morocco
e-mail: l.hachad@emsi.ac.ma

Z. Boulouard et al. (eds.), *AI and IoT for Sustainable Development in Emerging Countries*,
Lecture Notes on Data Engineering and Communications Technologies 105,
https://doi.org/10.1007/978-3-030-90618-4_32

ments, multilateral organizations, and industry associations have conducted strategic-foresight studies to help guide their long-term strategies. By recommending the execution of public policies surrounding digital transformation [2].

Digitalization is transforming the world in almost every aspect of life during the last few decades. In today's business world, information technology and digitalization have become increasingly significant as a result of the rapid growth in internet infrastructure. On the other hand, it can be claimed that the digitization of business models is enabled by technical improvements. In order to move their activities to digital environments, corporations rely on internet technology on a regular basis. A business network built in this way creates value for its stakeholders by taking advantage of today's technical architecture.

Society have the potential to benefit from digitalization in various development domains. For such development, it is required for the organizations to take measures for a digital future, to facilitate a digital enabling environment, and to enhance learning [3]. The digital transformation concerns several organisation's sectors such as Banking. Business undertaken or services given by a bank is characterized as banking. Banks are institutions that retain money for clients and make it available to them on demand. They also lend money to businesses and individuals. Investment and insurance services may be available in various nations. The Cambridge Dictionary defines banking as "the process of managing the money in your bank accounts." Transforming the financial services industry with digital technology has shaken up banking habits while bringing in new rivals [4].

To ensure the digital transformation, numerous emerging technology trends drive this revolution. The use of data science [5] in the banking sector is a valuable tool for predicting customer behavior and adapting strategic decisions based on the collection, analysis, and valuation of consumer data. In this work, we present our system that predicts loan repayment or default basing on customer's digital data using classification models. The efficiency of our methodology will be determined with three performance evaluation criteria for classifier comparison of accuracy, AUC and ROC. As a result, the banking based technology is more comfortable for the customers and saves time and effort. The services become more affordable and information becomes more readily available. The customer can also keep track of the transactions more easily with the help of automated banking services.

The remainder of this paper is organized as follows. The next section defines digital transformation, its technologies, and stages. The related works are presented in the third section. The research methods used for this study are presented in the following section. After that, we present the results and discuss the research's main findings. Finally, the conclusion rounds out the paper.

2 Theoretical Background

2.1 Digitalisation Versus Digital Transformation

One of the most significant changes in society today is digitalization, which encompasses a wide range of aspects of business and everyday life [6]. Digitization is defined as the use of digital technologies and data to generate revenue, improve business processes and creates an environment conducive to digital business, in which digital information plays a central role [6]. Digitalization is generally defined as the integration of digital technologies into business or social processes, with the aim of improving them.

Digitization has an impactful effect on the world. It alters the way companies interact with their customers and, in many cases, the way they generate revenue. With a business-oriented focus, Gartner defines digitalization as follows: “Digitalization is the use of digital technologies to change a business model and provide new revenue and value-producing opportunities; it is the process of moving to a digital business” [7]. To clearly define digitalization, we must first examine digitization. Some people confuse the two themes, giving them a similar meaning. However, the two have a distinct meanings. Digitization is the process of transforming physical data into digital data [8]. Enterprises have been digitizing for a long time. Examples include: converting handwritten or typewritten text into digital form, or scanning a report.

A clear distinction can be made between the two themes. Digitalization has a positive ring to it. However, digitization is a neutral process. Digitalization, on the other hand, depends on it. When physical data are “translated” into a digital format, the process is known as “digitization.” 1s and 0s make up digital information, which is information presented in digital form. In addition, they are not susceptible to distortion due to their nature. It is also possible to transmit it without any loss. The world’s communication networks rely on, store, and manage digital information.

Processes can be improved through the use of digital information, which is digitized. Digitization, for example, is the act of entering text into a digital document. Digitalization is making that document available via the cloud. Digitalization and digitization are not mutually exclusive concepts, but rather complementary ones.

In contrast, digital transformation is referred to as “the use of new digital technologies to enable major business improvements in operations and markets such as enhancing customer experience, streamlining operations or creating new business models” [9]. Digitalization dealt with technological innovations, that digital transformation affects every aspect of an organization (Fig. 1). Digital capabilities, business models, operational procedures, and user experience [10] (internal and external IT consumers) are the four aspects of digital transformation [11].

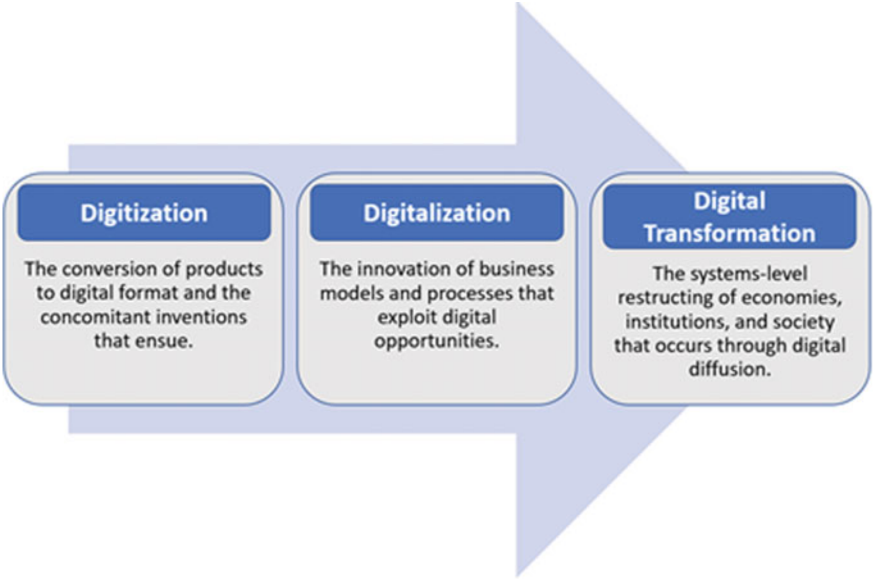

Digitization
The conversion of products to digital format and the concomitant inventions that ensue.
Digitalization
The innovation of business models and processes that exploit digital opportunities.
Digital Transformation
The systems-level restructing of economies, institutions, and society that occurs through digital diffusion.

Fig. 1 A framework for understanding digitalization

2.2 *Technological Aspect of Digital Transformation*

Nevertheless, the digital transformation goes further, because at the level of companies, in particular, it alters their economic models, as well as the employees' relationship to work. And it is possible to define the digital company through 4 pillars

- As a result of the digital transformation, time and space are no longer barriers;
- A digital company's telecommuting is facilitated by the ability to stay connected while traveling;
- The concept of the Internet of Things (IoT), where the physical and digital worlds meet;
- As a final note, let's not forget that the United Nations itself makes access to the Internet one of the development criteria.

Emerging technologies fuel digital transformation, a sort of economic transformation. This shift in technology's position within a company is what drives digital transformation. Technology is no longer just a support function that facilitates business processes. Today's technology allows for new, innovative business models, drives sales growth, and even provides a source of Competitive Advantage; Digital Transformation is driven by several emerging technology trends. Among the most notable are:

- **Mobile technology**: when used in conjunction with digital transformation technologies, mobile devices help achieve the goal of digital transformation and provide seamless interaction with customers at all points of contact with business [12].
- **Cloud computing**: an evolution of information technology, as well as an important business model for delivering information technology resources. It allows individuals and organizations on-demand access to a managed pool of IT resources, such as server storage and software applications that are scalable. Other key digital trends are powered by cloud computing [13].
- **Social media** [14].
- **Big data analytics**: with the ability to analyze a large volume of data, in a variety of formats, and at a high speed across a wide range of networks to support decision making and action taking [15–17].
- **Internet of Things (IoT)**, that allows new possibilities to explore business opportunities [18].

The technology's applicability is heavily influenced by the industry and the organization. As a result of these technological advancements, businesses have the ability to digitize, transform, and grow their companies [19].

2.3 Digitalisation Stages

Our economy is being substantially altered by digitalization. Basically, there are three stages to this (Fig. 2) [20].

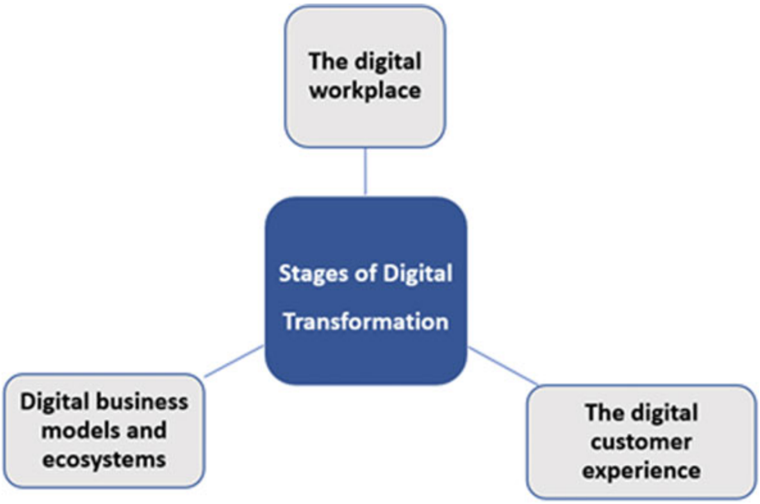

The digital workplace
Stages of Digital Transformation
Digital business models and ecosystems
The digital customer experience

Fig. 2 The stages of digitalisation [20]

The spread of IT change the form of workplace from a traditional to a modern or a digital one. Employees are disengaged from the physical attributes of the traditional workplace in these modern workplaces. Their flexibility in terms of time and space makes them a great choice. They could work from any location without ignoring their co-workers' collaboration and communication [21]. This changed also the user experience of the IT user. The digital transformation can also affect the customer, this is especially true in today's digitally networked environment, where customers are sharing their experiences with others and can move to a competitor in a matter of seconds. The reason for this is that a terrible customer experience will directly harm the brand perception of a firm and its sales. Customers' experiences across digital and conventional channels must be optimized holistically and individually by companies. It's also possible to have a "digital" business model if new digital technologies cause fundamental shifts in the way business is conducted and revenue is generated.

3 Related Work

With the development of the digital economy, new technologies appear that help the emerging countries keep the pace of the computational and communication economic operations progress with other developed countries [22]. These technologies have changed the way we do business in all industries, from health to education [23].

In the health sector, many operations have been digitalized as a result of the COVID-19 epidemic and the harsh measures are taken to stop its spread. Morocco has made substantial use of digital health to help handle the country's present health crisis. Digital health and telemedicine will be the cornerstone of health care in the future, according to the digital health strategies described in [24]. Digitalization is critical to the management and mitigation of the pandemic crisis as well as the future development of the health system in Morocco and on the African continent. A detailed examination of the Outpatient Appointment System (OAS), which is a gateway to the healthcare system, is provided by [25].

Incorporating digital technologies into education has a profound impact on society's modernization, encouraging growth and competitiveness through a more educated workforce and more employment. Kerroum et al. [26] attempt in their research to assess the current state of the Moroccan university and identify the most significant means, to facilitate the implementation of the digital transformation in H2C. Ahmad et al. [27] analyze the measures taken to ensure non-disruptive learning in the context of COVID-19. Because of this study, higher education institutions, and the Ministry of Higher Education will be able to compare their experiences with other faculties around the globe.

Bankers all across the world are coming to understand the consequences of banking digitalization, although they are still unsure of its nature and effects. Khanboubi and Boulmakoul [28] provides guidance on how financial institutions can benefit from digital transformation and explain how IoT is used in finance and examine the impact of digital trends and IoT on a traditional bank's procedural scheme. Peshkova and Zlobina [29] describes the structure of modern bank and bank services as transformed with cutting edge digital technologies, With speech and voice technologies banks can raise the levels of customer satisfaction, thus increasing customer loyalty and gaining a competitive edge, they can also change the character and structure of spending on organizing bank operation, and customer services transforming banking in general [30]. Explore the impact of digital transformation on banking governance. Abdulquadri et al. [31] has as its goal to use chatbots to transform business models, improve customer experience and increase financial inclusion in emerging markets.

Banking establishments optimize customer experience and customer service in a consistent manner in order to retain their clients and attract new prospects. And they're creating new products and services as well as expanding their existing ones. Their research and development services produce creative and digital ways to bank.

Today's bank customers are not as loyal as before and many rates highly digital experiences in their interaction with the bank, ready to switch to those banks that can offer them going completely online. With the technological aspects of digitalization banks can raise the levels of customer satisfaction, thus increasing customer loyalty and gaining a competitive edge, they can also change the character and structure of spending on organizing bank operation and customer services transforming banking in general.

The use of machine learning in the banking sector is a valuable tool for predicting customer behavior and adapting strategic decisions based on the collection, analysis, and valuation of consumer data. In the next section, we present our model that predicts loan repayment or default basing on customer's digital data.

4 Research Methodology

This research presents our system to predict loan repayment or default based on the digital data for a customer using classification models see Fig. 3. The methodology comprises of different steps which are as follows.

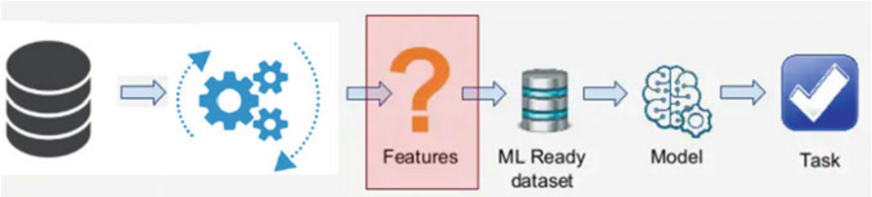


Fig. 3 Our system to predict loan repayment using machine learning

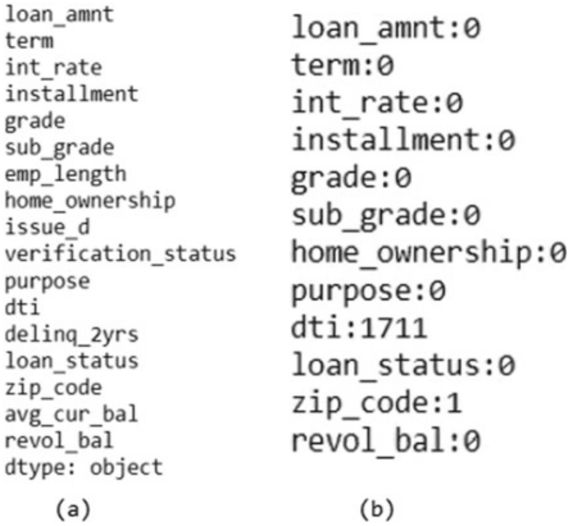
4.1 Data Cleaning

The dataset contains a large number of columns with null values. Users must know how many columns contain null values in order to eliminate those columns that do not meet a certain percentage threshold. Aside from that, there are variables that have a large difference in the number of normal and default categories, which will make it difficult to learn a model. Thus, we need to use the oversampling methods to replicate the observations from the minority class to balance the data, this may cause overfitting. Hence, the SMOTE method is used in this step. SMOTE is based on the following premise:

- Find random points within the nearest neighbors of each minority sample using Euclidean distance.
- Set for each points x_i in the minority class several points randomly from its k-nearest neighbors based on the unbalanced proportion of samples.
- For each sample, construct a new randomly selected neighbor x_n according to this formula:

$$x_n ew = x_i + rand(0, 1) * |x - x_n|$$

The original sample size of the minority class is eventually expanded to an ideal ratio by iterating each sample X. Therefore, New data are not the same as the existing data it does not have any overfitting problem (Fig. 4).



```
loan_amnt
term
int_rate
installment
grade
sub_grade
emp_length
home_ownership
issue_d
verification_status
purpose
dti
delinq_2yrs
loan_status
zip_code
avg_cur_bal
revol_bal
dtype: object
```

(a)

```
loan_amnt:0
term:0
int_rate:0
installment:0
grade:0
sub_grade:0
home_ownership:0
purpose:0
dti:1711
loan_status:0
zip_code:1
revol_bal:0
```

(b)

Fig. 4 **a** Columns with their respective datatype left after dropping those columns which did not meet the percentage threshold. **b** The dataset after cleaning null values in each column

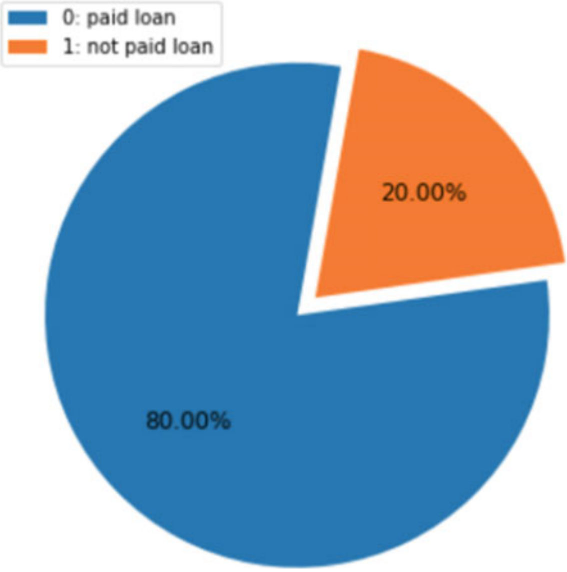
4.2 Feature Engineering

Missing values are a common occurrence in datasets, with some features having thousands of them. Even so, in loan prediction modeling, the exactness of a loan customer's information also impacts the credit assessment of that customer's creditworthiness. Customers with 100% complete information have a better credit rating or risk prediction than those with only 50% of complete information, according to the research. As a credit assessment feature, missing values are important. Throwing away or deleting such information could lead to the loss of important information and, as a result, erroneous predictions.

Therefore, 'month-rep' is our first new feature, which represents a user's monthly repayment expense as a percentage of his monthly income. More "month-rep" means more stress on the lender's debt, which increases the likelihood of default. As a second step, feature abstraction should be taken into consideration. For example, we encode loan statuses such as "Current", "Fully Paid", and "Issued" as 0 and "Default", "Charged Off", "In Grace Period", and "Late (16–30 days)" as 1, and "Late (31–120 days)" as 1. In Fig.5, we can see the status of the loans. There were 80% of samples with "normal" loan status, but only 20% percent of defaulted loans, indicating a serious imbalance in the dataset. This is followed by abstraction of "emp-length" and "grade," with the remaining features being one-hot encoded in the meantime.

We've already discussed removing all samples from the training set with more than 194 missing values (that is, samples whose discrete feature of missing values is five). This makes it difficult for the model to learn, or introduces noise that leads to

Fig. 5 Percentage of each loan status



overfitting of the model when missing values are present. It is, therefore, necessary to remove them.

Preprocessing of data results in the creation of a sorted feature. The original table contains 1045 numerical characteristic variables. They have 1045 dimensions when sorted in descending order. To reduce overfitting, sorted features help stabilize the model by preventing it from being overfitted by abnormal data.

4.3 Machine Learning: Predictive Modeling

In machine learning, patterns are predicted and identified, and then appropriate results are generated based on that knowledge. Algorithms that use machine learning analyze data patterns and draw conclusions from them. Using machine learning, a model will learn from each attempt and get better. Before a model can be evaluated, the data must be divided into training and test sets, respectively. As a next step, our model's predictions would need to include a selection of performance metrics. A borrower's likelihood of defaulting on a loan was examined. We used 5 algorithms for our modeling purpose: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), K Nearest Neighbour (KNN), Support Vector Machine (SVM).

- **K-Nearest Neighors (KNN) Classifier**: is an easy-to-use method that works well for a wide range of problems and datasets. Assuming that K is a positive integer and that N_0 is a test observation, then a KNN classifier determines K points (neighbors) in training data that are closest to the test observation (x_0). Also calculated is the conditional probability for class j, which is expressed in percentages.

$$Pr(Y = j|X = x_0) = \frac{1}{k}\sum_{i\in N_0} I(y_i = j)$$

  Last but not least, KNN classifies the test observation x_0 according to the Bayes rule.
- **Decision Tree (DT) Classifier**: It's a tree-structure that can be either binary or non-binary. This tree's non-leaf nodes represent tests of features; its branches represent the output of an attribute over a range of values, and its leaf nodes represent a categorical value. According to its value, an output branch will be selected until it reaches a leaf node. The decision result is then stored in the leaf node as a category of information.
- **Random Forests (RF) Classifier**: is a supervised learning algorithm that uses trees as building blocks to build more potent prediction models. It is a collection of unrelated decision trees. The Gini index was used as a selection metric in the decision tree, and the number of levels in each tree branch depends on the algorithm parameter d. The Gini Index for this attribute is calculated as:

$$G(X_i) = 1 - \sum_{j=1}^{J} Pr(X_i = L_j)^2$$

where G is the Gini index, the larger the value of G, the higher the uncertainty of data; Pr is the probability of a sample being selected.

- **Logistic Regression (LR) Classifier**: is a generalized linear regression analysis model that can be used for multivariate control. Due to its use of a sigmoid function, this model differs from other linear regression models in that it limits the output value range to [0, 1]. Landslide susceptibility and the independent variables are:

$$f(z) = \frac{1}{1 + \exp -z}$$

where $z = w_1x_1 + w_2x_2 + \cdots + w_Mx_M + b$ refers to a weighted linear combination model, and b indicates the intercept of the function; w_i denotes the correlation coefficient of the function.

- **Support Vector Machine (SVM) Classifier**: SVM is depicted as a supervised machine learning method that employs statistical learning theory and the structural risk minimization principle. Using a hyperplane, the SVM reformats the non-linear world into one that can be processed. SVM is a binary classifier in which the class labels contain only two values +1 or 1.

$$f(x) = sign(w, x) + b$$

where a training data set $S = (x_1, y_1), \ldots, (x_n, y_n)$ and $x_i \in R^n$ and $y{+}1, 1$. The algorithm is based on finding the hyper-plane which gives the maximum distance of separation between training samples.

5 Results and Discussion

5.1 Dataset

Table 1 shows a labeled dataset of 45,211 anonymous borrowers with two classes "0" and "1" and 18 columns ordered by date (from May 2008 to November 2010) in Kaggle's "Banking Dataset".[1] As a result of our experiments, instances with missing values are omitted from consideration. Eighty percent of the labeled data set is used for training, and twenty percent is used for testing. In both sets, there is a class ratio of about 13:1. Sixty-six hundred and eighty-eight negative instances were found on both the training and the test sets.

[1] https://www.kaggle.com/prakharrathi25/banking-dataset-marketing-targets.

Table 1 The detailed column descriptions in banking dataset

1	Age (numeric)
2	Job: type of job (categorical: “admin.”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar”, “self-employed”, “retired”, “technician”, “services”)
3	Marital: marital status (categorical: “married”, “divorced”, “single”; note: “divorced” means divorced or widowed)
4	Education (categorical: “unknown”, “secondary”, “primary”, “tertiary”)
5	Default: has credit in default? (binary: “yes”, “no”)
6	Balance: average yearly balance, in euros (numeric)
7	Housing: has housing loan? (binary: “yes”, “no”)
8	Loan: has personal loan? (binary: “yes”, “no”)
9	Contact: contact communication type (categorical: “unknown”, “telephone”, “cellular”)
10	Day: last contact day of the month (numeric)
11	Month: last contact month of year (categorical: “jan”, “feb”, “mar”, …, “nov”, “dec”)
12	Duration: last contact duration, in seconds (numeric)
13	Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14	pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15	Previous: number of contacts performed before this campaign and for this client (numeric)
16	poutcome: outcome of the previous marketing campaign (categorical: “unknown”, “other”, “failure”, “success”)
17	y: has the client subscribed a term deposit? (binary: “yes”, “no”)

https://www.kaggle.com/prakharrathi25/banking-dataset-marketing-targets

5.2 Validation Metrics

In this paper, we will determine the efficiency of our methodology with three performance evaluation criteria for classifier comparison of accuracy, AUC, and ROC.

- **Accuracy**: is the ratio of the number of correct predictions by the classifier to the total number of samples for a given test data set. The formula is as follow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **F1-score**: is also called a balanced F-Score, is defined as the balanced average of Precision and recall.

$$F1\text{-}score = 2 * \frac{precision * recall}{precision + recall} \tag{2}$$

- **ROC (receiver operating characteristic curve)**: is a popular graphic plot that illustrates the performance of a binary classifier system. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Therefore, if the curve is closer to the top left then the accuracy of the prediction is higher.

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

- **AUC value**: is a metric for binary classification that measures the accuracy of the model, ranging from 0.5 to 1. It represents the area under the curve (AUC) is the receiver operating characteristic (ROC) curve-based measurement of the alternative discrimination ability. Assume that the ROC curve is formed by the sequential connection of points with coordinates of $(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots (x_n, y_n)$AUC can be estimated as:

$$AUC = \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} + x_i) \cdot (y_{i+1} + y_i) \tag{5}$$

5.3 Experimental Process

By applying data preprocessing as the first step, we evaluate our methodology process for predicting loan amounts for customers. Feature selection is the next step after feature scaling, giving priority to features with high relevance to the target and removing irrelevant features can reduce the learning difficulty. Last but not least, we use a variety of classifiers to determine whether or not a borrower is likely to default on a loan. More clearly, 19,779 numerical characteristic variables are extracted from the original data table. Each of the 19,779 dimensions of the sorting features is then sorted in ascending order. Due to their robustness against abnormal data, the sorting features stabilize the model and reduce overfitting risks we use SMOTE. This is followed by an analysis of how well our proposed algorithm fares against other popular classifiers such as Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine (SVM). The performance of these algorithms is presented in Table 2 and Fig. 6. When comparing these algorithms, we can see that the Random Forest (RF) algorithm outperforms them both in terms of total accuracies as well as overall balance.

Table 2 Evaluation metrics comparison of the four techniques

Classifier	Accuracy (%)	AUC	F1-score	Recall
Random Forest	98	0.978	0.97	0.96
KNN	80	0.801	0.91	0.92
Decision Tree	95	0.967	0.95	0.94
SVM	75	0.789	0.81	0.85
Logistic Regression	73	0.794	0.83	0.84

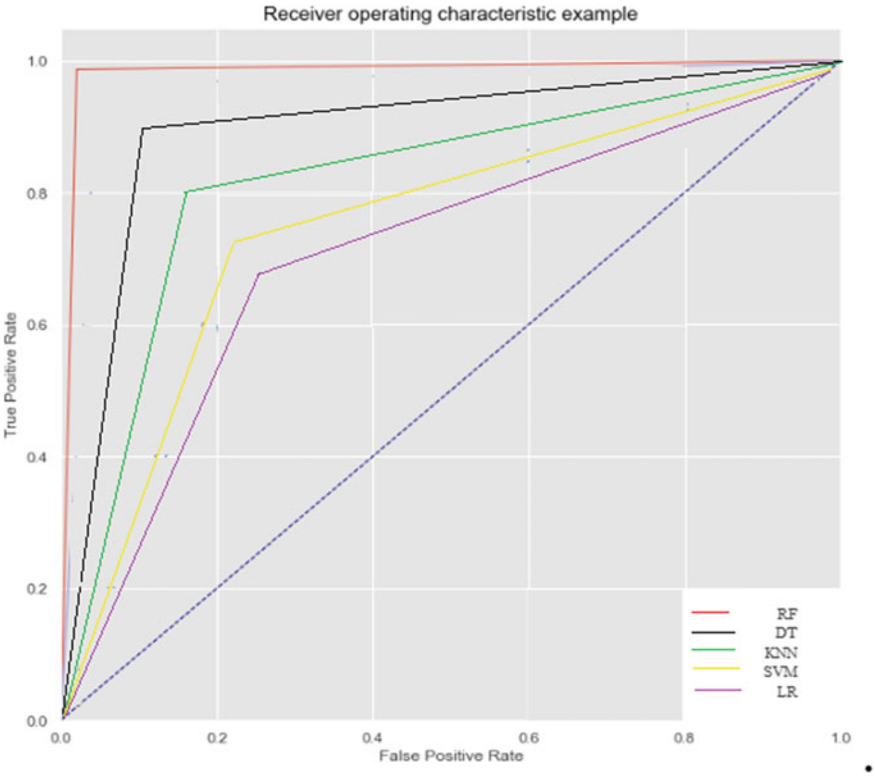


Fig. 6 ROC performance comparison of the four classifiers

5.4 *Quality of Results and Discussion*

The results of the customer prediction loan are studied and discussed in this section. The 'anytime, anywhere' availability of banking services is the major draw of technology-based banking. As a result, not only are customers more comfortable, but they also save time and effort. Services become more affordable as a result, as does the availability of information. It's also easier to keep track of your transactions with the help of automated banking services, leaving behind an easily traceable financial trail. In fact, the companies in the emerging countries benefit from better fund management, as well as mobile access to a variety of services, increasing efficiency. Other digital byproducts include the creation of new services such as warnings, alerts, and budgeting tools. Digital banks have an advantage over branch-based banks due to the removal of time, geography, and cost limitations.

6 Conclusion and Future Work

Digitalization is transforming the world in almost every aspect of life during the last few decades. In today's business world, information technology and digitalization have become increasingly significant. As a result, businesses are being forced to change their ways of doing business. Organizational value generation, strategy, and structure processes are greatly impacted by its implementation.

In this work, we focus on the impact of digitalization in the banking sector. We present a system that predicts loan repayment or default basing on customer's digital data using classification models. Classifier comparison accuracy, AUC, and ROC are the three performance evaluation criteria used to measure the efficiency of our proposed methodology. Finally, the results of the customer prediction loan are analyzed and discussed in order to conclude that banking-based technologies are more convenient for customers, as well as saving time and energy. Moreover, automatic banking services allow customers to keep a better track of their finances, as well. Thus, the information and services become more affordable as a result of technological advancement for the emerging countries.

In future works, we intend to use our methodology to build an ecosystem that's enabled by payments for online purchases for customers to predict loan repayment or default in record time. Moreover, to further improve the quality of the machine learning result, we will use different algorithms, which have better accuracy and has less computational cost, other than used in this research work.

References

1. Vial G (2019) Understanding digital transformation: a review and a research agenda. J Strateg Inf Syst 28:118–144
2. Ebert C, Duarte CHC (2018) Digital transformation. IEEE Softw 35:16–21
3. Rwigema P (2020) Digital technology and its relevance to political and social eco- nomic transformation. Case study of East African Community Region. Strateg J Bus Change Manage 7:1402–1436
4. Gouveia LB, Perun M, Daradkeh YI (2020) Digital transformation and customers services: the banking revolution. Int J Open Inf Technol 8:124–128
5. Eddamiri S, Zemmouri E, Benghabrit A (2021) Theme identification for RDF graphs based on LSTM neural recurrent network in the international conference on artificial intelligence and computer vision, pp 711–720

6. Luz Martın-Peña M, Dıaz-Garrido E, Sánchez-López JM (2018) The digitalization and servitization of manufacturing: a review on digital business models. Strateg Change 27:91–99
7. Bloomberg J (2018) Digitization, digitalization, and digital transformation: confuse them at your peril. Forbes. Retrieved 28 Aug 2019
8. Gobble MM (2018) Digitalization, digitization, and innovation. Res Technol Manag 61:56–59
9. Paavola R, Hallikainen P, Elbanna A (2017) Role of middle managers in modular digital transformation: the case of Servu
10. Demraoui L, Behja H, Abbou RB et al (2016) A case-based reasoning approach to the reusability of CWM metadata. In: 2016 third international conference on systems of collaboration (SysCo), pp 1–6
11. Henriette E, Feki M, Boughzala I (2015) The shape of digital transformation: a systematic literature review. In: MCIS 2015 proceedings, vol 10, pp 431–443
12. Schwertner K (2017) Digital transformation of business. Trakia J Sci 15:388–393
13. Sunyaev A (2020) Internet computing. Springer, pp 195–236
14. Ratten V (2020) Sport startups: new advances in entrepreneurship. Emerald Publishing Limited
15. Wang WYC, Wang Y (2020) Analytics in the era of big data: the digital transformations and value creation in industrial marketing
16. Dahaoui F-Z, Demraoui L, Louhdi MRC, Behja H (2021) Advances on smart and soft computing. Springer, pp 235–245
17. Eddamiri S, Zemmouri E, Benghabrit A (2020) Theme identification for linked medical data. In: International conference on artificial intelligence & industrial applications, pp 145–157
18. Ceipek R, Hautz J, De Massis A, Matzler K, Ardito L (2021) Digital transformation through exploratory and exploitative internet of things innovations: the impact of family management and technological diversification. J Prod Innov Manage 38:142–165
19. Tang D (2021) What is digital transformation? EDPACS 64:9–13
20. Châlons C, Dufft N (2017) The drivers of digital transformation. Springer, pp 13–22
21. Yalina N, Rozas I (2020) Digital workplace: digital transformation for environmental sustainability. IOP Conf Ser Earth Environ Sci 456:012022
22. Sturgeon TJ (2021) Upgrading strategies for the digital economy. Glob Strategy J 11:34–57
23. Reis J, Melão N (2021) The path to digital transformation: overcoming prejudice in the digital era with service operations. Int J Serv Oper Manag 39:81–97
24. El Otmani Dehbi Z et al (2021) Moroccan digital health response to the COVID-19 crisis. Front Public Health 9:690462
25. Bensbih S, Bouksour O, Rifai S (2019) On line appointment systems in a patient centric strategy: a qualitative approach in a case study for hospitals in Morocco. In: 2019 6th international conference on control, decision and information technologies (CoDIT), pp 1735–1739
26. Kerroum K, Khiat A, Bahnasse A, Aoula E-S et al (2020) The proposal of an agile model for the digital transformation of the University Hassan II of Casablanca 4.0. Procedia Comput Sci 175:403–410
27. Ahmad HAS, El Kharki K, Berrada K (2020) Agility of the post COVID-19 strategic plan on distance learning at Cadi Ayyad University. An opportunity towards a total digital transformation of the university. In: International workshop on higher education learning methodologies and technologies online, pp 199–213
28. Khanboubi F, Boulmakoul A (2019) Digital transformation in the banking sector: surveys exploration and analytics. Int J Inf Syst Change Manag 11:93–127
29. Peshkova GY, Zlobina O (2020) Digital transformation of banking with speech technologies. In: European proceedings of social and behavioural sciences EpSBS, pp 294–303
30. Figuigui B, Machrouh F (2020) Banking governance in the era of digital transformation. J Res Adm Sci 9:10–16. ISSN: 2664-2433
31. Abdulquadri A, Mogaji E, Kieu TA, Nguyen NP (2021) Digital transformation in financial services provision: a Nigerian perspective to the adoption of chatbot. J Enterpr Commun People Places Glob Econ