November 19, 2022

2022 WIESymp

3rd International Women In Engineering Symposium

WIESymp 2022

Organized IEEE Women in Engineering Sri Lanka Section

SUSTAINABLE TRANSFORMATION OF TECHNOLOGY”

PROCEEDING BOOK

ISSN : 2756-908X

Table of Contents

Message from Chair- IEEE Sri Lanka Section

It is my privilege and pleasure to convey this message on behalf of the IEEE Sri Lanka Section for the 3rd International Women in Engineering Symposium (WIESymp 2022) under the theme “Sustainable transformation of Technology”.

This is a timely topic to discuss amid the economic crisis of Sri Lanka after the post covid era. I am happy to see that the WIE symposium brings many speakers who talk about sustainable technological solutions and cutting-edge technological research. This will create a great platform to have networking opportunities that can ignite collaborative research among the women's engineering community.

Knowledge dissemination is a major part of any academic and this symposium has created opportunities for researchers both local and international to publish their multi-disciplinary research under the IEEE banner. IEEE WIE affinity group attached to IEEE Sri Lanka Section is very active in research, community work, as well as volunteering activities. Hence, the IEEE Sri Lanka section is always willing to collaborate with them. As the chair of, IEEE Sri Lanka Section, I would like to put my blessing to the symposium.

I would like to take this opportunity to congratulate all presenters who presented during the symposium. Further, I would like to thank the organizing committee of the WIE symposium 2022 headed by Dr. Akila Wijethunga.

I am confident that you will enjoy the technical program as it will be inspiring, valuable, and exciting.

Thank you

Prof Pradeep Abeygunawardhana
Chair,
IEEE Sri Lanka Section

Message from General Chair – WIESymp 2022

As the General chair, I warmly welcome you to the Third International Women Women in Engineering Symposium (WIESymp) 2022, which will be held entirely online on November 19th, 2022 which was organized by IEEE Women in Engineering (WIE) Sri Lanka Section. The organizing committee is delighted to have your participation and contribution in WISymp 2022. The research presentation tracks allow for virtual interaction, knowledge sharing, and the exchange of creative ideas with regard to potential and upcoming developments in a variety of engineering disciplines. The event was organized under the guidance of a reputed international advisory board and I take this opportunity to thank all our advisors Electronics and Telecommunications, Intelligent Systems and Robotics, Information and Communication Technology, and Power and Energy were the four main technical tracks that made up WIESymp 2022 Peer reviews of each paper were conducted by reputable national and international subject matter experts. After two rounds of comprehensive assessment, 70% of the applications were accepted to be presented at the symposium. I extend my sincere gratitude to the members of the organizing committee, technical program committee, paper reviewers, all volunteers, and the participants of WIESymp 2022. I also congratulate the authors of the papers that were accepted. They provided us with numerous contributions that assisted us perform this event successfully. I sincerely hope you will enjoy the conference and work with us in the future.

Dr Akila Wijethunge

General Chair, WIESymp 2022
Lecturer, Faculty of Technology, University of Sri Jayewardenepura

Message from the Technical Program Committee Chair

On behalf of the Technical Program Committee (TPC), we are proud to present you with an excellent technical program covering a wide range of topics focused on design, analysis, and innovations under the theme of ‘’Sustainable Transformation of Technology’’. We are pleased that you consider WIESymp to be a flagship conference and worthy of your time as an author and attendee.

This year, for the first time the conference called full papers covering diverse areas of Engineering, under four tracks, namely Electronics & Telecommunications, Intelligent Systems & Robotics, Information & Communication Technology, and Power & Energy.

The WIESymp 2022 received submissions from both local and international authors. All submitted papers underwent a rigorous and fair peer review process by subject domain experts, from both industry and academia. Measures were taken to ensure that each submission was peer-reviewed by at least two reviewers and two rounds of review cycles were carried out to provide an opportunity for the authors to improve their manuscript while addressing the reviewer's comments.

The Technical Program Committee comprises eminent researchers who made a timely contribution throughout the review process in identifying relevant reviewers, coordinating with them and using their expertise to finalize acceptance. The commendable service rendered by the panel of reviewers, laid the foundation for thorough scrutiny of the submission while elevating the quality of the symposium proceedings.

Based on the reviews and meta-reviews, the TPC co-chairs recommended accepting 14 papers, which are published in these proceedings. The recognition awarded by the IEEE Sri Lanka Section for the symposium is appreciated, as it encouraged us to deliver a high-standard symposium.

On behalf of the entire Organizing Committee, I believe that you will enjoy the WIESymp 2022 and look forward to seeing you once again in 2023. Finally, I wanted to thank all of you who have contributed to WIESymp as an author, reviewer, or TPC member.

Dr. Maheshi Buddhinee Dissanayake - TPC Chair/ WIESymp 2022

Message from the Publication Chairs – WIESymp 2022

On behalf of the publication committee, we would like to extend a warm welcome to all participants and delegates of the third International Women in Engineering Symposium (WIESymp) 2022. Continuing from the highly successful symposium we held in 2020, this year's session too will be held online as a virtual symposium.

This year's theme of ***Sustainable transformation of Technology*** is quite appropriate when the world has just witnessed the closure of the 27th Conference of the Parties to the United Nations Framework Convention on Climate Change. As Sri Lanka navigates the challenging economic environment, this year's theme holds high relevance.

We received 17 papers for the symposium and after careful review, 12 papers were selected for presentation at the symposium. We thank our reviewers immensely for the time spent in reviewing the manuscripts, as well as timely feedback. Based on the submissions we have pooled the papers into three tracks.

Furthermore, we express our sincere appreciation to all authors of papers of the symposium. It is the result of their generous contribution of time and effort on engineering and technology related research. The willingness to make an effort to share knowledge and thoughtful insights with the engineering and technology community is greatly appreciated which has made this conference proceedings possible.

We also expect to provide technical demonstrations, and numerous networking opportunities to jointly explore current and future research directions.

We hope the virtual symposium creates the ideal environment for providing great insights for all our participants and presenters.

Dr.PrabhaniLiyanage and Ms. Abarnah Kiruppananda
Publication Chairs/ WIESymp 2022

Keynote Speaker:
Speaker Profile – Prof. Pradeep Abeygunawardhana

Prof. Pradeep Abeygunawardhana is the present Chair of the IEEE Sri Lanka Section

He entered the University of Moratuwa in 1997 and graduated with BSc. in Electrical Engineering in 2002.

In 2004, He entered to the Graduate School of Science and Technology of Keio University, Japan as a master's student and he obtained his Master's and Ph.D. degrees in robotics from the Keio University in 2006 and March 2010 respectively. His research area was the non-linear control of two-wheel manipulators.

Upon completion of his higher studies, He joined to Sri Lanka Institute of Information Technology as a senior lecturer (Higher Grade) in May 2010. He was appointed as the head of the research center at the SLI IT in June 2011.

He successfully organized the SLIIT ROBOFEST 2010 and ROBOFEST 2011. Robofest is one of the premier robotics competitions in Sri Lanka.

Since March 2012, he has been working as a postdoctoral researcher at Kagawa University, Japan.

He also serves as the Director - Technology Diffusion - at the Information and Communication Technology Agency of Sri Lanka

His Interested research areas are Agricultural, Robotics, the Internet of Things, Artificial Intelligence, Multi-Robot Communication, and Corporate Control of robots.

Keynote Speech by Prof. Pradeep Abeygunawardhana

Sustainable Transformation of Technology

Transforming a product or a service to a new level through the applications of technology is called technology transformation. In the meantime, sustainability is the ability to transform their firm is critical to surviving and capitalizing on a new wave of business disruption. Sustainable Technology transformation is transforming a business through the applications of technology transformation which will have a long-term impact on the business.

Sustainability technology transformation requires a deep understanding of emerging technologies, strong engagement with stakeholders, and a series of investments tailored to each unique organization. Traditional firms willing to start their multiyear transformation journey toward sustainability must embrace three key principles:

1. Bold leadership. Lead with a bold vision that aligns your business model to a unique purpose.
2. Sustainable execution. Drive action by integrating sustainable execution deep into your organization.
3. Aligned stakeholders. Align internal and external stakeholders by delivering on your sustainability pledge.

Sustainable technology transformation mainly depends on three factors Collaboration, Technology diffusion, and infrastructure facilities.
Synergy is combining two things to get better output. Collaboration isworking together for a common purpose to achieve better output. Collaboration creates advanced technological solutions which cover different aspects. Technology diffusion is adopting technologies to the grassroots level assuring benefits for the user and making an impact. Technology diffusion works based on three factors Education, Research and Development, and commercial value. Without Infrastructure facilities, the development and deployment of technological solutions are not possible. Therefore, we need to have proper and required infrastructure facilities.

INTERNATIONAL WOMEN IN ENGINEERING SYMPOSIUM 2022

Sustainable Transformation of Technology

Agenda

Saturday, November 19, 2022 - 09:00 AM IST(UTC+05:30)

9.00am - 9.10am **:**Welcome Address –
General Chair, WIESymp 2022

9.10am - 9.25am **:**Keynote Address –
Prof Pradeep Abeygunawardhana,
Chair, IEEE Sri Lanka Section, 2022

9.25am - 9.30am **:** Vote of Thanks –
Secretary, WIESymp 2022

Organizing Committee

❖ **Conference Chair**	**:** Dr. Akila Wijethunga
❖ **Conference Secretary**	**:** Ms. Shashika Lokuliyana
❖ **Technical Program Committee Chair**	**:** Prof. Maheshi Dissanayake
❖ **Finance Chairs**	**:** Dr. Pubudu Jayasena
	: Ms. Warnakula Hippola
❖ **Publicity Chairs**	**:** Dr. Damayanthi Herath
	: Ms. Gresha Samarakkody
❖ **Local Organizing Chair**	**:** Ms. Sewwandie Nanayakkara
	: Ms. Upeksha Kudagamage
❖ **Publication Chairs**	**:** Dr. Prabhani Liyanage
	: Ms. Abarnah Kiruppananda
❖ **Awards Committee Chairs**	**:** Dr. Rasara Samarasinghe
	: Dr. Lasanthika Dissawa
❖ **Workshops and Special Sessions Chair**	**:** Dr. Hiruni Rupasingha
	: Ms. Heshani Mahalaksha

Advisory Committee

- Prof. Pradeep Abeygunawardena – Chair, IEEE Sri Lanka Section
- Mr. Deepak Mathur – Director of IEEE Region 10 (Asia-Pacific Region)
- Ms. Emi Yano – Chair, IEEE Region 10 Women in Engineering (WIE)
- Prof. Dileeka Dias – Department of Electronic and Telecommunication Engineering, University of Moratuwa
- Prof. Kumudu Perera – Department of Electronics, Faculty of Applied Sciences, Wayamba University of Sri Lanka
- Prof. Chinthaka Premachandra, Professor- Department of Electronic Engineering, Shibaura Institute of Technology
- Dr. Malka Halgamuge – Department of Electrical and Electronic Engineering, Faculty of Engineering and Information Technology, The University of Melbourne
- Prof. Saman Halgamuge – Department of Mechanical Engineering, School of Electrical,Mechanical and Infrastructure Engineering, The University of Melbourne
- Prof. Nihal Kularatna- School of Engineering, The University of Waikato

Technical Program Committee

- **Honorary TPC Chairs**

  Snr. Prof. Janaka Ekanayake, Chair Professor of Electrical and Electronic Engineering, Department of Electrical and Electronic Engineering, University of Peradeniya

- **Technical Program Committee Chairs**

  Prof. Maheshi Dissanayake, Professor, Department of Electrical and Electronic Engineering, Faculty of Engineering,
  University of Peradeniya

- **Track 01: Electrical, Electronics and Telecommunication**
  - Prof. (Mrs.) J.M.J.W. Jayasinghe, Professor, Department of Electrotechnology, Faculty of Technology, Wayamba University of Sri Lanka.
  - Dr. Kasun Hemachandra, Senior Lecturer, Department of Electronic and Telecommunication Engineering, University of Moratuwa.
- **Track 02: Intelligent Systems & Robotics**
  - Dr. Achala Pallegedara, Senior Lecturer, Department of Manufacturing and Industrial Engineering, Faculty of Engineering, University of Peradeniya.
  - Dr Malaka Miyuranga, Department of Materials and Mechanical Technology, Faculty of Technology, University of Sri Jayewardenepura.
- **Track 03: Information & Communication Technology**
  - Dr. Damayanthi Herath, Senior Lecturer, Department of Computer Engineering, Faculty of Engineering, University of Peradeniya
  - Ms. Abarnah Kirupananda, Senior Lecturer, Informatics Institute of Technology

List of Reviewers

- Prof. Chinthaka Premachandra. - Shibaura Institute of Technology
- Prof. Jeevani Jayasinghe. - Wayamba University of Sri Lanka
- Prof. Maheshi Dissanayake - University of Peradeniya
- Prof. Lidula Widanagamaarachchi - University of Moratuwa
- Dr Akila Wijethunge - University of Sri Jayewardenepura
- Dr Amirthalingam Ramanan - University of Jaffna
- Dr Damayanthi Herath - University of Peradeniya
- Dr Isuru Shanaka Lakmal - General Sir John Kotelawala Defence University
- Dr Hiruni Rupasingha - Sabaragamuwa University of Sri Lanka
- Dr Kasun Hemachandra - University of Moratuwa
- Dr Pasan Maduranga - University of Vocational Technology
- Dr Ruwan Ranaweera - University of Peradeniya
- Dr Shama Naz Islam - Deakin University
- Dr Shanaka Gunasekara - University of Peradeniya
- Dr Yohan Weerasinghe - National Engineering Research and Development Centre
- Ms Abarnah Kirupananda - Informatics Institute of Technology
- Ms Shashika Lokuliyana - Sri Lanka Institute of Information Technology

Session 01: Electrical, Electronics and Telecommunication		
Paper ID	**Paper Title**	**Corresponding Author**
14	Soil Quality Monitoring System for Smart Farming Based on Internet of Things (IOT)	A. A. Lakshitha
10	An IoT-Based Architecture for Remote Animal Health Monitoring	Thisura Rajapakse
12	Optimal Feature Selection using Genetic Algorithm for Cyber-attack detection in Internet of Things	Pawara Tharkana
17	Analysing Dynamic Line Rating of power cables using Electrical and Thermal Analysis	H.M.C.G.B Herath

Soil Quality Monitoring System for Smart Farming Based on Internet of Things (IOT)

A.A. Lakshitha[1], H.M.S.L. Dissanayake[2], L.A.B.C. Ranasinghe[3], G.S. Samarakkody [4]

[1,2,3,4] *Department of Electronic & Telecommunication Engineering, CINEC Campus, Malabe, Sri Lanka*
[1] *ashanlakshitha412@gmail.com* , [2] *dissanayakesasini97@gmail.com* , [3] *buddhibcr18@gmail.com* ,
[4] *gresha.samarakkody@cinec.edu*

Abstract— **The quality of soil is one of the crucial factors which affects the growth of crops. But the contribution for the soil quality in farming sector in Sri Lanka is very less. The main cause for this is the lack of a reliable and cost-effective product to detect the soil quality. As a solution for this shortage, this particular system which uses the concept of IOT is designed as a full package of ensuring soil quality. It is implemented with the aid of an electronics simulation software. The parameters such as humidity and temperature of the field (Through DHT11 sensor), the soil moisture (Through soil moisture sensor), PH value of soil (Through PH sensor) and Nitrogen, Phosphorous and Potassium content of soil(Through NPK sensor) were able to detect under this particular system. An alerting mechanism by which the technology of Global System for Mobile communications (GSM) is used here. The obtained readings were displayed on a Liquid Crystal Display (LCD) with the use of Arduino Uno and could convey the status of the farm to the user through a Short Message Service (SMS).**

Keywords— Internet of Things, Reliable, Cost-effective, Soil quality

INTRODUCTION

Sri Lanka which holds a remarkable history, reveals that the economy had been completely based on an agricultural system. Since such ancient times, agriculture has been considered as the backbone of the livelihood of Sri Lankans. Basically, farming was the major element which linked the economy, society, culture and the religion in past era. The primitive farming prevailed at that period consumed the maximum intervention of human labour, leading to a sustainable economy and, these traditional techniques of farming has developed over time. Hence, the farming sector in Sri Lanka today has taken a different direction due to the rapid development of technology. This technological transformation during the last few decades has influenced the introduction of new concepts such as smart farming and IOT but it seems these technologies have not been fully embraced by the Sri Lankan agricultural sector yet. Therefore, an attempt has been taken to design a system involving these sophisticated technologies, for farming in Sri Lanka.

This particular project titling, Soil Quality Monitoring System for Smart Farming Based on IOT is an approach to introduce the concept of smart farming along with the principles of IOT. The term “smart farming” is a sensor monitored and software managed concept which monitors the farming fields with the aid of the modern technology. This concept is used in farming in order to increase the quality and quantity of the yield while optimizing the required labour. With the influence of globalization, this concept has become much popularized within the farmer community. With smart devices, multiple processes can be activated at the same time, and automated services enhance product quality and volume by better controlling production processes. Smart farming systems also enable careful management of the demand forecast and delivery of goods to market just in time to reduce waste [1]. In parallel to that, the concept of IOT is mainly applied in implementing this system. IOT, in essence is a network of physical objects that are embedded with sensors, for the primary intention of exchanging data with other devices. Basic intention of accompanying the IOT tools in data collection systems, is to maximize the output with minimum human intervention. This is applied in a vast range of applications in order to provide solutions related to technology-based extremes.

The advantages listed with this concept are; high productivity, efficiency, versatility, ease of access, and many more. The principal concept of the IOT is that it is connected to every single object, with sensors that can respond to situations as they arise. The applications of the IOT are comprehensive and cost-effective [2]. In smart farming systems, as the sensors are used in the process of detecting real time environmental parameters, the user is required to respond immediately according to the sensed readings. Therefore, IOT can be considered as the best solution to get the optimum output. The basic intention of using the concept of IOT in this particular Soil Quality Analysis System is to obtain the status of the farm and to monitor the farm remotely. IOT technology helps better control agricultural processes to reduce production risks and enhances the ability to foresee production results, which helps farmers better plan and distribute product [1].

When considering the parameters required for the crop cultivation, the quality of soil is one of the foremost factors which affects the growth of crops immensely. The soil health acts as the basis for the productive farming. As the available nutrient content in the soil has a massive impact on the growth of crops, the quality of soil can be considered as a critical element in sustainable farming. The distribution of these soil components in a particular soil is influenced by the five factors of soil formation: parent material, time, climate, organisms, and topography (Jenny 1941). Each one of these factors plays a direct and overlapping role in influencing the suitability of a soil for agriculture [3].

With the mechanization of the fields, most of the activities within the field have been done with the least human intervention. This has affected for the soil quality analysis process such that primitive methods used in the typical farmlands have been in less practice. But the worst aspect related to this in present Sri Lanka is the unavailability of a reliable mechanism to detect the nutrients and the quality of soil. Similarly, this scarcity of reliable and precise mechanism to detect the quality of soil has affected to the less productivity and the malfunctioning of the sector.

Therefore, when considering the above facts, a sophisticated technology which is reliable and also affordable to the income level of the farmers is timely required as the farmers are incapable to go for technologies which require high knowledge and capital. By focusing on this timely requirement in farming sector, under this project, an attempt is taken to incorporate smart farming through IOT to create a positive impact on maintaining quality of soil.

The system implemented under this particular project titling Soil Quality Monitoring System for Smart Farming based on IOT is a software simulation-based project. This designed system is intended with the main aim of delivering a reliable, cost –effective mechanism to detect properties of soil and ensure the quality of soil in the farming fields in Sri Lanka. The most highlighting features of this system is that it is entirely acting as a full package of detecting properties of soil. The system discussed under this project was carried out through the Proteus simulation software. The specific sensors and other hardware components which were available in the Proteus platform have been used in the designing process. DHT11 temperature and humidity sensor, soil moisture sensor, pH sensor, NPK sensor, Arduino UNO board, GSM module, LCD display were used in the implementation process. In this implemented system, the DHT11 sensor was used in order to get the temperature and humidity readings of the farmland. Soil moisture sensor was used in the process of measuring the moisture of the soil. The pH sensor was used in determining the pH value of soil. Furthermore, the NPK sensor is used in determining the Nitrogen, Phosphorous and Potassium concentrations of soil. Interfacings of the sensors were done through Arduino and the real time readings which were sensed through the particular sensors, were depicted on the LCD display. Meanwhile, the data and the status of the farm were notified to the mobile phone of the farmer through a SMS. Though there are several other technologies of implementing, only the concept of IOT is utilized in this implemented system.

Simultaneously, rather than creating a mobile app and sending notifications via that app, this particular system has only concentrated on sending SMS alerts directly to the user by considering the shortage of using smart phones in remote areas in the country. Along with that, the system so designed has not incorporated the advanced technologies such as Wi-Fi based cloud applications in the alerting mechanism due to the less practicality for such technologies in rural village areas in Sri Lanka where farming is practiced. Thus, GSM has been utilized in the alerting mechanism as it is the best way of communication for this system. Based on the positive results obtained through this particular system, it is evident that this designed system to detect quality of soil is one of the most reliable approaches in this field.

APPROACH

The system developed under this particular project is a software-based implementation such that the Proteus 8 simulation software was entirely used in the process. The libraries available in the Proteus were used as the sensors to detect each and every input data. This designed system was implemented with the use of several sensors and other hardware components which are available in the Proteus simulation software. The particular sensors used in the implementation process were soil moisture sensor, pH sensor, NPK (Nitrogen, Phosphorus and Potassium) level tester and the DHT11 sensor. The components like Arduino Uno board, Arduino GSM module and 20*4 liquid crystal display (LCD) were used along with the sensors. Other than that, the virtual terminal was used, because as this is a software-based implementation, a virtual terminal is required to display the message sent by the GSM module to the mobile phone of the user. By using all the above-mentioned gadgets, the implementation was completed with the intention of analyzing the quality of soil in the farmland.

The sensors like DHT11 sensor, soil moisture sensor were already available in the Proteus 8 software. The DHT11 sensor was used to measure the humidity and the temperature of the field while the soil moisture sensor was used to measure the moisture of the soil. As the other sensors required for the implementation of this particular project were not available in the software, they were constructed with the use of the variable resistors. Therefore, the pH sensor which is to measure the pH value of soil and the NPK level tester which determines the Nitrogen, Phosphorous and Potassium concentrations of the soil were implemented by using variable resistors.

The basic functions carried out by the system are detecting the soil moisture, pH value of the soil, Nitrogen, Phosphorous and Potassium level of the soil, temperature and humidity of the farm, sending those details to the mobile phone of the user via a Short Message System (SMS) and finally depicting the sensed data on a LCD display which is placed at the farm .Because this was a software-based implementation, the output values were obtained through the software, and they were assumed as real time data.

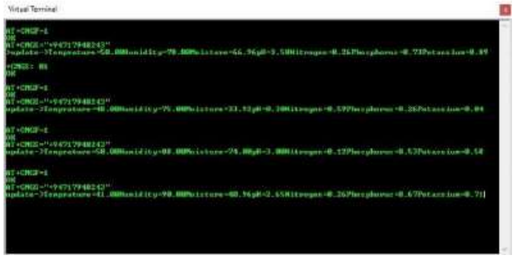
The connections between the individual sensors were made as follows.

The main four sensors namely soil moisture sensor, pH sensor, NPK level tester and DHT 11 sensor were connected to the Arduino Uno board. From the libraries available in the Proteus 8 software, the Arduino GSM module was selected and was connected to the control unit. Virtual terminal component was selected and its transmission pin (TXD) was connected to the receiving pin (PD0) of the Arduino Uno. The receiving pin (RXD) was connected into the transmitting pin of the GSM module respectively. With that, the receiving pin of the GSM module was connected into the transmitting pin (PD1) of the Arduino Uno board directly. The virtual terminal was used to display the message which received to the mobile phone of the farmer from the system. The 20*4 LCD display was connected to this system because it helps the farmer to get the information while working in the field.

Contrasting to the previous research works related to this area, in this project an advanced overall package is implemented in order to detect several soil parameters through one system.

RESULTS

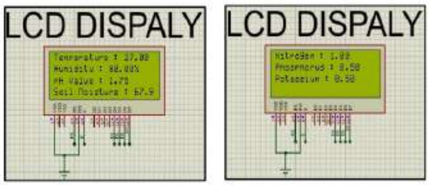
A short message (SMS) was received to the mobile phone of the user by indicating all the details including temperature of the farm, humidity of the farm, soil moisture content, pH value of soil, concentrations of Nitrogen, Phosphorous and Potassium respectively at one instance.



```
AT+CMGF=1
OK
AT+CMGS="+94717948243"
>update->Temperature=50.00Humidity=70.00Moisture=66.96pH=3.50Nitrogen=0.26Phosphorus=0.73Potassium=0.89
+CMGS: 86
OK

AT+CMGF=1
OK
AT+CMGS="+94717948243"
update->Temprature=40.00Humidity=75.00Moisture=33.92pH=0.30Nitrogen=0.59Phosphorus=0.26Potassium=0.04

AT+CMGF=1
OK
AT+CMGS="+94717948243"
update->Temprature=50.00Humidity=80.00Moisture=74.00pH=3.00Nitrogen=0.12Phosphorus=0.53Potassium=0.58

AT+CMGF=1
OK
AT+CMGS="+94717948243"
update->Temprature=41.00Humidity=90.00Moisture=40.96pH=2.65Nitrogen=0.26Phosphorus=0.67Potassium=0.71
```

Figure 1: Data Send to The Farmer Using Short Message System

All the above-mentioned readings namely temperature of the farm, humidity of the farm, pH value of soil, the soil moisture content, Nitrogen, Phosphorous and Potassium concentrations of the soil were depicted on the LCD display clearly**.**



```
Temperature : 27.00
Humidity : 80.00%
PH Value : 1.70
Soil Moisture : 67.9
```

```
Nitrogen : 1.00
Phosphorus : 0.50
Potassium : 0.50
```

Figure 2: Display the Measured Data on LCD

Other than having separate systems for measuring different parameters related to soil, under this particular project a strong and advanced entire package was able to develop which includes all the soil parameters in one system.

DISCUSSION

This particular project named Soil Quality Monitoring System for Smart Farming Based on IoT was designed with the main objective of introducing a reliable system to detect quality of soil along with the features such as soil moisture, soil pH level, Nitrogen, Phosphorous and Potassium concentration, and temperature and humidity of the farm. According to the system, the farmer can receive all information about soil quality parameters through a SMS notification and at the same time, the system will display the status of the farm along with soil details by using a LCD display. Soil health is the foundation for a productive farming. Fertile soil provides essential nutrients to plants. Not only the unavailability of a reliable mechanism to detect the soil quality, but also by considering the hardness for the farmers to measure all the soil quality determining parameters daily on a manual basis, the system implemented under this project is assigned to measure and provide these data on soil quality easily using IoT technology. The entire system was designed basically through Proteus simulation software and in the simulation process it was assumed that the sensor will provide real time values. The sensors sense the parameters and convert it into an electrical signal which is sent to the Arduino board. Through the implemented system it was able to send a SMS notification to the farmer's mobile through GSM module as well as was able to depict the status of the farm on the LCD display successfully. It is crystal clear that, the implemented system will be a smart solution for the soil quality analysis issue faced by typical farming sector in Sri Lanka.

CONCLUSION

This particular soil quality monitoring system was designed in order to alert the soil quality analysis parameters namely soil moisture content, soil pH level, respective Nitrogen, Phosphorous and Potassium concentrations of soil along with temperature and humidity of the farm to the mobile phone of the user through a SMS notification. This system does not require a smart phone compulsorily to convey the message to the farmer. Through this SMS alert, the farmer will be highly benefited as he can get an overall idea about the quality of soil of the farmland even without entering to the farmland. The user will be able to get steps and precautions to manage the farmland remotely by considering the factors like temperature, moisture content of the farm. Simultaneously, as there is a mechanism to represent the data obtained through the sensors on a LCD display, the farmer will be able to directly access the data on quality of soil at an instance where he is at the farmland. Therefore, he will not get any unnecessary burden to check the quality of soil manually as this system automatically calculates and delivers the soil quality parameters to the user. As this is a simulation-based system the implementation was carried with several assumptions. But as an attempt to monitor the soil quality in a farmland this designed system will be a solace for the issues of the farmer as the quality of soil is a critical factor which determines the growth of a crop and better yielding.

REFERENCES

C. Bernstein, "TechTarget, "TechTarget, June 2019. [Online]. Available: https://internetofthingsagenda.techtarget.com/definition/smart-farming. [Accessed 16 November 2021].

M. S. S. R. Razik Kariapper Ahmadh Rifai Kariapper, "Internet of Farming (IOF) and Internet of Things (IoT)," Journal of Information Systems & Information Technology (JISIT), vol. 3, no. 1, pp.23-35, 2018.

B. R. J. Sanjai J. Parikh, "Nature Education," 2012. [Online]. Available: https://www.nature.com/scitable/knowledge/library/soil-the-foundation-of-agriculture-84224268/. [Accessed 15 November 2021].

S. R. S. Dr.N. Suma," IOT Based Smart Agriculture Monitoring System" International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 2, pp. 177-181,2017.

G. R. R. S. Chatan Dwarkani M.." Smart Farming System Using Sensors for Agricultural Task Automation," in IEEE International Conference On Technological Innovations in ICT For Agriculture and Rural Development, Chennai, 2015.

An IoT-Based Architecture for Remote Animal Health Monitoring

KATD Rajapakse[1], MWP Maduranga[2], and MB Dissanayake [3]

[1] Department of Computer Engineering, General Sir John Kotelawala Defence University, Sri Lanka
[2] Department of Computer Engineering, General Sir John Kotelawala Defence University, Sri Lanka
[3] Department of Electrical & Electronic Engineering, Faculty of Engineering, University of Peradeniya, Sri Lanka

[1]36-ce-0003@kdu.ac.lk, [2]pasanwellalage@kdu.ac.lk, [3]maheshid@eng.pdn.ac.lk

Abstract- **There are over millions of animals used on farms for different purposes all over the world. Technological advancement especially in IoT has the potential to be used extensively in animal husbandry to improve livestock management and to observe its healthcare in order to improve the production volume of the farm. Animal health is one of the significant considerations for the daily production of a farm. Manually checking the health condition of each animal is inaccurate without experience. In this research, we have addressed some problems in animal health monitoring systems. In this work, we develop an Internet of Things (IoT) Based Remote Animal Health Monitoring System using wearable sensors. The proposed device can monitor the real-time heart rate and location of animals. We developed individual nodes that can communicate with the IoT server directly. The proposed system can be implemented for small-scale or large-scale animal farms. The primary goal is to develop a more reliable, accurate, and low-cost system that can be easily adopted in the developing world.**

Keywords—Internet of Things (IoT), Animal Health Monitoring.

INTRODUCTION

Technology made a massive difference in the agriculture sector when compared to the process of a few decades ago. With the help of sensors, machines, devices, and information technology, most of the processes became more accessible, more efficient, less time, and more human power-consuming. Applying innovations in farming makes an efficient and profitable environment for farmers to work. New technological tools and methods will make animal husbandry comfortable and easier to manage. Also, animal management decisions that need to be taken daily can be configured correctly using new technological applications. Human errors in decision-making and monitoring will directly affect product quality and profitability.

The production volume of an animal depends on factors such as genetic background, environment, diseases, feeding, climate, year, and season, which are identified as affected factors for production. Monitoring animal health and taking the necessary actions on time will keep the economy stable and the safety of the country's food supply. Animal disease outbreaks can be directly affected the country's economy due to trade halts, animal slaughters and subsequent disease eradication efforts. This can also affect public health, the stability of the agriculture sector, and global trade.

Most farms track their animal health manually. The manual health monitoring system in farms is the main reason reduce daily production. There are several drawbacks to a manual system. There is a clear lack of knowledge about identifying the effect of the disease on the animal on time. Late detection of disease will lead to a high amount of cost for medical treatments. The second drawback managing and keeping track of a large number of animals is very time-consuming and inherently labor intensive. Also, the manual system has highly subjective and the decision-making process is very often inaccurate and inconsistent. Currently developed some systems are very costly and some farmers cannot afford such a system. So, the goal is to develop a cost-efficient and accurate monitoring system.

This research is carried out to develop IoT based Animal Health Monitoring System to overcome several problems that farmers face during livestock. This system developed as individual nodes that farmers can purchase the only required number of units and can be affordable by small-scale farmers. Installation of the unit and linking with the account can be done by the farmer manually by following the instructions.

In this design, the concept of a base station is completely removed. Node is capable to communicate directly with IoT servers using GSM/GPRS technology. Because of direct communication, the distance is not limited. The system will automatically send an alert to the farmer through email notification if the system detects an abnormal health condition of an animal.

This paper is structured as follows, Section 1 gives an introduction to the research problem, and section 2 gives a literature review on the related works. The following sections contain the methodology, current results, conclusion, and future works.

RELATED WORK

The systems that are currently available address various types of issues. These systems used different technologies, parameters, and components. The below table summarises selected IoT systems from literature related to the problem analysed in this literature.

TABLE I
RELATED WORK SUMMARY

	Technology	**Problem**	**Solution**
A	IoT,	Detect	Using temperature,

[1]	LoraWAN, NFC, Arduino	diseases & insemination movement.	accelerometer, and GPS data develop a health monitoring system.
B	IoT, Raspberry-PI, Arduino, Machine Learning	Detecting diseased cows.	Using temperature and heart rate data develop a diseased cow detection system.
C [2]	IoT, Arduino, MATLAB Analysis Tool	Measure milk production of cattle.	Using temperature, humidity, heart rate, and rumination data and predict the milk yield of the cow.
D	Wireless Charging (Inductively Coupling)	Power consumption on wearable attaches to the cow.	Using the coupling method inductively charge wearable
E	IoT, Arduino, Raspberry-PI, Solar technology, Machine Learning	Detect the behavior of dairy cows.	Using temperature, heart rate, and accelerometer data, analyze cow behavior in three stages.
F [3]	IoT, Arduino	Activity detection of the cow.	Using acceleration, gyroscope, and GPS data develop an activity detection system.
G	IoT, Arduino	Detecting behavior monitoring of cows.	Using accelerometer data, develop a behavior detection system.
H	NB-IoT, Arduino	Monitor estruses the state of a cow.	Using temperature and accelerometer data develop an estruses monitoring system.

Most of these developed systems require a base station. Nodes will communicate with the base station and the base station will send collected data to the cloud server. In this method distance between node and base station will be limited. Most farmers send their animals out of the farm to find food for themselves. In this situation base station concept is practically unreliable because once the animal in out-of-range communication is lost. Also required extra components. Due to this system costs will increase. The accuracy of the components selected in some systems is low such as using DHT11 temperature to measure the body temperature of the animal. This sensor is mainly used to identify the temperature and humidity of the environment.

SYSTEM DESIGN

This system consists of 3 main sections. They are Wearable Smart collars, Cloud servers, and Web applications. Figure 1 will give an idea about the system architecture.

Smart collar: - This is a belt-type wearable Smart collar with sensors, placed around dairy cows' neck. The biosensors in the Smart collar are placed to make contact with the animal's body. The temperature sensor (MLX90614) and Heart rate sensor (Magene H64) are used to collect the body temperature and heart rate of the animal. Air quality sensor (MQ2) use to identify the environment where the animal moves from time to time. Atmega328p microcontroller is used in the Smart collar to connect sensors and communication module. SIM808 module is used to build the communication between the Smart collar and the IoT server. SIM808 comes with an inbuilt GPS, and this is used to track the animal's location, in real-time. The system is powered up by using an 11.1V Li-Po battery and using a Voltage transformer (LM2596) model to convert 11.1V to 7.2V. we used low power consumption modes to extend the lifetime of the device.

Cloud Server: - ThingSpeak IoT server is used as the IoT Cloud server to store and analyze the received data from the Smart collar. The server will receive data from the Smart collar through General Packet Radio Service (GPRS) technology. This server can store 8000 individual data points. We send one data set every six minutes of a time gap to keep one month of health data of an animal. According to the pre-defined algorithm, if it detects an abnormal behavior of an animal, the user will be notified about the health condition through an email. Cloud Server will be available 24/7 to retrieve data for the user from anywhere.

Web Application: – Using a web application, users can access individual animal profiles. Users can create a user account. The user data will be saved on Firebase with enabling Firebase authentication. The application will manage the history of health conditions. Also, users can check the location of the dairy cow through the application.

SIM808 module compatible with 2G quad band comes with internal TCP/IP stack to enable connecting internet via GPRS. This has Circuit Switch Data (CSD) speed up to 14kbps[4]. 2G is the widest spared network coverage in Sri Lanka[5]. Because of the island-wide coverage, communication nodes can be connected to the IoT server almost everywhere around the country.

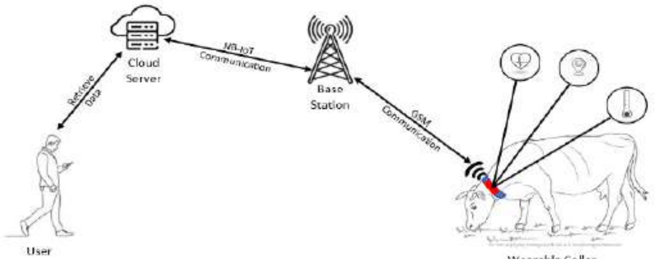


Figure. **1** System Overview

NOVELTY

In the proposed system, the Base station concept is completely removed. Without having any wireless router Smart collar will be directly communicate with 3G/4G mobile tower using SIM module. The Smart collar can communicate directly with IoT server and directly upload data using GPRS technology.

Also, in our design, contrary to other designs we measure the body temperature and heart rate of the animal. These two measures alone, help to diagnose the presence of infection in animals. Early diagnosis of potential infections would assist the farmers in better managing their herd and stopping the spreading of the disease through isolation of the infected.

Also, a gas sensor is included in this system to identify the air quality around the cattle. Maintaining good oxygen levels in the cowshed will help to increase productivity, while poor air quality indicators would again assist the famers to take preventive measures and improve the management of their livestock.

Furthermore, this system is energy efficient, cost effective and easy to use. Farmers can independently implement this system by following several steps.

EXPERIMENT AND RESULT

Once the communication between Smart collar and ThingSpeak IoT server is established, and data is uploaded to the server channel. A prototype Smart collar designed by the authors is used to capture heart rate, body temperature, air quality, and GPS location data from the animal, specially a cattle in this scenario. Collected data is retrieved from the IoT server to a web page to visualize the real-time variations. Figure 2 shows the parameters updated on the ThingSpeak channel in real-time.

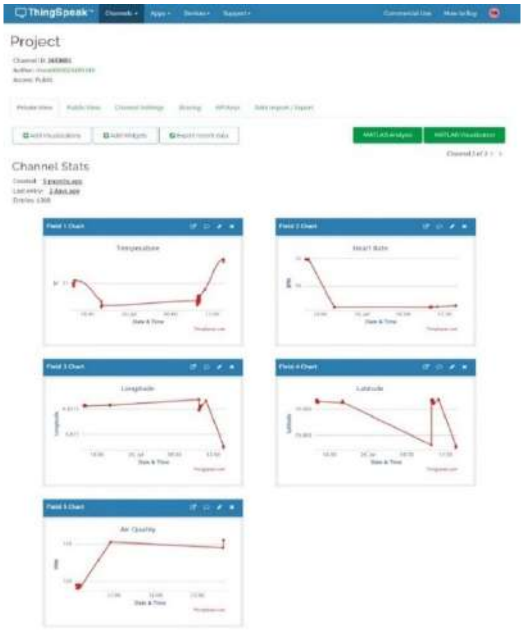


Figure. 2 Thing Speak Channel

Power consumption of this system is reduced in this system using inbuild functionalities and methods. Frequency of the microcontroller is reduced from 240MHz to 80MHz. Reducing the clock speed of the microcontroller furthermore will affect to the SIM808 GPRS communication. During the inactive period microcontroller will change to power down mode. **LowPower.h** library provide 6 different power saving modes and power down mode is best suitable for this system. Power supply to the Temperature and Air quality sensor will be managed using a transistor (IRFZ44N). Both sensors power up for 30 seconds to capture data and turn off until next cycle. In SIM808 after collecting GPS data, GPS function will be turn off using "AT" command until the cycle begins again. Once all the data is ready to send to IoT server GPRS function will be called.

FUTURE WORKS

In future works, the authors planned to build the prototype Smart collar units and collect real-time test data for a herd of cattle in a practical environment. Smart collar to Smart collar communication method needs to introduce to build a sensor network inside the farms. Although the initial deployment was specifically designed to cater for the cattle farms, the authors plan to extend the units for monitoring other animals as well. The web applications will be upgraded to monitor two or more categories of animals at the same time base on the requirement.

CONCLUSION

Animal products are one of the essential food items in the world. Technological adaptations in animal husbandry are still at the primary level. Implementing innovative devices and networks using state-of-the-art technology will help to increase local production. In most of farms, farmers manually check livestock health and due to the lack of knowledge this process is less efficient. Also, the manual process is very time-consuming and increase labour cost. With the help of IoT technology authors propose a cost-effective IoT Based Animal Health Monitoring System to improve production turnover while lowering the operation costs of livestock management.

REFERENCES

[1] F. Vannieuwenborg, S. Verbrugge, and D. Colle, "Designing and evaluating a smart cow monitoring system from a techno-economic perspective," *Jt. 13th CTTE 10th C. Conf. Internet Things - Bus. Model. Users, Networks*, vol. 2018-Janua, pp. 1–8, 2017, doi: 10.1109/CTTE.2017.8260982.

[2] V. Mhatre, V. Vispute, N. Mishra, and K. Khandagle, "IoT based health monitoring system for dairy cows," *Proc. 3rd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2020*, no. Icssit, pp. 820–825, 2020, doi: 10.1109/ICSSIT48917.2020.9214244.

[3] Z. Wang, C. Cao, H. Yu, and Y. Liu, "Design and Implementation of Early Warning System Based on Dairy Cattle Activity Detection," *2020 Int. Wirel. Commun. Mob. Comput. IWCMC 2020*, pp. 2186–2189, 2020, doi: 10.1109/IWCMC48107.2020.9148122.

[4] "SIM808 | SIMCom | smart machines, smart decision | simcom.ee." https://simcom.ee/modules/gsm-gprs-gnss/sim808/ (accessed Jul. 29, 2022).

[5] "Operator Watch Blog: Dialog Axiata: On its way to 5G after successful roll-out of commercial 4.5G." https://www.operatorwatch.com/2018/09/dialog-axiata-on-its-way-to-5g-after_67.html (accessed Jul. 29, 2022

Optimal Feature Selection Using Genetic Algorithm for Cyber-Attack Detection in Internet of Things

Pawara Tharkana[1], Maheshi B. Dissanayake[2]

[1,2]*Department of Electrical and Electronic Engineering, University of Peradeniya, Peradeniya, Sri Lanka*
[1] *pawara101dassanayake@gmail.com,*[2] *maheshid@eng.pdn.ac.lk*

Abstract— **Providing a reliable secure connection is an utmost important feature in networking. Security concerns, especially security breaches, are one of the key problems experienced by the Internet of Things (IoT) networks. In the research presented, the Message Queue Telemetry Transport (MQTT) protocol-created traffic is examined with the aim of automating the identification of potential security vulnerabilities in five categories. Brute force, denial of service (DoS), flooding, corrupted data, SlowITe assaults, and legitimate traffic are the five forms of cyberattacks taken into consideration. To simplify the analysis, first, the important features of the MQTT traffic data are extracted using the Genetic Algorithm. Next, widely utilised Machine learning algorithms, LGBM and Decision tree are applied to forecast cyber-attacks using the selected feature set. It is evident that each model's accuracy rose as the number of features increased to a certain point. On the other hand, it demands greater processing power to carry out the evaluation as the number of features increase.**

Keywords— MQTT, Decision tree, LGBM, Genetic Algorithm

Introduction

As we move toward the Digital Decade, numerous new technologies are being developed, and they all have the potential to drastically alter our way of life. The Internet of Things (IoT) links the physical and digital worlds so that linked items can report on their conditions and their surroundings. Large-scale advantages of the IoT include the industry's usage of IoT to assist with running factories; sensors in fields collect data that aids farmers in making better decisions; and the ability to outfit entire cities with sensors and monitors to transform into smart cities. It is crucial to take safety precautions against cyberattacks since IoT devices have access to sensitive user data [1],[3].

With the advancements in data analysis methods, these security problems can be tackled using Artificial Intelligence (AI)/Machine Learning (ML) models that rely on past data and utilise supervised learning, unsupervised learning, or reinforcement learning. These AI/ML approaches would eventually create a more secure cloud environment and increase the likelihood that the Internet of Things can reach its full potential [1]. A sizable dataset is needed to train an ML model for distinguishing effectively between legitimate and harmful actions.

In the work presented, we have utilised the MQTT protocol base dataset collected in an IoT network. The IoT datasets are massive by nature. The size of the dataset adopted for the study presented several difficulties in terms of the resources allotted. To assess the model's correctness and to facilitate manageable model training and tuning facilities, this experiment focuses on lowering the number of features that are pertinent. Genetic algorithm (GA) based feature selection was employed to identify the key features of IoT traffic which assist the intruder detection process. In overall, the paper present two novelties to the intruder detection system in IoT. First it proposes to use a reduce dataset to improve the computational power usage and the resource management such as storage. The second novelty of this experiment is the utilization of GA to find optimal features while using time and resource optimally for the supervised ML predictions.

Background Work

Research on predicting malicious behaviours in networks has been around since the 1990s [4]. With the advancements in data analysis techniques, such as ML, a great deal of literature has surfaced that research on the potential for utilising ML models to forecast cyber security breaches [4]. Analysing security breaches in IoT networks have been studied in several literatures [3],[4]. Many of these works utilise general intruder detection datasets such as KDDCUP99 dataset [2], while recent studies utilised datasets collected specially in IoT environments [3],[4]. While many literatures focus on categorising types of cyber-attacks [4], some attempt to identify key features within IoT traffic which help to identify cyber-attacks. [3],[4]. In this work, we analyse IoT traffic generated by MQTT protocol using state-of-the-art ML models, in order to categorise cyber-attacks. Also, using GA we attempt to identify key features (i.e., parameters) in MQTT traffic which helps to identify cyber-attacks.

Methodology

A. Data

This work aims to create a refined subset of the MQTT IoT dataset [4] using MQTT protocol base communications. The MQTT set was built using IoT-Flock. The particular data set used in this study is publicly available [3]. It consists of 3 sub-datasets, namely Randomly selected, reduced & augmented form. Data from legitimate traffic is randomly combined with the different malicious traffic data in randomly selected dataset to create the test set. The reduced form combines malicious traffic with legitimate traffic in 50:50 ratio to create the test set. In this study Augmented form is used for analysis. In the augmented form, the malicious traffic has been increased so that the sum of the traffic related to the attacks is equal to the legitimate traffic while each class of malicious traffic carries same amount of traffic. In this study, 5 types of attacks are predicted, based on the behaviour of the malicious traffic. Each incoming traffic consists of 33 parameters [4]. For this experiment the augmented form is used. The density of the selected dataset is shown in Figure 1.

B. Pre-processing, Feature selection & Models

The dataset used for this research consists of 6000000 entries of traffic ranging from both legitimate to malicious with 33 features for each entry [4]. Each entry belongs to one specific class out of the six listed below depending on the behaviour of the traffic. Hence, the classification models are trained for multi-class prediction problems. Further, the tasks of the research activity are extended to find an optimal number of features set out of a total of 33 features, with the objective of improving the resource utilisation as well as the model training time. Moreover, during pre-processing, the data set was separated into categorical columns and numerical columns. Then all categorical columns were encoded using the ordinal encoder. Then classification was performed under 6 classes:

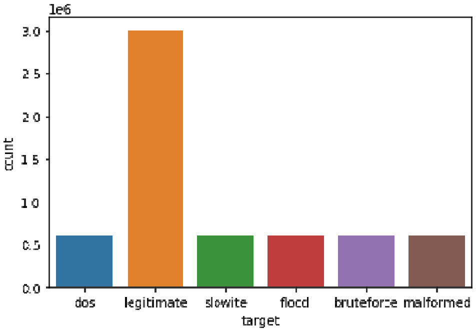


Figure 1. Dataset content

1 – (Denial of Service) Dos, 2 – slowIte, 3 – flood, 4 – brute force, 5-malformed and 6 – legitimate.

GA is a stochastic optimization algorithm. Genetic algorithms' mechanism, which simulates the natural evolution process to find the best answer, is exceedingly flexible, making it a great fit for the unpredictable and highly varied IoT scenarios. Its probabilistic natural elimination mechanism can save us from formulating rules in complex environments. Therefore, the subsets of variables selected by genetic algorithms are generally more efficient than those obtained by classical methods of feature selection.

During the experiment, at first the feature importance value for each of the 33 features has been derived using Genetic Algorithm feature selection. Then the original 33 features were ordered according to the feature importance. Next, a predefined number of features, specifically 5, 10, 15, 20 & 25 respectively, were selected considering the feature importance scores. Finally, these extracted features are used to train and predict cyber risks, using the LGBM & Decision Tree classifiers. The predictions were appeared in form of above-mentioned form (1 –Dos, 2 – slowIte, 3 – flood, 4 – brute force, 5-malformed and 6 – legitimate).

Results & Discussion

This section presents the results of genetic algorithm-based selection and LGBM & Decision tree classifiers-based prediction. All the experiments are repeated at least 4 times and the average output values are considered. All the experiments are carried out in a python environment and ML models are implemented using Sklearn, Light GBM libraries and genetic selection libraries [6].

A. Feature Selection

When utilising GA for feature selection it is essential to fine-tune its hyperparameters to obtain an accurate and cost optimised feature set [4],[5]. The optimised values of hyperparameters are listed below.

estimator=LinearRegression, cv=5, verbose=1, scoring="r2", n_population=20, crossover_proba=0.5, mutation_proba=0.2, crossover_independent_proba=0.5, mutation_independent_proba=0.05, tournament_size=3, n_gen_no_change=10, caching=True, n_jobs=-1

The values 40, 40, 40, 50, 50, are assigned to the parameter *n_generations* respectively for the number of features 5,10,15,20,25. Hence the extracted feature set the applied to the ML algorithm and then prediction is done.

B. Model fit and Hyperparameter tuning

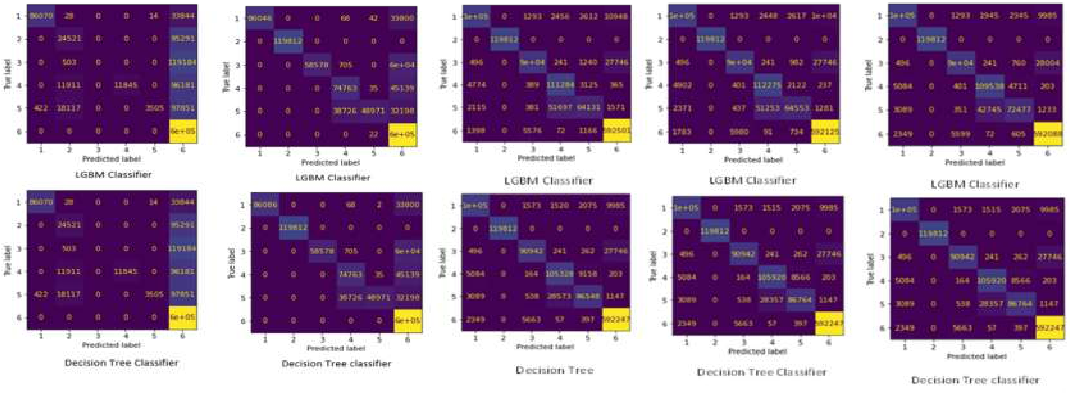
It is very important to tune the hyperparameters of both classifiers utilised. If not, the ML model may tend to overfit, underfit or lead to inaccurate predictions [6]. Parameter tuning takes a huge role in this experiment. Yet, the bulkiness of the dataset causes the hyperparameter optimization process to be time-consuming. The main hyperparameters of each classifier and its optimised values are listed below.

LGBM

learning_rate= 0.08, max_depth= 5, n_estimators=100, num_leaves= 31, boosting_type= 'gbdt', colsample_bytree= 1.0, reg_lambda= 0.5

Decision Tree

criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0



a) No of Features 5

b) No of features 10

c) No of Features 15

d) No of Features 20

e) No of Features 25

Figure 2. Confusion Matrices of predicted outputs

Also, the total dataset is segregated into 80% and 20% as train and test sets. Then the train set is used to train the model using "model fit". Then trained model is used to predict the possibility of a sample dataset belonging to one of 6 classes (1 –Dos, 2 – slowIte, 3 – flood, 4 – brute force, 5-malformed and 6 – legitimate) listed above, using the test dataset. Further, the performance model is evaluated using prediction accuracy, model prediction time, and confusion matrix. The confusion matrices for each selected feature set and for each classifier is shown Figure 2. Table 1 shows the summary of the outcomes obtained in this experiment. And the pictorial view of the above table observations is plotted in Figure 3.

According to the presented results, it is obvious that accuracy increases with the number of features and flattens after 15 features. On the other hand, the training time increases with the number of features. After 25 feature sets, the accuracy slightly gets lower compared to that of the 20 set, and it exhibits a very lengthy execution time and model training time.

One another important observation is that the Decision tree classifier has slightly more accurate predictions than LGBM classifier in higher dimensional data sets. When comparing the model training time, the decision tree has a very low model training time compared to LGBM classifier.

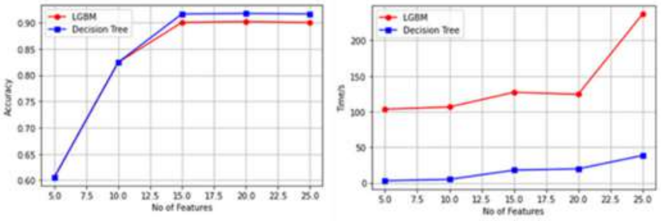
Predicting the Dos attacks resulted in a more accurate detection than the other attacks. As of the observations of confusion matrices, lower dimensional datasets have fewer false positive and false negative entries compared to 15,20 & 25-dimensional feature sets.

TABLE I SUMMARY OF EXPERIMENT OUTPUTS

Classifier	**LGBM**		**Decision Tree**	
No of features	Accuracy	Model training time/s	Accuracy	Model training time/s
5	0.6055	103.43902	0.6055	3.17241
10	0.8241	106.62445	0.8241	4.98449
15	0.9003	127.17778	0.9164	17.91887
20	0.9018	124.38311	0.9171	19.70952
25	0.9001	236.90764	0.9164	38.5628

CONCLUSION

This research analyses the possibility of classifying cyber-attacks using MQTT traffic data while utilising a minimum number of features. Hence, one of the key objectives of this experiment was to obtain the optimal number of features sufficient and necessary for the application using the Genetic algorithm (GA) based feature selection method for cyber-attacks in IoT networks. Using the GA, feature sets of 5,10,15,20 & 25 were obtained. It took a considerable number of resources and time to execute the GA algorithm. After training, prediction also required a high resource allocation. It was observed that a dataset with a dimensionality (no. of features) of 10 or 15 produced sufficient accuracy along with the allocation of time and resources.



a) No of Features vs Accuracy plot

b) No of Features vs Model training time plot

Fig 3. No of Features with accuracy & Model training time

REFERENCES

[1] Darley, O., Adenowo, A. and Yussuff, A., 2022. Machine Learning Intrusion Detection as a Solution to Security and Privacy Issues in IoT: A Systematic Review. FUOYE Journal of Engineering and Technology, 7(2), pp.148-156

[2] Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDDCUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defence Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6.

[3] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso, "MQTTset, a new dataset for machine learning techniques on MQTT," Sensors (Basel), vol. 20, no. 22, p. 6578, 2020.

[4] M. B. Dissanayake, "Feature Engineering for Cyber-attack detection in Internet of Things," International Journal of Wireless and Microwave Technologies, vol. 11, no. 6, pp. 46–54, Dec. 2021, doi: 10.5815/ijwmt.2021.06.05.

[5] K. Kanesamoorthy and M. B. Dissanayake, "Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm," Int. J. Mycobacteriol., vol. 10, no. 3, pp. 279–284, 2021

[6] Aurélien Géron, Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc., 2019.

Analysing Dynamic Line Rating of Power Cables Using Electrical and Thermal Analysis

H. M. C. G. B. Herath[1], Akila Wijethunga[2], and A. H. L. R. Nilmini[3]

[1,2,3]*Faculty of Technology, University of Sri Jayewardenepura, Sri Lanka*
[1]*chanukagayantha14@gmail.com,*[2]*akilawijethunge@sjp.ac.lk,* [3]*nilmini@sjp.ac.lk,*

Abstract— **At present, electricity demand is increasing due to rapid economic development and population growth. Also, there are tendencies to use renewable energy sources as well. The power flow capability of the existing power transmission lines is not sufficient to manage this increasing demand and this should be improved. It requires an unprecedented amount of capital investment. As an alternative solution, dynamic line rating can be used to operate the distribution networks optimally while maximizing the power generation of renewable energy sources. In the DLR approach conductor temperature plays a major role. In this research, conductor temperature will be simulated and analysed under different scenarios to identify the impact of different parameters. The results will be used to develop a coordinated application of dynamic line rating and demand-side management in a distribution network.**

Keywords— Dynamic Line Rating, Thermal Modelling, Overhead Cables.

INTRODUCTION

The electricity sector is growing rapidly, and the electricity demand is increasing daily. When the demand is increasing, the electricity network should be capable of distributing the correct level of electricity load to relevant areas. And also, there are tendencies and policies to ensure energy security through cleaner, secure, economical reliable supplies, thus providing convenient, affordable energy services to support socially equitable development. Most of the renewable resources are concentrated and often in remote locations. Therefore, development of the renewable energy resources is hindered by the bottlenecks in the distribution networks. New transmission lines or mechanisms to increase the power flow capability of the existing lines are required to overcome these bottlenecks. This procedure requires an unprecedented amount of capital investment.

As an alternative solution, power transmission based on dynamic line rating has been researched and basic methods are established. DLR applications use sensor networks that communicate between data processing centres and determine the DLR. A software environment is used to carry out the coordinated control of the renewable energy sources that take into account the dynamic ratings of the lines and other parameters in the distribution networks. Existing line capacity is determined conservatively based on the maximum allowable temperature and considering the worst-case scenario of weather and other conditions. By using a sensor network and a software environment it is possible to measure actual cable temperature and weather conditions to determine the practical allowable loading capacity of the power lines. Then the existing networks could facilitate renewable addition while minimizing the amount of capital investment required for a new line and substations.

MATERIALS AND METHODS

A. Dynamic Line Rating

Transmission lines are efficient and fast energy transmission channels and a key link for the safe operation of the power grid[1]. The current capacity of a power line is normally determined by the static rating of the line. This is defined as the maximum allowable value of current that can flow through transmission lines without adversely affecting the mechanical and electrical properties of the conductor[2]. DLR is a technology that can be used to improve the transmission efficiency and capacity of the existing power system without changing the system structure or breaking the current technical specification. It is an economical and feasible method to meet the increasing power demands and the need for new energy integration[1]. DLR reduces congestion on power lines, optimizes asset utilization, improves efficiency, and reduces costs. DLR permits increased solar and wind power integration while reducing curtailment for renewable energy sources and making power generation more cost-effective and cleaner.

DLR can be used to increase secure transfer capacity for transmission and distribution networks without building new lines. DLR determines the actual current-carrying capacity based on continuous measurements rather than a conservative assumption of weather conditions. Considering the fluctuation of the load curve and the intermittency of renewable energy, there is no need for the capacity-increase system of transmission lines to operate the whole time[4]. DLR technology can obtain the maximum current carrying capacity of transmission lines according to the DLR model by collecting or predicting the line environment and conductor status information. It improves the transmission efficiency and capacity of the transmission system without breaking the current technical regulations. DLR dynamically increase the transmission capacity of overhead lines (OHL) taking into account their thermal state and ambient conditions. Besides the main benefit, the increase of the OHL's transmission capacity, the DLR system optimizes the transfer of energy from renewable sources by predicting the production of energy that comes from these sources.

B. Approach to DLR

According to[3], there are two approaches to determining the DLR, which are the indirect and direct approaches.

Indirect DLR estimation is a prediction-based method using data from local weather stations or generating the data by using numerical weather modelling. The direct approach is based on monitoring line characteristics such as conductor temperature, line sag, the tension through the line and clearance to the ground. In recent years several mathematical methods have been proposed to predict the thermal rating of the conductor and to keep the maximum allowable current below safety limits[6]. The reliability of the DLR technology relies on the acquisition of key status information of transmission lines environment data and conductor status information. The main measurements include ambient temperature, solar radiation, wind speed, wind direction, conductor temperature, conductor sag, tension, vibration etc[7]. Devices installed on the transmission lines collect real-time conductor and environment information. The environment monitoring part is done using two common ways based on the device which is used. One way is by using weather stations to measure data at a specific point of the line. The other way is to get access to online data sources from satellites[8]. Direct monitoring of the line characteristics can offer more accurate measurements for the line ampacity associated with the weather measuring system[7]. Weather conditions, current intensity, conductor parameters, and direct measurements of the line's characteristics are considered input parameters. The reason behind direct measurements is to increase the accuracy of the DLR calculation in critical spans[3]. The device collects the data and sends it for processing and displaying through communication technology. In most cases, wireless communication methods like 4G, Wi-Fi and, Zigbee are used.

C. Conductor Temperature

DLR technology takes account of the transmission cable information and environment information to determine the maximum current capacity of the transmission line. The weather information includes ambient temperature, solar radiation, wind speed, and wind direction. Cable information includes the conductor temperature, conductor sag, tension, and vibration. The conductor temperature makes a great impact on the current-carrying capacity. It is also known as ampacity. The ampacity is defined as the maximum current, in amperes, that a conductor can carry continuously under the conditions of use without exceeding its temperature rating.

Regarding the conductor, several cable properties affect the conductor temperature which ultimately determines the ampacity. They are the cable diameter, material, and structure of the cable. Copper and Aluminium are materials that have high electrical conductivity and lower weight per unit volume. Aluminium has replaced copper because of its much lower cost and lighter weight. There are commonly used overhead power lines like AAC, AAAC, ACSR, ACAR and ABC. These cables come in different diameters and numbers of cables. Zebra, Racoon, Rabbit, Ant, and Gnat are some of the code words of those cables. Depending on the cable type, conductor temperature can be varied. Once the conductor cable is in operation, several parameters make impact the conductor temperature. They are current, voltage, wind speed, solar radiation, length of the cable and ambient temperature. Variations of the above parameters generate different surface and core temperatures of the cable.

D. Methodology

The thermal modelling and analysis characterize the inner heat exchanges among the cable layers and external heat exchangers with the environment under different conditions. The thermal model can be used to calculate the maximum operating temperatures that the cable can withstand and determine the steady state ampacity. The analysis can be carried out using several important input parameters such as current, voltage, irradiance, and wind speed. This allows for simulating different conditions and studying the thermal behaviour of the cable. The analysis provides valuable information such as the maximum current value that the inner conductor can carry continuously without exceeding the temperature limit values of the cable and the maximum time a cable can withstand an overload[9]. Cable static ratings are based on worst-case assumptions. Dynamic rating, which takes into account transient evolutions deliver increases of 5-20% in ampacity. Also, real-time change in environmental conditions strongly impacts the rate of heat dissipation from the cable. Electrical and thermal analysis of power cables can be used to determine the dynamic rating of the cables and implement DLR technology based on the results. The methodology of the proposed work begins with cable modelling. In this research, a simple cylindrical cable is used. In the modelling stage, cable diameter, current, voltage, wind speed, solar radiation and ambient temperature are set as the input parameters. Maximum and minimum temperature readings were set as the output parameters. The simulation was done for different cases to identify the impact of each parameter on the output temperatures.

Case 01: In the first case a 15mm diameter aluminium alloy cable was used in the simulation under the following conditions. Voltage 33kV, Environment temperature 280C, wind speed 1ms-1, convection rate 27.566 Wm-2 C-1. The current was changed from 100 to 200 and the maximum temperature of the conductor was observed.

Case 02: In the second case same 15mm diameter cable was used. Voltage 33000V, Environment temperature 280C, and Current was set to 200A. The wind speed was changed from 1ms-1 to 10ms-1 and the maximum temperature of the conduct was observed.

Case 03: The same 15mm diameter cable was used in the third case. Voltage 33000V, Environment temperature 280C, and Current was set to 200A. Temperature variation of the cable under four different radiation emissivity levels and ten convection rates were simulated.

RESULTS AND DISCUSSION

A. Case 1 Results

As shown in Figure.1, a minimum temperature of 28.980C was observed with a 100A current. A maximum temperature of 31.990C was observed when the 200A current was applied. The resulting temperature of the conductor increased by

3.010C when the current increased by 100A. According to the data obtained, a temperature increase can be seen when the applied current is high.

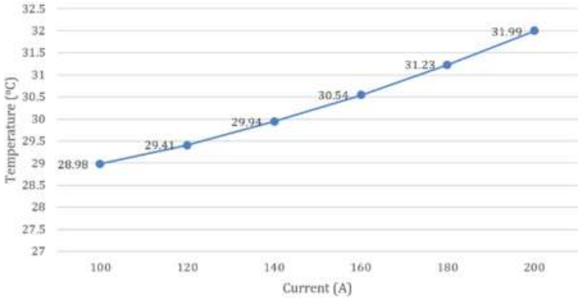


Figure. 1 Case 01 – Graph of current vs temperature

B. Case 2 Results

As shown in Figure.2, a minimum temperature of 29.230C was observed with a wind speed of 10ms-1. A maximum temperature of 31.990C was observed when the wind speed is 1ms-1. The resulting temperature of the conductor decreased by 2.760C when the wind speed varied between 1ms-1 to 10ms-1. High-speed wind increases the convection rate of the conductor and decreases the conductor temperature.

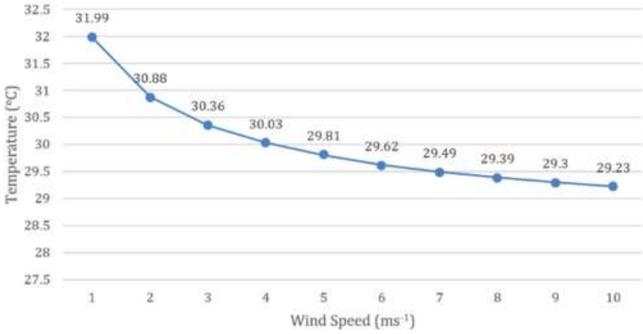
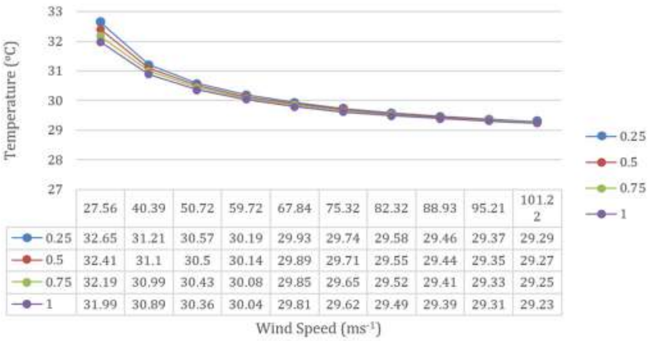


Figure. 2 Case 02 – Graph of wind vs temperature

C. Case 3 Results

As Shown if Figure. 3, the highest temperature of 32.65^{0}C was observed when the radiation emissivity and convection rate is low. The minimum temperature was observed when the convection is at its highest value and radiation emissivity is high. According to the obtained data, conductor temperature decrease can be seen at higher emissivity and convection rates.



	27.56	40.39	50.72	59.72	67.84	75.32	82.32	88.93	95.21	101.2 2
0.25	32.65	31.21	30.57	30.19	29.93	29.74	29.58	29.46	29.37	29.29
0.5	32.41	31.1	30.5	30.14	29.89	29.71	29.55	29.44	29.35	29.27
0.75	32.19	30.99	30.43	30.08	29.85	29.65	29.52	29.41	29.33	29.25
1	31.99	30.89	30.36	30.04	29.81	29.62	29.49	29.39	29.31	29.23

Figure. 3 Case 03 – Graph of radiation emissivity vs convection vs temperature

CONCLUSION

The paper focused on the dynamic line rating of power cables using electrical and thermal analysis. DLR technology can be used to increase the power flow capability of the existing lines. Conductor temperature makes a great impact on the current carrying capacity of the cable. And also, the dynamic line rating affects the thermal behaviour of the overhead cables. Ansys electrical and thermal analysis was used to simulate the impact of current, wind speed, radiation emissivity and convection on the conductor temperature. The simulation proved that the conductor temperature increases when the applied current increases. Also, the simulation proved that there is an inverse proportion between wind speed and the conductor temperature. The obtained results will be used to implement a coordinated application of dynamic line rating in a distributing network.

ACKNOWLEDGEMENT

RESEARCH GRANT NO- ASP/01/RE/TEC/2021/75

REFERENCES

[1] Y. Hou *et al.*, "Research and application of dynamic line rating technology," *Energy Reports*, vol. 6, pp. 716–730, 2020, doi: 10.1016/j.egyr.2020.11.140.

[2] M. Kosterec, "Determining the Current Capacity of Transmission Lines Based on Ambient Conditions Daljnovodov Na Osnovi Zunanjih Pogojev," vol. 10, no. 2, pp. 61–69, 2017.

[3] S. F. Hajeforosh and L. Abrahamsson, "Dynamic Line Rating Operational Planning : Issues and Challenges," no. June, pp. 3–6, 2019.

[4] S. Madadi, B. Mohammadi-Ivatloo, and S. Tohidi, "Dynamic Line Rating Forecasting Based on Integrated Factorized Ornstein-Uhlenbeck Processes," *IEEE Trans. Power Deliv.*, vol. 35, no. 2, pp. 851–860, 2020, doi: 10.1109/TPWRD.2019.2929694.

[5] M. Kabović, A. Kabović, S. B. Rakas, and V. Timčenko, "Improving the Accuracy and Time Interval of Predicting Ambient Parameters Applied to Dynamic Line Rating," *Eng. Proc.*, vol. 5, no. 1, p. 11, 2021, doi: 10.3390/engproc2021005011.

[6] B. P. Bhattarai *et al.*, "Improvement of Transmission Line Ampacity Utilization by Weather-Based Dynamic Line Rating," *IEEE Trans. Power Deliv.*, vol. 33, no. 4, pp. 1853–1863, 2018, doi: 10.1109/TPWRD.2018.2798411.

[7] E. Fernandez, I. Albizu, M. T. Bedialauneta, A. J. Mazon, and P. T. Leite, "Review of dynamic line rating systems for wind power integration," *Renew. Sustain. Energy Rev.*, vol. 53, pp. 80–92, 2016, doi: 10.1016/j.rser.2015.07.149.

Session 02: Intelligent Systems & Robotics		
	Paper ID Paper Title	**Corresponding Author**
11	Raspberry pi-based bearing fault diagnosis by bearing audio and vibration signal via cost-effective accelerometer	K. Jathursajan
13	Rapidly Manufacture-able Ventilator for Respiratory Emergencies	S. Ashan
8	Detection of Mosquito Breeding Areas using Semantic Segmentation	Pravina Mylvaganam
6	Lie Detection Based on Cues	Cassandra Jacklya Dakius

Raspberry Pi-Based Bearing Fault Diagnosis by Bearing Audio and Vibration Signal Via Cost-Effective Accelerometer

Kanakasuntharam Jathursajan[1] and Akila Wijethunge[2]

[1,2] *Department of Materials and Mechanical Technology,University of Sri Jayewardenepura,Homagama 10200, Sri Lanka*

[1] *kanaga.jathu@gmail.com,* [2] *akilawijethunge@sjp.ac.lk*

Abstract— **The feasibility of the raspberry pi to diagnose both localized and distributed faults by the audio signal of the bearing via a cost-effective microphone and/ or the vibration signal via a cost-effective accelerometer is assessed. The envelope analysis is experimented on raspberry pi and it took less than 160 ms to diagnose localized faults. Artificial Neural Networks (ANN) and Convolutional Neural Network (CNN) models trained by Mel Frequency Cepstral Coefficient (MFCC) feature of bearing audio and/or Fast Fourier Transform (FFT) feature of vibration signal were assessed on raspberry pi and it required at most 180 ms to diagnose distributed faults.**

***Keywords—** fault diagnosis, machine learning, raspberry pi*

Introduction

The contribution of bearings in rotating machinery is essential for ideal operation, consistency, and efficiency. Numerous techniques have been investigated to diagnose bearing faults so far. Most of the methods are based on motor current signature analysis, vibration monitoring, temperature measurement, and acoustic emission measurement [1]. Although most of them guarantee reasonable results those overall expenses limit the intermediate industries to use for non-critical and less expensive applications since the budget of the system might be greater than the budget of the monitored element itself [2]. Acoustic analysis has drawn attention recently and it has been suitable for many applications like speech recognition, voice recognition, speech emotion recognition, and acoustic event monitoring and, hardly in the industrial environment for condition monitoring purposes [3]. In this study, the feasibility of the raspberry pi to diagnose both localized and distributed faults by the audio signal of the bearing via a cost-effective microphone and the vibration signal via a cost-effective accelerometer is assessed.

Localized faults on the bearings are representative of spalls, pits, or localized damage that appears on raceways and rolling elements [1]. A great aspect of localized or single-point faults is that they produce a characteristic frequency that can be calculated from the speed and geometry of the bearing [1]. For localized faults, in this study, the audio signal of the bearing is analyzed by the envelope analysis for the detection of localized faults such as inner race, outer race, ball, and misalignment faults. Envelope analysis of the audio signal and the vibration signal via a cost-effective accelerometer is done in raspberry pi by applying Kurstogram for identifying the filtering band and Hilbert transforms for extracting the envelope of the audio and vibration signals. To assess the localized faults in the bearings, fault-related frequency components that are produced by the sudden occurrence of bearing faults are needed to be known [4]. Therefore, the mathematical equations to find defects in the outer race and inner race of bearings are related to the fault-related frequency given by Equations (1) and (2), respectively, and the frequency related to misalignment is stated as Equation (3). Using the calculation of the BPFO (Ball Pass Frequency of the Outer race), BPFI (Ball Pass Frequency of the Inner race), and misalignment components, it is possible to locate the fault-related frequency components in the frequency spectrum and ensure the existence of bearing faults on the outer race and inner race as well as the misalignment [4].

$$\boldsymbol{BPFO} = \frac{N}{2}\boldsymbol{f}\left(\mathbf{1} - \frac{BD}{PD}\boldsymbol{\cos\theta}\right) \tag{1}$$

$$\boldsymbol{BPFI} = \frac{N}{2}\boldsymbol{f}\left(\mathbf{1} + \frac{BD}{PD}\boldsymbol{\cos\theta}\right) \tag{2}$$

$$\boldsymbol{f} = \frac{RPM}{60} \tag{3}$$

Fault-related frequency components are calculated in terms of the number of rolling elements (balls), N; the ball diameter, BD; the pitch diameter, PD; the contact angle, θ; and the rotational frequency, f, [4].

Distributed faults are regarded as bearing surface defects spread over a large area therefore two or more rolling elements can be located in the faulty area at a given moment [1]. The distributed defects are mainly surface roughness, waviness, off-size rolling elements, etc [1]. Since distributed faults produce a complex and arbitrary signal at any measure it results in an increasing challenge of diagnosing distributed faults in a rolling bearing using the audio signal of the bearing since audio usually has a low Signal to Noise Ratio (SNR) because of its ease of mixing ability to other background noises as well as using vibration signal via cost effective accelerometer [1]. Mel Frequency Cepstral Coefficient (MFCC) features are extensively utilized in many auditory applications such as voice recognition, speech emotion recognition, and acoustic event monitoring since MFCC features characterize approximated human auditory system [5]. Therefore, deep learning techniques such as ANN and CNN trained using the MFCC feature to distinguish between healthy bearing and distributed faulty bearing faults are evaluated and compared with the effect of fusing Fast Fourier Transform (FFT) of vibration signal via a cost-effective accelerometer with the audio signal on raspberry pi. Here, the performances of ANN and CNN on the raspberry pi to diagnose distributed bearing faults are evaluated and compared.

Compared to other methods, there are only a few works that have been done to diagnose bearing faults based on bearing audio signals or vibration signals via a cost-effective accelerometer. However, in [1], both localized and distributed faults were diagnosed by an audio signal, wherein localized faults were analyzed by envelope analysis utilizing Kurstogram, Hilbert transform, and FFT, and distributed faults were analyzed by ANN and CNN. Moreover, in [2], the authors have done a great job that they diagnosed both types of faults by bearing audio signal as well as compared the results with vibration signal via a cost-effective accelerometer, wherein localized faults were diagnosed by envelope analysis and distributed faults were diagnosed by ANN and CNN by the MFCC feature extracted from the audio signal and FFT feature extracted from vibration signal.

Although these studies examine the feasibility of the bearing audio signal and vibration signal via a cost-effective accelerometer, those lack in deploying the proposed diagnosing methods on micro-controllers/ processors which would be cost-effective. Since audio signal and vibration signal via cost-effective accelerometer are selected to cut prices of diagnosing the system, the feasibility of diagnosing bearing using audio and vibration signal is analyzed by raspberry pi for cost-effectiveness in this study.

The rest of the article is organized as follows: First, we present the method. Then we state the results. Finally, we discuss the results and findings followed by the conclusion.

METHOD

The audio signal of the bearing is sensed by the microphone connected to the soundcard. Here, the soundcard is used as the interface between the mic and the raspberry pi (Raspberry pi 4B with 4GB RAM). The audio signal via USB interface makes it possible to connect multiple mics to the raspberry pi. However, the audio signal is sensed via only one mic in this work but this method will be beneficial in future works. The I2C interface is used to interface between raspberry pi and the LCD to reduce the complexity of installing hardware and to reduce the number of wires from 16 to 4. Figure 1 shows the implemented system for real-time fault diagnosing of bearing.

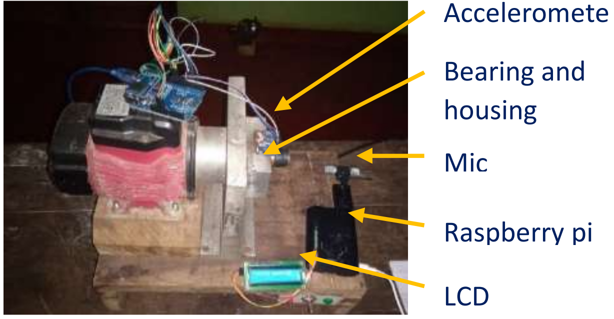


Figure 1: Implemented system

Data used in [2] is utilized in this work, wherein a total of 16 bearings of different types such as deep groove ball bearing (6205E) and needle roller bearing (25 38 15 CS) turning at different speeds was acquired, bearing audio signals were recorded by “Ipad-7th gen” at about 10cm distance from the bearing for the duration of each of the 60s at 22050 Hz sampling frequency, the ADXL335 accelerometer connected to Arduino UNO was used to acquire vibration at an approximated sampling rate of 22050Hz along the z-axis and, the speed of the motor was measured by a DM6236 digital tachometer.

A. *Distributed faults*

The ANN and CNN models accessed in [2] are deployed in raspberry pi here. There, the ANN and CNN were assessed while changing domains of training data. In the first type of model, the MFCC feature of bearing audio signal of distributed faulty bearing and healthy bearing are utilized to train the ANN and CNN. Samples of bearing audio signals with and without industrial background noises are utilized to train the model to generalize the model to be strong for unseen background noises. In the second type, FFTs of the vibration signals were utilized to train the ANN and CNN. In the third type, the MFCC of the bearing’s audio signal and the FFT of the vibration signal were fused and utilized to train the ANN and CNN.

There librosa library for extracting features such as MFCC, tensorflow runtime interpreter library for loading and using the tensorflow lite model generated in PC, and sounddevice library for recording audio via USB interface were imported. The tensorflow lite model is loaded and used to predict rather than the actual model because of the low computation power of the raspberry pi than the PC, for fast inference. Extracted MFCC features from the recorded audio signal for 2 s by mic connected to raspberry pi via soundcard are inputted to the lite model and the model outputs the state of the bearing. The vibration signal that had been acquired by Arduino before for prediction is stored in Raspberry pi and used to predict. The predicting process is stopped when the keyboard interrupt (Ctrl + C) for raspberry pi is generated.

B. *Localized Faults*

Experimented envelop analysis as in [2] deployed on raspberry pi for diagnosing localized faults consists of the following steps. There, first, the time-domain signal is filtered by a filtering band resulting from the Kurstogram because spectral kurtosis is higher for impulsive signals and zero for white noises since localized faults are impulsive. Since the localized faults are seen as amplitude modulated (AM), Kurstogram is used to discover the carrier frequency band of AM faulty signal. Then, the analytical signal of the Hilbert transform of the filtered signal is derived to develop the envelope of the faulty signal. Since localized faults are AM, the *FFT of the enveloped signal gives the fault frequencies of the rolling bearing.*

The filtering band derived from PC was used to filter the signals. Hilbert transform and Fast Fourier transform were performed by python installed in Raspberry pi. The results were displayed on a PC monitor connected by an Ethernet cable.

RESULTS

TABLE II: TIME TAKEN TO FEATURE EXTRACTION AND PREDICTION

Domain(s) - Model	Time to feature extraction	Time to prediction
MFCCs of the audio signal of bearing - ANN	88.92 ms	0.13 ms
FFT of vibration signal of bearing - ANN	52.32 ms	0.15 ms
Fusion of MFCCs of the audio signal of bearing and FFT of vibration signal of bearing - ANN	127.74 ms	0.18 ms
MFCCs of the audio signal of bearing - CNN	85.44 ms	0.23 ms
FFT of vibration signal of bearing - CNN	50.23 ms	0.22 ms
Fusion of MFCCs of the audio signal of bearing and FFT of vibration signal of bearing – CNN	129.98 ms	0.29 ms

Table I shows the feature extraction time and prediction time of the models in Raspberry pi for distributed faults.

TABLE III: DURATIONS OF LOCALIZED FAULTS

Domain	Filtering time	Hilbert transform time	Fast Fourier transform time
Audio signal	11.43 ms	100.24 ms	37.22 ms
Vibration signal	10.23 ms	108.78 ms	35.90 ms

Table II shows the durations for the filtering, Hilbert transform, and Fast Fourier transform processes.

DISCUSSION

A. *Distributed faults*

The feature extraction periods of systems with ANN are the same as that of CNN since ANN and CNN models are assessed concerning the same features. Although periods of extracting MFCC features of the audio signal are a little bit higher than that of extracting FFT of vibration signal, the periods of extracting features of fusion models are far away comparably than extracting period of MFCC or FFT since fusion model uses both MFCC and FFT features. However, periods of feature extraction of fusion models are not the sum of periods of feature extraction time of MFCC and FFT, which is higher than that of fusion models, therefore, resulting in a tiny grace. The prediction periods of vibration signal-based models and audio signal-based models are almost the same. However, the fusion models require a higher period than the audio signal-based models or vibration signal-based models. Despite that, the prediction time of CNN models was higher than the prediction periods of ANN models.

B. *Localized faults*

The localized faults could be diagnosed successfully by the method implemented on raspberry pi. It took longer periods for Hilbert transform than that of filtering time or time for FFT. The periods of the audio signal and vibration signals are almost the same since the same method is used to analyze both audio and vibration signals.

CONCLUSION

It was feasible to diagnose both localized and distributed bearing faults based on audio and/or vibration signals via a cost-effective accelerometer by Raspberry pi. It took less than 150 ms to extract features and less than 30 ms to predict for all the models on the raspberry pi to diagnose distributed faults. At most 160ms was taken to diagnose localized faults on raspberry pi.

REFERENCES

[1] K. Jathursajan and A. Wijethunge, “Diagnosing Localized and Distributed Bearing Faults by Bearing Noise Signal Using Machine Learning and Kurstogram”, *Adv. Technol.*, vol. 2, no. 2, pp. 139–150, May 2022.

[2] K. Jathursajan and A. Wijethunge, “Diagnosing localized and distributed faults of rolling bearing using Kurstogram and machine learning algorithms using bearings audio signal in comparison with vibration signal”, *Mercon.*, Jul 2022

[3] A. Altinors, F. Yol, and O. Yaman, “A sound based method for fault detection with statistical feature extraction in UAV motors,” *Applied Acoustics*, vol. 183, p. 108325, Dec. 2021, doi: 10.1016/J.APACOUST.2021.108325.

[4] Saucedo-Dorantes, J.J.; Delgado-Prieto, M.; Ortega-Redondo, J.A.; Osornio-Rios, R.A.; Romero-Troncoso, R.D.J. Multiple-fault detection methodology based on vibration and current analysis applied to bearings in induction motors and gearboxes on the kinematic chain. Shock. Vib. 2016, 2016, 5467643.

[5] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, “Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques,” CONIELECOMP 2012 - 22nd Int. Conf. Electron. Commun. Comput., pp. 248–251, 2012, doi: 10.1109/CONIELECOMP.2012.6189918.

Rapidly Manufacture-able Ventilator for Respiratory Emergencies

Ashan Padukka[1]and Akila Wijethunge[2]

1.2Department of Materials and Mechanical Technology, Faculty of Technology, University of Sri Jayewardenepura, 10200, Sri Lanka.

pdsashan@gmail.com[1], akilawijethunge@sjp.ac.lk[2]

Abstract; **This paper evaluates the procedures and experiments followed to design a rapidly manufacture-able, Low-cost fully synchronized medical ventilator that is designed to perform in an emergency situation. Studies about the use of different sensors, different actuators different design methods, and controlling algorithms of ventilators were followed to develop the rapid manufacture-able and low-cost ventilator. IOT techniques were introduced to the prototype to change and observe the parameters of the device. That proved the importance of implementing IOT techniques on medical devices in situations where the disease has high spread rates.**

Keywords: mechanical ventilation, open source, covid, ARDS, low cost mechanical ventilator

INTRODUCTION

The respiratory system is one of the most important multi-organ system in the human body. The human brain can't survive more than four minutes without supplying oxygen to the brain. As a result of different diseases like Acute Respiratory Distress Syndrome (ARDS) and Asthma, the functionality of the lungs can be weak or malfunction. That's the point where mechanical ventilators come into the role [1].

This research was motivated by the insatiable demand across ventilators that was raised as a result of the COVID-19 pandemic [2]. Anyway, humans passed the most risky time period of COVID-19 with the help of vaccines, but it's possible to expect the spreading of viruses or other types of infections in the future too. And most of the time highly spread viruses that have pandemic properties will be based on the respiratory system. Because respiratory system-based infections can spread at a higher rate than others. If the rate of spreading of some pandemic reaches over the speed of manufacturing ventilators with respect to demand in the world, the vacuum that induce through the high insistence for the ventilators will not be able to fill till the time passed. Therefore the potential of developing ventilators in a short period of time with minimum cost will be much important in the future. This research representation suggest a formal method to follow to full fill the particular task. Also an advanced IOT techniques were introduced in to the system to control the device remotely. This newly inspired technique may help to protect the device users/staff from the diseases which have high spread rates.

METHODOLOGY

A. Experimental and Development Procedure

The functionality of the “Medulla” prototype was much more advanced. It was designed to operate under six ventilation modes. The parameters and settings of the system could be adjusted via input devices over IOT based environment. Advanced safety systems and alarm systems were also integrated into the prototype. A battery backup setup was developed to function even under loss of power. Addition for that, pre-compressed air and oxygen supply were used to deliver the breath smoothly. Pressure sensors were used to detect the pressure variances and take feedback from the patient. 1L (liter) test lung was used as a model of the patient's lung to operate the machine.

B. Selection of transducers and system components

To fulfill the requirements of different modes, transducers and the controlling system should have minimum specifications to deliver the best-operating properties of the system (ex: resolution of sensors). While fulfilling those requirements, the particular sensor should be much more common in an industry than the special and biomedical grade components that are used to develop an original systems. By considering all those stuffs, transducers and controlling platform were selected.

C. Focused Parameters

To operate a Ventilator, its settings should be defined by the user (anesthesiologist), it can follow by defining different parameters of the machine. Those parameters were based on interrelated relations with each other. (ex: Changing the cycle time will affect the respiration rate). Each of those parameters' coloration and their mathematical and functional behaviors were observed and defined. The program of MCU was based on those derived equations and variables.

D. Used techniques of input and output signal Conditioning

Gain Adjustment -
MEMS and NEMS-based sensors are working under a range of millivolts/microvolts, such sensors required signal amplification before the signal transforms to the MCU. In such a case Op-amps were used in form of inverting or non-inverting amplification mode. Similarly, when transferring processed signals into actuators, the output signal needs to be conditioned (ex: Voltage/Current amplification). In such a case suitable amplification techniques were used. In this case, the PWM controller and mos-fet drivers were used as Output signal conditioners.

Offset elimination - Most of the time offset elimination is required when connecting the sensors to the MCU. In this case, Operational amplifiers were used in differential mode for offset elimination.

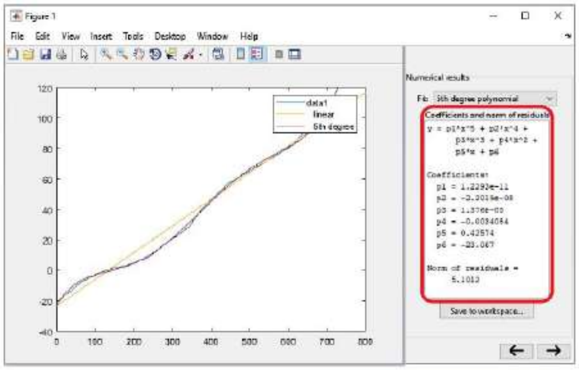
Linearization - Gain adjustment and offset elimination practices are very common techniques that using in signal conditioning. Those methods can perform externally from the MCU. But linearization requires arithmetic capabilities to perform. Usually, linearization techniques belong to the MCU.

When the rate of change of input differs from the rate of change of output signal of the system, it's necessary to follow linearization techniques to make a linear relationship between the input and output. By obtaining a close fit function to the obtained data, the linearization process can perform. The poly-fit function of Matlab was used to find the equation that fit to the set of data. When the degree of a polynomial function is increased, the accuracy of the function may also increase. The below table 01 shows the different analog read values obtained by the MCU by subjecting the sensor to different pressure rates.

TABLE I
ANALOG READ VALUES THAT OBTAINED RESPECT TO THE CHANGE OF PRESSURE

Term	Sensor Reading / Analog Read (X)	Pressure (cmH2O) (Y)	Term		Sensor Reading / Analog Read (X)	Pressure (cmH2O) (Y)
01	05	-19.5	10		184	3.50
02	39	-13.25	11		239	8.75
03	58	-8.75	12		277	16.75
04	82	-5.25	13		340	27.75
05	101	-2.75	14		378	40.75
06	110	-1.50	15		443	56.75
07	120	-0.50	16		610	81.75
08	130	0.00	17		714	112.25
09	137	0.50				

The above set of data was executed with Matlab poly-fit function to generate the best fit polynomial function as shown in the below figure 02.



```
y = p1*x^5 + p2*x^4 +
      p3*x^3 + p4*x^2 +
      p5*x + p6

Coefficients:
  p1 = 1.2292e-11
  p2 = -2.2019e-08
  p3 = 1.376e-05
  p4 = -0.0034084
  p5 = 0.42574
  p6 = -23.067

Norm of residuals =
     5.1012
```

Figure 01: Natural and fitted graphs of results

Data 1 curve represents the original data while the 5th-degree curve represents the fitted function. According to the fitted curve, the generated equation according to the numerical results was;

$$y = -23.06 + 0.425x + (-0.003x^2) + 1.376e\text{-}5x^3 + (-2.201e\text{-}8x^4) + (1.229e\text{-}11x^5) \quad (1)$$

By substituting the sensor read value for x, it's possible to take a much more accurate value for the current pressure. The same technique was used to obtain the data from all the sensors and operate actuators.

RESULTS AND DISCUSSIONS

The design of the system based on the supporting way to conduct manufacturing in minimum amount of time. Ventilators have the capability of working in different modes. In different scenarios, it's required to use different ventilation techniques to treat patients under various health conditions. So the experiments were followed to make the functionality of the prototype under five modes of ventilation. Accordingly the developed system could be able to operate under;

Pressure control Ventilation – PCV allows the practitioner to control ventilator pressure throughout the cycle in order to generate the pressure necessary.
Volume Control ventilation – VCV mode ensures the delivering predefined volume to the patient. The pressure limit ensures that no peak pressure could harm the lungs
Pressure support ventilation – Pressure support ventilation is a spontaneous mode of ventilation. The patient initiates every breath and the ventilator delivers support with the preset pressure value
CPAP (Continuous Positive Airway Pressure) – The ventilator may provide continuous positive airway pressure to keep the open alveoli.
Synchronized Intermittent Ventilation – In the SIMV mode, the patient is allowed to take additional breaths in between the mechanical breaths. The ventilator gives full breath, according to VC or PC only if the patient triggers a breath in a triggering window. The size of intermittent breath can be large or small, depending upon the patient's ability. The patients detect the patient's spontaneous breathing and wait until the patient exhales before delivering another mechanical breath.

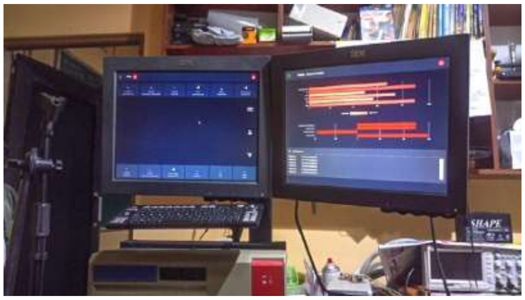


Figure 02: Finalized Prototype

A. Alarm testing

The alarm system is a very important sub-system in a ventilator that helps to communicate the errors and malfunctions, over limitations, or any incident that affects the system or patient. The system was tested and validated to indicate the alarm features under;
-Malfunction of Oxygen/Air supply to the ventilator (Not present/Overpressure/Under pressure)
-Too low O_2 Concentration
-Malfunction of O_2 Sensor
-Too high/ too low minute volume ventilation
-Battery operation/capacity malfunction

-Leak detection
-Lower pressure limit / Upper-pressure limit exceeded
-Communication or technical problems

B. Implemented web-based controlling techniques

This research was motivated by the effect of the COVID-19 pandemic. It is a disease that has a high spread rate and is capable of easily spreading from person to person. The ability to spread via droplets or airborne particles increased the risk of transmitting the virus. Therefore capability to control the system via a remote controlling platform gave much value to the system by improving the security of management of the health sector. The remote controlling capability of the system is enchased by Angular and React frameworks. The communication between MCU and the server was followed via serial communication. The remote controlling function was well performed as shown below

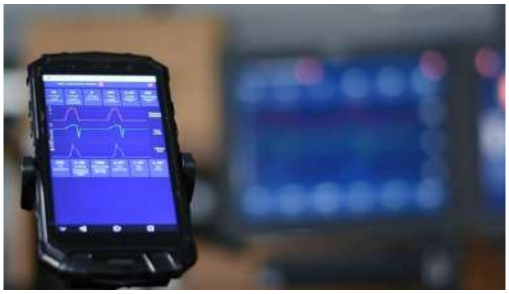


Figure 03 : Controlling the Ventilator over IOT device

Comparison in between Siemens Servo 300 Commercial Ventilator and Medulla Ventilator

Below tabular form of commercial ventilator, Siemens Servo 300 and the prototype of the Medulla ventilator will generate an adequate intention about the gap in between those two systems.

TABLE II
FRAILTIES OF THE MEDULLA VENTILATOR OVER SIEMENS 300 VENTILATOR

Siemens 300	**Medulla**
System components were made from biomedical grade materials (ex – non-latex breathing circuit)	The system was not made from biomedical materials (HDPE, ABS, PLA, Acrylic, Latex rubber) Low degree of quality materials and components were used
Reliable system components were used	Unreliable, Low grade MCU with low clock speed
Medical grade transducers were used	Low cost, General transducers were used
High resolution of input and output	Low resolution of input and output
High degree of aesthetic design	External design was based on readymade components and substitutions
Stated values of physical parameters are much accurate (Parameter values shows in the display has closer to its real world physical dimensions)	Elevated error in between the status parameter values and physical values

TABLE IVII
SOUNDNESS OF THE MEDULLA VENTILATOR OVER SIEMENS 300 VENTILATOR

Siemens 300	**Medulla**
Based on Conventional parameter settling system (Variable potentiometers and buttons were used)	Based on virtual and digitalized data gathering system based on HMI techniques
Remote control operations were not allowed	IOT techniques were introduced to the system to implement the remote controlling operation
Small scale display with Low resolution (Only two curves out of three (pressure/flow/volume) could see at a time	All three curves (pressure/flow/volume) could be seen at single display simultaneously with respectively high resolution output.

CONCLUSIONS

After spending lots of effort and time on the implementation, the objectives of the research could be validated. Some of the objectives could be well defined while others could be partially defined. Manufacture-ability of synchronized ventilator could be validated based on its concepts of operation. But not at the level of commercial. As discussed, the ventilator is a machine that overtakes the tasks of human internal organs which has a higher risk of the lives if the system subjected to malfunction. So much advanced technology and reliable components should be used to improve the machine up to the commercial position. Also it is necessary to using medical-grade components and materials to improve the quality of the outcome to reach that level. That much-advanced procedure may require a high level of financial requirement that is difficult to achieve in this state. The cost was the most disruptive barrier that stand across the procedures of advancement.The objective of developing mathematical equations related to the functionality of each step/mode could be achieved. The functionality of MCUs depends on arithmetic logic status, there for the success of this prototype was led by those defined relations and equations. As a new implementation in the field, the effort taken to introduce IOT techniques on the machine helped for reaching into the success. Variables, parameters, and information of the system could be observed, edit and change via the developed IOT-based system. This newly inspired technique may help to protect the device users/staff from the diseases which have high spread rates. The authors request to follow the links below to find out the visual functionality of the systems

Primary prototype;
(*https://www.youtube.com/watch?v=s9IrEIePjwU&ab_channel=AshanPadukka*)
Medulla prototype;
(*https://www.youtube.com/watch?v=3Yh3ZVDyzVU&ab_channel=AshanPadukka*)

REFERENCES

[1] Mechanical Ventilation - Robert L.Chatburn - Respiratory Care Department University Hospitals of Cleveland Associate Professor Department of Pediatrics Case Western Reserve University Cleveland, Ohio
[2] Ventilator Stockpiling and Availability in the US - Amanda Kobokovich, MPH - Johns Hopkins Bloomberg School of Public Health - September 3, 2020

Detection of Mosquito Breeding Areas using Semantic Segmentation

Pravina Mylvaganam[1], and Maheshi B. Dissanayake[2]

[1,2]*Department of Electrical and Electronic Engineering, University of Peradeniya, Peradeniya, Sri Lanka*
[1] *pravina.m@eng.pdn.ac.lk*, [2] *maheshid@eng.pdn.ac.lk*

Abstract— **The combination of deep learning (DL) and convolutional neural networks (CNN) with image analysis to locate stagnant water will play a crucial role in the dengue control process. This paper aims to automatically segment stagnant water areas in aerial images, acquired by a drone camera, using the latest CNN semantic segmentation method (SegNet). To enhance the effectiveness of our system and as the solution for the lack of dataset, we utilise two different datasets with high domain feature correlation. In our project, pre-training is first done on a large generalised dataset with areas of water, and then the trained model with trained weights is retrained using a task-specific dataset. It should be noted that the conditional distribution of the labels is similar for both datasets. The performance of the SegNet was evaluated with pixel accuracy and dice score. The model exhibited an accuracy of 80% and a dice score of 91%, indicating that our proposed method is efficient to segment water in RGB aerial imagery.**

Keywords—Semantic Segmentation, aerial images, water retaining objects

INTRODUCTION

Epidemics of dengue fever have been recognized for more than 200 years [1]. One modeling estimate indicates 390 million dengue virus infections per year, of which 96 million (67–136 million) manifest clinically [2]. Another study on the prevalence of dengue discloses that 3.9 billion people are at risk of infection with dengue viruses. Despite a risk of infection in 129 countries, 70% of the burden is in Asia [2]. The principal mosquito vector of dengue and urban yellow fever is Aedes aegypti, which breeds in water that has been collected in natural and artificial containers around human habitations. Water storage tanks, flower pots, garden fountains, bird baths, fridge trays, water dispenser trays, broken cisterns, discarded bottles and tires, tins, coconut shells, etc are all possible sites for mosquitoes to breed. The eggs can survive up to one year in dry containers and hatch when water is available. Therefore, keeping neighborhoods clean and free of receptacles that attract dengue-carrying mosquitos is vital. In addition to vector control measures, World Health Organization is working closely with the Ministry of Health Nutrition and Indigenous Medicine to control the spread of dengue, by specifically collaborating in reviewing dengue control and prevention activities at district and national levels. Even though they have taken plenty of dengue control activities, some limitations also include identifying mosquito breeding sites in inaccessible places, like rooftops and overhead water tanks. We proposed a system for identifying possible breeding places with UAV-based aerial images to cope with this limitation.

An Unmanned Aerial Vehicle (UAV), commonly known as a drone, is a type of aircraft that operates without a human pilot onboard. Recent technologies have allowed for the development of many different kinds of advanced unmanned aerial vehicles used for various purposes. As control technologies improved and costs fell, their use expanded to many non-military applications. These include forest fire monitoring, aerial photography, product deliveries, agriculture, policing and surveillance, infrastructure inspections, and science.

Image segmentation is an essential component in many visual understanding systems. Segmentation plays a central role in a broad range of applications, including medical image analysis, autonomous vehicles, video surveillance, and augmented reality to count a few. Instance segmentation extends the semantic segmentation scope further by detecting and delineating each object of interest in the image (e.g., partitioning of individual persons). Compared to other techniques such as object detection in which no exact shape of the object is known, segmentation exhibits pixel-level classification output providing richer information, including the object's shape and boundary. A recently emerged semantic segmentation method, which incorporates CNN structure is SegNet. Semantic segmentation makes it easier to understand images because it segments images into semantically significant objects and assigns each part of predefined labels. Thereby, different objects from remotely captured images can be extracted simultaneously. SegNet method is applied in several remote sensing applications [4]. (Du et al., 2018) [4] exploited the SegNet technique to classify and extract cropland in high-resolution remote sensing images, showing that the proposed approach efficiently obtained accurate results (98%) for the segmentation task.

Although several previous works are available in the public domain for identifying water pooling sites [5], they were designed to address a different task than ours. Mettes et al. developed a robust water detection algorithm for videos. They have proposed a methodology to detect water using spatial and temporal dynamics of water [5]. However, this approach only discusses the identification of water, which spreads in a considerably large area, such as pools and ponds, but does not analysis the applicability of this technique for small-scale water pooling areas.

Our research object is to detect possible water retention areas using aerial images. To find a solution to our problem, first, we have created a dataset using locally collected high-resolution images taken from a UAV operated at low altitudes. These characteristics of the dataset bring more clarity to the task as well as novelty. Furthermore, it helps deep learning models to learn specific features which assist to make accurate and precise detection of even small water retention areas.

With this objective, this paper aims to automatically segment stagnant water areas in RGB aerial imagery using the SegNet semantic segmentation method. We experimented with our custom dataset captured by a drone camera. Since our task-specific dataset is of a smaller size, we used domain adaptation for this task to enhance the efficacy of our system. In our proposed solution, pre-training of SegNet is first carried out on a generalized large custom dataset. Later, the trained model is sub-trained on another dedicated dataset collected locally without retraining from scratch.

The rest of the paper is organized as follows. Section 2 and Section 3 presents the methodology adopted in this study and discusses the results obtained in the experimental analysis respectively. Finally, Section 4 summarizes the main conclusions.

METHODOLOGY

In this paper, we proposed a semantic segmentation model for identifying mosquito breeding sites that contain water efficiently and automatically. At the initial stage of the research, two different datasets were acquired using images captured with a drone camera during a clear sunny day. The model is generalised to all geographical locations with the image capturing condition set as sunny day, as the training was done using universal dataset. For the pre-training purpose, a total of 2,668 images, belonging to one class: water areas, were collected and a total of 300 images of water retaining objects, such as water storage tanks, flower pots, garden fountains, bird baths, water dispenser trays, broken cisterns, discarded bottles and tires, tins, and coconut shells, were captured as a second dataset. Even though the image resolution is 1080 x 1900 pixels for the image samples of the dataset, we evaluated the input images at 256 x 256 during the experiments. Both datasets have been annotated using the LabelMe Software. The third and second column in Figure. 3 shows examples of original and labeled images of the datasets respectively. Next, the first dataset was divided into a training set and a validation set with a 70:30 ratio respectively, while the second dataset was split with an 80:20 ratio respectively.

A. Semantic Segmentation

In this application, the CNN-based SegNet architecture was first trained to segment the pixels into water area and background on a large generalized dataset. SegNet consists of a symmetrical encoder-decoder followed by a pixel-wise classifier as shown in Figure. 1 [3]. SegNet has 13 convolutional layers in the encoder/contraction and the decoder/expansion part, and the designed network consists of 29,443,142 trainable parameters. We chose the hyperparameters to obtain high segmentation accuracy under a low learning rate. The model was trained over 100 epochs with batch sizes of 8 and an adadelta optimizer. Once the model is trained using the large dataset, the trained weight files are migrated to the encoder, which is sub-trained using the smaller yet customized (specific) dataset. This approach is known as transfer learning-based domain adaptation in deep learning.

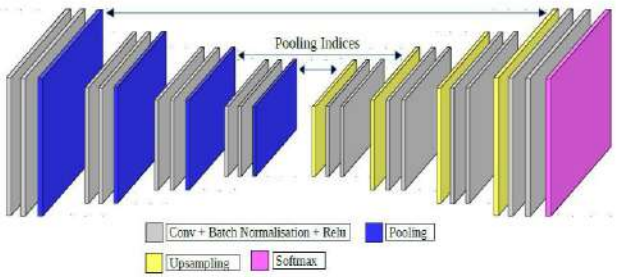


Figure. 1 Proposed SegNet architecture

The same aforementioned CNN SegNet was used for the sub-training with the second dataset. The validation dataset was used to determine the learning rate, which defines how the weights are adjusted in the CNN during training to reduce the risk of overfitting. The model was trained over 100 epochs with the softmax classifier and Google Collab with GPU was used for the simulations. Finally, the test images captured using a drone camera were tested using the trained model to evaluate the model performance.

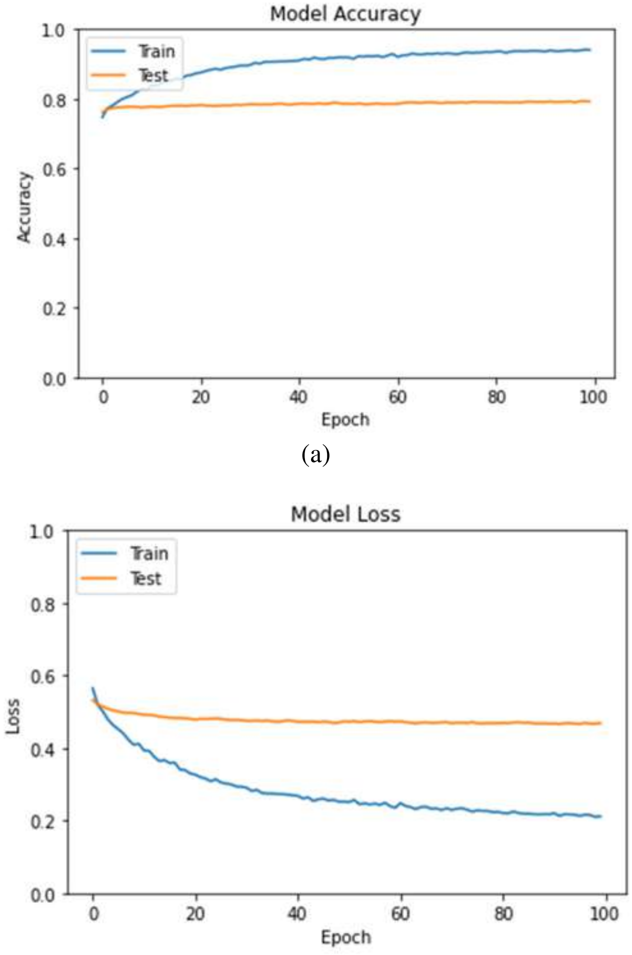
RESULTS AND DISCUSSION

The performance of our proposed SegNet model is evaluated on RGB test images captured by a drone camera. When we input RGB aerial images, the trained network showed a pixel accuracy of 80% and a training accuracy of 94%. The pixel accuracy shows the percentage of the pixels that were correctly classified. Training time and predicted time were 20 minutes and 8 sec respectively. Furthermore, model accuracy and loss variation through the 100 epochs during the final training is shown in Figure. 2. In addition, Fig. 3 shows samples of RGB test images and segmented outputs of the SegNet model.

Although the loss variation shown in Figure 2 (b), shows room for further improvement, the sample test outputs in Figure. 3 are aligned with the expected results. Furthermore, accuracy pixel estimates indicate that the method used is efficient to segment water in drone camera images with a small amount of dataset with the help of transfer learning approaches.

Furthermore, the Dice score, mathematically expressed as shown in Eq (1), is used to quantify the performance of the proposed image segmentation method. The average dice score of 91% indicates that the pixel-wise degree of similarity between the model predicted segmentation mask and the ground truth is high.

$$\text{Dice coefficient} = \frac{\text{2*True positive}}{\text{2*True positive+False positive+False negative}} \tag{1}$$



(a)

(b)

Figure. 2 (a)Training and validation accuracy variation of SegNet model; and (b) Training and validation loss of SegNet model

Conclusion

The main objective of this research is to propose a mechanism to identify possible water retention areas in inaccessible places. Especially on rooftops. With this objective, we proposed a system to automatically detect potential Aedesaegypti breeding sites using a Deep Learning algorithm to help public health agents to combat its reproduction. We present a semantic segmentation method (SegNet) to automatically segment water in imagery acquired by a drone camera to identify the mosquito breeding sites. The overall performance, with an average accuracy of 80% and average dice score of 91%, indicated that the SegNet method is an efficient approach to segment the water area in images.

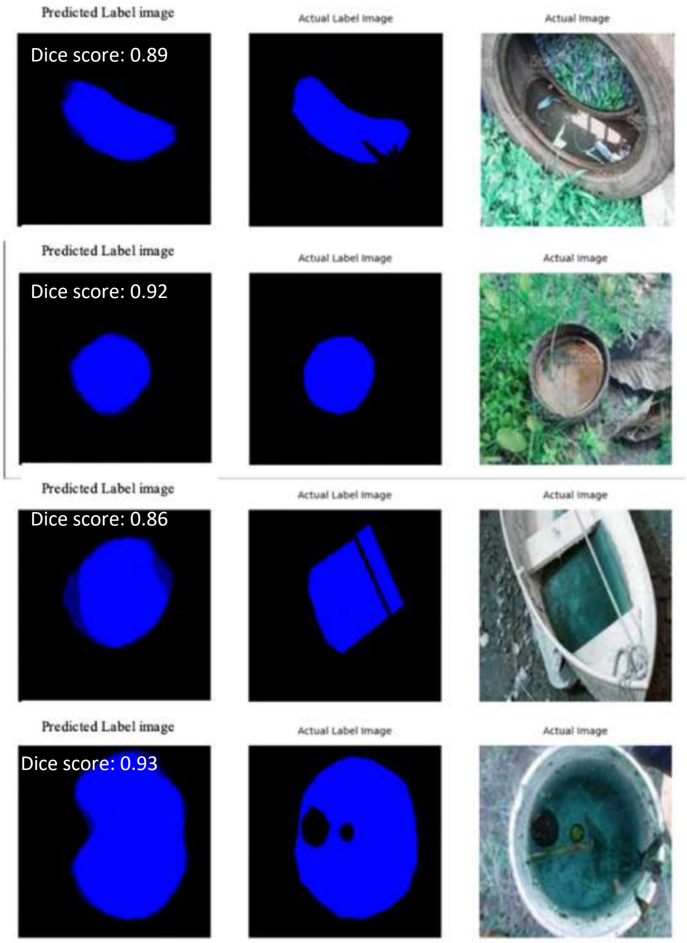


Figure. 3 Sample original RGB images; Actual labeled images; and Segmented outputs from the SegNet model

References

[1] Siler JF, Hall MW, Hitchens AP, "Dengue: its history epidemiology, mechanisms of transmission, etiology, clinical manifestations, immunity, and prevention," *Philippine J Sci. 1926*, vol. 29, no. 1-2, pp. 1-304.

[2] Bhatt, S., et al., "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, April 2013.

[3] Badrinarayanan, V., Kendall, A., Cipolla, R., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. on Pattern Anal. and Mach. Intel.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[4] Du, Z., Yang, J., Huang, W., Ou, C., "Training SegNet for cropland classification of high-resolution remote sensing images," in *Proc. AGILE Conf.*, 2018.

[5] P. Mettes, R. T. Tan, and R. C. Veltkamp, "Water detection through spatio-temporal invariant descriptors," *Comput. Vis. Image Underst.*, vol. 154, pp. 182–191, Jan. 2017.

Lie Detection Method Based On Cues

Cassandra Jacklya Binti Dakius[1], Saavethya Sivakumar[2] and Rushikesh Bodhe[3]

[1,2,3]Department of Electrical and Computer Engineering, Curtin University Malaysia, CDT 250, 98009 Miri Sarawak, Malaysia
[1]700040366@student.curtin.edu.my,[2]saaveethya.s@curtin.edu.my

Abstract— **Many people have long been fascinated by deception, particularly the capacity to determine whether or not someone is telling the truth. In many circumstances, such as criminal investigations, court trials, or even small white falsehoods told, identifying lies within a person is more important. This should be done in the least invasive and covert manner possible to avoid unrealistic and phoney actions. The purpose of this study is to present an overview of lie detection based on eye features. Previous researchers have always gathered biological signals or data for any deception detection, but little to no research has been conducted on ocular characteristics. This is performed by conducting an electronic online search on well-known databases such as Google Scholar, IEEE, SpringerLink, Scopus, and ScienceDirect, focusing primarily on content published in the twenty-first century to avoid outdated information.**

Keywords— Deception detection, Eye tracker, Blink detection, Facial features, Eye properties, Cognitive Load; Lying.

INTRODUCTION

Historically, lie detection has always relied on methods such as polygraph testing and responses of the biological portion of the human autonomic nervous system (ANS) during interrogation. Because their meanings are so similar, we use the terms lying and deception interchangeably in this work. Although lying has a negative connotation, it is not surprising that many people still engage in the practise for a variety of reasons - to avoid punishment, protect the feelings of others, for entertainment, to hurt others, and so on - in which practitioners and even ordinary people have been intrigued by the behaviour of people who lie.

While many research on lie detection are undertaken, the majority of them are based on biological aspects such as galvanic skin reaction GSR, fMRI, and EEG recorded P300 event linked potentials, where the invasive nature of the experiment resides. We concentrated on the review of lie detection using eye features - average fixation time, saccade count, revisits, glances count, gaze direction, blinking frequency, and many more - with the advantage of requiring little to no interpersonal skills. Furthermore, employing ocular characteristics enables remote and non-invasive diagnostics while saving time on calibration, testing, and data processing. With the flexibility to switch to non-invasive approaches, results can be collected more accurately, with the goal of preventing people from doing rehearsed activities that lead to distortion of results.

This paper mostly reviews work from the 21st century with just little inclusion of material from earlier periods. A researcher in the field of lie detection proposed certain cues to deceit detection such as blink rate, gaze direction, and eye movement AU (action units) recognition, with the eyes being one of the most expressive areas of the human face conveying a variety of information including cognitive workload and attention. This points the way forward for future study on blink parameters in order to implement the proposed method for high precision and accuracy in detecting falsehoods.

METHODOLOGY

The study employs a quantitative methodology to highlight the need of having non-invasive methods for deceit detection as well as offer a prospective eye parameter as one of the deception detection cues. In this case, available web data was leveraged to generate results and create a prototype model. The materials collected are pre-recorded movies from a previous experiment, which resulted in the creation of a database known as the Silesian Database. According to the study, participants who made more mistakes blinked at a higher rate than those who made less mistakes.

A total of 320 videos were gathered from the Silesian database [6], and the first stage was to develop a programme that can count the number of blinks based on the required eye aspect ratios (EAR) to establish whether the subject's eyes in the video are open or closed for blink detection. The diagram below depicts the visual depiction of the procedures required to detect blinks in videos. The program is coded purely on Google Colab as it has the sufficient resources
to accomplish the objectives.

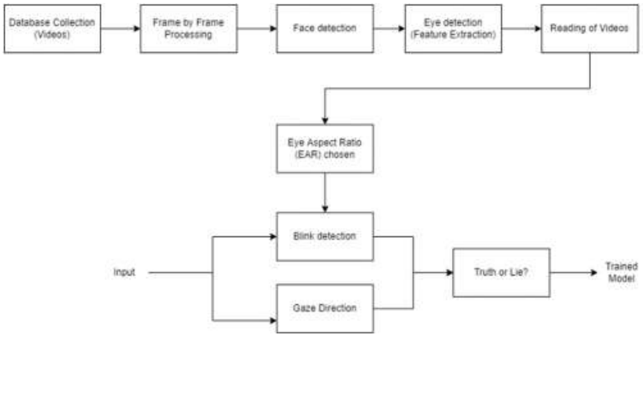

Database Collection (Videos)
Frame by Frame Processing
Face detection
Eye detection (Feature Extraction)
Reading of Videos
Eye Aspect Ratio (EAR) chosen
Blink detection
Input
Gaze Direction
Truth or Lie?
Trained Model

Figure.1: Methodology

A. Face and Eye Detection

Face localization was initially performed exclusively on photographs as a test run to confirm that the code worked properly before moving on to apply the software to video files.

Facial critical points for detection are required before detecting the eyes, whether in real-time or pre-recorded videos. In this scenario, a pre-trained network contained within the dlib library and capable of detecting the '68 important points' was utilised to assist in the marking of the points on the faces.

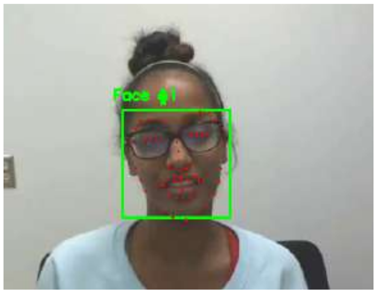


Figure. 2: Face Localization of a Subject

The computer effectively detects a person's face in the figure above while also marking the 68 important places of facial detectors.

Using the knowledge gained from doing facial landmark detection on video, we created an application that can detect blinks.

B. Blink Detection

According to [5], there is a correlation between the width and height of the coordinates that indicate the eye and blink detection. Each eye has six (x,y) coordinates that begin at the left corner of the eye and proceed clockwise around the rest of the eye.

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}$$

Figure 3: Eye Aspect Ratio Equation

This is how the eye aspect ratio (EAR), as indicated in the above graphic, exists during the blink detection process. The EAR is used to detect whether or not the individual is lying based on the number of blinks specified as a threshold. The facial landmark locations are represented by p1 through p6 in the equation.

The specified EAR value is 0.25. A number of EAR thresholds were employed to discover the best most accurate value in determining whether or not a blink occurred, with a value of 0.25 (more specifically, 0.2468) demonstrating the maximum accuracy of 96.85 percent.

C. Gaze Direction

Because it is impossible to identify fraud just based on blinking rates, three new parameters have been added: left, right, and centre. NLP established the notion that gaze direction can be utilised as a lie detector. Specifically, people are more likely to look to the left while reminiscing about the past than to the right when lying. As a result, this ocular parameter was also gathered for extra characteristics.

D. Audio

Before the evaluation of the model, the audio was extracted from the videos using the library ‘moviepy’ and then used a feature extraction technique which is the Mel-Frequency Cepstral Coefficient (MFCC). It adjusts the frequency to better fit what the human ear can hear (humans are better at identifying small changes in speech at lower frequencies) which was developed through a series of subjects. MFCC are often derived characteristics from speech signals for use in recognition tasks. MFCC are thought to represent the filter in the source-filter model of speech (vocal tract).

TABLE I: CONVERT TO MFCC

```
def mp3tomfcc(file_path, max_pad=20):
   audio, sample_rate = librosa.core.load(file_path,
res_type='kaiser_fast')
   mfcc = librosa.feature.mfcc(y=audio, sr=sample_rate,
n_mfcc=20)
   pad_width = max_pad - mfcc.shape[1]
   if (pad_width > 0):
     mfcc = np.pad(mfcc, pad_width=((0, 0), (0, pad_width)),
mode='constant')
   else:
     mfcc = mfcc[:,0:max_pad]
return mfcc
```

The code above was used on the 319 videos of the subjects to obtain the 20 coefficients from the MFCC technique which has important features that will be used in the final evaluation for results.

E. Final Touch

Once all the features have been collected, the data are placed into one database collection through the Google Spreadsheet to allow for easier learning through the models and obtain the features as inputs into the machine learning model.

RESULTS

TABLE II: FIRST RESULT OBTAINED

Gaze Direction, Blinking Rate and Audio	
Model	Accuracy
MLP	50.00%
RNN	54.69%
LR	45.31%
RF	52.08%
SVM	58.33%
SGD	75.00%
DTC	37.50%

TABLE III: SECOND RESULT OBTAINED

Blinking Rate and Audio	
Model	Accuracy

MLP	56.25%
RNN	56.25%
LR	46.87%
RF	40.62%
SVM	59.37%
SGD	72.91%
DTC	47.91%

TABLE IV: THIRD RESULT OBTAINED

Gaze Direction and Audio	
Model	Accuracy
MLP	50.00%
RNN	59.38%
LR	46.87%
RF	51.04%
SVM	63.54%
SGD	71.87%
DTC	40.62%

The comparisons were made from various models to allow the distinction of the usefulness and effectiveness for the model accuracy.

DISCUSSION

Multiple models were trained on the dataset in this project to compare the accuracy of the models. Although the accuracy did not significantly improve to at least 60% for both MLP and RNN at first, it is planned to investigate several alternative suitable models that may be better at identifying lies. In this case, from the results shown, a total of 7 machine learning models were used to aid in detecting lies.

The findings demonstrate that after data normalisation, the accuracy of most of the models improved dramatically, breaching the 70% mark for one of the models, Stochastic Gradient Descent (SGD). SGD is stochastic in nature, which means it chooses a "random" instance of training data at each step and then computes the gradient, which makes it considerably faster because there is much less data to edit at once. As clarification, the data has been pre-processed to be normalized to be in the range between -1 and +1 for a much accurate evaluation of the models.

The results have been obtained in three different categories. Firstly, the category of using eye as the only parameters (blinking rate) together with audio, another with just gaze direction and audio and the last having all the different parameters as the input features to evaluate the most suitable and feasible model for lie detection.

The Support Vector Machine approach comes next in the model accuracy level since it generates significant accuracy while requiring low computational power. In terms of consistency, the next most efficient and effective model for this lie detection approach is Recurrent Neural Networks (RNN), which has a relatively high accuracy of more than 50% for each category provided in the results section. Recurrent neural networks can be used to represent time-dependent and sequential data issues. RNNs have internal memories that allow previous inputs to impact future predictions. Knowing what the preceding words were, for example, makes it much easier to predict the next word in a sentence with more accuracy.

CONCLUSION

In conclusion, models can be employed in the detection of deception using non-invasive approaches such as those used in this work. Eye characteristics, coupled with auditory analysis, are two significant elements that can physically reveal whether a subject is telling the truth or not. Stochastic Gradient Descent, Support Vector Machine and Recurrent Neural Networks proved to be the main potential model for lie detection and this can further developed into tuning of parameters for better accuracy.

REFERENCES

[1] A. Lanat`a, A. Armato, G. Valenza, and E. P. Scilingo, “Eye tracking and pupil size variation as response to affective stimuli: A preliminary study,” in 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, IEEE, 2011, pp. 78–84.

[2] R. Hooda, V. Joshi, and M. Shah, “A comprehensive review of approaches to detect fatigue using machine learning techniques,” Chronic Diseases and Translational Medicine, 2021.

[3] A. R. Bhamare, S. Katharguppe and J. Silviya Nancy, "Deep Neural Networks for Lie Detection with Attention on Bio-signals," 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2020, pp. 143-147, doi: 10.1109/ISCMI51676.2020.9311575.

[4] Marcolla, Felipe & Santiago, Rafael & Dazzi, Rudimar. (2020). Novel Lie Speech Classification by using Voice Stress. 742-749. 10.5220/0009038707420749.

[5] C. Dewi, R.-C. Chen, X. Jiang, and H. Yu, “Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks,” PeerJ Computer Science, vol. 8, e943, 2022.

Radlak, M. Bozek, and B. Smolka, “Silesian deception database: Presentation and analysis,” in Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, 2015, pp. 29–35.

Session 03: Information & Communication Technology		
	Paper ID Paper Title	**Corresponding Author**
19	Clustering Human Personality Based on Persons' Behavior	K.M.G.S Karunarathna
9	Utilizing Noise as an Attack Independent Measure for Representing Privacy in Logistic Cumulative Noise Addition	U. H. W. A. Hewage
15	Predicting Social and Economic Impact of Social Entrepreneurs Using Machine Learning Algorithms	K.V.K.C.Gamage
16	Machine Learning Approach to Predict the Job Satisfaction of Freelancing Jobs in Sri Lanka.	H.R.I.E. Ranasinghe

Clustering Human Personality Based on Persons' Behaviour

K.M.G.S. Karunarathna[1], M. P. R. I. R. Silva[2], R. A. H. M. Rupasingha[3*]

[1,2,3]*Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Sri Lanka*
[1] *gayathrisarangika599@gmail.com*, [2]*rangikaishani75@gmail.com*, [3]*hmrupasingha@gmail.com*

Abstract— **According to a person's unique thought, emotions and behaviour patterns, can identify the personality traits. This study's goal is to cluster human personalities according to their behaviours. The supervisor, the commander, the inspector, the doer, and the idealist are the major five behaviourally oriented personality types that were the focus of this study's secondary data collection. After the pre-processing seven clustering algorithms applied; namely for Expected Maximum, Hierarchical clustering, Simple k-means, Canopy, Filtered Cluster, Make Density Based Cluster and Farthest First. Using classes to cluster method, the Hierarchical cluster approach categorized data successfully with good accuracy, precision, recall, and f-measure scores. The Mean Squared Error and Root Mean Squared Error also show the Hierarchical cluster algorithm's lowest error value. The evaluation findings demonstrate that the performance of the Hierarchical clustering method was superior to that of the other six clustering algorithms.**

Keywords— Personality, Clustering, Machine Learning, Human behaviour

INTRODUCTION

A person's personality is what makes them different from others and makes them stand out from other people. It can be characterized as a combination of a person's traits and outward appearance, including their way of thinking, feeling, acting, communication and having physical features. A strong personality is essential to a person's success in this cutthroat society.

Our study uses five personality types to automatically categorize personality features based on their behaviours. This is made feasible by the diverse actions of the people, namely the supervisor, the commander, the inspector, the doer and the idealist. The major goal of this study is to develop a model that can automatically determine a person's typical personality type based on their behaviour.

When consider about the related work regarding to this study, the fuzzy clustering approach based on fuzzy statistics is used to cluster learners based on their personality and learning strategy data obtained from an online system [1]. And also, the transitive closure approach was implemented in MATLAB, and the results were examined. Personality categorization using the k-means clustering technique is described in this study [2].

In this paper, they established the learner model as the foundation for categorizing learners. And used K-means to suit the need for learner grouping after analysing several standard clustering techniques. The results of the experiment indicate that the approach of learner categorization based on personality clustering accurately represents the distribution of genuine features [3].

This study is intended to provide a complete assessment of current advances in User Behaviour Analysis, covering the fundamental applications, chosen ML techniques, and data types used. A representative sample of 127 publications was discovered according to particular traits, and rated using a reputation score that allows the importance of any contribution in the area to be measured [4].

When we consider the research gap, there has been recent study on a variety of personality traits; however, the same clustering techniques were not employed in previous studies. Some have just compared two algorithms or utilized very few data. However, in our study, we raised the amount of data and the number of various clustering methods to seven in order to gain better comparison and outcomes.

PROPOSED APPROACH

The steps of the proposed approach are shown in Figure.1. According to that, there are three main steps of the approach such as data collection, data pre-processing, and clustering.

A. Data Collection

This was done using the secondary data set, which was received from the Kaggle website [5]. 1000 data points from the secondary data set were chosen for this study based on different five actions of the people, namely the supervisor, the commander, the inspector, the doer, and the idealist. These five actions are used as the five target variables. For taking the result we used 17 attributes like you are sentimental, you are struggling with deadlines, you avoid leadership goals, you often feel overwhelmed, you avoid making phone calls, you enjoy participating in group activities etc.

B. Pre-processing

Data pre-processing was done using the Waikato Environment for Knowledge Analysis (WEKA) data mining software.

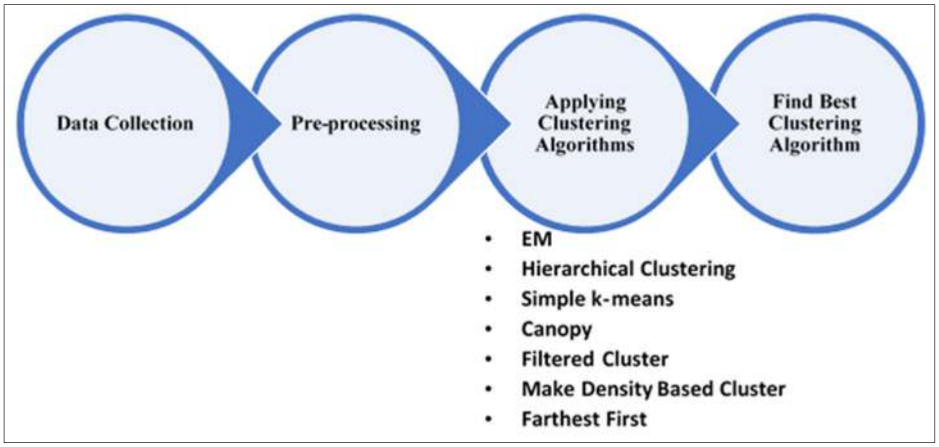

Data Collection
Pre-processing
Applying Clustering Algorithms
Find Best Clustering Algorithm
• EM
• Hierarchical Clustering
• Simple k-means
• Canopy
• Filtered Cluster
• Make Density Based Cluster
• Farthest First

Figure. 1 Proposed approach

16 personality types and 62 traits were included in the data collection. The top 17 attributes are chosen after the attributes are ranked using the information gain ranking algorithm. They are,

- "You often end up doing things at the last possible moment"
- "You have always been fascinated by the question of what if anything happens after death"
- "You find it easy to empathize with a person whose experiences are very different from yours"
- "You rarely second-guess the choices that you have made"
- "After a long and exhausting week, a lively social event is just what you need"
- "You enjoy going to art museums"
- "You often have a hard time understanding other people's feelings"
- "You like to have a to-do list for each day"
- "You avoid making phone calls"
- "In your social circle, you are often the one who contacts your friends and initiates activities"
- "You are still bothered by mistakes that you made a long time ago"
- "You feel more drawn to places with busy, bustling atmospheres than quiet, intimate places"
- "You often feel overwhelmed"
- "You would pass along a good opportunity if you thought someone else needed it more"
- "You struggle with deadlines"
- "You feel confident that things will work out for you"
- "Personality" (target variable)

In addition, depending on their personality traits, the personality types were condensed to 5 types.

C. Applying Clustering

Using the WEKA data mining tool, the clustering process is applied to the pre-processed data set. The prediction model is constructed in order to distinguish the personality traits. The methods for Expected Maximum (EM), Hierarchical clustering, Simple k-means (SKM), Canopy, Filtered Cluster (FC), Make Density Based Cluster (MDBC) and Farthest First (FF) were all applied to the data set.

EXPERIMENTS AND RESULTS

These comparison results were obtained on a computer running Microsoft Windows 10 with an Intel Core i5 processor and 4.0GB of RAM. And for the clustering process, we used the "Classes to cluster" method in WEKA 3.8.5 tool and apply it for the all the collected data.

In this "Classes to cluster" method, first, the WAKA tool did not take the class attribute and generate the relevant clusters. When the testing is run it assigns classes to clusters. The majority value of the class attribute within each cluster is used for this process.

A. Accuracy of the clustering algorithms

The result of seven clustering algorithms were compared in terms of accuracy. Figure. 2 shows it.

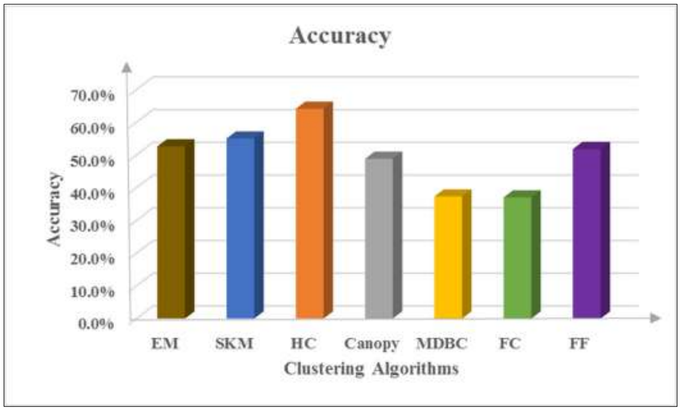


Figure 2 Accuracy of clustering algorithms

As shown in Figure. 2, we have obtained the accuracy, considering seven cluster algorithms, the EM, SKM, hierarchical clustering, Canopy, MDBC, FC, and FF algorithms show 53%, 55%, 64%, 49%, 37%, 37%, and 52%, respectively. The hierarchical clustering algorithm showed the highest accuracy.

TABLE I. RESULT OF PERFORMANCE MEASUREMENTS & ERROR

Clustering Algorithms	**Canopy**	**EM**	**Simple k-means**	**Hierarchical Cluster**	**Make Density Based Cluster**	**Filtered Cluster**	**Farthest First**
Precision	0.161	0.293	0.209	0.635	0.328	0.32	0.230
Recall	0.161	0.269	0.209	0.635	0.820	0.80	0.230
F-measure	0.161	0.280	0.209	0.635	0.469	0.457	0.230
MAE	0.0512	0.0474	0.045	0.0359	0.0627	0.063	0.0482
RMSE	0.2262	0.2177	0.2121	0.1894	0.2503	0.2509	0.2195

B. Results of precision, recall, f-measure and error rates

The precision, recall, f-measure results obtained using the (1), (2), (3). Here, *Ps*, *Px* and *Pstx* stand for the total number of relevant members that make up a particular cluster, the total number of members that make up a particular cluster, and the total number of relevant specified-cluster members that make up the corpus, respectively.

$$Precision = \frac{P_s}{P_x} \quad (1)$$

$$Recall = \frac{P_s}{P_{stx}} \quad (2)$$

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

Using the following (4) and (5), we then determined the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for each of the seven-clustering algorithm. Here, *Pvx* stands for the current labelled values based on the outcomes, *Mvx* stands for the anticipated result and *T* stands for the total number of forecasted values.

$$MAE = \frac{1}{T}\sum_{x=1}^{T} |\ p_{vx} -\ M_{vx}\ | \quad (4)$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{x=1}^{T}\left(p_{vx} - M_{vx,}\right)^2} \quad (5)$$

According to (1), (2) and (3), we calculated the precision, recall and f-measure of seven algorithms. The outcomes are shown in Table 1. According to the experiment results, highest precision, recall, and f-measure are presented at 0.635, 0.635, and 0.635 respectively in the Hierarchical clustering algorithm. Furthermore, we considered the error rates, using (4) and (5), we were able to obtain the MAE of 0.0359 and RMSE of 0.1894 as the lowest error rates in Hierarchical clustering algorithm.

Based on the all the evaluation results proved that the Hierarchical clustering algorithm has a better performance. This result is supported intuitively by considering the nature of the collected data (amount of data and number of attributes), when comparing the other six clustering algorithms. Personality traits vary from person to person, therefore identifying specific personality traits is a challenging task. That is a major challenge faced in clustering as it involves human characteristic clustering.

CONCLUSION

The proposed method determined the persons' behaviours according to their personal characteristics. Their way of thinking, feeling, acting, communication also obtained. This information was obtained from the Kaggle website as secondary data. The main objective of this research is identifying the personality type according to their behaviours and characteristics. For that we put out a clustering algorithm. After pre-processing, we used seven clustering algorithms to identify the most effective algorithm. Namely for EM, Hierarchical clustering, Simple k-means, Canopy, Filtered Cluster, Make Density Based Cluster and Farthest First.

Based on the entire data set, the Hierarchical clustering algorithm shown 64.1% as the highest accuracy. The Hierarchical clustering algorithm outperforms the other six algorithms in terms of accuracy overall. The Hierarchical clustering also obtained minimum MSE and RMSE values. Hierarchical clustering exhibits the best values for precision, recall and f-measure. Based on the final results we can apply the hierarchical clustering approach to cluster the human personality based on their behaviour with the best performance.

In order to further improve the classification process' accuracy, further study will entail a larger data set and an examination of the different classification algorithms including the Ensemble Learning algorithm.

REFERENCES

[1] F. Tian, S. Wang, C. Zheng, and Q. Zheng, "Research on e-learner personality grouping based on fuzzy clustering analysis," *Proc. 2008 12th Int. Conf. Comput. Support. Coop. Work Des. CSCWD*, vol. 2, pp. 1035–1040, 2008, doi: 10.1109/CSCWD.2008.4537122.

[2] A. Talasbek, A. Serek, M. Zhaparov, S. M. Yoo, Y. K. Kim, and G. H. Jeong, "Personality Classification by Applying k-Means Clustering," *2020 Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2020*, pp. 421–426, 2020, doi: 10.1109/ICAIIC48513.2020.9065244.

[3] D. Jin, Z. Qinghua, D. Jiao, and G. Zhiyong, "A method for learner grouping based on personality clustering," *Proc. - 2006 10th Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2006*, pp. 1420–1425, 2006, doi: 10.1109/CSCWD.2006.253206.

[4] A. G. Martín, A. Fernández-Isabel, I. Martín de Diego, and M. Beltrán, "A survey for user behavior analysis based on machine learning techniques: current models and applications," *Appl. Intell.*, vol. 51, no. 8, pp. 6029–6055, 2021, doi: 10.1007/s10489-020-02160-x.

[5] A. Mehta, "Personality Classification Data: 16 Personalities."https://www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt (accessed May 27, 2022).

Utilizing Noise as an Attack Independent Measure for Representing Privacy in Logistic Cumulative Noise Addition

U.H.W.A. Hewage[1], R. Sinha[2], and R. Pears[3]

[1,2]School of Engineering Computer and Mathematical Sciences, Auckland University of Technology, New Zealand
[3]College of Engineering, University of North Texas, USA
[1] waruni.hewage@aut.ac.nz, [2] roopak.sinha@aut.ac.nz, [3] russel.pears@unt.edu

Abstract— **Privacy preservation of data plays a significant role as organizations world-wide use data for different purposes. Measuring privacy provided by a privacy preservation method requires considerable attention. Most methods currently being used are based on background knowledge of the data. However, an attacker equipped with the knowledge of original data is not always valid. In this study, we investigate the possibility of using noise variance as a measure to represent privacy where no background knowledge about data is available. We employ a noise addition-based perturbation method called logistic cumulative noise addition and Area Under the Curve as core components. The proposed approach can be used as an attack independent method to represent the privacy.**

Keywords—noise variance, privacy, area under the curve, logistic cumulative noise addition, attack independent

INTRODUCTION

Privacy-Preserving Data Mining (PPDM) performs data mining tasks without directly accessing the original data values[1]. This is achieved by converting the original dataset to another form that hides the original data's actual values, providing privacy for the original data. This process is called data perturbation [2]. The objective of perturbation methods is to increase data privacy while maintaining data mining tasks' accuracy [1], [3].

Measuring privacy is the most critical task after applying a perturbation to original data. Privacy can be defined in many ways, considering the environment it has been used. A more generic definition for privacy is proposed by [4], which is "the degree of uncertainty according to which original private data can be inferred". Most privacy-measuring metrics assume that some background knowledge of the original data is known to the attacker and performs attacks based on this knowledge. We think that this assumption is overrated as it is not always possible for an attacker to have some knowledge of original data. An attacker can be someone entirely new to that specific set of perturbed data who does not know about the original data. In that case, it is helpful to have a method to get an idea about the privacy of the dataset without performing attacks on the perturbed dataset.

Let us investigate attacking/data reconstruction methods that measure privacy. Most of them are based on some background knowledge of the original data. For example, Known Input/output attacks, distribution attacks, Independent Component Analysis (ICA) based attacks, Distance inference attacks and MAP attacks [1] can be considered. Reference [5], which proposes a method to set limits on privacy breaches, is the only method we could find to represent the privacy of noise addition-based methods without using the knowledge of the original data.

The impact of the perturbation method on privacy and noise addition has been discussed in many works. Privacy provided by the noise has a direct relationship with the total noise variance added to the dataset because when we increase the noise variance, the privacy of the dataset also increases. In [6], data owners specify a noise constraint S where random noise up to S should be added to ensure privacy is preserved. The signal-to-noise ratio (SNR) has been used in [7] in the data perturbation context to achieve optimal data utility while preserving privacy. Authors have defined SNR as the variance of original data over the variance of noise which is again a metric of privacy based on the noise. The authors of [1] have experimentally proved that their data reconstruction method works quite well with small noise variance. It is harder to recover data when the noise variance is high. All these works imply that the level of privacy directly correlates with the variance of the noise added.

Traditionally, the privacy provided by a perturbation method has been measured using noise variance. However, later research works such as [5] and [3] argue that noise variance alone is not an adequate indicator of privacy. Privacy also depends on the original data distribution and other parameters of the perturbation method. We agree that noise variance is not sufficient to measure privacy when the distribution of data changes. However, the fact that noise variance has a considerable impact on privacy also cannot be ignored. Suppose we assume that the data distribution is consistent, which is valid for most of the databases/traditional datasets. In that case, noise variance significantly impacts the level of privacy.

We propose an attack-independent method to represent privacy, considering the properties of the perturbation method. To achieve this task, we use the Logistic Cumulative Noise addition (SRW) [2], the most recent development of noise additive-based perturbation methods. This work is significant due to two reasons. The first reason is that performing attacks on data based on background knowledge is not always valid, as the attacker can be someone new to the data. The second reason is that the noise variance alone cannot accurately represent privacy, as other factors affect that. Therefore, this work provides a novel approach to representing privacy by capturing other properties of the perturbation method together with the noise variance. This attack-independent approach does not require any background knowledge of the original data. The remainder of this paper has been organized as follows. Section II provides an overview of the proposed methodology to represent privacy. Results and Discussion are explained in Section III, and Section IV outlines the conclusions.

PROPOSED METHODOLOGY

This work aims to propose an attack independent method which is not based on the background knowledge of original data or its distribution to represent privacy using Logistic Cumulative Noise Addition (SRW) [2]. Suppose we can capture other characteristics of the perturbation method together with the total noise variance added to the data. In that case, it is sufficient to give an idea about the expected privacy level assuming data distribution does not change over time. To achieve this, we used the concept of Area Under the Curve (AUC) in the context of SRW. It captures not only the total noise variance added but also other behaviours of the perturbation method.

A. Logistic Cumulative Noise Addition (SRW)

The SRW is a cumulative noise addition method that is combined with cycle-wise noise addition to controlling the maximum noise level added to data [2]. The dataset is virtually divided into cycles which are defined by the cycle size and noise is added in cycles. The variance of the noise added is decided using applying the logistic function to each cycle. From this method, we have control over the noise addition rate (*k*) and the maximum noise level (*L*). The logistic function can be defined as in (1) and Figure. 1 represents the logistic curve.

$$f(x) = \frac{L}{1+ e^{-kx}} \qquad (1)$$

In SRW, noise variance changes throughout the cycle. This behaviour provides more privacy than using a constant noise variance throughout the perturbation process. We also can change *k* appropriately and it decides how fast the curve reaches the maximum level. The final noise variance produces in each independent step is f(x)*σ^2 where σ^2 is a small noise variance value used to control the noise level. All these behaviors and parameter values should be considered when proposing an efficient method to represent privacy.

B. Representing Privacy Using AUC

Area Under the Curve (AUC) is an interesting concept that has been used in different areas such as medicine and signal processing. It has been used to measure the total amount of drug exposure as a function of time and used to distinguish the total noise from the signal. Therefore, AUC can be used to model the total amount of some parameter as a function of another parameter. This concept can be adapted to the SRW environment since the area under the logistic curve allows to measure the total amount of noise variance added as a function of data records. Moreover, it indirectly captures the noise addition rate, the maximum nose level, and the behaviour of the logistic curve, making the concept of AUC is ideal to represent all the aspects of the SRW perturbation method.

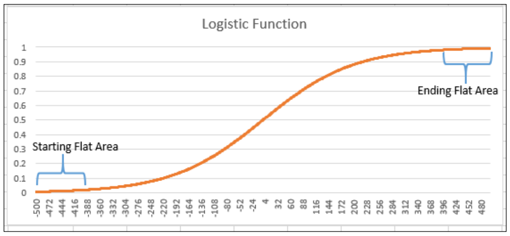


Fig. 1 Logistic Curve 1

Calculating the AUC of the logistic curve is straightforward. We can integrate the logistic function from lower bound *(lb)* of cycle size *(cs)* to upper bound *(ub)* of cycle size to calculate AUC of logistic curve (if *cs* = a, then *lb* = -a and *ub* = a).

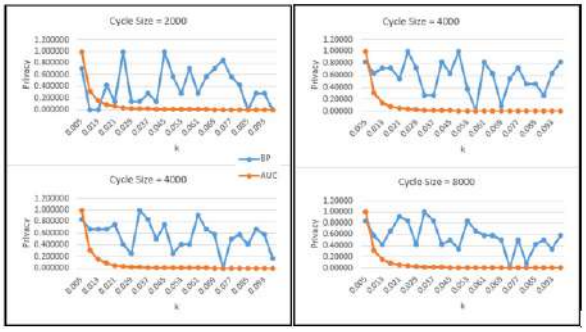


Figure. 2 AUC and BP comparison of AReM (left) and Electricity $\int_{lb}^{ub} f(x) = x + \frac{1}{k}\ln\left|1 + e^{-kx}\right| + c$ (right) datasetsauc $= f(x)_{ub} - f(x)_{lb}$ (2)

Where c is a constant. Equation 2 calculates the total noise variance from the logistic curve. But we should cooperate σ2 to get the total overall noise variance as we generate the noise from a Gaussian distribution with mean zero and variance of (f(x)*σ2).

$$F(x) = \frac{1}{1+ e^{-kx}} * \sigma^2; F(x) = \frac{\sigma^2}{1+ e^{-kx}} \qquad (3)$$

Integral of *F(x)* is: $\int_{lb}^{ub} F(x) = \sigma^2 x + \frac{\sigma^2}{k}\ln\left|1 + e^{-kx}\right| + c$ (4)

And using (2), we can write (4)

as $AUC = [f(x)_{ub} - f(x)_{lb}] * \sigma^2$ AUC (5)

According to the above proof, the total amount of noise variance added can be measured by multiplying the AUC of the logistic curve with additional noise variance. But, as we add noise cumulatively, AUC should be considered cumulatively. That means AUC for every data record should be added to the AUC of all the subsequent data records.

$$\text{Total AUC} = A_1 + (A_1 + A_2) + (A_1 + A_2 + A_3) + \cdots + (A_1 + A_2 + \cdots + A_i) + (A_1 + A_2 + \cdots + A_i + \cdots + A_n) \qquad (6)$$

A_i – AUC up to the data record i

Equation 6 gives the total noise variance added within one cycle of SRW and it also indirectly captures other properties and behaviors of the perturbation method. With the support of the above proof, we can define privacy in general for all the noise addition-based perturbation methods.

Definition (Privacy): *The percentage of protection applied to data concerning the total amount of noise variance added, given the noise addition rate (k) in a noise addition-based environment.*

EXPERIMENTS

We conducted experiments for different *k* values of logistic function and calculated the privacy using AUC. Finally, we normalized all the AUC values to bring them into the same range for ease of understanding. Additionally, we calculated the Breach Probability (BP) to measure privacy and compared it with AUC privacy values to see if there is any relationship. BP was calculated using MAP attacks [1]. We conducted the perturbation experiments for two datasets (AReM from UCI and Electricity from OpenML). We recorded the total noise variance added to the dataset by calculating AUC for 24 different noise addition rates (*k*) and four different cycle sizes of the logistic curve. This range of *k* values was selected according to the details provided in [2], maintaining the ideal shape of the logistic curve to get the maximum privacy benefits. Calculated AUC values for the AReM dataset have been displayed in Table I. (Note that AUC values for the Electricity dataset also showed a similar trend.)

THE BEHAVIOR OF AUC FOR DIFFERENT *K* VALUES

k	**AUC for Different Cycle sizes**			
	1000	***2000***	***4000***	***8000***
0.005	333.683	1112.863	4084.760	15945.275
0.009	284.715	1029.270	3994.933	15855.369
0.013	266.519	1008.430	3974.033	15834.469
0.017	258.768	1000.443	3966.046	15826.482
0.021	254.915	996.565	3962.168	15822.604
0.025	252.747	994.395	3959.998	15820.434
0.029	251.411	993.059	3958.662	15819.098
0.033	250.531	992.179	3957.782	15818.218
0.037	249.920	991.568	3957.171	15817.607
0.041	249.480	991.127	3956.730	15817.166
0.045	249.151	990.799	3956.402	15816.838
0.049	248.900	990.547	3956.150	15816.586
0.053	248.703	990.351	3955.954	15816.390
0.057	248.546	990.194	3955.797	15816.233
0.061	248.419	990.067	3955.670	15816.106
0.065	248.315	989.963	3955.566	15816.002
0.069	248.228	989.876	3955.479	15815.915
0.073	248.156	989.803	3955.406	15815.842
0.077	248.094	989.742	3955.345	15815.781
0.081	248.041	989.689	3955.292	15815.728
0.085	247.995	989.643	3955.246	15815.682
0.089	247.956	989.604	3955.207	15815.643
0.093	247.921	989.569	3955.172	15815.608
0.097	247.891	989.539	3955.142	15815.578

Looking at the AUC calculated for all the *k* values, we can see a similar trend for all four-cycle sizes. When k increases, AUC decreases, indicating that the total noise added to the data also decreases. This behaviour is expected and can be explained using the shape of the logistic curve. When k increases starting and ending flat areas of the logistic curve also becomes lengthier (See Figure. *1*). That means the period we add the noise in its lowest variance (closer to zero) also increases, reducing the total noise variance added to the dataset. This implies that when the total noise added to data is low, privacy provided by the perturbation method is also low. Figure. 2 displays the graphs comparing AUC and BP for cycle sizes 1000 and 4000 for AReM and cycle sizes 4000 and 8000 for electricity datasets. Min-max normalized values of both measures have been used for understandability.

Overall, we do not observe any strong relationship between AUC and BP for both datasets. BP fluctuates throughout all the k values, while AUC shows a decreasing trend. Trying to recover original data records from perturbed records using known I/O pairs and measuring the success rate of recovery when calculating BP is the primary reason for this. Hence, it considers the properties of the original data and assumes that the attacker knows some of the original data records and their perturbed counterparts. Nevertheless, AUC gives a privacy measurement depending on the perturbation method, assuming that the attacker does not know about the original data. If we carefully observe the behaviour of BP curves of both datasets, we can see a slightly decreasing trend though there are fluctuations throughout. This is a good sign indicating that the properties of the perturbation method (calculated using AUC) successfully capture the privacy trends without considering the properties of the original data.

In summary, the results of the experiments show that the total noise variance calculated using the AUC of the logistic curve has a direct effect on privacy. This can be further clarified from the fact that the behavior of AUC retrieved from the experiments can be justified from the behavior of perturbation method using logistic curve. Additionally, experiments do not display a strong relationship between BP and AUC. But a slightly similar pattern can be inferred.

CONCLUSIONS

In conclusion, we find that the total noise variance calculated using AUC is a justifiable measure to represent the privacy of SRW. It can be considered an attack-independent method to represent privacy. This measure also captures other properties of the perturbation method, such as noise addition rate and maximum noise variance allowed. The use of total noise variance as a measure of privacy can be extended to other noise addition-based methods, such as additive noise and multiplicative noise, effectively if it is possible to capture the properties of the perturbation method. The proposed approach can be considered valid as it shows the same trend as privacy using BP. BP assumes the availability of background knowledge, while AUC only considers the properties of the perturbation method. Incorporating the generic properties of the dataset with the proposed privacy measure is a possible future avenue as it is essential when original data distribution changes.

REFERENCES

[1] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 37–48, 2005, doi: 10.1145/1066157.1066163.

[2] U. H. W. A. Hewage, R. Pears, and M. A. Naeem, "Optimizing the Trade-off Between Classification Accuracy and Data Privacy in the Area of Data Stream Mining," *International Journal of Artificial Intelligence*, vol. 20, no. 1, pp. 147–167, 2022.

[3] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving datamining algorithms," *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 247–255, 2001, doi: 10.1145/375551.375602.

[4]C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining -Models and Algorithms*. USA: Springer US, 2008. doi: 0.1007/978-0-387-70992-5.

[5]A. Evfimievski and J. Gehrke, "Limiting Privacy Breaches in Privacy Preserving Data Mining," 2003.

[6]K. Liu, C. Giannella, and H. Kargupta, "An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining," in *Springer Verlag.*, 2006, pp. 297–308. doi: 10.1007/11871637_30.

[7]S. Virupaksha and V. Dondeti, "Subspace based noise addition for privacy preserved data mining on high dimensional continuous data," *J Ambient Intell Humaniz Comput*, no. 0123456789, 2020, doi: 10.1007/s12652-020-01881-8.

Predicting Social and Economic Impact of Social Entrepreneurs Using Machine Learning Algorithms

K.V.K.C. Gamage[1], K.S. Ranasinghe[2] and R.A.H.M. Rupasingha[3*]

[1,2,3]*Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Sri Lanka*
[1]*kavicrai@gmail.com,* [2]*ksranasinghe@ssl.sab.ac.lk,* [3]*hmrupasingha@gmail.com.*

Abstract—**Individuals who start their carrier path and try to address any social issue those kinds of entrepreneurs are called social entrepreneurs. Creating social or environmental changes and gaining profit are their major goals. The primary aim of this research is to predict the impact of social entrepreneurship on social and economic elements within Sri Lanka by using classification algorithms. To build this classification model, we collected 400 data from social entrepreneurs. After pre-processing, five algorithms such as Naïve Bayes, Decision Tree (J48), Support Vector Machine (SVM), Multilayer Perception (MLP), and Random Forest were used. Final evaluation was taken based on the values of the accuracy, precision, recall, f-measure, and error rates. It showed that Naïve Bayes is the best classifier to build this prediction model.**

Keywords—Machine Learning, Prediction, Social Entrepreneurship, Social and Economic Impact

INTRODUCTION

Identifying the social and environmental challenges and answering them with sustainable business solutions could explain as social entrepreneurship. Social enterprise's primary focus is social gains, while social entrepreneurs are keen on the final outcome of social entrepreneurship, which is to solve challenges that exist within the society in which the organization operates. Social entrepreneurship has various definitions to elaborate on the concept; however, its social aim is common among all explanations.

Primarily this research has a few key research purposes, such as predicting the impact of social entrepreneurship on social and economic elements within Sri Lanka, identifying the dimensions of social entrepreneurship, and identifying the best possible classification algorithm for this approach. Among most of the countries which have embraced this modern business model, Sri Lanka is relatively new and slowly experiencing the benefits generated by social entrepreneurship. However, there is a significant lack of a prediction model to measure the impact of social entrepreneurship on social and economic elements.

Key literature which had a direct link to this study was a study conducted in the Malaysian [1] context which predicted the social impact of social entrepreneurship. This research has designed a blueprint, which has helped the author to make a prediction with a high level of accuracy.

Further, [2] research focus on the social and economic benefit gained when conducting a social enterprise. The findings of this research will focus on identifying, what is the impact of social entrepreneurship on the variables derived from sociaeconomic elements. These social and economic elements are divided into six measuring variables, namely; social mission, social innovation, social networking, financial returns, social sustainability, and economic sustainability.

A social entrepreneurship's social worth will rise by addressing any social issue, and having a clear social mission will help it stand out from other organizations. Social innovativeness, also known as innovation, is a prerequisite for finding creative solutions to social problems. Any firm that satisfies a need will be labelled as social entrepreneurship. The complicated yet fundamental needs of people will be addressed by these social advances. These specific social innovations are in the form of goods, services, or business methods.

Through social networking, businesses that compete with them can establish relationships with customers. These people and groups are connected, and it is well recognized that businesses cannot thrive on their own without social networking. Profits are a major motivation for any company. It has been noted that there is an increasing trend of interest in business that generates social advantages while generating financial rewards. Most studies have not yet identified the breadth of social sustainability. Thus, social sustainability is defined as a combination of many elements such as income, healthcare, access to goods and services, employment, social equality, etc. in article, which defines a common term. Economic in general terms outlines how continuous economic growth is achieved while favourably effecting social, environmental and cultural factors within a social ecosystem.

When it comes to the research gap there is no research directly relevant to the proposed approach. Key literature which had a direct link to this study was a study conducted in the Malaysian [1] context which predicted the social impact of social entrepreneurship. This research has designed a blueprint, which has helped to predict with a high level of accuracy. Further, this research focus on the social and economic benefit gained when conducting a social enterprise. But, the proposed approach used five key ML algorithms namely Naïve Bayes, Decision Tree (J48), Support Vector Machine (SVM), Multilayer Perception (MLP), and Random Forest for predicting, and those algorithms are not used in the previous approaches.

MATERIALS AND METHODS

The research architectural framework is being drafted in Figure 1.

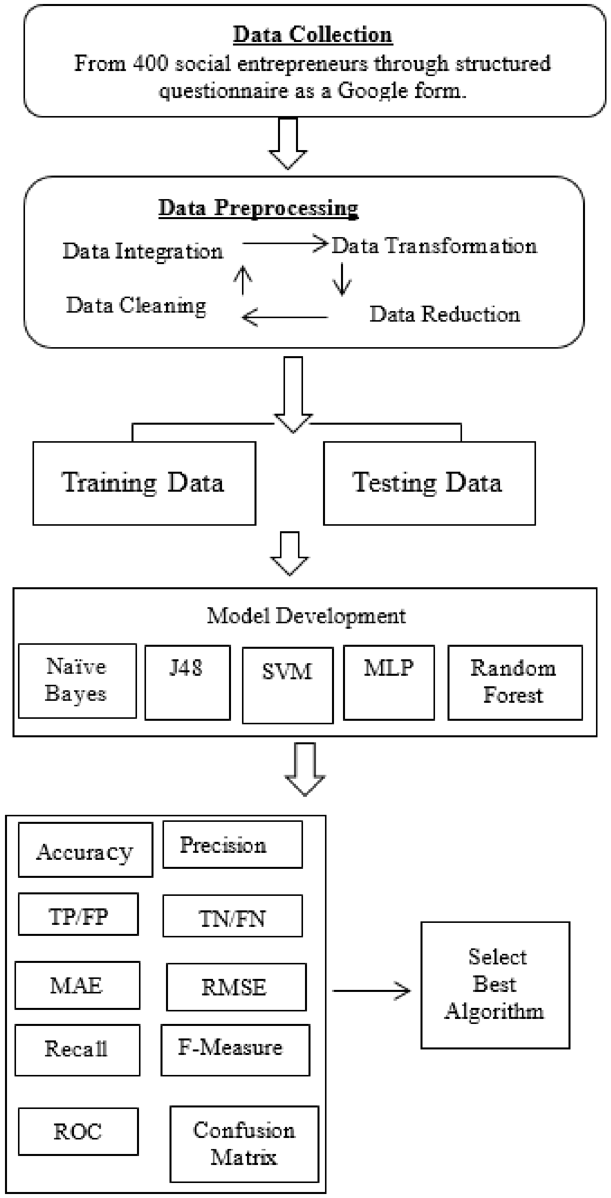


Figure 1 Architectural Framework

A. *Data Collection*

In this research, our target population is registered Social Entrepreneurs in Ratnapura and Hambantota District. The dataset includes social entrepreneurs from around the nation because the unit of analysis chosen for the current study is Social Entrepreneurs in Sri Lanka. Due to the significant number of social entrepreneurs in Ratnapura and Hambanthota District, the study was mostly done in that regions. According to information obtained from a list of registered social entrepreneurs, the target audience for this study would be made up of social entrepreneurs who are registered in the both districts Sri Lanka. We used 400 social entrepreneurs by convenience sampling method. We use the structured questionnaire as a form of a google form to collect the data [3]. The google form was shared among those registered users to collect the data. To validate the data, got support from a few registered social entrepreneurs and government entities like the Department of small business development, Sabaragamuwa provincial council, Ministry of rural economic affairs, etc.

B. *Data Pre-Processing*

Thereafter data set is being preprocessed under data integration, data transformation and removing incomplete data. This procedure identifies missing and irrelevant data, which is then updated, changed, or eliminated. To discover which attributes are most affected, attributes are ordered using hyper parameter tuning in the WEKA tool.

C. *Classification*

Then the classification model was built based on five machine learning algorithms namely Naïve Bayes, decision tree (J48), SVM, MLP, and Random Forest [4]. Based on the literature, these five algorithms are the most accurate and effective algorithms among others and we selected them for our approach as well. Further, we have used the percentage split as the statistical method to dived the data into training and testing partitions.

III. RESULTS AND DISCUSSION

This evaluation used Microsoft windows 10 on a personal computer with processor Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz and 8 GB RAM. The WEKA 3.8.6 tool was used to train the data set and run each machine learning algorithm. Finally test set is being evaluated under accuracy, precision, recall, f-measure, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), ROC area, [5].

Accuracy is the value generated after number of positive instances divided by the number of total instances and it calculated using (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Precision gives out the total positive predictive value out of the predictive values and it calculated using (2).

$$Precision = \frac{True\ Positive}{True\ Positive +\ \ False\ Positive} \qquad (2)$$

Recall value simply gives the sensitivity and its processed values or the completeness of the selected values and it calculated using (3).

$$Recall = \frac{True\ Positive}{(True\ Posituve + False\ Negative)} \qquad (3)$$

F-measure value gives out the combined measure of precision and recall and it calculated using (4).

TABLE I. RESULTS OF CLASSIFIER EVALUATION MATRICES

Algorithm	Accuracy	Precision	Recall	F - Measure	MAE	RMSE
Naïve Bayes	90 %	0.906	0.9	0.901	0.0982	0.309
J48 – Decision Tree	86.25 %	0.862	0.863	0.862	0.2022	0.3298
SVM	88.75%	0.891	0.888	0.888	0.1956	0.2793
MLP	81.25%	0.828	0.813	0.814	0.1964	0.4077
Random Forest	87.5 %	0.877	0.875	0.875	0.1852	0.2899

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (4)$$

Mean absolute error is mean value of individual prediction errors of the data set. Relative square error value is reduced to the predictor size with the same dimensions. Following (5) is being used to calculate the MAE value and (6) is used for RMSE value. The sample size is *n*, the real value is *Y1*, and the anticipated value is *X1*.

$$MAE \ = \frac{1}{n}\sum_{i=1}^{n} | \, Y_i - \, X_i| \qquad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2} \qquad (6)$$

As shown in Table 1, first we compare two test options percentage split with 80% and 66% to find out the best test option for our prediction. With the results, we can see the best test option is a percentage split (80%) for this prediction. After finding the test option all the evaluation matrices are evaluated by this percentage split method.

As shown in Table 2, each individual classifier shows good accuracy levels, and when comparing these five algorithms Naïve Bayes shows the best accuracy level at 90%.

As per the results better precision, recall, and f-measure values are shown by the Naïve Bayes algorithm. And also comparatively lower MAE and RMSE error values are shown by the Naïve Bayes algorithm. When we compare all classifier evaluation results of each algorithm we can see that Naïve Bayes shows the best results out of all.

IV. CONCLUSION

This research has forced it's attaining in predicting the impact of social entrepreneurship on social and economic elements within Sri Lanka. For this data was collected through a Google form, where such data are pre-processed and fed into classification algorithms such as Naïve Bayes, decision tree, random forest, SVM, and MLP. Here 400 entrepreneurs participated as resource providers. Based on the results we can say Naïve Bayes is the best classification algorithm for this prediction. And also analytical findings of this research show that social entrepreneurship has a direct and positive impact on social and economic elements. Social entrepreneurship will improve the social mission, social innovation, social networking, financial returns, social sustainability, and economic sustainability and improve the impacts the social and economic elements.

In future work we are planning to improve this prediction model using the ensemble learning method by combining these five individual algorithms.

TABLE II. RESULTS OF PERCENTAGE SPLIT METHOD

Type	Percentage Split (66%)	Percentage Split (80%)
Naïve Bayes	88.2353 %	90 %
J48 Decision Tree	85.2941 %	86.25 %
SVM	87.5 %	88.75%
MLP	83.8235 %	81.25%
Random Forest	86.7647 %	87.5 %

REFERENCES

"The future of social entrepreneurship: modelling and predicting social impact | Emerald Insight", *Emerald.com*, 2022. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/INTR-09-2020-0497/full/html. [Accessed: 21- Mar- 2022].

A. Javed, M. Yasir and A. Majid, "Is Social Entrepreneurship a Panacea for Sustainable Enterprise Development?", *Pakistan Journal of Commerce and Social Sciences*, vol. 131, no. 01-29, pp. 4,5,6, 2019.Available: http://hdl.handle.net/10419/196185. [Accessed 22 March 2022]

G. Brancato, S. Macchia and M. Murgia, Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System, 1st ed. 2022, p. 142.

S. Neelamegam and .. Ramaraj, "Classification algorithm in Data mining: An Overview", *International Journal of P2P Network Trends and Technology (IJPTT)*, vol. 4, no. 2249-2615, pp. 370,371,372,373, 2013. Available: http://www.ijpttjournal.org/. [Accessed 30 July 2022].

Ž. Vujovic, "Classification Model Evaluation Metrics", *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021. Available: 10.14569/ijacsa.2021.0120670 [Accessed 21 July 2022

Machine Learning Approach to Predict the Job Satisfaction of Freelancing Jobs in Sri Lanka

H.R.I.E.Ranasinghe[1], K.S. Ranasinghe[2] and R.A.H.M. Rupasingha[3*]

[1,2,3]*Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Sri Lanka*
[1]*erandihr47@gmail.com,* [2]*ksranasinghe@ssl.sab.ac.lk,* [3]*hmrupasingha@gmail.com*

Abstract— **Freelancing has become one of the major business field in the world since few years, and majority of younger people have the opportunity to be a freelancer. In COVID-19 pandemic situation freelancing jobs became more popular. Also, freelancing is a one of the best ways to foreign currency inflow for the developing countries to increase the foreign currency reserves. This study was conducted on primary data collected from freelancers in Sri Lanka. After pre-processing, we develop a model for predicting the job satisfaction of freelancers in Sri Lanka through Machine Learning approach. We conducted a comparative study among Naïve Bayes, Support Vector Machine (SVM), Decision Tree (J48), Random Forest and Multilayer Perceptron (MLP) classification algorithms and the decision tree algorithm shows the best results compared to other algorithms. In here, the best algorithm selected based on accuracy, precision, recall, f-measure and error rate.**

Keywords— Freelancing jobs, Job satisfaction, Machine Learning, Prediction

I. INTRODUCTION

Freelancing has become a major business field in the world since twenty years. This is a field in which the majority of the younger people work [1]. Job satisfaction refers to a person's sense of fulfilment at work, which serves as a motivator to continue working. Millions of people around the world have lost their jobs, temporarily or permanently, during the COVID-19 pandemic [2]. Furthermore, some countries are facing several financial crises because of the low foreign exchange. For this situation, freelancing is a proper solution for all job seekers.

When considering existing knowledge of this topic, there were no previous studies relevant to this study. So, it is important to do this type of study for decision makers to make decisions related to these freelancing jobs. Further there were no any prediction or analysis of job satisfaction of freelancers on freelancing jobs in Sri Lanka or any other countries using.

The main purpose of this research is to create a model to predict the job satisfaction of freelancing jobs in Sri Lanka using Machine Learning algorithms.

Here, data was collected online through a Google form and the collected data is processed via the Waikato Environment for

Knowledge Analysis (WEKA) data mining tool using five classification algorithms, namely Naïve Bayes, SVM, Decision Tree (J48), Random Forest and MLP. We used supervised learning algorithms for this study because the classification problem is successfully solved by supervised learning, which has been applied to the classification with highly encouraging results [3]. The evaluation process was mainly conducted using accuracy, precision, recall and f-measure.

When it compared with all these algorithms, not only accuracy level and Recall, precision and f-measure but also need to concern the mean absolute error (MAE) and root mean squared error (RMSE) to compare the evaluation results. Furthermore, we have considered confusion matrix as well.

II. MATERIALS AND METHODS

Data Collection: In this study, we collected 240 primary data samples from Sri Lankan freelancers through a questionnaire with help of social media. Because in this decade most people frequently use various kinds of social media platforms. We have gathered the samples using selected social media platforms by sharing a Google form. Facebook, WhatsApp, LinkedIn, Instagram, and Messenger were the selected social media platforms. Here we used the convenience sample for selecting respondents.

Data Pre-Processing: The gathered data have been pre-processed and stored in a proper Comma Separated Values (CSV) file format before applying them to the classification algorithms. In the beginning, the data set contained 27 attributes, but it was reduced to 25 attributes (dependent variable and 24 independent variables). First we removed empty value columns and then attributes are ranked using the information gain ranking algorithm in Weka tool. Here low ranked attributes were removed. Selected attributes shown in below. We mainly categorised all those attributes as personal details and freelancing job satisfaction details. We used the WEKA data mining tool for data pre-processing.

Independent Variables

- Gender
- Age
- Education level
- Working Platform
- Years of Service
- Service type
- Reason for choose freelancing
- Family Status
- No. of Projects
- Full-time or part-time

- Income level
- Complete work on time
- Regular customers
- Work at convenient hours
- Receive good feedback
- Easily find the possible works
- Work & personal life balance
- Enough facilities to do freelancing
- Enough Communication skills
- Satisfaction of current income
- Alone or as a team on freelancing
- Continue freelancing in the future
- Spend time on freelancing

Depend Variable

- Job Satisfaction

Classification: The data set is separated into two parts called training and testing before applying to the algorithm. Then the data set is classified using Naïve Bayes, SVM, and Decision Tree (J48), Random Forest, and MLP classification algorithms through the WEKA tool. These classification algorithms have chosen based on literature review relevant to the prediction based studies. Research framework of this study is illustrated in Figure 1.

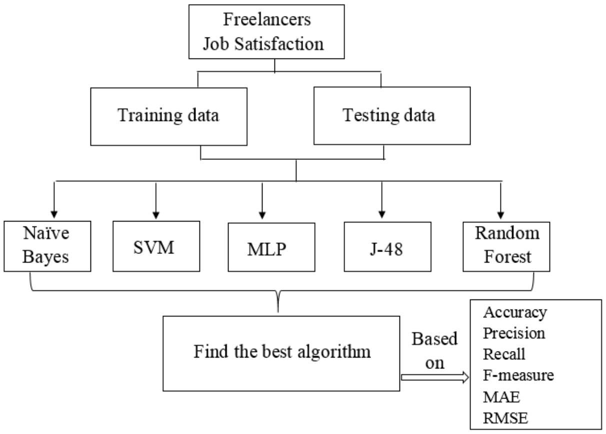


Figure 1: Research Framework

The environment for the experiment used Microsoft Windows 10 on a PC with Processor Intel® Core (TM) i5-1135G7CPU @ 2.40GHz, RAM 8.0GB. The WEKA 3.8.6 data mining tool is used as the training and testing environment. Above six classification algorithms were used to find the best algorithm for the prediction task. Mainly we consider the accuracy level because it is the most important measure for evaluating the model's accuracy.

Precision defines the number of positive class predictions that actually belong to the positive class. It's calculate using below (1).

$$Precision = \frac{True\ Positive}{True\ Positive +\ \ False\ Positive} \qquad (1)$$

Recall is a metric that measures the amounts of accurate positive predictions among all possible positive predictions. It's calculate using below (2)

$$Recall = \frac{True\ Positive}{(True\ Posituve + False\ Negative)} \qquad (2)$$

F-measure value gives out the combined measure of precision and recall and it calculated using (3).

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (3)$$

Error is essentially the difference in absolute terms between the true or actual values and the predicted values where n= no of observations, Y_i= actual value, X_i=predicted value. MAE calculation is shown in (4).

$$MAE\ = \frac{1}{n}\sum_{i=1}^{n} |\, Y_i -\ X_i| \qquad (4)$$

The standard deviation of the errors that happen when a prediction is made based on a dataset is known as RMSE where n= no of observations, Y_i= actual value, X_i=predicted value. RMSE calculation is shown in (5).

$$RMSE\ = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2} \qquad (5)$$

Precision, recall, f-measure, ROC, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and confusion matrix also used to build the prediction model.

Table 1 illustrates the accuracy, confusion matrix, precision, recall, f-measure, of the analyzed algorithms.

Through the overall results we can conclude the decision tree algorithm gives the highest accuracy level which is 92.5%. When compared with other algorithms, MLP shows a 90.4% accurate level as the second highest accuracy level.

Confusion matrix is important to evaluate the precision, recall, f-measure and ROC area [4].

TABLE I. ACCURACY, CONFUSION MATRIX, PRECISION, RECALL, F-MEASURE EVALUATION RESULTS

Classification Algorithm	**Confusion Matrix**				**Accuracy**	**Precision**	**Recall**	**F-measure**
	TN	**FP**	**FN**	**TP**				
Naïve Bayes	105	6	20	105	89.17%	0.898	0.892	0.892
SVM	100	11	17	112	87.92%	0.885	0.883	0.883
J48	104	7	11	118	92.5%	0.926	0.925	0.925
Random Forest	97	14	11	118	89.58%	0.896	0.896	0.896
MLP	100	11	12	117	90.42%	0.904	0.904	0.904

Those measures are calculated by using the True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) values shown in the table 01.

According to Table I, it illustrates the results of precision, recall and f-measure values as an average. In precision the highest value, 0.926 is for decision tree while the other algorithms shows the lower rates than it. In recall, 0.925 and f-measure, 0.925 also highest values in decision tree when compared with other algorithms. As an overall, decision tree shows the highest rates. Meanwhile, SVM shows the lowest rates.

Further, it needs to be concern about the error rates as well. In this analysis, we concern the MAE and RMSE to compare the evaluation results. Based on the evaluation results, J48 shows lowest error value. Random Forest shows the highest error rate. So, as an overall result J48 showing lowest error rate. The result is illustrated in Figure 2.

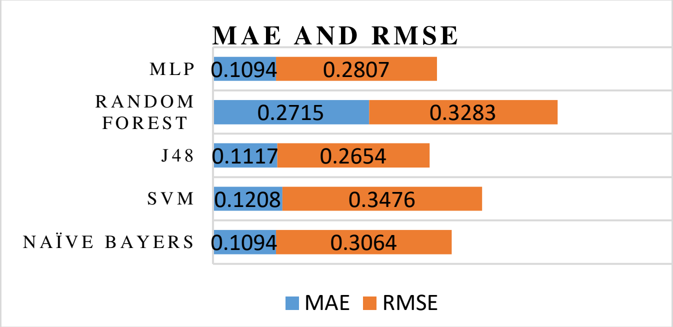


Figure 2: MAE and RMSE chart

For the analysis the *k*-fold Cross-Validation is the statistical method used in evaluating the algorithms by dividing the data in to two segments called testing set and training set. This is mainly used to build the models in prediction based research. We used 5-folds cross validation and 10-folds cross validation. Predominantly 10-folds cross validation obtained best performance and J48 showed it as 92.5 in percentages.

Iv. CONCLUSION

As many Sri Lankans tend to do freelancing as their carrier, it is become important to conduct a prediction based research on job satisfaction of freelancers.

Because of that this research implemented a model for predicting the job satisfaction of freelancers in Sri Lanka. Data for this study gathered from primary data sources and the pre-processing is done through the WEKA data mining tool. We used five classification algorithms such as Naïve Bayes, SVM, Decision Tree (J48), Random Forest, and MLP for data Processing. When considering results of these five algorithms, Decision Tree (J48) is provided higher accuracy (92.5%) than other four algorithms. The maximum results of precision (0.926), recall (0.925) and f-measure (0.925) also represented by the decision tree algorithm. Further, minimum error rate which is MAE and RMSE also provided. It is 0.1117 and 0.2654 respectively. So, decision tree is the best model for this study.

In this study job satisfaction was measured based on factors of the independent variables. The dependent variable is job satisfaction and it was directly asked from the respondents when collecting the data. But, after implementing the proposed model by a decision tree algorithm we can predict the freelancers' job satisfaction. If we want to know the job satisfaction of the new freelancer, we can ask the same questions (independent variables) from him/her and then apply those answers to the model. Then based on those answers model will predict the satisfaction (dependent variable) of that person.

We are planning to collect more and more data on the same field and improve the performance of the prediction model using ensemble learning approach in the future. This prediction results would help future freelancers, governments and relevant authorities to take decisions.

REFERENCES

[1] W.A.K., Exporting Services: Identifying Freelancers as Entrepreneurs to Boom Sri Lankan Economy, 2018.

[2] V. shiyani, "Job Satisfaction: Meaning, Definition, Importance, Factors, Effects and Theories," *Business management Ideas,* 2022.

[3] E. A. a. M. A. Akcayol, "A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques".

[4] J. Brownlee, Machine Learning Mastery, 3 January 2020. [Online]. Available: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/.

[5] "www.guru99.com," [Online]. Available: https://www.guru99.com/confusion-matrix-machine-learning-example.html.

3rd International Women in Engineering Symposium WIESymp 2022

2022 WIESymp

“Sustainable transformation of Technology”

Organized by

IEEE Women in Engineering

Wie

Sri Lanka