



SMART COMPUTING & SYSTEMS ENGINEERING

International Research Conference

Colombo, Sri Lanka | 24th September 2020

PROCEEDINGS



Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka

**University of Kelaniya
Sri Lanka**

PROCEEDINGS

International Research Conference on
Smart Computing and Systems Engineering
(SCSE 2020)

24th September 2020

Department of Industrial Management,

Faculty of Science, University of Kelaniya, Sri Lanka

© University of Kelaniya, Sri Lanka
Proceedings of the
International Research Conference on Smart Computing and Systems Engineering
SCSE 2020

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the publisher.

ISSN 2613-8662

Published by

Department of Industrial Management

Faculty of Science, University of Kelaniya

Sri Lanka

Contents

	<i>Page</i>
Editorial Board	v
Programme Committee	vii
Organizing Committee	ix
Keynote Speeches	
Professor Darshana Sadera	xi
Professor Janaka Wijayanayake	xiii
List of Papers	xvii
Smart Computing	1
Systems Engineering	142
Industry Forum 1.0	223

Editorial Board

International Research Conference on Smart Computing and Systems Engineering 2020

Chief Editor : Dr. Suren Peter

Committee : Dr. (Mrs.) Annista Wijayanayake

Dr. (Ms.) Dilani Wickramaarachchi

Dr. Chathura Rajapakse

Dr. Ruwan Wickramarachchi

Ms. Hiruni Niwunhella

Mr. Janaka Senanayake

Programme Committee

International Research Conference on Smart Computing and Systems Engineering 2020

Prof. Darshana Sedera	Southern Cross University in Gold Coast, Australia
Prof. Janaka Wijayanayake	University of Kelaniya, Sri Lanka
Prof. Setsuya Kurahashi	University of Tsukuba, Japan
Prof. Athula Ginige	Western Sydney University, Australia
Prof. Takano Terano	(Retired) Tokyo Institute of Technology, Japan
Associate Prof. Masakazu Takahashi	Yamaguchi University Management of Technology, Japan
Dr. Muhammed Badruddin Khan	Al Imam Mohammad ibn Saud Islamic University(IMSUI), Riyadh, KSA
Dr. Lalith Goonatilake	(Former Director) Trade Capacity Building, UNIDO
Dr. Niroshinie Fernando	Deakin University, Australia
Dr. Irvan Mhd	Tokyo Institute of Technology, Japan
Dr. Raj Prasanna	Massey University, New Zealand
Dr. Prem Samaranayake	Western Sydney University, Australia
Dr. Malathi Sajeewani	University of Melbourne, Australia
Dr. Jeewanie Jayasinghe	University of Ruhuna, Sri Lanka
Dr. Antonio Hyder	Hackers and Founders Research, United States
Dr. Julian Nanayakkara	(Retired) University of Kelaniya, Sri Lanka
Dr. Annista Wijayanayake	University of Kelaniya, Sri Lanka
Dr. Suren Peter	University of Kelaniya, Sri Lanka
Dr. Ruwan Wickramarachchi	University of Kelaniya, Sri Lanka
Dr. Shantha Jayalal	University of Kelaniya, Sri Lanka

Dr. Keerthi Wijayasiriwardhane	University of Kelaniya, Sri Lanka
Dr. Chathura Rajapakse	University of Kelaniya, Sri Lanka
Dr. Dilani Wickramarachchi	University of Kelaniya, Sri Lanka
Dr. Ajantha Athukorala	University of Colombo School of Computing, Sri Lanka
Dr. Dhammika Elkaduwa	University of Peradeniya, Sri Lanka

Organizing Committee

International Research Conference on Smart Computing and Systems Engineering – 2020

Conference Chair **Dr. (Ms.) Dilani Wickramaarachchi**

Track Chairs **Dr. Shantha Jayalal**
 Dr. Julian Nanayakkara
 Dr. Chathura Rajapakse

Conference Secretary **Ms. Hiruni Niwunhella**

Members **Dr. Suren Peter**
 Dr. (Mrs.) Annista Wijayanayake
 Dr. Ruwan Wickramarachchi
 Dr. Keerthi Wijesiriwardana
 Mr. Buddhika Jayawardana
 Mrs. Mahikala Niranga
 Mr. Janaka Senanayake

Keynote Speech

‘Digital First’ Thinking and the Caveats of Digital Transformation

Professor Darshana Sedera

*Deputy Dean, School of Business and Tourism and Director of Digital Enterprise Lab
Southern Cross University, Gold Coast – Australia*

*Adjunct Professor
Swinburne University of Technology, Melbourne – Australia*

darshana.sedera@scu.edu.au

The recent advancements in technologies such as cloud, wearables, internet-of-things, mobile, social media, business analytics and artificial intelligence purported to provide unprecedented opportunities to innovate and develop new business models. The rise and growth of digital organizations such as Uber, Airbnb, and Alibaba.com are strongly attributed to the advancements in modern technologies and the consumerization of information technology. While such examples epitome the characteristics such as availability, accessibility, scalability, and ease-of-use and ease-of-deployment of modern technologies, contemporary organizations face a new set of challenges in innovation with modern IT. The scientific research conducted show that organizations are still struggling to derive the full potential of the modern technology portfolio. The keynote will outline what strategies are needed for the ‘digital first’ thinking in organizations and how to prepare organizations for the new digital era.

In a similar vein, societies and individuals too are being impacted by the modern technologies. There is strong and recurring anecdotal commentary on the negative impact of social media, wearables, and other digital devices. Moreover, social media has been linked to the propagation of falsified information, violence, and anti-social behavior. Latest insights that are published in international research outlets, arising from the results of three scientific investigations will be shared. Therein, the notions of ‘digital perfectionism,’ ‘digital hoarding’ and ‘social media addiction’ as emerging mental disorders will be introduced.

Impact of COVID-19 on Digital Transformation

Professor Janaka Wijayanayake

*Department of Industrial Management
University of Kelaniya, Sri Lanka
janaka@kln.ac.lk*

Whether public or private, large or small, different types of Industries have been influenced by the digital revolution during the last few decades. As need of individuals and societies evolves, it is absolutely necessity that industries retool, upskill and strategize to remain competitive and relevant to the clients they serve. This reality is the primary reason that conference such as these are necessary, so educators, researchers and industry leaders can find strategies to become more agile to face the challenges of the future.

In this information age, information technology is an integral part of the growth and development of businesses and nations. In today's digital society, we are not only global citizens but also 'netizens.' Our lives are characterized by increased automation, utilizing the Internet of Things (IoT), artificial intelligence, machine learning, data science and a whole lot of new and emerging technologies. We are in a data driven culture where data is one of the most important strategic assets. With each browse of the internet, we are leaving behind digital footprints, that websites can track our activity and predict our behaviour. From our search patterns on the Internet to our social media posts and online transactions, we are constantly sharing data about ourselves which is being collected to design a unique experience tailored specifically for us. Data and data analytics are driving digital transformation, creating monetization opportunities, radically changing customer experience and reshaping industries. At a time when technology is changing at an exponential rate, digital transformation is not a choice, it is an inevitability. The difference between winners and losers can be as simple as one who embrace the digital transformation and one who does not embrace the digital transformation in their organization.

Digital transformation is the integration of digital technology into all areas of an industry, government or society as a whole. It's also a cultural change that requires organizations to continually challenge the status quo, experiment, and introduce new way to do things using the technology. This sometimes means changing long-standing processes within organizations in favour of relatively new practices that are more efficient and effective. Digital transformation closes the gap between what customers already expect and what organizations actually deliver. This involves different thinking, the encouragement of innovation and new business models, incorporating digitization of assets and an increased use of technology to improve the experience of your organization's employees, customers, suppliers, partners and stakeholders. Digital transformation can help to develop and grow businesses, create opportunities, generate jobs and reduce social inequalities and support economic growth.

The world is now going through its own digital revolution. The COVID-19 pandemic has further reinforced the importance of digital transformation of organizations using technology as an enabler. Technology supported nearly every change triggered or accelerated by the crisis. Emerging technologies that enable automation, artificial intelligence, connected devices and equipment, large volume of data transfer, and collaboration tools, are champions in these turbulent times, helping businesses recover from the initial impacts of COVID-19 and set a strong foundation for the future growth.

The impact of the COVID-19 pandemic is affecting every industry, but in very different ways. Through this unprecedented change in the way we engage, some industries are experiencing strong growth, while others are on the verge of collapse without significant intervention. Following the shutdown, when human contact was limited, organizations that were able to remotely connect with employees and customers and manage current operations demonstrated a level of resilience that increased trust in the business. According to 2020 Global Emerging Technology Survey Report by KPMG International and

HFS Research, 57% of executives say that COVID-19 has significantly changed their organization's strategic priorities for emerging technologies, 59% of executives say the pandemic has created an urgency to accelerate their digital transformation initiatives, 56% of executives agree that cloud migration has become an absolute necessity in such situations, and 80% of companies are investing or planning to invest in emerging technologies face the crisis.

Although digital transformation will vary widely based on organization's specific challenges and demands, there are a few constants and common themes among existing case studies and published frameworks that all business and technology leaders should consider as they embark on digital transformation. Those elements are customer experience, operational processes, culture and leadership, workforce enablement through teleworking and digital technology integration.

Customer experience

Before a digital transformation can truly begin, organizations must change their focus from product-focused to customer-focused. The driving force behind technology decisions should be customers, and the goal should be to make their lives easier instead of making things easier for the organization. For example, a bank established a Twitter account to answer client complaints quickly, helping customers avoid going physically to a branch, and retailers offering home shopping with the option to receive products by mail or at a store. Crucially, a customer focus should be the basis for all digital transformation decisions.

Operational processes

Although transformed customer experiences are the most visible and arguably the most exciting aspects of digital transformation, organizations are also realizing benefits from transforming internal and external processes through process digitization. This requires a change in thinking about how an organization delivers its products and services. Digitization can be powerful in improving the speed and reliability of the internal and external processes from how fast products are manufactured, to efficiency of order fulfilment and delivery of product and services to customers. To fully leverage the transformation, organizations need to look at how the supply chain can be digitized and improved too. By automating processes, organizations can become more agile, more responsive to changes in demand, and better able to increase and sustain profitability.

Culture and leadership

The most effective change management efforts are aligned with the modern, dynamic business environment. Organizations need to break down internal silos to create a cohesive organization that embraces change. Change is hard, no matter how much it will benefit the organization. One of the most common reasons digital transformations fail, is because employees do not embrace the change and support it. A successful digital transformation starts from the top with leaders who drive employees towards the vision. This involves engaging all employees, including executives and leaders to get them on board with the new digital vision. Every executive and leader must play a role in championing digital change and synthesizing the digital transformation with the organization's, long-term goals.

Workforce enablement through teleworking

When focusing on digital solutions, organizations need to consider their internal staff. Assessing their competence and knowledge in applying digitization to their work, obtaining employee feedback and providing state of the art technology solutions that empowers employees are key, areas which have to be addressed in operationalizing the plan. Traditional commute

to work may not be a necessity for all employees. They could be allowed to work from home several days per week. The collaboration and networking tools allow employees to communicate with anyone in the organization from wherever they are working from. The tools that virtualize individual work would become powerful enablers for knowledge sharing. After COVID-19, many industries that formerly prohibited teleworking are now allow employees to work from home. From bankers, engineers, to teacher working from home have become normal and work-life as we have known it, has changed for most of us.

Digital technology integration

A streamlined data strategy is required for a successful digital transformation. All data systems need to work together and be integrated into the company's internal processes. Integrated advanced digital technology can improve business processes in several ways. For example, big data analytics can help in-bound logistics run more smoothly by tracking product movements, the cloud can be used to create uniform business processing platforms, and mobile platforms can enable employees to perform their work anytime, anywhere and on any device. When using standardized integrated platforms, it is also much easier to globally source processes, which can lead to substantial cost reductions.

Impact on Higher Education

Universities will always be the founding pillar of higher education. Despite the impact of COVID-19 and the havoc it may have faced, we cannot underestimate or forget the resilience universities have shown throughout human history. Forced by the sudden shuttering of their physical campuses, universities had to take their classes online and ensured continuity and normality. This, of course, led to an immediate disruption in the lives of many. It also highlighted the inequalities between students those who have access to resources and technology and those without access to these devices.

However, we should consider this pandemic as a blessing that forces the universities to develop new strategic and tactical measures to face the future. We can now see that COVID-19 has redirected and amplified the actions of universities across the world toward digital transformation, reshaping and changing their interests into guaranteeing short-term operational continuity, while ensuring long-term institutional viability through adoption of technology.

Digital transformation means transforming an organization's core business to better meet customer needs by adopting technology. In education, the customer is students and it could also be faculty, staff, and alumni. As an example, transforming the student experience could be

- Facilitating online registration via any device on scalable cloud-based student information systems;
- providing different types of online learning options;
- converting courses to blended models;
- using technology to monitor student progress; and
- partnering with Massive Open Online Course to provide digital certificates to enhance career opportunities.

We need to understand that mere, uploading of academic content on Learning Management System (LMS) does not help virtual learning. Therefore, universities now in a virtual learning transition should refocus and emphasize capacity building within the university to facilitate virtual learning environment. Instructional design, multimedia production and data analytics are vital in digital transformation of university education. Faculty from all disciplines need to be motivated, guided and trained, as their courses and programs are reconfigured and adapted to a new environment.

Developing a virtual culture for universities will require open leadership with an innovative mentality. Enforcing changes are not enough unless they are followed by an internal culture of action. However, it would not be wise to abandon conventional, face-to-face interaction model of universities entirely. What seems to have become a consensus view is that the universities should shift face-to-face classes towards blended and a hybrid mode. When we transform into blended and hybrid mode, we

need to change evaluation methods, support material, assessment systems, as well as technological tools that enable virtual interaction in class which can to be done without leadership support.

Digital transformation is an ongoing process, Change and venturing into the unknown can be difficult, but the benefits that come from creating customer-focused, digital organization can be lasting. Perhaps this can be the best impact on businesses and education in the age of COVID-19. Ladies and gentlemen, the time is right not just for digital transformation but digitally aware leaders, who are bold and strategic to evolve with the times. I wish for you a successful conference.

LIST OF PAPERS

Smart Computing	1 - 141
Systems Engineering	142 - 222
Industry Forum 1.0	223 - 241

Theme 1: Smart Computing

SC-01	Towards detecting morning surge from sleep self-evaluations Masakazu Takahashi, Noriyuki Sugahara and Masashi Shibata	1
SC-02	Machine translation system to convert Sinhala and English Braille documents into voice K. Sasindu Anuradha and Samantha Thelijjagoda	7
SC-03	Grammatical error detection and correction model for Sinhala language sentences H. M. U. Pabasara and S. Jayalal	17
SC-04	Vision-based automatic warning system to prevent dangerous and illegal vehicle overtaking Mahinsa Athree and Anusha Jayasiri	25
SC-05	Dengue mosquito larvae identification using digital images W. D. M. De Silva and S. Jayalal	31
SC-06	A modified cognitive complexity metric to improve the readability of object-oriented software Thilini Jayalath and Samantha Thelijjagoda	37
SC-07	Contextual assistant framework for the Sinhala language D. D. S. S. Dasanayaka and N. Warnajith	45
SC-08	Blockchain-based distributed reputation model for ensuring trust in mobile ad-hoc networks P. P. C. Peiris, Chathura Rajapakse and B. Jayawardena	51
SC-09	Design of an auto disconnecting regulator and a safety switch to prevent domestic gas leakages R. M. I. U. Rajapaksha, P. S. H. Perera, P. K. D. M. Nandasena, P. S. T. K. Gunarathna, P. P. D. Gihan Kanishka, A. L. A. K. Ranaweera and S. R. D. Kalingamudali	57
SC-10	Anomaly Detection in Cloud Network Data Tharindu Lakshan Yasarathna and Lankeshwara Munasinghe	62
SC-11	Minimization of fraudulent activities in land authentication through Blockchain-based system L. W. D. C. Jayabodhi, Chathura Rajapakse and J. M. D. Senanayake	68

SC-12	Smart electricity monitoring and analysing an IoT system with a mobile application A. I. R. Fernando and M. D. R. Perera	75
SC-13	A microtransaction model based on blockchain technology to improve service levels in public transport sector in Sri Lanka S. A. Jayalath, Chathura Rajapakse and J. M. D. Senanayake	82
SC-14	IoT based animal classification system using convolutional neural network L. G. C. Vithakshana and W. G. D. M. Samankula	90
SC-15	Source credibility analysis on Twitter users Malith Wijesekara and Gamage Upeksha Ganegoda	96
SC-16	Identifying religious extremism-based threats in Sri Lanka using bilingual social media intelligence Aneesha Fernando and Thareendra Keerthi Wijayasiriwardhane	103
SC-17	Affective gaming in real-time emotion detection and music emotion recognition: Implementation approach with electroencephalogram Pradeep Kalansooriya, G. A. D Ganepola and T. S. Thalagala	111
SC-18	Impact of metacognition and age group on contemporary video game interface and gameplay design Nandunie N. Harischandra, Lahiru A. Jayakody and Tiroshan Madusanka	117
SC-19	Detecting human emotions on Facebook comments E. J. A. P. C. Chathumali and Samantha Thelijjagoda	124
SC-20	Keyword extraction from Tweets using NLP tools for collecting relevant news Thiruni D. Jayasiriwardene and Gamage Upeksha Ganegoda	129
SC-21	Sentiment classification of Sinhala content in social media Pradeep Jayasuriya, Sarith Ekanayake, Ranjiva Munasinghe, Bihara Kumarasinghe, Isuru Weerasinghe and Samantha Thelijjagoda	136

Theme 2: Systems Engineering

SE-01	Ecosystem mapping technology to understand innovation challenges in MSME Sector: An analysis of the handicraft sector in Sri Lanka Chathura Rajapakse	142
SE-02	Contribution of reflection in language emergence with an under-restricted situation Kense Todo and Masayuki Yamamura	147
SE-03	Profiling purchasing behavior of Generation Z Nadeesha Kahawandala, Suren Peter and Hiruni Niwunhella	155
SE-04	Initiating customer relationship measurement at small and medium enterprises with low computational cost Tharika Chalani Weerakoon and Kapila Tharanga Rathnayaka	161
SE-05	Identification of factors and classifying the accident severity in Colombo - Katunayake expressway, Sri Lanka M.A.K. Kushan and N.V. Chandrasekara	166
SE-06	Industry 4.0 readiness assessment for apparel industry: A study in the Sri Lankan context Erandika Lakmali, Kasuni Vidanagamachchi and Julian Nanayakkara	174
SE-07	A solution to overcome speech disorder of patients using Brain Neuron EEG Signals J. A. D. T Jayawickrama and Samantha Thelijjagoda	182
SE-08	Extending use-case point-based software effort estimation for Open Source freelance software development Dharshitha Srimal Senevirathne and Thareendhra Keerthi Wijayasiriwardhane	188
SE-09	A survey on applying machine learning to enhance trust in mobile adhoc networks G. M. Jinarajadasa and S. R. Liyange	195
SE-10	Predicting examination performance using machine learning approach: A case study of the Grade 5 scholarship examination in Sri Lanka U. M. Ranawaka and Chathura Rajapakse	202
SE-11	Intelligent changeover solution for a domestic hybrid power system H. M. R. M. Herath, K. H. S. Tharindu, K. D. M. Fernando, P. M. K. Alahakonn and R. Hirshan	210

SE-12	Novel computational approaches for border irregularity prediction to detect melanoma in skin lesions D. V. D. S. Abeysinghe and S. Sotheeswaran	216
-------	---	-----

Theme 3: Industry Forum 1.0

IF-01	Relationships between climatic factors to the paddy yield in the North-Western Province of Sri Lanka Lasini Wickramasinghe, Jeevani Jayasinghe and Upaka Rathnayake	223
IF-02	Aspect-based sentiment analysis on hair care product reviews Malki Kothalawala and Samantha Thelijjagoda	228
IF-03	Demystifying the concept of IoT enabled gamification in retail marketing: An exploratory study Shalini Jayasooriya, Tharindhie Alles and Samantha Thelijjagoda	234

Towards detecting morning surge from sleep self-evaluations

Masakazu Takahashi*
Graduate School of Innovation and
Technology Management
Yamaguchi University, Japan
masakazu@yamaguchi-u.ac.jp

Noriyuki Sugahara
Graduate School of Sciences and
Technology for Innovation
Yamaguchi University, Japan
b003wc@yamaguchi-u.ac.jp

Masashi Shibata
Graduate School of Sciences and
Technology for Innovation
Yamaguchi University, Japan
g501wc@yamaguchi-u.ac.jp

Abstract: This paper aims to analyze the blood pressure transition during sleep. Morning surge is a sudden increase in blood pressure from awakening or before awakening. Morning surge is implicated in cardiovascular diseases, such as Stroke, Angina Pectoris, and Myocardial Infarction. Morning surge has been detected mainly by the ABPM (Ambulatory Blood Pressure Monitoring) method, which measures blood pressure for 24-hours. Since the ABPM method cannot distinguish awakening and sleep automatically, their alternative method is forcibly delimiting time or manually processing based on behaviour records. Therefore, it is necessary to capture the blood pressure change under clear sleep separation. This paper employs two sleep criteria for accurate blood pressure during sleep.

Keywords: Blood pressure, Clustering, Hypertension, Machine learning, Unsupervised learning

I. INTRODUCTION

Blood pressure indicates the pressure that blood applies to blood vessels. The heart contracts in sending blood. The pressure applied to the blood vessel at this contract time is called the SBP (Systolic Blood Pressure). SBP shows the highest value in mmHg. On the contrary, when the blood returns to the heart, the heart expands. The blood pressure at this time is called DBP (Diastolic Blood Pressure) with the lowest score.

Hypertension exhibits excessive state in which the pressure is exerted on the vessel. The continuation of hypertension imposes a burden on blood vessels, and the risk of cerebral hemorrhage occurring due to the rupture of blood vessels in the brain is increased. Further, the heart gradually increases as it attempts to submit blood flow, cardiac hypertrophy. As the cardiac hypertrophy progresses, the movement of the heart strikes and the function of delivering blood to the whole-body decreases. Thus, hypertension causes various diseases.

When the early morning blood pressure rises sharply before or after waking up early in the morning, it is called Morning surge. Morning surge is thought to be involved in the pathology of cardiovascular disease independent of normal blood pressure fluctuations.

Morning surge has been detected mainly by the ABPM (Ambulatory Blood Pressure Monitoring) method. The ABPM measures blood pressure for 24 hours every 15 minutes. However, both hypertension standards and normotensive pressure standards by the ABPM are not specified. Furthermore, there is no unique method for

efficiently separating daytime and nighttime. The separation between daytime and nighttime in the ABPM is made by the fixed time and merged from the record of patients' behaviour.

In this paper, data analysis is made for morning surge detection by capturing the distinction between sleeping and awakening. The rest of the paper is organized as follows. Section 2 discusses the backgrounds of the research and related work; Section 3 briefly summarizes the gathered data from the target clinics; Section 4 describes the analytics of the data and presents analytical results; and Section 5 gives concluding remarks and future work.

II. RELATED WORK

The Ministry of Health, Labor and Welfare, Japan defines hypertension as a disease that increases blood pressure and increases the risk of stroke or heart disease [1]. It is estimated that the number of Japanese patients with hypertension is approximately 43 million. As blood pressure rises, it gradually inflicts damage to the brain and heart. Therefore, if we ignore a high blood pressure conditions, the probability of causing a stroke or myocardial infarction and causing sudden death increases. It is estimated that as the SBP rises by ten mmHg, the risk of stroke increases by about 20% and the risk of angina and myocardial infarction by about 15% [2]. Therefore, the blood pressure value is one of the essential indicators that can evaluate the risk of stroke onset or myocardial infarction.

Hypertension Treatment Guidelines 2014 was the world's first, which stated that in case of a blood pressure gap between the examination room and home, home blood pressure should be given priority in hypertension diagnosis. Further, it defines reference values of the blood pressure by the measurement conditions as follows:

In hospital: more than 140 mmHg in SBP or more than 90 mmHg in DBP

In-home: more than 135 mmHg in SBP or more than 85 mmHg in DBP

Hypertension is divided into essential hypertension and secondary hypertension categories. Approximately 85% to 90% is an essential one, and causes of blood pressure increase cannot be identified due to multiple causes such as genetic predisposition, lifestyle, weight, stress, and other factors. The rest is identified as secondary hypertension, which is caused by certain diseases. Secondary hypertension can be identified, and healing and improvement can be expected by appropriate diagnosis and treatment [3].

Blood pressure changes during sleep also increases when the sympathetic is actively working due to stress. Hypertension is followed by hardness and fragility in arteries and becomes arteriosclerosis. Moreover, hypertension causes various diseases in the brain and heart. The report says that the higher the blood pressure, the higher the risk of death from cerebral cardiovascular diseases such as stroke and myocardial infarction [4]. Conventionally, diagnosis and treatment of hypertension have been carried out based on occasional blood pressure in the outpatient examination. In 2009, the JAMP (Japan Ambulatory Blood Pressure Prospective Study) was launched [5]. It was an epidemiological study that measures blood pressure for 24 hours using the ABPM Study (Ambulatory Blood Pressure Monitoring). From this study, the importance of blood pressure fluctuation during sleep was found. Furthermore, blood pressure fluctuation was classified as follows [6].

Coat hypertension: Elevation of blood pressure only at the outpatient examination.

Morning surge: Early-morning boost in which blood pressure rises sharply in early morning awakening or before arousal.

Non-dipper: Does not show nocturnal depression observed in normotensive subjects.

Following findings are made such as the blood pressure fluctuation, the morning surge of elderly hypertension becomes a stroke risk independently of the 24-hour blood pressure level and morning surge has become a cardiovascular event risk [7] - [12].

One of the features of the ABPM method is a number of measurements [13]. The average blood pressure values with many numbers of measurements well reflect the patient's blood pressure condition [14]. The ABPM with many blood pressure measurements can predict cardiovascular organ disorders, and the prognosis of patients was reported [15].

The sixth report of the Joint Committee of the United States in 1997 (JNC-VI) and the WHO / ISH hypertension guidelines in 1999 has presented the normal range of ABPM [16] [17].

The reference values are also presented for blood pressure at daytime and at the sleeping time [18]. However, the criteria for hypertension and normotensive pressure are not specified. Furthermore, there is no unified separation method for separating daytime and nighttime so far. Currently, it is processed by the integration with the methods and activity record delimiting the time to wake-up and sleep at a fixed time.

J-MUBA (Japanese Multicenter Study on Barnidipine with the ABPM), was to evaluate the antihypertensive effect of barnidipine by 24-hour blood pressure measurement (ABPM) using a portable automatic sphygmomanometer. They separate the daytime and sleeping from the activity records manually. As a result, there is almost no difference between the two methods, and these two methods are applied for day and night separation [19].

In the ABPM method, measurement is frequently performed in the daytime and low frequency in the night. Therefore, in order to calculate the average value of the 24-

hour blood pressure, the correction of Equation (1) is required.

$$24\text{-hour ABPM average} = (\text{Daytime blood pressure average}) \times (\text{Arousal hour}) + (\text{Nighttime blood pressure average}) \times (\text{Sleeping time}) / 24 \quad (1)$$

The definition of early morning hypertension is divided into broad and narrow interpretation shown by Equation (2) to Equation (6) [20].

Broad interpretation:

$$\text{Morning Hypertension} \stackrel{\text{def}}{=} [85,135] \quad (2)$$

Narrow interpretation:

Definition 1

$$\text{Morning Blood Pressure} \stackrel{\text{def}}{=} [85,135] \quad (3)$$

$$\text{Bedtime Blood Pressure} \stackrel{\text{def}}{=} [85,135] \quad (4)$$

Definition 2

$$\text{ME Average} \stackrel{\text{def}}{=} [85,135] \quad (5)$$

$$\text{ME Difference} \stackrel{\text{def}}{=} [85,135] \quad (6)$$

Where ME Average and ME Difference are expressed by Equation (7) and Equation (8).

$$\text{ME Average} = (\text{Morning Blood Pressure} + \text{Bedtime Blood Pressure}) / 2 \quad (7)$$

$$\text{ME Difference} = (\text{Morning Blood Pressure} - \text{Bedtime Blood Pressure}) \quad (8)$$

Accordingly, classification based on blood pressure is shown in Fig. 1, and the morning surge group is also indicated by Equation (9) to Equation (12) [7].

Morning surge group:

$$\text{Daytime surge (in SBP)} \geq 45\text{-}55 \text{ mmHg} \quad (9)$$

$$\text{Morning blood pressure} = (\text{Average blood pressure for two-hours after waking up}) \quad (10)$$

$$\text{Nighttime lowest blood pressure} = (\text{Average two-hour blood pressure including nighttime lowest pressure}) \quad (11)$$

$$\text{Daytime surge} = (\text{Morning blood pressure}) - (\text{Nighttime lowest blood pressure}) \quad (12)$$

Prehypertension clusters are defined as Masked Hypertension, where both clinical blood pressure and blood pressure average are not high, but midnight surge comes from sleep apnea syndrome and morning high blood pressure [21]. Therefore, capturing accurate blood pressure fluctuation makes it possible to predict the vascular risk of the brain, heart, kidney, and other organs. The primary issue of blood pressure measurement at night, such as the ABPM is clear identification of sleeping and awakening.

In this paper, the analytical results based on the nighttime blood pressure measurement using equipment allow each person to operate and measure the wake-up time clearly before sleeping and awakening is described. Furthermore, the two types of sleep level surveys are merged with this result.

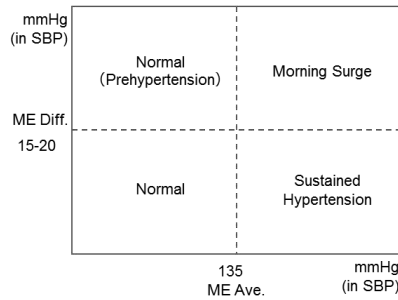


Fig. 1. Hypertension classification by average and difference score in SBP

III. EXPERIMENTAL CONFIGURATION

For the measurement method of nocturnal blood pressure, a blood pressure monitor is lent to a patient diagnosed as hypertension, and blood pressure is monitored for a week. Fig.2 shows the examination and blood pressure data collection cycle.

With the Sphygmomanometer, blood pressures before sleeping and immediately after waking up are measured at regular intervals by button operation before going to bed and after waking up. In this trial, it was made every hour at night. Table I indicates the result of gathering data.

In the ordinal hypertension examination, physicians focus on the value of SBP. Therefore, Table II shows the patients' experimental summary in SBP. Since the living cycle differs from each patient to patient, the nighttime blood pressure was taken from 2 o'clock to 5 o'clock in the nighttime analysis.

As for the insomnia determination method, the PQSI (Pittsburgh Sleep Quality Index) and the AIS (Athens Insomnia Scale) are popular [22] [23]. These are intended to determine the sleep quality from the questionnaire.

The PSQI is a self-written questionnaire development to evaluate sleep and its quality. The AIS is a worldwide standard insomnia assessment method created by WHO (the World Health Organization) with the World Project on Sleep and Health project.

The answers to the subjective sleep level of the patient were merged in addition to the objective sleep level in AIS to investigate the relationship between the blood pressure transition and the sleep level during the night.

The sleep level scale of HNC was evaluated in 7 scales. This scale was made after collecting the blood pressure record for one week, and the scale is from 1 as "well slept" to 7 as "insomnia".

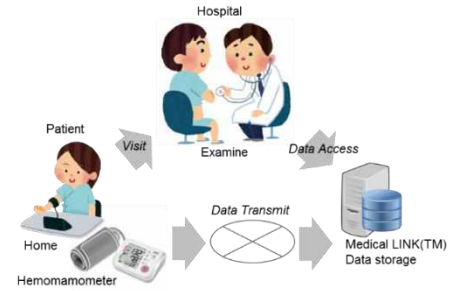


Fig. 2. Examination flow

TABLE I. GATHERING DATA ATTRIBUTES

Term	2015/11/1-2017/07/01
Number of Clients	140 clients
Number of Records	20,102 (104 records were excluded for machine error)
Record Information	ID, Date, Hour, SBP, DBP, Pulse, Temperature, Time Flag, ASL Sleep Level, Original Sleep Level
Hemomamometer	HEM-8723A-ND

TABLE II. EXPERIMENTAL SUMMARY IN SBP

Hour	N	Ave.	Med.	S.D.	Min.	Max.
E	5,208	121	120	16	73	216
23	7	90	90	9	78	101
0	19	103	107	10	88	116
1	275	112	113	15	65	151
2	2,460	114	113	16	65	202
3	2,440	114	112	15	62	179
4	2,362	114	113	15	71	188
5	2,004	116	114	15	68	187
6	73	115	113	17	81	170
Total	19,998	119	118	16	62	216

ASL	Insomnia	12	14(11.4%)	6(4.9%)				
		10						
		9						
		8						
		7						
	6							
	Suspect	5	19(15.4%)	3(2.4%)				
		4						
	Normal	3	61(49.6%)	20(16.3%)				
		2						
		1						
0								
		1	2	3	4	5	6	7
		Normal			Insomnia			
		HNC Sleep Level						

Fig. 3. Sleep relation between ASL and HNC

The sleeping level was set as the NHC sleep level. Both the ASL and the HNC are defined by the following Equation (13) and Equation (14).

$$\text{NHC} \stackrel{\text{def}}{=} [1, 7] \quad (13)$$

$$\text{ASL} \stackrel{\text{def}}{=} [0, 10] \quad (14)$$

The number of patients satisfying Equation (15) amount to 123 patients.

$$\text{HNC} \cap \text{ASL} \quad (15)$$

In the ASL scale, scales 4 and 5 are suspected of insomnia. In the HNC, Scale 4-7 is insomnia suspect. Fig. 3 shows the relationship between two sleeping levels. 49.6% of the patients had proper sleep levels for both the ASL and the HNC. 4.9% of patients who both became insomnia in both indices are also partly stressful against nighttime measurement.

As a result, there is no better sleep during the trial.

IV. EXPERIMENTAL RESULT

Table III indicates the experimental summary in SBP from night to morning. Then, potential morning surge patients were extracted. Based on Equation (7) and Equation (8), two patients could be identified as morning surge patients, whose sleep levels are indicated in Table. IV. These patients responded that they were sleeping well by self-assessment, but there is doubt of insomnia.

Fig. 4 and Fig. 5 are indicated time-series transition of blood pressure.

In Fig. 4, This patient was scored 0 in ASL, and the score of Equation (9) was 61. In Fig. 5, This patient was scored 4 in ASL, and the score of Equation (9) was 47.

Thus, by integrating multiple sleeping levels and observation of night blood pressure, it is possible to discover patients with a possibility of morning surge.

Then, clustering was performed only with the transition of nocturnal blood pressure separately from multiple sleep indices. Clustering is a dividing method of a set of classification objects into subsets such that internal cohesion and external isolation are achieved [24].

The clustering method can be classified into a hierarchical method such as the shortest distance method and a non - hierarchical method such as the k-means method. This time, the Ward's method was selected.

This paper, the 103 patients with a score of 19 or higher who replied to the ASL and the HNC were subjected to clustering based on the average of the period.

Both Table V. and Fig. 6. show the SBP blood pressure transition of five clusters.

TABLE III. EXPERIMENTAL SUMMARY IN SBP

Hour	E	2	3	4	5	M
Ave.	119	110	110	110	112	126
S.D.	20	19	19	19	18	19
Min.	79	65	74	71	76	92
Max.	168	148	179	161	161	174
Med.	118	110	110	110	112	125
Mod.	121	110	111	111	112	118
N	1,254	564	564	546	423	1,258

TABLE IV. EXPERIMENTAL SUMMARY IN SBP AND SLEEP LEVELS

Id	E	M	ME Ave.	ME Diff.	NHC	ASL	Eq. (9) Score
(a)	117	160	138	44	1	0	61
(b)	126	135	130	9	1	4	47

TABLE V. AVERAGE SBP BY CLUSTER

Cluster	E	S2	S3	S4	S5	M	N
1	125	112	111	109	111	133	17
2	115	111	111	111	112	119	31
3	111	97	99	99	103	118	17
4	131	119	118	121	121	137	23
5	126	129	132	130	128	128	15

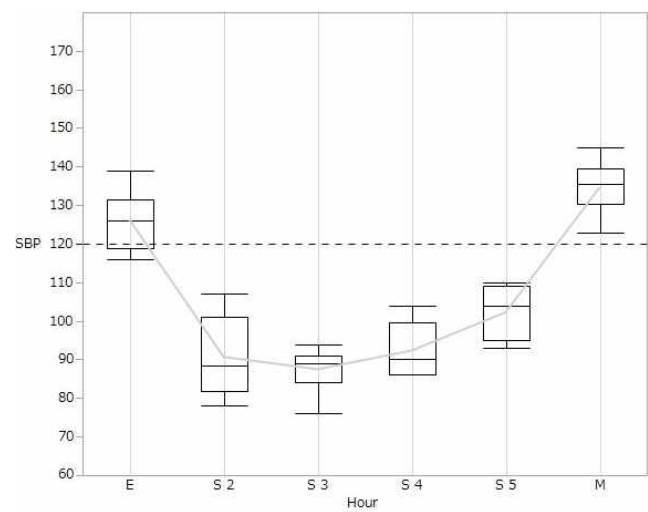


Fig. 4. SBP transition (a)

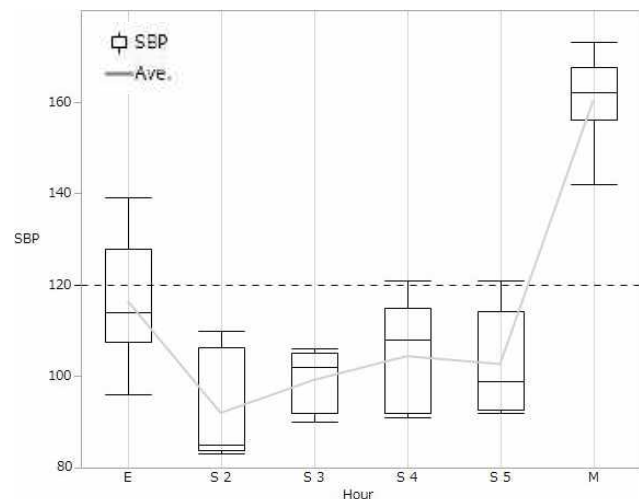


Fig. 5. SBP transition (b)

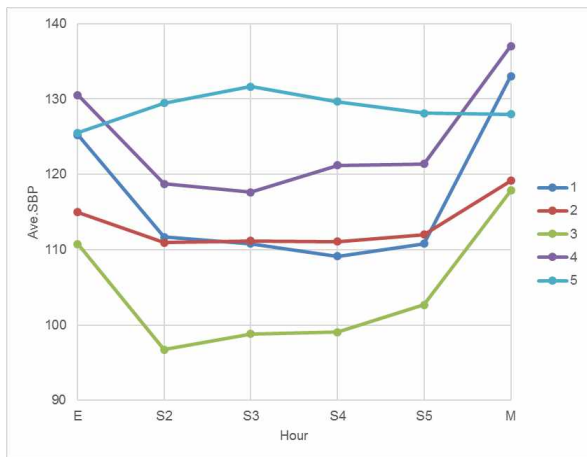


Fig. 6. SBP transitions by cluster

Cluster 1 can be thought of as a morning surge that blood pressure rises early in the morning, and Cluster 5, since their blood pressure did not fall below 120mmHg even once the middle of the night, it can be considered that they could not sleep at that time.

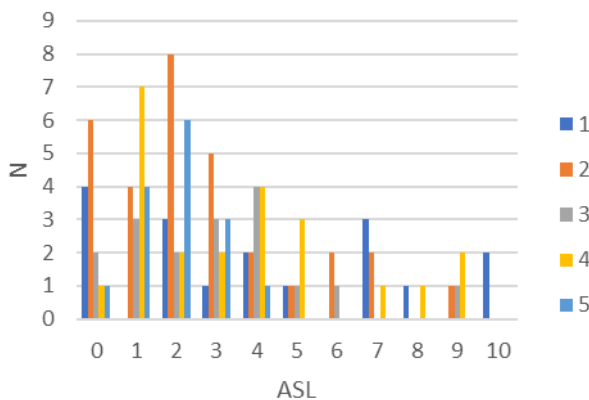


Fig. 7. ASL score by the cluster

Fig. 6 shows the ASL score of each cluster. Among those who were classified into cluster 1, there were patients elevated blood pressure in the early morning even though they were able to sleep from the objective evaluation with the low-scored ASL.

Then, the Distribution of ASL scores for each cluster is considered. Fig. 7 indicates the ASL score distribution for each cluster. In cluster 1, there were 4 patients with an ASL score of 0, No patient with ASL score of 1 and 3 patients with ASL score of 2. When the relationship of the transition of SBP and the ASL score is considered, unconscious insomnia based morning surge is able to be detected.

V. CONCLUDING REMARKS

This paper presented investigating for analyzing blood pressure transition from the clinical records, which aims to classify the patients' blood pressure surge type.

We described research backgrounds, related work, research method, proposed a method and analytical results.

Concerning the screening of suspicious blood pressure surge patients, it depends on the heuristic knowledge of

cardiology doctors, not only the observation of blood pressure. Since to predict the accurate diagnosis of hypertension, both the subjective and objective sleep level was focused not only on the transition of the blood pressure. The analytical results suggested that potential hypertension clusters were extracted. As a result, it is possible to classify the transition patterns. This finding will assist in practical cardiology therapy.

Our future work includes; 1) parameter tuning for the accurate classification, 2) examine another machine learning method such as an xgboost, 3) generate new transition patterns with meta-heuristic algorithms. This work will require practical experiments and further survey studies.

REFERENCES

- [1] K. Shimamoto et al., "The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2014)," *Hypertens. Res.*, vol. 37, no. 4, 2014, pp. 253–253.
- [2] Ministry of Health, "A Basic Direction for Comprehensive Implementation of National Health Promotion," 2012 [Online]. Available: http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryo_u/kenkou/kenkouinippon21/en/kenkouinippon21/index.html. [Accessed: 08/01/2020].
- [3] K. Yatsu, T. Umehura "Management of Secondary Hypertension.," *Heart View*, 2015, pp.350-354.
- [4] A. Fujiyoshi et al., "Blood pressure categories and long-term risk of cardiovascular disease according to age group in Japanese men and women," *Hypertens. Res.*, vol. 35, no. 9, 2012, pp. 947–953.
- [5] "Japan Ambulatory Blood Pressure Prospective Study: JAMP study," UMIN-CTR Clinical Trial. [Online]. Available: https://upload.umin.ac.jp/cgi-openbin/ctr_e/ctr_view.cgi?recptno=R000023397. [Accessed: 05-03-2020]
- [6] K. Kario, "Time for focus on morning hypertension: Pitfall of current antihypertensive medication," *American Journal of Hypertension*, vol. 18, no. 2. 2005, pp. 149–151.
- [7] K. Kario et al., "Morning surge in blood pressure as a predictor of silent and clinical cerebrovascular disease in elderly hypertensives: A prospective study," *Circulation*, vol. 107, no. 10, 2003, pp. 1401–1406.
- [8] P. Gosse, R. Lasserre, C. Minifé, P. Lemetayer and J. Clementy, "Blood pressure surge on rising," *J. Hypertens.*, vol. 22, no. 6, pp. 1113–1118, Jun. 2004, doi: 10.1097/00004872-200406000-00011.
- [9] H. Metoki et al., "Prognostic significance for stroke of a morning pressor surge and a nocturnal blood pressure decline: The Ohasama study," *Hypertension*, vol. 47, no. 2, 2006, pp. 149–154. doi: 10.1161/01.HYP.0000198541.12640.0f.
- [10] "JALS -Japan Arteriosclerosis Longitudinal Study." [Online]. Available: <http://jals.gr.jp/en/index.html>. [Accessed: 05-03-2020]
- [11] A. Amici et al., "Exaggerated morning blood pressure surge and cardiovascular events. A 5-year longitudinal study in normotensive and well-controlled hypertensive elderly," *Arch. Gerontol. Geriatr.*, vol. 49, no. 2, 2009, pp. e105–e109. doi: 10.1016/j.archger.2008.10.003.
- [12] Y. Li et al., "Prognostic value of the morning blood pressure surge in 5645 subjects from 8 populations," *Hypertension*, vol. 55, no. 4, 2010, pp. 1040–1048. doi: 10.1161/HYPERTENSIONAHA.109.137273.
- [13] J.-M. Mallion, J.-P. Baguet, J.-P. Siché, F. Tremel, and R. De Gaudemaris, "Clinical value of ambulatory blood pressure monitoring," *J. Hypertens.*, vol. 17, no. 5, 1999, pp. 585–595. doi: 10.1097/00004872-199917050-00001.
- [14] S. Trazzi, E. Mutti, A. Frattola, B. Imholz, G. Parati, and G. Mancia, "Reproducibility of non-invasive and intra-arterial blood pressure monitoring: Implications for studies on antihypertensive treatment," *J. Hypertens.*, vol. 9, no. 2, 1991, pp. 115–119. doi: 10.1097/00004872-199102000-00003.

- [15] Y. Imai, "Prognostic significance of ambulatory blood pressure.," *Blood Press. Monit.*, vol. 4, no. 5, 1999, pp. 249–256. doi: 10.1097/00126097-199900450-00007.
- [16] E. J. Roccella, "The Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure," *Archives of Internal Medicine*, vol. 157, no. 21, 1997, pp. 2413–2446.
- [17] WHO-ISH, "1999 World Health Organization-International Society of Hypertension Guidelines for the Management of Hypertension. Guidelines Subcommittee," *J Hypertens*, vol. 17, no. 2, 1999, pp. 151–183. doi: 10.1097/00004872-199917020-00001.
- [18] JCS Joint Working Group, "Guidelines for the clinical use of 24 hour ambulatory blood pressure monitoring (ABPM) (JCS 2010)-Digest version," *Circulation Journal*, vol. 76, no. 2, 2012, pp. 508–519.
- [19] I. Kuwashima, "Japanese Multicenter Study on Barnidipine with ABPM," *Japan Medical Journal*, 3973, G1-G6, 2000.
- [20] K. Kario "Clinician's Manual on Early Morning Risk Management in Hypertension," Science Press, 2004, pp. 1-68.
- [21] K. Kario, "Preceding linkage between a morning surge in blood pressure and small artery remodeling: An indicator of prehypertension?," *J. Hypertens.*, vol. 25, no. 8, 2007, pp. 1573–1575. doi: 10.1097/HJH.0b013e32825fea8a
- [22] D. J. Buysse, C. F. Reynolds, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research," *Psychiatry Res.*, vol. 28, no. 2, 1989, pp. 193–213, doi: 10.1016/0165-1781(89)90047-4.
- [23] C. R. Soldatos, D. G. Dikeos, and T. J. Paparrigopoulos, "Athens Insomnia Scale: Validation of an instrument based on ICD-10 criteria," *J. Psychosom. Res.*, vol. 48, no. 6, 2000, pp. 555–560. doi: 10.1016/S0022-3999(00)00095-7.
- [24] S. Everitt Brian, *Cluster analysis*. John Wiley & Sons, 2011.

Machine translation system to convert Sinhala and English Braille documents into voice

K. Sasindu Anuradha*
Faculty of Graduate Studies & Research
Sri Lanka Institute of Information Technology, Sri Lanka
sasindu071@gmail.com

Samantha Thelijjagoda
SLIIT Business School
Sri Lanka Institute of Information Technology, Sri Lanka
samantha.t@slit.lk

Abstract: Reading Braille documents are a time-consuming and labor-intensive task. A blind person should touch every Braille letter by his or her fingers. Therefore, high sensitivity in fingers and memorizing every Braille letters are key factors in Braille reading. Due to enhancement in technology, several Optical Character Recognition (OCR) systems have been introduced for different languages in different parts of the world. However, in Sri Lanka, there are no systems that extract Sinhala or English Braille characters using OCR and convert those Braille codes to sound output. The main purpose of the research is to create a system that extracts both Sinhala and English Braille segments from a given Braille document and makes the Braille content into voice. At the beginning Embossed Sinhala or English Braille image which took from webcam or a high-resolution phone will use for image processing techniques such as gray scaling, thresholding, erosion, and dilation. Erosion and gray scaling help to eliminate the noise and the noise-free image used to detect contour. After pre-processing, the image using an OpenCV library do the segmentation and character extraction. Braille character recognition has done by taking binary to decimal equivalent numbers. Before generating voice output, recognized Braille letters should convert to corresponding language letters (either Sinhala letter or English) and mapping English letters can directly generate English words but in Sinhala need an additional step called Unicode Mapping to generate a Sinhala word. When a Braille document is in high quality, the system will reach for 95 percent accuracy level and generally, results have around 85.30 percent success rate. This software provides many usability characteristics to increase simplicity when it's using OpenCV and related technologies. Even teachers can use this to improve their teaching terminologies and explain things more clearly in their lessons.

Keywords: Braille letters, Image processing, OCR, Segmentation, Unicode mapping

I. INTRODUCTION

Information Technology is everywhere. In the last decade, IT has pernitrate to every corner in society. In this research, the study concentrates on visual impairments and investigates the possibility to support this community with modern information technology while developing an application. Visual impairment is a significant health problem in Worldwide. Blindness is the condition of lacking visual perception due to physiological or neurological factors.

There are a lot of visually impaired people in our community. These people face difficulties with their regular actions such as reading, walking, driving, socializing and writing. In this modern world, blind people spend more time when they are reading Braille documents by hand and

especially for researchers it is a huge disadvantage. This application uses image processing algorithms to take the Braille document (.JPG, .png or.Pdf) as input and Convert those embedded Braille letters to voice would be a great help to them in their education as well as day to day lives.

II. LITERATURE REVIEW

A French educator named Louis Braille developed a system for blind people to read and write which was known as the Braille system. Braille Coding System is a method that is widely used by visually impaired people to read and write. This system uses patterns of raised dots to make the characters on paper, helping visually impaired people read and write with single touch instead of vision. It is a way for blind people to actively participate in social activities and literature. Braille system was first developed in the nineteenth century. Its characters are six-dot cells, with two columns and 3 rows. Any of the dots may be raised above the paper plain thus giving 64 possible characters. Although Braille cells are used worldwide, the meaning of each of the 64 cells depends on the language that they are being used to depict. The concept of Braille was universally adopted [1].

There is an article called "Braille Translation" written by 3 authors, Farhan Bodale, Uddhav Bhide, and Dilip Gore [2]. Braille Characters are extracted from the given text message. This extracted Braille Character is compared to entries and the system returns the related patterns. However, visually impaired people a little bit hard to use the suggested system as it is. It is necessary to convert the Braille code in any local language so that blind people can hear the audio chaos without the effort of reading with touch sensations. In the article entitled "Development of a Bangla Character Recognition (BCR) System for Generation of Bengali Text from Braille Notation" [3], the development of Bangla letters of the Braille code is achieved. The characters go through a translation engine that consists of a suitable processing algorithm to convert into Bangla Character and any local language extraction of Braille script can be possible. According to the language rule table, the input Braille segment will convert into Bangla characters. The concept of translation rule table has been described by Alasdair King in his paper titled "Text and Braille Computer Translation". Considering these pre-existing methods, the paper called "Braille Code Conversion to Voice in Malayalam" by Bindhu K Rajan and Anjitha V [4] describes the Malayalam language extraction from Braille characters and then the message could be read as audio using proper conversion from text to speech. Their database is made for the standard representation of alphabets. A scanned image will be given

as system input and it undergoes several steps such as image enhancement, Image Filtering, Image segmentation, Braille character extraction and finally Translation.

The research paper "Braille Character Recognition Using Find Contour and Artificial Neural Network by Joko Subur, Tri Arief Sardjono and Ronny Mardiyanto uses Open Cv library for image processing and artificial neural network to identify braille letters. This system supports only the English language and it generates plain text from the given English Braille document [5]. The paper "Optical Braille Translator for Sinhala Braille System: Paper Communication Tool Between Vision Impaired and Sighted and Sighted Persons" by Talagalage Don Sanka Harshana Perera and W.K.I.L Wanniarachchi again use OpenCV library to image processing and binary equal decimal value mechanism to generate text output from given braille document. This system supports Sinhala language and directly convert Braille letters to Sinhala letters [6]. In this research analysis, there is no preprocessing technique to generate words from Sinhala letters and as a result of this output, the text is hard to understand. "Sinhala Braille Translator" by N.M.T De Silva and S.R. Liyanage is another research analysis that uses OpenCV library to image processing [7]. Like previous research, they use binary equal decimal value mechanisms to generate text output from a given braille document and use a preprocessing technique called "Normalization" before generating words from letters. As a result of this additional step output text which is generated by giving Sinhala, Braille image is more meaningful.

III. METHODOLOGY

The system consists of mainly 3 main steps. In the beginning, Braille image should be processed in the way it is convenient to read and it is done by using methods in OpenCV library, which is open source and freely available on their official web site. Then the image is read by the system and uses several mathematical formulas to extract and identification of Braille characters. The final step is generating audio input using extracted Braille letters. It has several sub-steps such as match six points composed Braille

letters to corresponding Sinhala or English letters and generates words. English words can directly map to sentences, but in Sinhala need an additional step called normalization before generating Sinhala sentences. In English, there are several libraries to convert English sentences to voice, but in Sinhala needs a manual process to voice generation.

A. Capturing Braille images

There are several ways to capture a Braille image and this system highly recommended a scanned image.

- Scan a Braille image
- Use Webcam
- A high-resolution image which is taken by mobile phone

When using a web camera to better accuracy in Braille reorganization needs 650 pixels by 500 images with the distance between web camera and the Braille image should be +/- 15cm. When taking images using a mobile phone, the phone should be flatted and the distance between the image and the camera should be less than 6cm.

B. Image Processing

With the original Braille image, it is hard to identify Braille letters so there are compulsory image processing steps. The main idea underneath image processing is a clear recognition of Braille letters for the next process [5].

C. Image processing using OpenCV

OpenCV is called an Open Source Computer Vision library feature primarily aimed toward a real-time PC vision. The library is cross-platform and free to be used underneath the ASCII text file BSD license. It is initially written in C++ programming language and its primary interface is in C++, however, it still retains a less comprehensive although in-depth older C interface. Currently, it supports Python, Java and Matlab/Octave. The system uses OpenCV library methods for mainly Object Identification, Segmentation, and recognition [8].

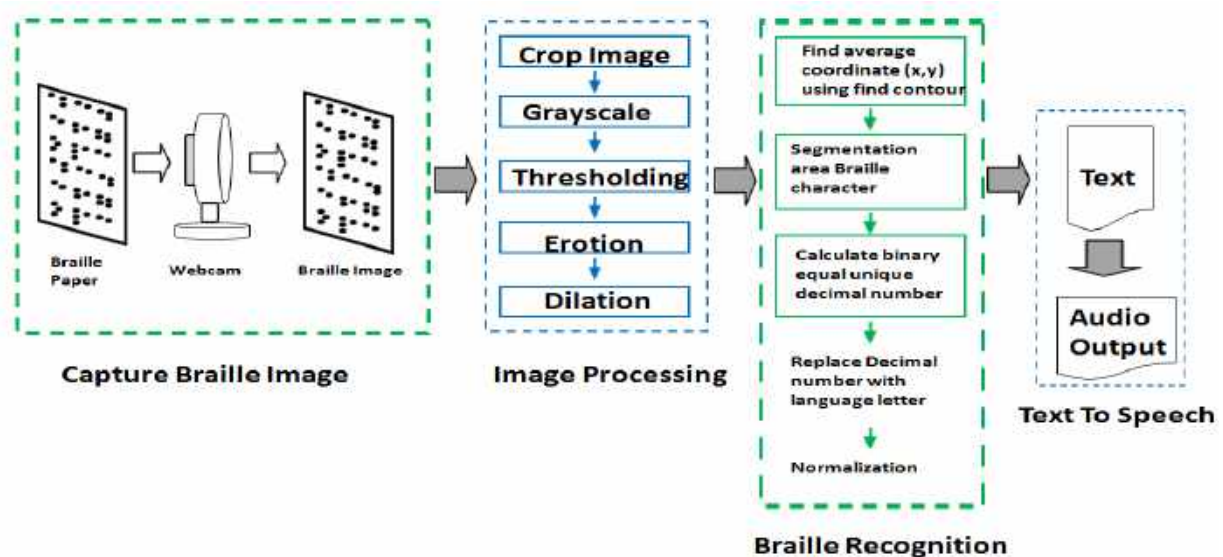


Fig.1.System overview

D. Crop the given Braille image

To achieve better accuracy in Braille reorganization, we need to remove the edges which are not wanted. The entire process is done manually and OpenCV provides below Java method (command) to image cropping [5].

```
Rect rectCrop = new Rect (cropCenterX,  
cropCenterY, cropWidth, cropHeight);
```

CropCenterX and cropCenterY arguments determined the initial point x and y coordinates and cropWidth and cropHeight suggest the cropped image width and height.

E. Gray scaling the image

The original Braille image contains an RGB color scale and the main purpose of this step is to convert the original image RGB color scale into Grayscale images. A grayscale picture is kind of a black and white or gray monochrome; it contains a variety of shades of gray. OpenCV method for the Gray scaling mention bellow and it contains 3 parameters [5].

```
Imgproc.cvtColor(originalImage,imgGrayscale,  
Imgproc.COLOR_BGR2GRAY);
```

The original image represents the giving Braille image and imgGrayscale is the equivalent grayscale image for original image. Imgproc.COLOR_BGR2GRAY is the static attribute that gives a command for OpenCV for gray scaling.

F. Thresholding the image

This process helps to convert grayscale image to binary image; an image that has the color range of 0-255. Before segmentation thresholding should be done [5].

```
threshold(src, dst, thresh, maxval, type)
```

The original method signature looks like above and it has five parameters. Parameter src represents a math object of an input image and dst represents the result, math image object after the method execution. The thresh variable contains a double value of the threshold and maxval is a double variable that holds the value that to be replaced when the pixel value exceeds the threshold value. The final parameter is an integer character describing the kind of threshold to be practiced.

G. Erosion

In this analysis, the erosion method is applied to reduce white area dots and eliminates noise from the Braille picture that emerged from the thresholding method [5].

```
erode(src, dst, kernel)
```

Similar to previous image processing methods, src denote to a math object of an input image, and dst denote the result math image object after the method execution. The parameter kernel represents a Mat object of the kernel.

H. Dilation

According to this research, the dilation process applied to expand white dots on image Braille.



Fig. 2. Original Scanned Image



Fig. 3. Output of grayscale method.



Fig. 4. Output of thresholding method



Fig. 5. Output of erosion method. (Noise removal)



Fig. 6. Output of dilation method. (large Braille dots)

```
dilate (src, dst, kernel)
```

The dilate method signature consists of 3 parameters and src and dst represent the input Mat image object and output Mat image object. The parameter kernel represents a Mat object of the kernel. Below show the results images of image processing steps.

I. Braille reorganization

The next major step is locating each and every center coordinates of Braille dots. It also has several steps. Normally Braille dot covers a circular area and OpenCV provides a method to find position coordinates of each dot.

findContours(Mat image, List<MatOfPoint> contours, Mat hierarchy, int mode, int method, Point offset)

Above method uses the algorithm called CITE: Suzuki85 for retrieves contours from a binary image. The contours are a valuable tool for shape review and object detection and identification. FindContours method has 6 parameters; it is a static method in Imgproc class. Image represents the output image of image processing step, contours is a list of MatOfPoint which saved detected contours as points of vectors, hierarchy includes information regarding the image topology, mode determines the Contour retrieval tone and, in this research, uses CHAIN_APPROX_SIMPLE as retrieval tone and finally offset is an optional parameter which helpful if the contours are obtained from the image ROI.

J. Mechanism of findContours method

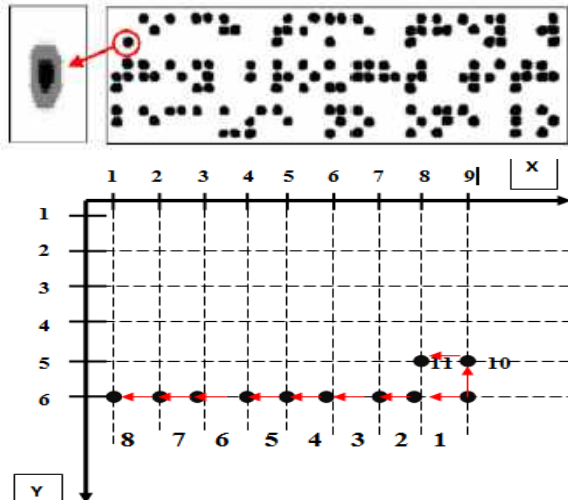


Fig. 7. Dot identification by findContours method

FindContours method identifies Braille dots from the rightest bottom corner and traverses through the x-axis up to the top in the way shown in figure x. Every black dot point consists of several members and those members can be identified from this process. All members have unique X and Y coordinates and it will help to find the center point of Braille dots. If members of a Braille point printed out it would look like bellow.

- 1) X ->: 180.0 Y -> 276.0
- 2) X ->180.0 Y ->277.0
- 3) X -> 181.0 Y -> 277.0
- 4) X -> 181.0 Y -> 276.0

Member count: 4

K. Center X and Y Coordinates of a Braille Dot

By analyzing the above information, it can be recognized that the white dot is one point there were four or three members, or it can be concluded that there are four to three white pixels. Therefore, the need to calculate average XY coordinates of every member of a white dot to determine the center point of a Braille dot.

After the pixel coordinate data of every member of the white dot concerned need to find minimum X and Y coordinates as well as maximum X and Y coordinates. For that following formulas can be used. [5]

$$\text{Center X} = \frac{(\text{Minimum X} + \text{Maximum X})}{2} \quad \text{01}$$

$$\text{Center Y} = \frac{(\text{Minimum Y} + \text{Maximum Y})}{2} \quad \text{02}$$

L. Find the average middle of rows and columns

After finding the center XY coordinate values of all the Braille dots next step is determining the average middle values in each and every column and row. The main reason for calculating the average middle values of each column and row is column to column and row to row X and Y coordinate values change significantly. Therefore, results have better accuracy in terms of Braille identification. There is a mathematical formula to calculate average X and Y coordinates and it mentioned bellow.

AvgXk stands for average x coordinates for each column, Centerxn denotes for summation of all centers X coordinates for each column, k denotes for column number and n is the number of dots in a column. Therefore, the 3rd equation obtained average X coordinates in each column.

$$\text{AvgXk} = \sum_{k=1}^k \sum_{n=1}^n \frac{\text{AvgXk} + \text{CenterXn}}{n} \quad \text{03}$$

$$\text{AvgYb} = \sum_{b=1}^b \sum_{n=1}^n \frac{\text{AvgYb} + \text{CenterYn}}{n} \quad \text{04}$$

AvgYb stands for average y coordinates for each row, CenterYndenotes for summation of all centers Y coordinates for each row, b denotes for the row number and n is the number of dots in a row. Therefore, the equation number 4 obtained average Y coordinates in each row. [5]



Fig. 8. Average column indexes

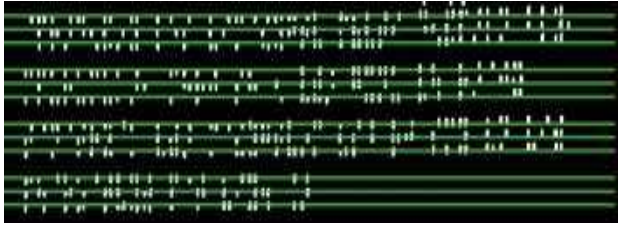


Fig. 9. Average row indexes

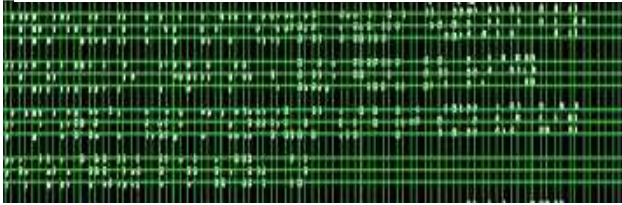


Fig. 10. Average row and column indexes

M. Braille character segmentation

The segmentation scenario is the most significant step in this research project because Braille characters have 2 columns and 3 rows. When determining the area of Braille of 2 columns and 3 rows based on average X and Y coordinates the Segmentation process comes into the action. There is a mathematical equation for the Braille characters segmentation process, and it is mentioned below. [5]

$$S = \sum_{b=1}^b \sum_{k=1}^k [t = t + 1] \left(\sum_{h=b}^{b+3} \sum_{d=k}^{k+2} ([avgXd][avgYh]) \right)$$

— 05

The letter S determined the area of each Braille letter and b and k stands for row index and column index, respectively. The letter t stands for the value counter of a Braille figure, the letter h represents the counter of 3 rows in every Braille sign and d means 2 column counters. The variables avgX and avgY denotes for average X coordinate value and average Y coordinate value respectively. Fig 11 highlights the outcome of the segmentation method and the segmented region of one Braille letter can already be labeled with a blue border. After the segmentation method, every segment area managed to do a letter identification process.

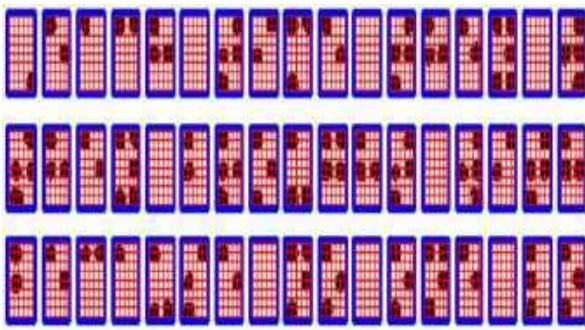


Fig. 11. Braille letter segmentation

N. Braille character identification

The next major step is Braille dot identification in the segmentation areas and consists of several steps. In this research analysis, the average X and Y coordinates are considered and need further Braille character verification process as an additional step. According to Fig 12 Braille dot not situated in the exact average X and Y coordinate value. It may shift to 1-pixel right, left, top or bottom and need to check whether adjacent pixels have a white color pixel. According to this research check one pixel in major 4 directions (top, bottom, left and right) whether it contains RGB value of the white color [9].

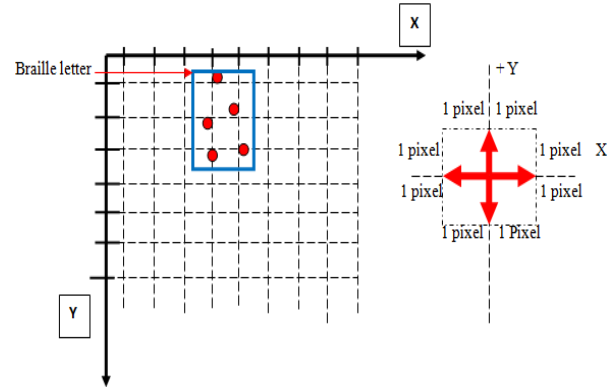


Fig. 12. Pixel Loop

There is a method in OpenCV to get the RGB value of a pixel which is in given X and Y coordinates and it looks like below.

getRGB(int x, int y)

The letter x and y denote the average x and y coordinates respectively, and it returns 32 bits integer value. Normally pixel is considered as the smallest unit in an image. Braille dots appear in white and use pixel loop to check whether the adjacent pixel has white color RGB values or not. If it has a white color RGB values, it registered as a Braille dot otherwise omit it.

O. Calculate Binary equal decimal values

A Binary letter has 6 points and the way those points presences or absent generate a range of mixtures. Every Braille letter has a unique 6-bit binary number and after getting the 6-bit binary number, then convert it to a decimal number. For that, there is an arithmetic operation to generate the binary equal decimal number. According to fig 13 6-point notation for single Braille, the letter can create 63 different combinations and all most all the languages' alphabets in the world have less than 63 letters.

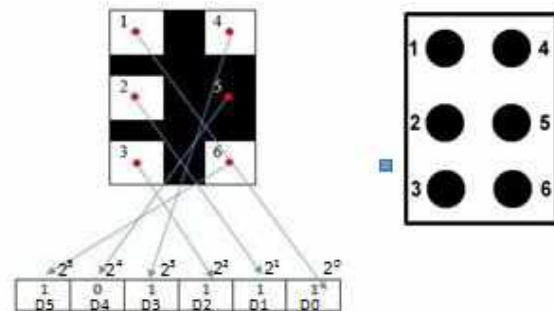


Fig. 13. Braille letter equals 6-bit Binary value

$$\begin{aligned}
 \text{Decimal number} &= (D0*20) + (D1*21) + (D0*22) + \\
 &\quad (D0*23) + (D0*24) + (D0*25) \\
 &= (D0*1) + (D0*2) + (D0*4) + \\
 &\quad (D0*8) + (D0*16) + (D0*32) \\
 &= (1*1) + (1*2) + (1*4) + (1*8) + \\
 &\quad (0*16) + (1*32) \\
 &= 1 + 2 + 4 + 8 + 0 + 32 \\
 &= 32
 \end{aligned}$$

Therefore 6 points for the Braille letter mechanism can apply all the languages in the world and it has universally accepted. In this research project mainly focus on English and Sinhala Braille reorganization [6].

TABLE I. NUMERIC BINARY EQUAL DECIMAL VALUES

Braille Notation	Binary Equal Decimal Value	Number
⠠	01	01
⠡	03	02
⠢	09	03
⠣	25	04
⠤	17	05
⠥	11	06
⠦	27	07
⠧	19	08
⠨	10	09
⠩	26	00

TABLE II. SINHALA CONSTANT LETTERS BINARY EQUAL DECIMAL VALUES

Braille Notation	Binary Equal Decimal Value	Sinhala Letter
⠠	01	අ
⠡	28	ආ
⠢	55	ඇ
⠣	59	ඈ
⠤	10	ඉ
⠥	20	ඊ
⠦	37	උ
⠧	51	ඌ
⠨	17	එ
⠩	34	ඒ

Braille numbers and some English Braille letters have the same Braille symbols and there is another mechanism to distinguish English Braille and numeric Braille. Braille has a special indicator for numbers, it shows in fig 14. When it is in the Braille document it means that on that point onwards all letters have numeric values until finding a space or letter indicator symbols which show in fig 15 [10].

TABLE III. ENGLISH BINARY EQUAL DECIMAL VALUES

Braille Notation	Binary Equal Decimal Value	Number
⠠	01	A
⠡	03	B
⠢	09	C
⠣	25	D
⠤	17	E
⠥	11	F
⠦	27	G
⠧	19	H
⠨	10	I
⠩	26	J
⠪	05	K
⠬	07	L
⠮	13	M
⠰	29	N
⠲	21	O
⠴	15	P
⠶	31	Q
⠸	23	R
⠺	14	S
⠼	30	T
⠾	37	U
⠿	39	V
⠻	58	W
⠽	45	X
⠿	61	Y
⠻	53	Z



Fig. 14. Numeric indicator



Fig. 15. Letter indicator

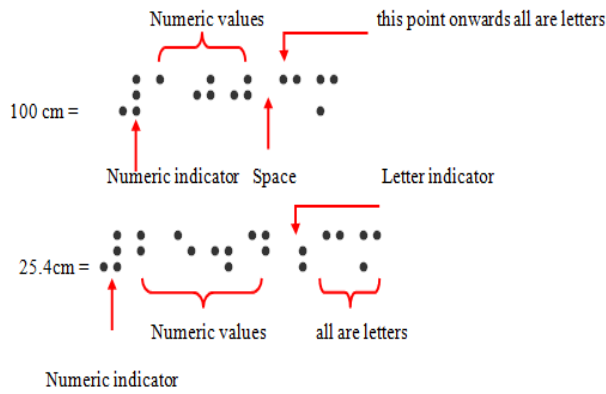


Fig. 16. Example

P. Sinhala text normalization

This step only needs for when converting Sinhala Braille document to voice. For English Braille does not need normalization because from decimal values can create words, but in Sinhala after generating Sinhala letters from decimal values need an additional step to create words. Eg1 shows a text before normalization and can't get a proper meaning at once. Eg2 shows a text after normalization and it can easily understand. In practically Eg1 shows the text a blind person initially identifies when reading a braille document and from his brain, he normalizes it into a more readable format which shows in Eg2. Therefore, normalization is a compulsory step in this research analysis and needs to maintain normalization tables for normalizing [7].

TABLE IV. SINHALA TEXT NORMALIZATION

Letter 01	Letter 02	Normalized Letter
ක	ආ	කා
ක	ඇ	කැ
ක	ඈ	කෑ
ක	ඉ	කී
ක	ඊ	කී
ක	උ	කු
ක	ඌ	කු
ක	එ	කෙ
ක	ඒ	කේ
ක	ඔ	කො
ක	ඌ	කෝ

ක	ඔ	කො
ක	ක	කේ
ක	උ	කා
ක	ඌ	කා

Normal text:

Eg1: පර ආද ඒ ශ ර් ය ල එස ව්‍යුමධයගන ව
පඤ්චන ඉ ද ඒ ශ ප ආ ල න බල ය

After normalization

Eg2: ප්‍රාදේශීය ලෙස විමධයගනව පැවති
දේශපාලන බලය

Q. Sound outputs

For English Braille, there are open source libraries to convert English text to voice and in this research analysis, the library called Freetts used for text to voice translation. For Sinhala it is a manual process and similar to fig 17 all audio outputs save in one folder. All the paths of the sound clips save in a HashMap and when converting Sinhala text to sound read one letter at a time and match with Hashmap. Then find the sound clip address corresponding to Sinhala letter and play it.

```
public static HashMap<String, String> sinhalaSounds = new HashMap<String, String>();
sinhalaSounds.put("ක", "src/main/resources/static/audios/ki.mp3");
sinhalaSounds.put("ආ", "src/main/resources/static/audios/kiu.mp3");
sinhalaSounds.put("ඇ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඈ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඉ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඊ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("උ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඌ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("එ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඒ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඔ", "src/main/resources/static/audios/a.mp3");
sinhalaSounds.put("ඌ", "src/main/resources/static/audios/a.mp3");
```

Fig. 17. Sound clips and their path in a HashMap

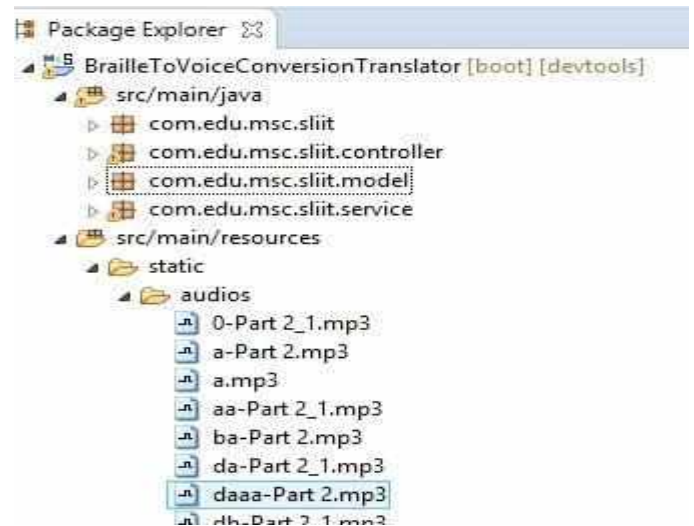


Fig. 18. Sound folder

IV. RESULTS AND DISCUSSION

The system has one single interface and the project is deployed as a web application. The outcomes of the system can be explained using an illustration and fig 19 shows the only interface of the system. There is a Choose File option to upload a Braille image and uploaded a Braille image display like in fig 20. The next step is to scan the image and when the client clicks the Scan Image button system generates a text document corresponding to the Braille image as an intermediate outcome.

Binary equal decimal values

[56, 50, 0, 0, 25, 55, 29, 62, 0, 19, 13, 37, 0, 39, 20, 0, 30, 10, 3, 17, 29, 0, 14, 28, 46, 5, 0, 1, 29, 37, 39, 0, 13, 17, 23, 62, 0, 25, 10, 61, 37, 53, 37, 0]

[5, 16, 23, 41, 10, 5, 28, 8, 23, 13, 10, 5, 0, 14, 13, 28, 26, 61, 5, 0, 28, 23, 8, 13, 24, 61, 0, 1, 39, 13, 0, 39, 47, 61, 17, 8, 29, 0, 8, 5, 23, 10, 50, 0]

[15, 51, 50, 0, 60, 3, 25, 26, 26, 0, 15, 13, 53, 39, 8, 30, 0, 59, 30, 62, 0, 25, 10, 39, 17, 61, 10, 50, 0, 17, 13, 0, 61, 37, 27, 61, 34, 0, 14, 10, 62, 0, 13, 0]

Replacing those decimal values with letters

[ඒ, ., ., ., ද, ඇ න, ට, , හ, ම, උ, , ව, ඊ, , න, ඉ, බ, එ, න, , ස, ආ, ධ, ක, , අ, න, උ, ව, , ම, එ, ර, ට, , ද, ඉ, ය, උ, ණ, උ,][ක, 16, ර, ෂ, ඉ, ක, ආ, p, ර, ම, ඉ, ක, , ස, ම, ආ, ජ, ය, ක, , ආ, ර, p, ම, හ, ය, , අ, ව, ම, , ව, ශ, ය, එ, p, න, , p, ක, ර, ඉ, .,][ප, උ, ., , null, 2, 4, 0, 0, , ප, ම, ණ, ව, p, න, , ඇ න, ට, , ද, ඉ, ව, එ, ය, ඉ, ., , එ, ම, , ය, උ, ග, ය, ඒ, , ස, ඉ, ට, , ම,]

Normalization

දැනට හමු වී
නිබෙන සාධක
අනුව දමරට
දියුණු
කෘෂිකාර්මික
සමාජයක
ආරම්භය අවම
වශයෙන් කිරී.
යු. 2400
පමණවත්
අතට දිවයි.
එම යුගයේ සිට
ම

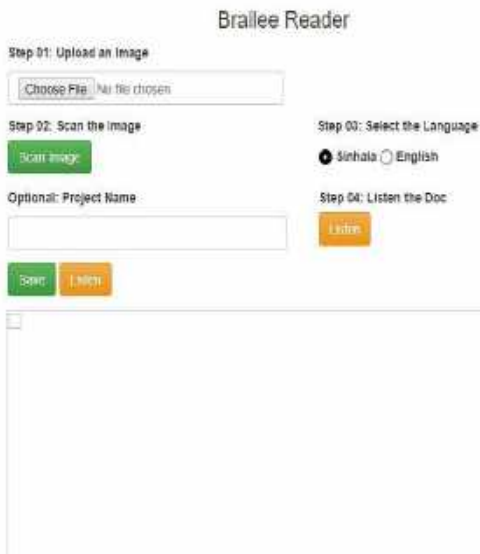


Fig. 19. System interface



Fig. 20. Uploaded Braille Image



Fig. 21. Original Braille Image



Fig. 22. Output of Image Processing

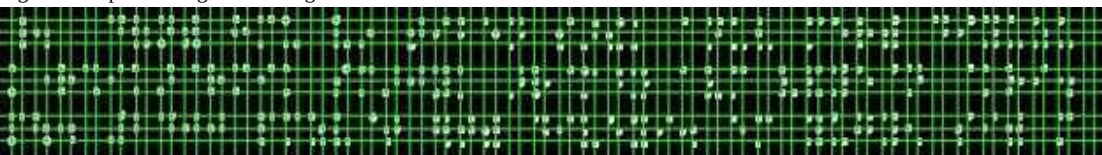


Fig. 23. Find Average X and Y Coordinates.

Results which are generated by the system can be classified as below. It was subjected to identification of 6 braille documents and 3 of them are Sinhala and the rest of them is English. Each Braille paper has a standard size of 8.5" x 11 printed on a single side. Three different English short stories have been embossed in lines on three English Braille papers and the three Sinhala Braille papers have been embossed with 3 pages in Grade 11 Social Studies book. The Sinhala letters such as “ඌ, ඹ, ඹ, ඹ, ඡ, ඤ, ධ, ඵ, ද, ට” seldom appeared in Sinhala documents and thus such letters could not be captured in test samples.

TABLE V. LETTERS IDENTIFICATION

Lett er	No of occurrence	Correct identification	Accuracy Percentage
අ	35	29	82.85
ආ	90	81	90.0
ඇ	41	37	90.24
ඈ	08	06	75.0
ඉ	162	140	86.41
ඊ	20	16	80.0
උ	76	68	89.47
ඌ	08	07	87.5
ඍ	78	66	84.61
ඎ	41	36	87.80
ඏ	0	0	0.0
ඐ	27	23	85.18
එ	10	08	80.0
ඒ	0	0	0.0
ක	89	80	89.88
ඛ	5	04	80.0
ග	42	35	83.33
ඝ	2	01	50
ඞ	0	0	0.0
ඟ	0	0	0.0
ච	1	1	100.0
ඡ	0	0	0.0
ඣ	7	6	85.71
ඤ	1	1	100.0
ඦ	0	0	0.0
ට	42	35	83.33

ඵ	0	0	0.0
බ	16	13	81.25
භ	0	0	0.0
ඈ	14	12	85.71
ඉ	42	37	88.09
ඊ	0	0	0.0
උ	64	57	89.06
ඌ	10	8	80.00
ඍ	152	135	88.81
ඎ	0	0	0.0
ඏ	34	26	76.47
ඐ	01	01	100.0
එ	25	20	80.0
ඒ	02	01	50.0
උ	102	79	77.45
ඌ	95	84	88.42
ඍ	78	69	88.46
ඎ	53	47	88.67
ඏ	106	88	83.01
ඐ	19	14	73.68
එ	04	03	75.0
ඒ	75	66	88.0
උ	56	51	91.07
ඌ	08	06	75.0
ඍ	0	0	0.0
A	145	138	95.17
B	23	20	86.95
C	39	32	82.05
D	79	70	88.6
E	245	201	82.04
F	33	29	87.84
G	32	29	90.62
H	125	103	82.4
I	87	75	86.20
J	07	06	85.71
K	20	16	80.0
L	60	47	78.33

M	36	32	88.88
N	104	83	79.80
O	105	85	80.95
P	24	20	83.33
Q	02	02	100.0
R	95	82	86.31
S	107	85	79.43
T	162	135	83.33
U	43	35	81.39
V	18	14	77.77
W	42	35	83.33
X	04	03	75.0
Y	42	35	83.33
Z	02	01	50.0

TABLE VI. NUMERIC IDENTIFICATION

Number	No of occurrence	Correct identification	Accuracy
01	3	3	100.0
02	1	1	100.0
03	3	3	100.0
04	0	0	0.0
05	0	0	0.0
07	2	1	50.0
08	2	1	50.0
09	3	2	66.66

Convert Sinhala and English Brille document is a new research area and very little research has conducted so far. The main technologies used to detect Braille letters are open source as well as inexpensive. The system is tested by different sizes and different qualities of Sinhala and English braille documents and scored more than 85 percent accuracy. Conversion of identified braille letters to corresponding voice is straightforward and in Sinhala play letter by letter but in English, it plays words at once. The quality of the image is the most significant factor because it affects Braille identification and when the image is in high-quality accuracy will reach up to 95 percent. If the braille document is old and heavily used it would reduce the accuracy and the system accept single side braille printed documents.

However, hardware components utilized by the system is less and no need for expensive and complicated devices. The system runs on top of any platform and tested in different browsers and different operating systems.

V. CONCLUSION

This system is capable of extracting Braille letters from Braille image and converts it into the corresponding language voice. Initially, the process of extracting Braille letters is associated with several image processing steps such as Crop, Grayscale, Thresholding, Erosion, Dilation, and segmentation. Then extracted Braille letters are decoded into Sinhala characters and then through the normalization process reorder the Sinhala characters into legible Sinhala text. Finally, legible Sinhala text is converted into sounds like the final outcome. In the case of the English Braille image, the normalization process is not applied and directly converts to legible English text.

The strategy is a successful, cost-effective, quick and straightforward method to recognize dots in Braille papers. It does not require costly or complex hardware. The current system gains around 85.30 percent accuracy level in both languages and even teachers can use this to improve their teaching terminologies and explain things more clearly in their lessons.

REFERENCES

- [1] Braille Alphabet, [online] Available at: <http://www.who.int/countries/lka/en>. [Accesses: 02.06.2019].
- [2] Farhan Bodale, Uddhav Bhide, Dilip Gore, "Braille Translation". International Journal of Research in Advent Technology, E-ISSN:2321- 9637 Vol.2, No.4, April 2014.
- [3] Santanu Halder, Abul Hasnat, Amina Khatun, Debotosh Bhattacharjee, Mita Nasipuri "Development of a Bangla Character Recognition (BCR) System for Generation of Bengali Text From Braille Notation". International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol- 3, Issue-1, June 2013.
- [4] Bindhu K. Rajan and Anjitha V., "Braille Code Conversion to Voice in Malayalam", International Conference on Communication and Signal Processing, April 6-8, 2017, India
- [5] Joko Subur, Tri Arief Sardjono, Ronny Mardiyanto "Braille Character Recognition Using Find Contour and Artificial Neural Network", JAVA, International Journal of Electrical and Electronics Engineering 19, Vol.14, Number 1, April 2016.
- [6] T. D. S. H. Perera and W. K. I. L. Wanniarachchi "Optical Braille Translator for Sinhala Braille System: Paper Communication Tool between Vision Impaired and Sighted Person", The International Journal of Multimedia & Its Applications (IJMA), Vol.10, No.1/2/3, June 2018.
- [7] N.M.T De Silva and S.R Liyanage, "Sinhala Braille Translator", International Journal of Trend in Research and Development, Vol.3(4), ISSN 2394-9333www.ijtrd.com.
- [8] Open CV Documentation, [online] Available at: <https://docs.opencv.org/2.4.13.7/modules/core/doc/intro.html>. [Accesses: 02.06.2019]
- [9] How to get and set pixel value in Java, [online] Available at: <https://www.dyclassroom.com/image-processing-project/how-to-get-and-set-pixel-value-in-java> [Accesses: 08.06.2019]
- [10] The Braille Alphabet, [online] Available at: <https://www.pharmabraille.com/pharmaceutical-braille/the-braille-alphabet> [Accesses: 09.06.2019].

Grammatical error detection and correction model for Sinhala language sentences

H. M. U. Pabasara*

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
pabaup@gmail.com

S. Jayalal

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
shantha@kln.ac.lk

Abstract: As the national language of Sri Lanka, the greater part of the exercises at most of all the services are completed in Sinhala whereas it is imperative to guarantee the spelling and syntactic accuracy to convey the ideal significance from the perspective of automated materials with the unavailability of resources even though there are enough amount of available materials as hard copy and books. With the high multifaceted nature of the language, it sets aside extensive effort to physically edit the substance of a composed setting. The necessity to overcome this problem has risen numerous years back. But with the complexity of grammar rules in morphologically lavish Sinhala language, the accuracy of the grammar checkers developed so far has been contrastingly lower and thus, to overcome the issue a novel hybrid approach has been introduced. Spell checked Sinhala active sentences being pre-processed, separated nouns and verbs were analyzed with the help of a resourceful part-of- speech-tagger and a morphological analyzer and alongside the sentences were sent through a pattern recognition mechanism to identify its sentence pattern. Then a decision tree-based algorithm has been used to evaluate the verb with the “subject” and output feedback about the correctness of the sentence. To train this decision tree, a dataset consisting of 800 records which included information about 25 predefined grammar rules in Sinhala was used. Finally, the error correction was provided using a machine learning algorithm-based sentence guessing model for the three possible tenses. Conducted research results paved the way to identify the sentence pattern, grammar rules and finally, suggest corrections for identified incorrect grammatical sentences with an acceptable accuracy rate of 88.6 percent which concluded that the proposed hybrid approach was an accurate approach for detecting and correcting grammatical mistakes in Sinhala text.

Keywords: Grammar checking, Hybrid approach, Part-of- speech-tagger

I. INTRODUCTION

Sinhala is used as the first language by approximately 16 million Sinhalese people in Sri Lanka [1] [2], where it is one of the national and official languages, alongside the Tamil language. Communication and correspondence in a majority of state departments, private institutions and other offices are conducted mainly in the Sinhala language to cater to the needs of the end user. A large section of the general population has no clear idea on how to utilize the language accurately [3]. It is essential to utilize the language accurately to safeguard the planned significance where the normal significance can be distinctive because of the inaccuracy of grammar. When thinking about the composed setting, Sinhala is an intricate language that contains many spelling and syntax rules where the rightness of the formal composing thoroughly relies upon

these well-characterized rules [4]. It is imperative to guarantee the spelling and syntactic accuracy to convey the ideal significance from the perspective of automated materials with the unavailability of resources even though there are enough amount of available materials as hard copy and books for the Sinhala language. On the other hand, with the high multifaceted nature of the language, it sets aside extensive effort to physically edit the substance of a composed setting. The necessity of an automated system to play out this assignment has risen for the Sinhala language numerous years back. Well-created grammar checking applications are accessible for dialects like English, Tamil and Chinese and plenty of many other languages [5]. But a morphologically rich language like Sinhala lacks precise grammar checking tools mainly because of the unavailability of enough resources [6]. Natural languages like Sinhala are not constrained by a specific syntax and have a wide range of vocabulary; hence, Sinhala grammar checkers must also have an extensive dictionary of most words with their complete meanings and part-of-speech usage [7].

Spell checking, on the other hand, is also one of the key areas to be considered when developing a grammar checker for a resource-poor language. Out of the many types of research that have been carried out in the context, the data-driven open-source approach proposed by Wasala et al. [8] [9], which claims to be able to test and correct spelling errors in the Sinhala sentences is one of the popular. Further attempts in the context like the approach proposed by Jayalatharachchi et al. [10] to obtain an alliance between two algorithms [8] [10] have been then further extended by Subhagya et al. [11]. This implies that fair focus has been given for the development of a proper spell checker for the Sinhala language. Construction of relatively inexpensive spell checkers without deep linguistic knowledge for the Sinhala language has been carried out in different dialects with several approaches like n-gram statistics, data-driven approaches while considering the morphological richness of the language and unavailability of enough digital footprint for the language and thus still being able to identify and correct many of the language's common spelling errors with results showing promising output at a reasonable precision rate [8-11]. Availability of such accurate spell checkers for Sinhala language and to shrink the scope of this research study to a feasible level while considering the complexity of the Sinhala grammar rules and sentence structures, this study was conducted to address the knowledge gap in detecting and correcting grammatical mistakes in already spell checked Sinhala written text with the help of possibly available and publicly accessible useful resources while providing a novel

hybrid approach aligned with the machine learning algorithms.

II. BACKGROUND

A. Sinhala language

Sinhala is a language belonging to the globe-spanning language tree, Indo-Aryan and is the native language to approximately 16 million Sinhalese people in Sri Lanka [1] [2]. Also, the language is being used as an auxiliary language approximately by another 4 million people belonging to other ethnic groups living in Sri Lanka [12], where it is one of the national and official languages, along with Tamil. There is also a considerable composition of Sinhala speakers, writers, authors and people who are using the language for many other purposes all over the world. Nevertheless, the closest relative languages to Sinhala existing in the world currently are the Maldives and Dhivehi [13]. The writing system of the Sinhala language is a descendant of the ancient Indian Brahmi script [14] and is thus a member of the Aramaic script family [15] by extension.

B. Grammar checking tools

The natural language processing field is relatively new in the Asian language and plenty of apparatuses are yet to be created. One of these is an accurate grammar checker [16-17]. When all is said and done, rule-based and AI strategies have been utilized for creating sentence structure checkers [18]. Since the field of regular language preparing for Asian dialects is constrained, alongside the intricate syntax, it exhibits a few issues when building up a sentence structure checking framework for the Sinhala language.

Despite the fact that there are syntax checking devices accessible for different dialects, for example, English, the inaccessibility of a technique to check the sentence structure of Sinhala language at constant is as yet not pervasive. The unpredictability of the Sinhala syntax appeared to be the primary explanation behind this. Since a corpus-based vocabulary for the Sinhala language has been executed [19], it can very well be used to helpfully channel meaning and importance from the human language. Engineers and Data Scientists are eligible to prepare and structure complex processes such as relationship extraction, interpretation, grammar sentence structure checking, subject decomposition, etc while using NLP [20], pursuing an information-driven measurement-based methodology considerably at ease. A particular strategy to arrange the finish of the sentence should be actualized to break down a contribution by distinguishing each sentence. Furthermore, the framework should be executed such that it offers proposals for the end clients by recognizing the example of the given information.

III. RELATED WORK

The complex composition of a lot of spelling and grammar rules paves the way for the Sinhala language to be more complicated. All the more, it is difficult to implement all these rules according to a specific order nevertheless, perform all the grammar rules in one application [21].

Even with all these difficulties, there have been very few attempts in developing automated grammar checkers for the Sinhala language within recent decades. “Spell and Grammar Checking Tool for Sinhalese, අකුරු හේරුව - සියලුම සුඛසක් කරනු රීසියයෙහි”, developed by Abeyrathne et

al. [4] is one such system. The fact that there are already existing systems to facilitate spell checking has been taken into consideration, and thus the system has been developed for providing spell-checking and grammar checking. Even though there are other existing spell checking systems such as spellchecker.net and “මධුර”, to provide accurate spell checking functionality a spell checker has been implemented in a data-driven approach in “සුඛස”, which is also already an existing system. A corpus-based Sinhala lexicon has been studied to establish the initial step of spellchecking in this study. A timely concept proposed for morphologically lavish languages like Sinhala, where the words extracted from the corpus have been labeled with parts-of-speech categories defined has been used when conducting the research. With the help of an approach based on n-gram statistics which is a well-known data-driven approach, checking and correcting spelling errors in Sinhala in the study has been conducted which is relatively inexpensive to construct without deep linguistic knowledge [4]. Another related project carried out by the developers: Akila Gunarathna, Pathmasri Ambegoda, Nirasha Katugamapala, Dhanushka Bandara with the guidance of the supervisors: Professor Gihan Dias, Dr. Sanath Jayasena from the Department of Computer Science and Engineering, Faculty of Engineering, University Of Moratuwa has been able to implement an open-source spelling and grammar checking tool for just simple three words sentence, with the name of the project as, “මහරුවණ” [22]. Since the system has been developed using a rule-based approach, further improvements to that system have become more complicated and thus inevitable due to the heavy number of complex grammar rules, composite noun phrases changing their meaning according to the place of application in a sentence, variety of grammatical features embedded in the language structure, etc. in the Sinhala language [23]. But all these attempts for developing grammar checkers for the Sinhala language have been with the scope of checking grammatical correctness only in simple 3-words Sinhala sentences. Nonetheless, the researches have been only focused on using single approaches or a comparison of a few approaches separately. Also, usage of a higher amount of grammatical features has been at a minimum level in most of all the researches. Yet, the accuracy levels have also been in contrast revolved around lower values.

When considering the other languages similar to Sinhala language Tamil, is also considered as a resource-poor language closely related to the Sinhala language while acting the role of the other official language in Sri Lanka [24].

Nonetheless, because of the presence of larger Tamil speaker populations worldwide, especially in countries like India, a considerable amount of research and tools are available for NLP work related to the Tamil language [25]. Consequently, it is, therefore, reasonable to recognize that Sinhala and Tamil NLP endeavors help assist one another. Additionally, relativity along with alternative mechanisms have been helpful to bridge other languages like Japanese with the Sinhala language [26] [27]. Similarly, in research disbursed for the Kannada language, agreement of noun and verb phrase in those sentences has been sculptured [28]. Noun phrases have been categorized into three subcategories as adjective-noun, noun, and pronoun, but only the gender and the number have been considered as grammatical features of the language. The resemblance of a similar scenario in the

Sinhala language; verb agreement with the subject of the sentences in number and gender. Hence, the verb suffix is extracted to examine masculinity, femininity, and plurality of the verb. In this research along with Python programming language context-free grammar (CFG) has been used to implement the grammar rules [21]. Another related research that has been carried out by Sagar et al. [29] elaborates on the generation of a context-free grammar (CFG) for simple sentences in the Kannada language [29]. Here, both a Top-Down Parser and a Bottom-Up Parser have been used and compared their reliability for the grammar checking process. The main two conflicts have been identified by the authors occurring during the usage of Bottom-Up parser; Shift-Reduce and Reduce-Reduce. Therefore the Top-Down parser has been implicitly selected as the more suitable parser to be used. A Computational Grammar for Bengali has been implemented by Khan and Khan using the Head-Driven Phrase Structure Grammar (HPSG) formalism [30].

A process synchronic linguistics for a language may be a possibly helpful resource for polishing off varied language process activities like testing synchronic linguistics, machine translation, and question respondent. Similar to the situation in most South Indian languages [26] [31], among alternative grammatical options, Sinhala may be an extremely incurred language with 3 gender forms and 2 range forms [32]. To develop a computational model, it is required to understand the language well. Research has been conducted to describe the event of a feature-based CFG for non-trivial sentences in Sinhala [21]. Originally, ensuing synchronic linguistics covers a big set of Sinhala as represented by very popular synchronic linguistics hard copies [23] [32]. To manufacture the suitable, dissect tree(s) related to sentences input to the system, a parser has been conjointly implemented victimizing the NLTK toolkit during this research. Generally, the synchronic linguistics conjointly detects then rejects ungrammatical sentences during the conduct. 200 sample sentences have been acquired via Sinhala grammar textbooks of primary grades for testing purposes in the study. 10 sentence structures have been selected and have been accustomed to style the synchronic linguistics. The accuracy of the process has been limited to 60% according to test subjects. Implementation of a computational model of grammar for the Sinhala language has been considered in another research carried out by Hettige and Karunananda [33]. Usage of morphological and grammar analysis of Sinhala has been thought of during this research study to shapely employ Context-Free descriptive linguistics and a Finite State electrical device (FST) that also handles solely straightforward sentences that contain eight constituents [33]. Computational models for grammatical error detection have been evolved from simple rule-based models to deep learning models [34]. Statistical models have shown to perform well in many natural language understanding tasks [35] [36]. Applying statistics requires a clear understanding of the language and the language structure [37]. However, this requires a large amount of knowledge about the language, which can be hard to achieve on languages that are not popular, or which have complex structures like Sinhala [38].

The Sinhala language has a limited digital footprint when compared to other popular languages. Therefore, using Sinhala for natural language processing tasks are harder [35]. Natural language processing tasks for the Sinhala language has been conducted using many techniques [39]. The rule-

based approach was a very common approach used in simple Sinhala natural language tasks [40]. The statistical models show the potential in achieving better natural language models for Sinhala. With technological improvement, most recent developments in NLP related researches have been, using machine learning algorithms [41-43]. These models have been used in many industrial language tasks since the models are better. However, statistical models fail to capture all features of language models [37]. Machine learning algorithms do not require any feature selections but are capable of achieving higher accurate results compared to statistical methods and rule-based models [41] [43-44]. Machine learning algorithms have been applied as a comparison technique for grammatical error detection in very few researches so far [4]. A machine learning algorithm-based model, however, requires a larger dataset to learn. Therefore, languages like Sinhala with smaller digital data presence fail in producing better results while only using a machine learning algorithm-based approach [41]. Thus, the necessity of a novel approach arises to earn more accurate and efficient results in the context of grammar mistake detection and suggestion of corrections for a morphologically lavish language like Sinhala.

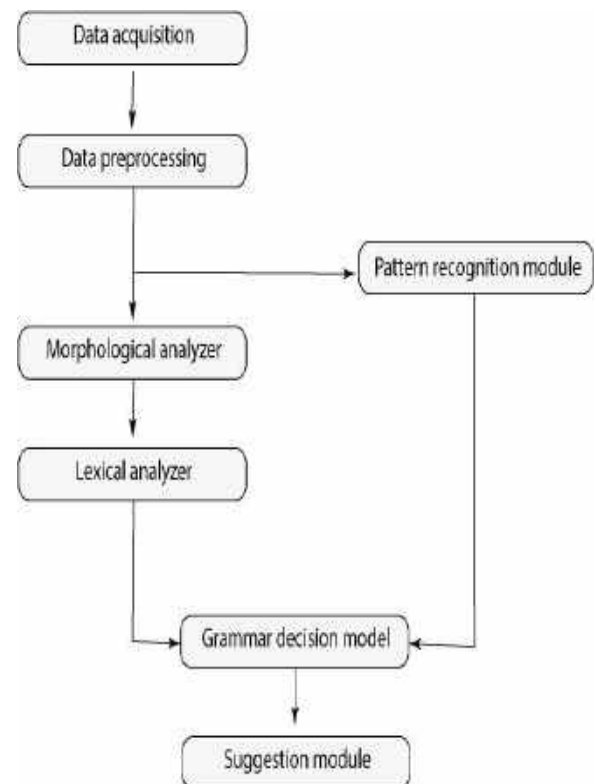


Fig. 1. Proposed and implemented a methodology for the study

IV. METHODOLOGY

The commencement via a thorough literature review of the research area deprived that applying only one approach has been unable to output with more accurate results and thus, this study has been carried out using a novel hybrid approach while combining traditional rule-based approach along with the machine learning algorithm-based approach, for grammatical error detection and correction in Sinhala text to identify the possibility of having a high potential of achieving comparative results. Since a proper lexical analyzer has not been implemented yet for the Sinhala language at present, engaging with only a rule-based approach could be inappropriate when considering the lack of possible resources, unlike other resource-available languages. Thus when examining most appropriate approaches for detecting and correcting grammatical mistakes for morphologically lavish

languages like Sinhala, the severe complexity of the language, difficulty of defining a fixed set of grammar rules due to containing a lot of spelling and grammar rules and availability of enough resources for implementing such approaches were taken into consideration in this research.

The study has focused on detecting the grammatical mistakes in active sentences in Sinhala text where sentences in written text format were checked sentence-wise and word-wise respectively and the nouns and verbs were separately analyzed with the help of a resourceful part-of-speech (pos) tagger [7] and a morphological analyzer [45]. To uplift the performance of the grammar checker, both the pos tagger and the morphological analyzer used in this study have been further modified accordingly to upgrade the error detection and correction mechanism to uplift the available Sinhala

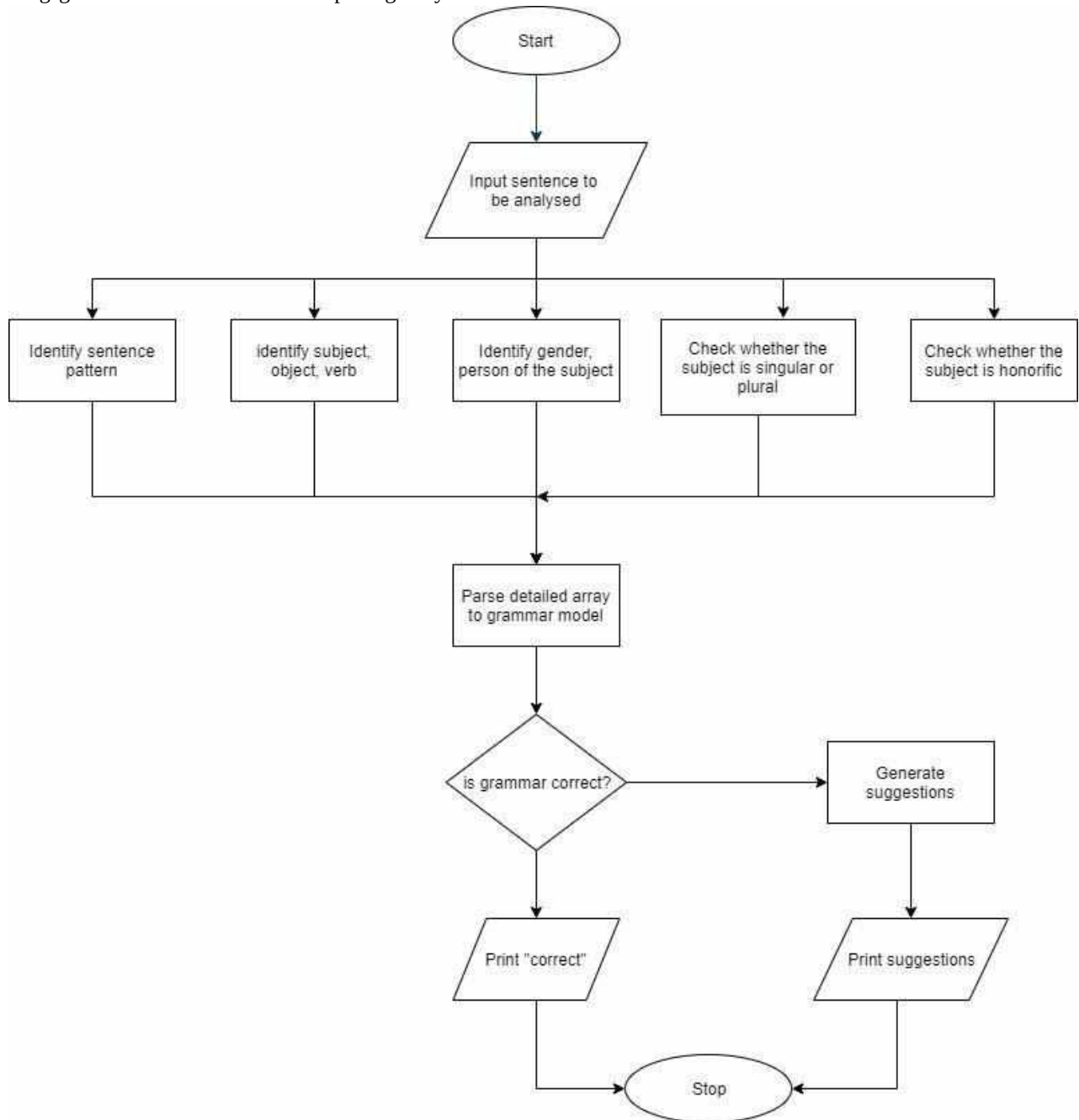


Fig. 2. The algorithm used for grammar mistake detection and suggestion of corrections

digital footprint to a higher level. Parallel to this phase, the given sentence was then sent through a pattern recognition mechanism that was developed to identify the sentence pattern of the given sentence.

Finally, the system which comprised of a hybrid novel approach combining a system developed using traditional rule-based approach and decision-tree; a machine learning algorithm-based model was able to detect the correctness of the given sentence and provide corrections for erroneous grammatical sentences. Specifically, here the consideration has only been on already spell checked sentences [46] for the detection for grammatical mistakes which helped compress the scope to a feasible level and provide more accuracy for the study. Fig.1 depicts the main approach which has been implemented in this study, and it has been further elaborated below.

A. Data acquisition

A data set of around 1050 Sinhala active sentences with no spelling errors without a specific number of word-range were gathered from the web using online Sinhala newspapers via Python scripting and they were categorized according to their tenses and grammatical person. Spelling check has already been done before with the help of an existing accurate spell checker for the Sinhala language; “සිංහල | අකුරු විනිස පිරික්සුම” [46]. This data set was then thoroughly validated with the help of a group of Sinhala language experts from different dialects and a data set containing around 800 sentences with about 25 pre-defined sets of grammar rules and about 32 sentence patterns were finally separated according to their grammatical feature and those were recorded and used as the training dataset.

For further testing purposes, the remaining 250 sentences were used and they were then pre-processed using NLTK toolkit and Python programming language and a data set containing around 180 correct grammatical sentences and around 60 incorrect grammatical sentences were classified while eliminating 10 unidentified sentences due to lack of proper sentence structure in those sentences.

B. Data pre-processing

Separation of sentences in the paragraph, separation of words from spaces in a sentence, removing unwanted characters from the sentences were been carried out to pre-process the data input into the system.

C. Morphological and lexical analysis

Then the identification of subject and verb through the already pre-processed sentences were been carried out using an existing pos tag system [7] and the tag of each word was identified and stored separately with the corresponding word for further reference. A comprehensive, multi-level pos tag set for Sinhala was being used here which has been developed by the Center for National Language Processing, University of Moratuwa while considering its accuracy level [7]. This pos tagger has been further modified in this study, to increase the performance of the grammar checker for example with newer sub-tags for the gender of the common nouns used in the available pos tagger. Then a morphological analyzer [46] was used to separate the root and the suffix from the identified verb and separated root and suffix of the verb was further carried over to the next phase with the corresponding noun acquired from the above lexical analysis. Due to lack of access to most

of the accurate morphological analyzers for Sinhala [16], on the notion that the above mentioned morphological analyzers work, another existing considerably accurate morphological analyzer which is commonly known as “polyglot” [45] has been used with further modifications to increase the performance of the grammar checker accordingly in this study which is a natural language pipeline that supports the Sinhala language in this study. Further modifications were necessary for the morphological analyzer because even though it provided accurate noun breakdowns, it provided several suffixes for the verb giving ambiguity when selecting the correct suffix to be concatenated to the verb root for a given verb in a correct grammatical sentence.

D. Pattern recognition module

Alongside the above process been carried out, the given sentences were then passed through a pattern-recognition mechanism that has been developed using a regular expression based algorithm, to identify the sentence pattern of the sentence input into the system. Around 32 different sentence structures have been identified and used in the pattern-recognition mechanism to detect grammatical mistakes in a variety of sentence patterns to expand the error detection mechanism for a higher possible range of sentence patterns.

E. Grammar decision model

Grammar rule-based model developed for identifying the grammar rules and sentence pattern-recognition mechanism for identification of sentence structures developed in this research that was used for the training purpose of the model was created with the thorough referencing through many Sinhala hard copies and literacy books, and then the produced data set was cross-checked with the help of a panel of experts consisting Sinhala language experts from several divisions including Sinhala teachers from different schools, language experts from different dialects. The finalized data set was then further directed to a set of language experts from the Department of Sinhala, Faculty of Humanities at the University of Kelaniya. All these processes were carried out with the comprehensive usage of the Delphi method [47] to acquire an accurate and non-bias data set to accomplish the tasks carried out in this research study.

F. Novel hybrid approach combining rule-based and machine learning algorithms (Suggestion module)

To elaborate in detail, the general flow once a sentence was entered into the system has been, first, each word was assigned with a tag by the pos tagger.

Eg: - In the sentence, “මම ගමට ගියෙමි”, it worked as below.

```
{
  "sentence": "කමල් සමග මම පාවිච්ඡි කරමි",
  "is_correct": false,
  "suggestions": {
    "present": "කමල් සමග මම පාවිච්ඡි කරමු",
    "future": "කමල් සමග මම පාවිච්ඡි කරන්නෙමු",
    "past": "කමල් සමග මම පාවිච්ඡි කළෙමු"
  }
}
```

Fig. 3. Machine-learning algorithm and grammar rule-based corrected sentence suggestion

- Pronoun -> මම - PRP
- Verb -> ගියෙමි - VFM
- Common Noun -> ගමට - NNC

The array which consisted above details were then passed through the pattern-recognition mechanism to identify its sentence pattern. According to the identified sentence pattern, lexical analyzed data and morphologically analyzed verb, the grammatical error detection were then carried out with the help of the developed grammar-rule data set using a decision-tree classification machine learning algorithm. This has been used to evaluate the verb with the “subject” and output feedback about the correctness of the sentence. Fig.2.shows the flow chart for the algorithm established in the study. The system has considered around 25 predefined sets of grammar rules in this phase while considering the performance as a key aspect. Accordingly, for grammatically incorrect sentences, the system has predicted suggestions using the decision-tree classification-based machine learning algorithm [43] for all three tenses. Fig.4.shows a sample data set that has been used for the decision tree algorithm in this study. Then the suffix-engine would try to concatenate given suffix and root of the verb and create a meaningful sentence to the user while finally providing the user with a correctly spelled and grammatically sound sentence.

V. RESULTS AND DISCUSSION

subject	tense	person	gender	animate	number	verb_root	is_correct	
2	මම	අතීත	උත්තම	නොමැත	ප්‍රාණවාව	ඒක	එමි	1
3	මම	අතීත	උත්තම	නොමැත	ප්‍රාණවාව	ඒක	ඒය	0
4	මම	අතීත	උත්තම	නොමැත	ප්‍රාණවාව	ඒක	එමු	0
5	අපි	වර්තමාන	උත්තම	නොමැත	ප්‍රාණවාව	බහු	අමු	1
6	අපි	වර්තමාන	උත්තම	නොමැත	ප්‍රාණවාව	බහු	අමි	0

Fig. 4. Sample data set used for the decision tree algorithm in the study

The data set consisted of around 800 records that were used to train the decision tree in the grammar model in the study was acquired with the thorough referencing through many Sinhala hard copies and literacy books [23] [32] [48-49], and then cross-checked with the help of a panel of experts consisting Sinhala language experts from different dialects including Sinhala teachers of different schools and finalized it by further directing it to a panel of language experts with the comprehensive usage of the Delphi method [47] to acquire an accurate and non-bias data set. This data set included data about 25 predefined grammar rules in Sinhala. It consisted variety of grammatical features including person, tense, gender, number, honorific feature, animate feature, etc. This data set also comprised of a broad range of subjects from most common pronouns to deterministic pronouns and even extended to several uncommon proper nouns as well. This trained model then gave us an accuracy score of 88.6%.

A system identified incorrect grammar sentences were automatically passed to the decision-tree classification-based machine learning sentence correction algorithm to suggest correct sentences. It then suggested a suitable verb root and then added it to a matching verb suffix and gave the output as shown in Fig.3.

The data set used for further testing the developed system consisted of around 240 sentences with a non-bias set of sentences with both correct and incorrect grammar. To test the performance of the classification model a confusion matrix was used and the results obtained have been elaborated in Fig.5 below.

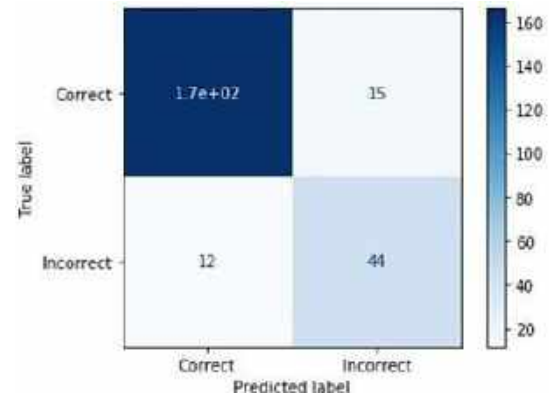


Fig. 5. Confusion matrix obtained to analyze the performance of the developed model

During the testing phase system gave the following results.

Input correct sentences = 181, System identify only 166 sentences

Input incorrect sentences = 56, System identify only 44 sentences.

- Accuracy = $[(166+44)/237] * 100\% = 88.6\%$
- Recall = $[166 / (166+15)] * 100\% = 91.71\%$
- Precision = $[166 / (166+12)] * 100\% = 93.25\%$

$$F\text{-measure} = [(2 * 0.9171 * 0.9325) / (0.9171 + 0.9325)] * 100\% = 92.47\%$$

According to the above-calculated results that were obtained using the confusion matrix, the model has given higher recall, higher precision with a considerably high amount of accuracy for the given test cases. Nonetheless, F-measure which calculates the harmonic mean instead of arithmetic mean has precisely output with an accurate result using both the precision and recall values. Hence, a considerable amount of accurate results were achieved through the proposed and implemented a novel hybrid approach with machine learning techniques.

Most of the researches that have been carried out so far in the context of grammatical error detection for the Sinhala language have been only using 3-words sentences and have only covered the scope of error detection for selected features in the Sinhala language like number, gender, tense and person where features like volition has been omitted in most of all the scenarios. One important aspect that has been achieved in this study was the elimination of the limitation of grammar checking for only 3-words sentences. This has been upgraded to more than 3-words sentences in this study. Nonetheless, in this research as special accomplishments, sentences consisting of post-positions, conjunctions like, “ද, හා, නි, හහ”, grammatical features like volition, deterministic pronouns like, “අයෙක්, කෙනෙක්, ඇතැමෙක්” and few proper nouns were taken into consideration. To elaborate more on the

study, given the sentence, “මම ගෙදර කමු”, here the word gives a meaning of “I eat the house”, which is semantically incorrect. But, the algorithm suggested in this study could escape this problem and the suggestion of meaningful sentences was also achieved as a secondary output of the study. So, this algorithm could be used to suggest a correct verb in any given word processing software.

Even though the main intention of this study was to propose and develop a grammatical error detection and correction mechanism using a novel hybrid approach, this research study was able to contribute to further developing and modifying the digital footprint for the Sinhala language as well. While accomplishing the task of grammar detection and correction, necessary changes and modifications were done accordingly for the existing pos tagger [7] and the morphological analyzer [45] to contribute to future researches that would be carried out on resource-poor Sinhala language.

VI. CONCLUSION

This research has been carried out to introduce a hybrid approach for identifying grammatical errors and suggesting corrections for the Sinhala written sentences and to identify whether there is a possibility of having a high potential of achieving comparative results compared to other relative approaches. With the gained results through the study, it was able to identify the sentence pattern in the given sentence and then identify its grammar rules. Further, using machine learning algorithms it was able to suggest corrections for the sentences identified as incorrect grammatical sentences with an acceptable accuracy rate. Therefore with all the achieved results from this study, it can be concluded that the hybrid approach is an efficient and accurate approach for detecting and correcting grammatical mistakes in written text formats for the Sinhala language. Further, the research work carried out in the context confides that the proposed novel hybrid approach could be possibly extended towards other emerging technological aspects as well with better promising results.

VII. FUTURE WORK

As suggestions for future work, the dataset must be developed and optimized further to identify more errors mostly in third person format and also the sentence pattern-recognition model could be further optimized to identify more sentence patterns to increase the performance of the established system. For this purpose, using a Deep leaning based extension for the used rule-based pattern recognition algorithm might be a success. To identify the gender of a given third-person noun, the rule-based approach was used in this study. But as third-person format consists of huge variety and number of words, using a separate dictionary for this is also not a practical solution. Therefore, a strong recommendation is suggested to focus more on overcoming this issue in future work. The existing morphological analyzer was used after a few modifications to overcome the ambiguity issue occurred when separating some verb-roots. Hence, a further modification for the existing morphological analyzer or development of a proper and accurate morphological analyzer is necessary for the future to uplift the Sinhala language-based research.

Also for future implementations, it can be possibly recommended to use emerging technologies like neural network systems and try out statistical models like

mathematical model text classification to compare the accuracy of the hybrid system implemented in this research study which was developed by combining rule-based and machine learning algorithms to improve the development of Sinhala language-based researches in the future.

ACKNOWLEDGMENT

The authors would like to extend their heartfelt gratitude and acknowledgment to all the language experts, Sinhala teachers from different schools and especially, the Sinhala Department in the University of Kelaniya for their immense support and encouragement they gave throughout the development phase of the data sets.

REFERENCES

- [1] Statistics.gov.lk. (2017). Census of population and housing 2011. [on - line] Available at: <http://www.statistics.gov.lk/PopHouSat/CPH2011/in-dex.php?fileName=pop42gp=Activitiestpl=3> [Accessed 6 Jun. 2019].
- [2] Ethnologue. (2017). Sinhala. [online] Available at: <https://www.ethnologue.com/language/sin> [Accessed 16 Jul. 2019].
- [3] Anon, (2020). [online] Available at: <https://groundviews.org/2011/10/24/reflections-on-issues-of-language-in-sri-lanka-power-exclusion-and-inclusion/> [Accessed 13 Nov. 2019].
- [4] L. Abeyrathne, S. Edirisinghe, R. Premachandra, A. Warsha, N. De Silva and S. Thelijagoda, “Spell and grammar checking tool for Sinhalese: අකුරු සසර්ග - සියලුම සුඛසක් කරනු ලබන සෙයින්”, 2018.
- [5] I. Wijesiri, M. Gallage, B. Gunathilaka, M. Lakjeewa, D. Wimalasuriya, G. Dias, R. Paranavithana, and N. De Silva, “Building a wordnet for Sinhala,” in Proceedings of the Seventh Global WordNet Conference, 2014, pp. 100–108.
- [6] B. Hettige, A.S. Karunananda, “A morphological analyzer to enable English to Sinhala machine translation,” in Proceedings of the 2nd International Conference on Information and Automation (ICIA2006), Colombo, Sri Lanka.
- [7] S. Fernando, S. Ranathunga, S. Jayasena, G. Dias, “Comprehensive part-of-speech tag set and SVM based POS tagger for Sinhala,” In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016) (pp. 173-182) (2016, December).
- [8] A. Wasala, R. Weerasinghe, R. Pushpananda, C. Liyanage, and E. Jayalatharachchi, “A data-driven approach to checking and correcting spelling errors in Sinhala,” Int. J. Adv. ICT Emerg. Reg, vol. 3, no. 01, 2010.
- [9] R. A. Wasala, R. Weerasinghe, R. Pushpananda, C. Liyanage and E. Jayalatharachchi, “An open-source data driven spell checker for Sinhala,” International Journal on Advances in ICT for Emerging Regions (ICTer), 3(1), pp.11–24, 2011.
- [10] E. Jayalatharachchi, A. Wasala, and R. Weerasinghe, “Data -driven spell checking: the synergy of two algorithms for spelling error detection and correction,” in International Conference on Advances in ICT for Emerging Regions (ICTer2012). IEEE, 2012, pp. 7–13.
- [11] L. G. B. Subhagya, L. Ranathunga, W. H. A. Nimasha, B. R. Jayawickrama, and K. L. Mahaliyanaarchchi, “Data-driven approach to Sinhala spellchecker and correction,” in 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2018, pp. 01–06.
- [12] Statistics.gov.lk. (2013). [online] Available at: <http://www.statistics.gov.lk/PopHouSat/PDF/Population/p9p11%20Speaking.pdf> [Accessed 3 Jun. 2019].
- [13] N. Mohamed, (2005). Note on the early history of the Maldives. Archipel, Vol. 70(1), pp.7-14.
- [14] P.T. Daniels and W. Bright, “The world’s writing systems,” Oxford University Press on Demand, 1996.
- [15] R. Salomon, “Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages,” Oxford University Press, 1998.

- [16] N. De Silva, "Sinhala text classification: observations from the perspective of a resource poor language," 2015.
- [17] V. Verma and S. Sharma, "Comparative analysis of grammar checkers of various Asian languages," *International Journal of Computer Sciences and Engineering*, Vol 6(10), pp.697-700, 2018.
- [18] A.Vernon, "Computerized grammar checkers 2000: capabilities, limitations, and pedagogical possibilities," *Computers and Composition*, Vol 17(3), pp.329-349, 2000.
- [19] A.R.Weerasinghe, Dulip Herath, Viraj Welgama, "A corpus-based Sinhala lexicon," In *Proceedings of the 7th Workshop on Asian Language Resources*, Singapore, Aug 2009.
- [20] R. Sangal, S. Bendre and U. Singh, (2003). *Recent advances in natural language processing*. Mysore: Central Institute of Indian Languages.
- [21] C. Liyanage, R. Pushpananda, D.L. Herath and R. Weerasinghe, "A computational grammar of Sinhala," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 188–200, Springer, 2012.
- [22] Findbestopensource.com. Maharavana - Spelling and grammar checker for Sinhala. [online] Available at: <https://www.findbestopensource.com/product/maharavana> [Accessed 21 May 2019].
- [23] J. Gair and J. Paolillo, "Sinhala," Munchen: LINCOM Europa, 1997
- [24] J. Gair and B. Lust, "Studies in South Asian linguistics," New York: Oxford University Press, 1998.
- [25] N. Saharia, S. Sharma and J. Kalita, "Stemming resource-poor Indian languages", *ACM Transactions on Asian Language Information Processing*, Vol 13(3), pp.1-26, 2014.
- [26] A. Herath, Y. Hyodo, Y. Kawada, T. Ikeda, and S. Herath, "A practical machine translation system from Japanese to modern Sinhalese", *Gifu University*, pp. 153–162, 1994.
- [27] A.B. Kanduboda, "The role of animacy in determining noun phrase cases in the Sinhalese and Japanese languages," *Science of words*, vol. 24, pp. 5–20, 2011.
- [28] B.M. Sagar, G. Shobha, and Ramakanth Kumar, "Solving the noun phrase and verb phrase agreement in Kannada sentences," *International Journal of Computer Theory and Engineering*, Vol. 1, No. 3, August 2009.
- [29] B.M. Sagar, G. Shobha and Ramakanth Kumar, "Context Free Grammar (CFG) analysis for simple Kannada sentences," in *Proceedings of the International Conference [ACCTA-2010] on Special Issue of IJCCT*, Vol. 1 Issue 2, 3, 4., 2010.
- [30] Naira Khan and Mumit Khan, "Developing a computational grammar for Bengali using the HPSG formalism," in *Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT '06)*, 2006.
- [31] A. Grant, "Studies in South Asian linguistics: Sinhala and other South Asian languages (review)," *Language*, Vol 77(3), pp.639-639, 2001.
- [32] D. Chandralal, "Sinhala," Amsterdam, the Netherlands: John Benjamins Pub. Co, 2010.
- [33] B. Hettige, A. Karunananda, "Computational model of grammar for English to Sinhala machine translation," in *Proceedings of the International Conference on Advances in ICT for Emerging Regions*, 2011.
- [34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.
- [35] K. Senevirathne, N. Attanayake, A. Dhananjani, W. Weragoda, A. Nugaliyadde, and S. Thelijjagoda, "Conditional random fields based named entity recognition for Sinhala," in *IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, 2015, pp. 302- 307.
- [36] E. Cambria and B. White, "Jumping NLP curves: a review of natural language processing research [review article]," *IEEE Computational Intelligence Magazine*, vol. 9, pp. 48-57, 2014.
- [37] C. D. Manning, C. D. Manning, and H. Schütze, "Foundations of statistical natural language processing," MIT Press, 1999.
- [38] J. Jayakody, T. Gamlath, W. Lasantha, K. Premachandra, A. Nugaliyadde, and Y. Mallawarachchi, "Mahoshadha," the Sinhala tagged corpus based question answering system," in *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*, 2016, pp. 313-322.
- [39] S. Herath, S. Ishizaki, T. Ikeda, Y. Anzai and H. Aiso, "Machine processing of Sinhala natural language: a step toward intelligent systems," *Cybernetics and Systems*, 22(3), pp.331-348, 1991.
- [40] P. Antony, "Machine translation approaches and survey for Indian languages," *International Journal of Computational Linguistics Chinese Language Processing*, Volume 18, Number 1, March 2013, vol. 18, 2013.
- [41] N. Madi, R. Al-Matham and H. Al-Khalifa, "Grammar checking and relation extraction in text: approaches, techniques and open challenges," *Data Technologies and Applications*, 53(3), pp.373-394, 2019.
- [42] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, pp. 261-266, 2015.
- [43] "Machine learning algorithm for learning Natural Languages," *International Journal of Managerial Studies and Research*, Vol 4(2), 2016.
- [44] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv: 1410.3916*, 2014.
- [45] Polyglot.readthedocs.io. (2019). Welcome to polyglot's documentation! Polyglot 16.07.04 documentation. [online] Available at: <https://polyglot.readthedocs.io/en/latest/index.html> [Accessed 17 Sep. 2019].
- [46] Speller.subasa.lk.(2013) [online] Available at: <http://speller.subasa.lk/spellerweb.py> [Accessed 14 Nov. 2019].
- [47] W. Weaver, "The Delphi method," Syracuse, N.Y.: Educational Policy Research Center, Syracuse University Research Corp, 1970.
- [48] K. Jayathilake, "Nuthana Sinhala Vyakaranaye Mul Potha," Pradeepa Publications, 34/34, Lawyers' Office Complex, Colombo 12, (1991)
- [49] A. M. Gunasekera, "A Comprehensive Grammar of the Sinhalese Language," New Delhi, India: AES Reprint, 1986.

Vision-based automatic warning system to prevent dangerous and illegal vehicle overtaking

Mahinsa Athree*

Department of Statistics and Computer Science
University of Kelaniya, Faculty of Science, Sri Lanka
mahinsaathree@gmail.com

Anusha Jayasiri

Information Technology Centre
University of the Visual and Performing Arts, Sri Lanka
anusha.j@vpa.ac.lk

Abstract: The purpose of this research is to implement a system that can be a help for drivers to drive vehicles safely and reducing accidents. Most of the accidents have occurred when overtaking. The driver has to consider the distance between two vehicles and the speed of the front vehicle. If there is another vehicle moving the same or opposite direction, it must be considered as well. Also, traffic signs, solid line, and a broken line must be considered. Reducing the errors which can be happened by the driver when deciding all these stuff in instinct is the goal of the “Automated Vehicle Overtaking System”. This research has been used three pipelines, such as lane line detection, traffic sign detection, and vehicle detection. The input video stream from the front view camera of the vehicle is transferred through the above pipelines. The system gives the final output frame with confirmation that ‘is it safe to overtake now’. The driver can recognize the danger by alarm sound from the system. This system has a Python Application Programming Interface (API). Using that API, any device or application can access this system. Finally, the main goal of this research is to implement a system that can predict dangers in overtaking and saves the lives of passengers and vehicles.

Keywords: Flask, Image Processing, Object detection, OpenCV, Python API

I. INTRODUCTION

As the number of vehicles on the road increases exponentially, road accidents are also increasing proportionately. Although road rules, traffic signs, etc. play an important role in ensuring the safety of vehicles on the road, recent statistics show the number of accidents on the road is unacceptably high.

Different types of traffic signs are used for better traffic management. Danger warning signs, regulatory signs, directional informative signs, traffic light signals, and road markings are the categories of traffic signs used in Sri Lanka. Besides, Sri Lankan roads are marked with solid yellow lines, solid white lines, solid double yellow lines, solid double white lines, broken white lines and broken yellow lines.

Vehicle manufacturers are also adding more and more safety features to their products to make them safer as well as a way of competing with each other. As a result of these efforts, modern vehicles are much safer than previous generations. Even with these technologies still, numerous accidents take place. Many accidents occur when drivers attempt to overtake vehicles.

The purpose of this research is to warn drivers when overtaking through a camera-based system. Front view and rearview cameras are used in vehicles can introduce an additional layer of safety. Front view cameras, generally

known as dashcams are very useful to see a full view through the windscreen. Detected frames are shown on a display of the vehicle or drivers can view the video stream through a mobile device. Some dashcams can connect other devices via Wi-Fi, Bluetooth, and 4G.

It is better for drivers if there is no traffic, no pedestrians, no traffic lights and straight lanes, but it is not the reality. The driver has to take care of all of the above-mentioned stuff. If there is a pedestrian crossing traffic sign, the driver has to drive carefully. Because there can be pedestrians on the road. If there is a traffic light sign, the driver has to follow those signal lights. The driver has to pay attention to traffic signs, road markings, vehicles in front of his vehicle and vehicles behind as well.

When overtaking, the proposed system warns drives based on the speed of the vehicle, speed of the vehicle in front, the distance between two vehicles, road marking and traffic signs.

II. RELATED WORK

Hadi et al reviewed image processing tools that can be used for implementing vehicle detection and tracking systems. Also, this review contains what analysis tool is more effective when implementing the above-mentioned applications. Most effective preprocessing techniques can be identified by this review [1].

Ali et al introduced the “Safe passing System of Vehicle” system that facilitates to overtake safely with a microcontroller circuit. For this purpose, a measuring system sensing element was employed. The measuring system sensing element detects the gap and existence of a car. The signal is regenerated into visual through different colour light-emitting diode lightweight by a microcontroller circuit [2]. Image processing or computer vision techniques for detecting speeds of vehicles have not been used in this work.

Gholami et al proposed a new algorithm that uses the digital video, image processing and computer vision to automatically detect vehicle speed. The captured traffic videos are collected with a stationary camera that is mounted on a freeway. Since the foreground of an image represents the moving objects only, the speed of a vehicle is calculated by removing the background and extracted the foreground of the image sequence. The speed of the vehicle in each frame is calculated using the position of the vehicle in each frame. [3]. Redmon et al had implemented the YOLOv3 that has a new classifier network better than others. Bounding boxes are predicted utilizing dimension clusters as anchor boxes by YOLO9000. This network calculates four coordinates for each bounding box such as x, y, w, h. According to the paper x and y are coordinates of the centroid of the bounding box. Also, w

and h are the width and height of the bounding box [4]. The proposed system prefers to use the YOLOv3 network because its accuracy is 51.5 mean Average Precision (mAP) and speed is 45 frames per second (fps). Those values are better than other networks' accuracy and frame speed.

Bhaskar et al introduced an automatic vehicle investigation system using videos recorded from stationary cameras close to roads. Primarily image processing and object detection were used for implementing this model, furthermore, the Gaussian mixture model and blob detection methods were used for implementing this algorithm. Morphological techniques were used to remove noise from input frames. Final calculations were done by detected objects and regions [5].

Rosebrock calculated the distance between the camera and the detected object using Triangle similarity. Initially, an object was kept at a known distance from the camera and then taking a photo of it.

$$F = (P \times D) / W \quad (1)$$

The focal length of the camera can be calculated using (1). P , D , and W are apparent width in the image, known distance and actual width respectively.

$$D = (W \times F) / P \quad (2)$$

When the focal length of the camera is known, distance calculation can be determined by (2). As mentioned earlier, W is the actual width of the object. F is the calculated focal length. P is the apparent width of the object in pixels. Finally, the actual distance is calculated [6].

Low et al presented a system that identifies a road as well as the left and right boundaries of the road. A vision sensor was utilized to capture the front view frames of the vehicle. Edge detection algorithms and Hough transform algorithms were used to process the frame sequence. Using Hough transformation, the boundaries of the road can be drawn according to the edges that were recognized by Edge detection [7].

III. METHODOLOGY

This research consists of four main stages. Every single input frame passes these steps to extract important information and use those data to have the final result. The approach is shown in Fig. 1.

A. Road marking detection

As the first step, a pipeline has been implemented to detect lines on the road such as a solid line and a broken line in both yellow and white colours. The "Road marking detection" pipeline starts with input frames from the front view camera of the vehicle and ends with the final road marking detected frames.

The method of "Road marking detection" used several image processing techniques such as Grayscale, Gaussian Blur, Canny edge detection, Region selection, Hough line transformation and weighed image methods.

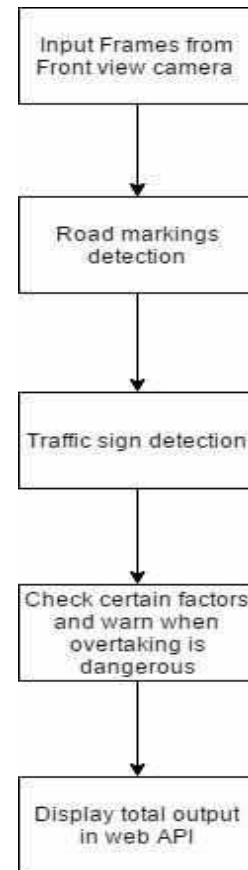


Fig. 1. The Overall approach of the research

The grayscale method is used to convert a multichannel image into a single-channel image. Because of the blurring, an image reduces the noise and unnecessary details of an image, Gaussian Blur method is used for the blurring of the Grayscale image. The Canny Edge detector is used to detect the edges of the images which is to remove the unnecessary information and extracted only the important data of an object [8]. The following Fig. 2 shows the edge detected and original images.



Fig. 2. Canny edge detected image and the original image

The region selection method was used to extract important and relevant regions of the frame concerning the road. Once the unwanted regions of the frame have been removed, the processing power can be reduced. Fig.3 shows an example of a region selection of a road. The size of the region was selected by changing the values of the bottom left, top left, bottom right, and top right vertices.



Fig. 3. Lane region of the image after region selection

Hough line transformation method was used to detect line segments in the frame. “HoughLinesP” is the method that was used in this system to detect lines in frames which is a default method in OpenCV. The system used the two attributes minimum line length and maximum line gap for detecting solid lines on the road but not for detecting broken lines. For better performance of the system the “Min line length” value is increased and the “Max line gap” value is decreased. AddWeighted function in OpenCV by default also used to add weighted features into the original frame. After applying all the above-mentioned methods, a detected road marking of a road is shown in Fig.4.



Fig. 4. Detected lanes in a road

B. Traffic sign detection

The detected road markings image is the input for this stage of the process and a traffic sign detected image is the output. Several subsidiary steps are involved in the process. The system must have Haar cascade files for each traffic sign that required to detect. The methodology for implementing the cascade file is presented in Fig.5.

To train the Haar cascade files for traffic signs, positive images and negative images were used. The images which are needed to be detected in the positive image list and any images other than the subject image are in the negative images list. In this research, 65 positive images and 1975 negative images for each traffic sign were used to train the system.

When a positive image set was created, there were different sizes of images. Therefore, all the images were resized into 50*50 size. As negative images contained images of various sizes and useless images, those useless images were removed and selected images were resized to 100*100 size. A vector file was created using the list of positive image samples with some additional information for training tasks.

As the next step, by giving negative and positive data, the classifier was trained using the OpenCV function Traincascade utility.

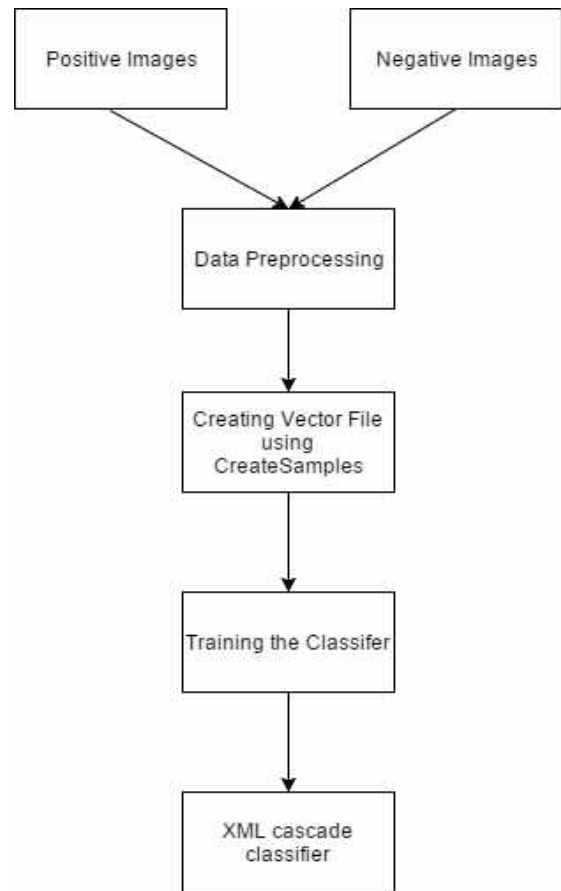


Fig. 5. Training of Haar Classifier

After generating the cascade Extensible Markup Language (XML) file successfully, it could be used to detect the traffic sign that was targeted when training the cascade file. There could be true detection and false detection. True detection needs to be higher than false detection to have efficient detections.

$$\text{True Detection} = \frac{\sum_{i=1}^N \frac{h_i}{c_i}}{\sum_{i=1}^N} \quad (3)$$

$$\text{False Detection} = \frac{\sum_{i=1}^N \frac{f_i}{N}}{\sum_{i=1}^N} \quad (4)$$

True detection and False detection were calculated from the two equations (3) and (4) given above. In (3), c_i is the number of call numbers in test images and h_i is the number of positive hit occasions. In (4), N is the number of test images and f_i is the false rate of each test image. A hit rate is several true detected call numbers in an image and a false rate is several false detected call numbers [9]. An example of a traffic sign detected image is shown below in Fig.6.



Fig. 6. Detected traffic sign

C. Vehicle detection

Vehicle detection is used to detect vehicles that are travelling in front of the subject vehicle. These detected vehicles can be either incoming vehicles or frontier vehicles that travel the same way as the subject vehicle heading. The algorithm is capable of not only detecting vehicles but also calculating distances and speeds of the vehicles relative to the subject vehicle.

Obtaining a blob from the image, loading the blob image to network, drawing bounding box and distance calculations are the main functions of the vehicle detection process.

Image frames were preprocessed to make correct predictions from the deep neural network. Mean subtraction, scaling and optionally channel swapping were performed. When implementing the network requires a weight file, a configuration file, and a text file. This text file contains all the classes that the system needs to detect.

You Only Look Once (YOLO) algorithm was used for the real-time object detection because this approach is faster and accurate. The coco data set was used in pre-trained weight files.

Drawing a bounding box was the most important part of this vehicle detection set up. The width of the bounding box was used when calculating the distance between two vehicles. Bounding boxes were represented by the four items (x, y, w, h), where x, y are the top left corner, and w, h are the width and height respectively. The detected vehicle image is shown below in Fig.7.



Fig. 7. Vehicle detection with a bounding box

Triangle similarity was utilized to calculate the distance between the detected vehicle and subject vehicle as mentioned in equation (1). After calculating the distance between two vehicles, the distance difference between frames could be calculated. Finally, the speed of the vehicle is calculated using the frame rate.

D. Web API development

Python Web API was selected as the front end for this research because it could be accessed by any application or device. Firstly, two Hypertext Markup Language (HTML) pages were implemented as the Home page and Demo page. The Home page includes a form that contained spaces for the focal length of the front view camera and web API Uniform Resource Locator (URL) which locates video stream from the camera.

The real-time video streaming by processing the input video stream from the front view camera was included in the Demo page. An alert was generated as an alarm sound when there is any danger present while overtaking another vehicle. Also, the distance from the subject vehicle to the detected vehicle and speed of the detected vehicle are shown with the real-time video stream.

The session concept was used when passing input data from the Home page to the Demo page. Also, the render template option was utilized to pass input values when submitting the button pressed by the user.

The frames that belonged to the input video transferred through road marking detection, traffic sign detection and finally, vehicle detection pipelines. Then Output frames were resized to fit into the HTML page and frames were encoded to bytes. Then those bytes were included in a NumPy array and later transformed into a byte array. Finally, the stream with context method streaming was performed.

In the flask app.py, template rendering was used to connect two HTML pages. The session was used to transfer data between two HTML pages. Stream with context method was used with mime-type as 'multipart/x-mixed-replace' to stream the processed video output. The image of the web API is shown in Fig.8 below.



Fig. 8. Python web API video demonstration page

IV. RESULTS

The automated Vehicle Overtaking System has four main sections. Even though the final section (Python web API) displays only the output of the system, the other three sections provide predictions and detections for the outcome.

In the Road markings detecting section, the main goal is to detect the solid lines on the road and informing the driver 'not to overtake at this moment'. The system does not detect broken lines on the road. If there is a broken line, the system does not generate any alert sign because it is safe to overtake the vehicle in terms of lane line rules. The proposed system with broken lines is shown in Fig. 9.



Fig. 9. Road with broken lines

If there is a solid line, then the system detects it and gives an alert to the driver that does not overtake now. The proposed system with a solid line is shown in Fig.10.



Fig. 10. Solid line detection

The accuracy levels of road markings detection in straight roads and bent roads are shown in the following table.

TABLE I. ROAD MARKINGS DETECTION ACCURACY

Road Feature	Total Frames	Detected Frames	Accuracy Percentage
Bend Road	102	88	86.27
Straight Road	110	102	92.73

The second section of the system addresses the traffic sign detection. The system detects pre-trained traffic signs such as pedestrian crossing, no entry and etc. In this study, the system was trained on traffic signs which are relevant for overtaking. If the system detects a 'trained traffic sign', the system warns the driver. The proposed system of a traffic sign detected occasion is shown in Fig.11.

Testing was done for two traffic signs such as pedestrian crossing and stop. 15 frames were used for two cascade classifier files respectively. Cascade Trainer GUI was used to testing tasks [10]. The accuracy of traffic sign detection is given in Table II.



Fig. 11. Traffic sign detection

TABLE II. ACCURACY OF HAAR CASCADE CLASSIFIERS

Cascade.xml file	Correctly detected frames	False detected frames	Detection rate (%)
Pedestrian crossing	12	3	80
Stop	14	1	93.33

The final step of the proposed system is vehicle detection and predicting overtaking possibilities according to certain conditions. All the vehicles on the frame are detected and then the distances from the camera to those vehicles were calculated and displayed on the frame. Calculated distances were used to generate alerts on vehicle overtaking. A clear road occasion is shown below.



Fig. 14. Clear road and safe

If the distance between two vehicles is less than 15 meters then the system starts to check availability for overtaking. Also, if there is an oncoming vehicle then distance should be 30 meters or more between oncoming vehicle and front vehicle. If all the conditions are satisfied, the system considers it is safe to overtake and give an alert to drivers as "safe" displayed on the video stream. Otherwise, the proposed system alerts the driver as "do not overtake" displayed and the reason is also displayed. The following Fig.15 shows an overtake occasion that alerts the driver as not to overtake with the reason.



Fig. 15. Vehicle detection & predicting

This proposed system was used previously trained weight files and configuration files of the YOLOv3 network for vehicle detection. YOLOv3 has the most accuracy and speed among other networks. YOLOv3 has 57.9% accuracy in mAP and also its accuracy is measured as 0.5 Intersection Over Union (IOU). The performance of yolo3 is graphed in the below image [11].

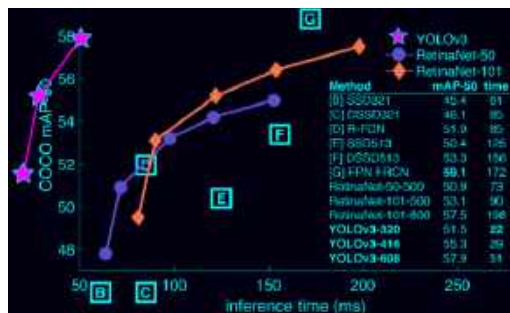


Fig. 16. Performance of yolo3 compared with other detectors.

V. CONCLUSION

The proposed system can generate alerts to warn drivers when overtaking based on road markings, traffic signs and distances between vehicles.

The method of triangle similarity was used in distance calculating. The perceived width of the vehicle was used as the width of the bounding box. There can be an error in distance calculation if the bounding box is not fitted to the detected vehicle. It can cause an error in the speeds of vehicles as well. The system was better able to detect road markings in straight roads rather than bend roads.

On some occasions, there can be inaccurate detections when the system performs traffic sign detection on input frames. As an example, when the system detects school and pedestrian crossing traffic signs, there can be wrong detections. The reason for that can be both two traffic signs include persons. More intensive training of the system on traffic signs may address this issue.

Also, while the subject vehicle is overtaking another vehicle, the front view camera can be focused on the other lane. Until the vehicle comes back to the previous lane, some detections are not valid. As an output of this model, it gives an alarm sound when there is a danger. A beep sound was embedded in the video stream that comes after processing input frames. There can be a very small time gap (less than second) between actual danger time and beep sound time. This

error can be fixed by introducing a hardware component for this task.

If a vehicle is integrated with an early warning system such as one proposed it can detect dangerous and illegal actions while driving, and generate alerts. Python, OpenCV, and Flask are the main technologies used for the development of this system. Any device or application can access this system using the Python API.

To utilize commercially, the accuracy of the distance calculation, speed detection, traffic sign detection and road markings detection needs to be improved further. Also, the system needs to be trained for all the traffic signs. The efficiency of the system can be increased by optimizing algorithms.

Ideally, all the output data must be converted into sound notifications as the drivers are not able to read the information displayed while driving.

The system assumed widths of all vehicles as a constant value. The research work can be further improved by adding capabilities to detect the width of the vehicle as well. This will result in more accurate predictions.

REFERENCES

- [1] R. Hadi, G. Sulong and L. George, "Vehicle Detection and Tracking Techniques: A Concise Review", *Signal & Image Processing : An International Journal*, vol. 5, no. 1, pp. 1-12, 2014. Available: 10.5121/sipij.2014.5101.
- [2] A. Ousuf, N. Ali, S. Anwar And T. Das, "Development of a Safe Overtaking System according to Bangladesh Highway Road", 4th International conference On Mechanical Engineering And Renewable Energy Icmere, 2017.
- [3] Gholami, Arash & Dehghani, Abbas & Karim, Mohamed, "Vehicle speed detection in video image sequences using CVS method", *International Journal of Physical Sciences*. 5,2010
- [4] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement", *arXiv preprint arXiv:1804.02767*, 2018.
- [5] Bhaskar, Prem & Yong, Suet, "Image processing based vehicle detection and tracking method." "1-5. 10.1109/ICCOINS.2014.6868357,2014
- [6] A. Rosebrock,2019 "Find distance from camera to object using Python and OpenCV." [online] PyImageSearch. Available at: <https://www.pyimagesearch.com/2015/01/19/find-distance-camera-objectmarker-using-python-opencv/> [Accessed 26 Nov. 2019].
- [7] Low, Chan & Zamzuri, Hairi & Mazlan, Saiful, "Simple robust road lane detection algorithm." "1-4. 10.1109/ICIAS.2014.6869550,2014
- [8] P. Prabhu, "Digital image processing techniques – a survey", *Golden Research Thoughts*, 2016.
- [9] Kanburoğlu, A. Buğra & Tek, F.A Haar, "Classifier Based Call Number Detection and Counting Method for Library Books.",10.1109/UBMK.2018.8566314,2018
- [10] Amin,"Cascade Trainer GUI - Amin", [Online]. Available: <http://amin-ahmadi.com/cascade-trainer-gui/>. [Accessed: 23- Dec- 2019]
- [11] J. Redmon. (2019). YOLO: Real-Time Object Detection. [online] Pjreddie.com. Available at: <https://pjreddie.com/darknet/yolo/> [Accessed 18 Dec. 2019].

Dengue mosquito larvae identification using digital images

W. D. M. De Silva*

Department of Industrial Management,
Faculty of Science, University of Kelaniya, Sri Lanka
dumidumadura@gmail.com

S. Jayalal

Department of Industrial Management,
Faculty of Science, University of Kelaniya, Sri Lanka
shantha@kln.ac.lk

Abstract: Dengue is one of the highest spreading mosquito-borne diseases in tropical and subtropical regions all over the world. This disease is mainly spread by the mosquito vector called 'Aedes'. In Sri Lanka, the number of infected patients reported is increasing, and it has become a public health problem. Health Inspectors are using different methods to reduce the spread of this viral disease and one of the main methods used is the fumigation by identifying the Aedes Larvae breeding locations. Currently, this identification is done manually by the specialized health inspectors and it is totally observer-biased and consumes a considerable amount of time, which could lead to false decision making and inefficient identification. The purpose of this research is to build an automated computational model to identify Aedes Larvae in real-time with more accuracy and convenience. Even though there are good results in previous researches done in Convolutional Neural Networks (CNN) on Aedes Larvae identification, the method of capturing Larvae Images is a bit complicated since they have used a Microscope lens of amplification capacity 60-100 times to get the magnified images. In this research, we propose the method of identifying Aedes mosquito larvae with a digital amplification of 8-12 times without using any microscope lenses attached, using ResNet50 CNN. The proposed model will identify the mosquito larvae by their genus 'Aedes' or 'non-Aedes' using a digital photo taken by a smartphone or camera in the upside of the larvae body. Hence it would help Health Inspectors, even the general public on identifying Aedes Larvae more efficiently, accurately and conveniently than the traditional method. This study shows that the trained model can identify images of Aedes and Non-Aedes Larvae separately with an accuracy of 86.65%. Furthermore, with using pre-processing techniques, the accuracy level can be enhanced to 98.76% for magnified images.

Keywords: Aedes larvae, Convolutional Neural Networks (CNN), Dengue, Larvae classification, Mosquito

I. INTRODUCTION

Dengue is one of the most dangerous and rapidly spreading mosquito-borne viral diseases in the world. Sri Lanka is one of the countries that has a higher risk of this viral disease being spread. For the year 2018, 51659 and for the year 2019, 92930 suspected dengue cases were reported (Fig. 1) to the Epidemiology Unit from all over the island [1].

Dengue (DENV) is spread from a bite of the female mosquitoes (Aedes aegypti and Aedes albopictus). The Aedes mosquito becomes infected when it bites a person who is infected with the dengue virus. Aedes aegypti has evolved into an intermittent biter and prefers to bite more than multiple persons during his feeding period.

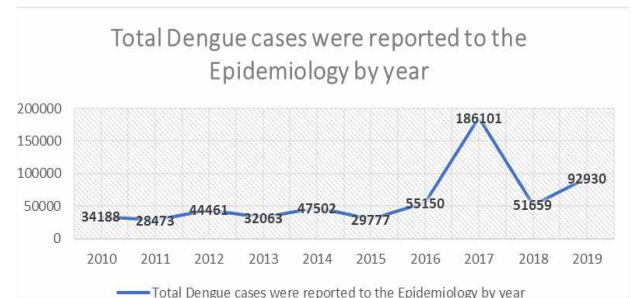


Fig. 1. Total Dengue cases reported in Sri Lanka from 2010 to 2019

This mechanism has made Aedes aegypti a very highly efficient epidemic vector mosquito. Once the mosquito is infected, it remains infected for its life span. So, it is difficult to control breeding after the mosquito starts flying. [2].

Aedes aegypti is known as a holometabolous insect. This means that the insects go through a complete metamorphosis with an egg, larvae, pupae, and adult stage. The adult life span can range from two weeks to a month depending on environmental conditions. Larval development is temperature-dependent. It can remain in the larval stage from 3 days to months, so long as the water supply is enough. [3]. So, the best stage to identify the mosquito before spreading is the larval stage since it lives as a larva for more than 4 days to months as discussed. There is a possibility to identify the larvae as Dengue-borne (Aedes) Larvae by observing the morphological key features such as the body parts, hairy or non-hairy, color etc. The easiest larvae to identify by its genus is Anopheles (Fig. 2) since it has no siphon tube in its tail. The Culex Larvae and Aedes Larvae are a bit similar from their siphon but if we observe closely, we can identify that the Aedes Larva has got a 'barrel sized' short siphon, where Culex Larva has got much longer thin siphon and has a lighter color compared with Aedes[4].



Fig. 2. Different mosquitos at larvae stage

Nowadays technology plays a vital role in almost every field but till today traditional methodologies are being used for identifying mosquitoes in both adult and larval stages in Sri Lanka.

PHI's (Public Health Inspectors) investigate Aedes Larvae breeding places, and they have the authority to fine the victims who welcome mosquitoes to breed. Those inspectors check the larvae from the naked eye and make the decision as to whether it is an Aedes Larva or not. This is totally observer biased. At some points, they take samples to their lab testing for further clarification, where they will enlarge the larvae using a microscope and identify the Genus. This is time-consuming and can be observer biased since observer inspects the siphon size of the larvae which is barrel size or longer. These inspections can make false identifications since the observer checks so many samples for a day. People get fined for the decisions taken by the observer which is totally biased for him/herself. Considering the zoological perspective, many researches try to study dengue mosquito spread by analyzing patterns and breeding places. For that purpose, they have to go to the field and collect larvae samples with the GPS location tagged. All the samples have to be tested back in a laboratory using a microscope with the guidance of an expert. This is also time-consuming and may damage the samples during transport. To eliminate these issues Information Technology can come forward to automate the identification of the larvae, in order to make it more effective and efficient. So, the main purpose of this research is to develop a model using image classification techniques such as deep learning, image processing, which will be able to identify mosquito larvae accurately and make the decisions effectively and easily. The major objective of this study is to design a model to identify Aedes Mosquito at the larval stage using image classification techniques with a smartphone camera. (using digital magnification).

II. PREVIOUS WORK

There are few researches done for identifying mosquitoes at the larval stage using Information Technology. The first study presented in [5] used texture descriptors, such as Local Binary Pattern (LBP), 2D Gabor filters and Co-Occurrence matrix, to extract characteristics and classify them using a Support Vector Machine (SVM) and K-means method using low-level features, such as Co-occurrence Matrix (CM), Local Binary Pattern (LBP), and Gabor filtered features (FG2). The results showed that SVM achieves higher accuracy, but with the research published by [6], they have shown that using AlexNet CNN they could archive better performance than SVM. In [7] they have used VGG16 CNN and with the data augmentation, they were able to archive higher performance. With the proposed system it will be able to identify the Aedes larva quickly in the fieldwork using an equipment(microscope) attached to the camera. In the study [8] they have compared different CNN architectures such as VGG16, VGG19, ResNet50, and InceptionV3) in classifying mosquito larvae images based on their classes (Aedes and Non- Aedes) as the previous studies. This study has collected enhanced images using a microscope and using image collection platforms. From the results, they have come up with the conclusion that ResNet50 is the best model to identify mosquito larvae and implement it to mobile devices or the web. The major drawback that can be seen compared with the proposed system is all those systems use a magnifier (Fig. 3) to get an enlarged system of

the larvae. So, the magnifier is a must, if a person needs to get benefited through their system.

III. METHODOLOGY



Fig. 3. Previous work is done using a microscope

The methodology of this research is to implement a suitable classification model to identify mosquito larvae using digital images taken from smartphones or digital cameras without any lenses attached. This model should be accurate and efficient. Since early identification of mosquito larvae was done using 60x-100x (Fig. 3), enlarged images were taken with the help of microscope lenses. Those microscope lenses would be costly and difficult to find in the Sri Lankan market for everyone.

Therefore, this study is to identify whether it is possible to classify mosquito larvae only using digitally zoomed images. For that, a thorough comparison is done between images collected with 60x-100x enlarged images using an equipment attached to the camera, and digitally zoomed images taken without any equipment attached under the same condition.

As shown in Fig. 4 these images will be sent through pre-processing steps to find out the best suited pre-processing technique that can be used to increase the accuracy of Aedes mosquito larvae identification. The main steps of the methodology are as shown in Fig. 5.

According to the previous research findings mentioned, this study has used the ResNet50 Model of CNN for the classification of mosquito larvae since it has performed as the best in previous studies.

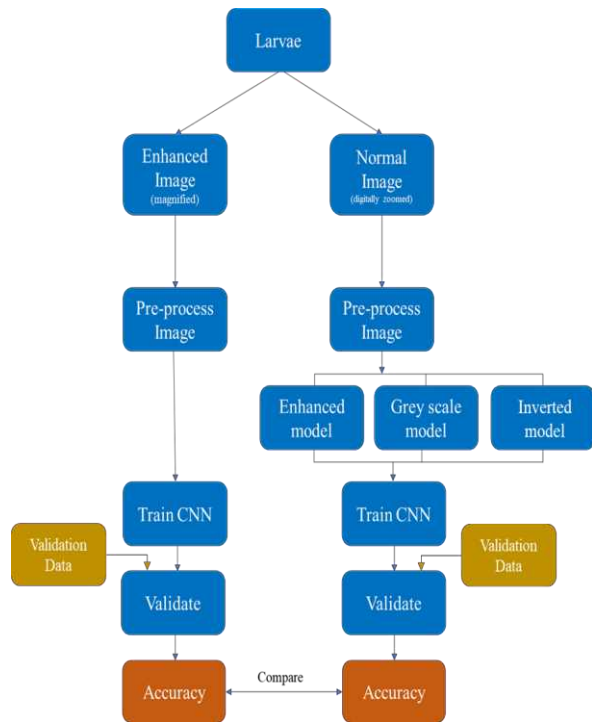


Fig. 4. Overall Methodology Process

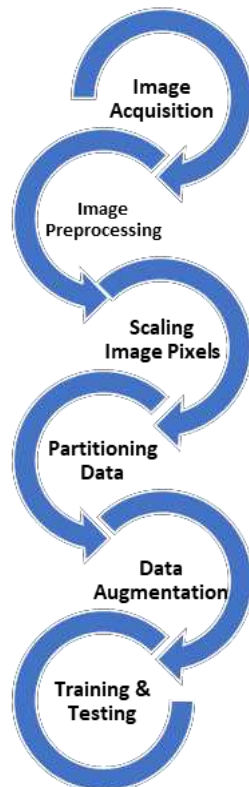


Fig. 5. Methodology Process

A. Image Acquisition



Fig. 6. Types of Images Collected

In this phase, sample larvae images were collected, which are required to train and build the classification model. There are 19 genera of mosquitos available in Sri Lanka and Aedes is a one popular genus available in Sri Lanka [9]. Out of those, the most available 3 genera were selected and categorized into Aedes and Non-Aedes. Each of the mosquito samples was selected from different regions in Sri Lanka where mosquito diseases are most populated like Gampaha, Colombo, Kalutara, since high dengue cases are reported in these districts in 2019 according to the statistical data. [1] Since the target is to identify Aedes mosquito larvae out of other larvae, these images are labelled as Aedes and Non-Aedes with the expert guidance [8]. Larvae from all 4 stages are captured in a white background. These images are taken using 12 mega-pixels mobile phone digital camera with 8x digitally zoomed and with x60 magnifier attached. We collected 90 images of Aedes and 70 of Non-Aedes images without using any microscope and 130 of Aedes and 108 of Non-Aedes Images using x60 microscope for both training and testing the classification model. 10 random images from every 4 types were chosen to validate the model (See Fig. 6). Images were taken from different angles, under different environmental and lighting conditions. The reason is, if this will be developed for a mobile application, the possibility of uploading larvae images using mobile phone will be very high. Then the mobile application should identify the larvae vector accurately. The images were saved in standard JPG format. Images shown in Fig. 6 are the categories of mosquito larvae that were collected under the data acquisition. Since the target task is a binary classification i.e. to find whether it was an Aedes Larvae or Non-Aedes Larvae, the collected data were categorized into two classes which are 'Aedes' and 'Non-Aedes'

B. Image Pre-processing

After collecting the images, image processing was done to improve the image quality. Then the images were resized into 224x224 pixels because the processing takes more time when the image size is too large.

As the next step to reduce the background of the images we implemented two automatic cropping algorithms which are 'crop to fit' and 'crop to square' methods. Crop to fit method will identify the larvae body max-contour area and crop the image into it where 'crop to square' will identify the max-contour axis and crop to it by creating a square (Fig.7).

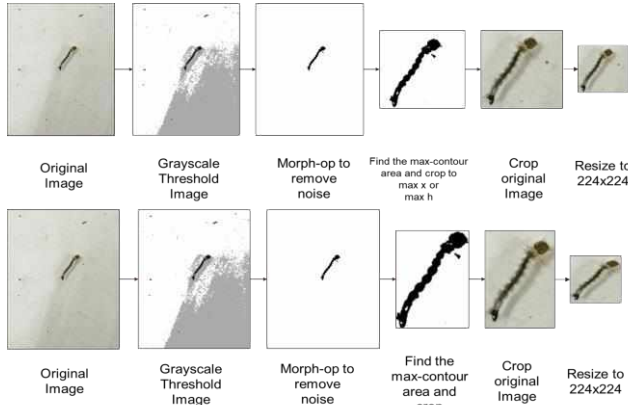


Fig. 7. Cropping Methods

CNN uses convolution layers to extract the patterns of the images in source image pixels. Therefore, enhancing image features may increase the CNN learning [10]. First, the original image is enhanced with added contrast and brightness to improve the larvae of the image from the background. Then the image is cropped for the target to avoid distraction from the background for the training. The same dataset with 3 channels (RGB) is converted to 2 channel grayscale data set and those grayscale images are bit-wise inverted as a further step. Those 3 types of pre-processed data are used to validate the accuracy and efficiency of the proposed approach because one of the research objectives is to find the most suitable pre-processing technique for identifying mosquito larvae.

I. Grey-scale model

Greyscale color pattern is a set of grey shades without apparent color. Black is the darkest shade available, and white is the lightest shade possible. This range of the models is measured from 0-1. Black is represented with 0, and white with 1.

II. Inverted Grey-scale model

Inverted Grey-scale color model is an inverted range of shades of grey without apparent color. The darkest possible shade is white, and the lightest possible shade is black. This model range is represented from 0 - 1. Black is represented by 1 and white is represented by 0.

III. Enhanced model

An Image can be enhanced by Histogram Equalization method. Histogram Equalization is a computer image processing method used to improve image contrast. This is achieved by distributing the most regularly distributed intensity values widely, i.e. stretching out the intensity range of the image. This can usually increase the global contrast of images when its usable data is represented by close contrast values. This allows for areas of lower local contrast to gain a higher contrast. which means the gap between rich intensities and poor intensities increases with this model.

C. Scaling image pixels

Data is inserted into CNN as a matrix of numeric values. Therefore, image pixel intensities are converted into numeric values in the range of 0-1. As the first step of scaling

image pixels, all the pixel intensities are converted into numeric values. Then those numeric values are divided by 255 to range the intensities to 0-1.

D. Partitioning data

In this study, data is divided into training, testing and validation datasets. 10 random images from each class will be taken into the validation dataset where those will be used for validating the model trained. The remaining data will be split into training and testing datasets with 80%, 20% of percentages (Fig. 8) [11].

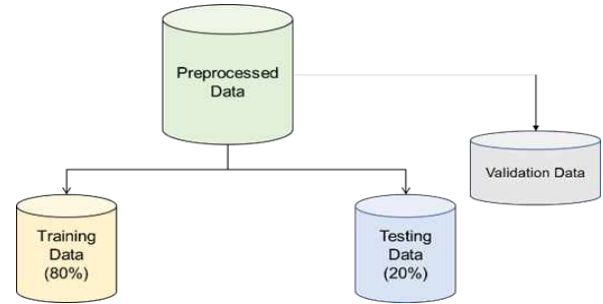


Fig. 8. Data Split Procedure

E. Data augmentation

Data augmentation is a technique to artificially produce new data for training using available data. Image data augmentation is done by applying domain-specific methods and techniques to the example from the training data that creates new and different training examples. This is one of the most well-known mechanisms in augmentation of data and involves creating transformed versions of images in the training dataset, that belong to the same class as the original image.

Transforms are done with a pool of operations from the field of image manipulation, such as width-shift, height-shift, flips, zooms, rotate and many more.

The intention is to expand the training dataset with new augmented images which means, variations of the training set will be high that is likely to be seen by the model. For example, a vertical flip of a picture of a larva may make sense, because the image could have been taken when the larvae was in its upside-down position (See Fig.9).

As per the study [12], they have shown that using a different variety of algorithms for data augmentation can bring huge potential to improve training data-hungry deep learning algorithms such as CNN.



Fig. 9. Data Augmentation

F. Training and Testing

Training testing dataset size and the method of training classification have to be determined. There are many different approaches for supervised learning, such as SVM, ANN and CNN. Out of them, literature has shown that the CNN method performs best when it comes to image classification. Therefore, CNN is used as the classification method in this study. Basically, CNN trains its network by feeding its nodes with images' intensities and identifying their highly specific features. In the training phase, we used the CNN with the architecture called ResNet50. It was decided to use this architecture because of the network configurations and because this model has scored best in terms of performance, lower size, training time and accuracy [8]. Furthermore, ResNet50 has been ranked as the best model to be employed into a mobile application or a web application. As mentioned earlier, out of 160 images 128 were added to the training, and for the zoomed image collection out of 238 images, 190 images were added. After the training and testing using CNN, each model's accuracy is calculated according to the following formula, with the validation dataset.

$$\text{Accuracy (\%)} = \frac{\text{avg Aedes}(p) + \text{avg Non-Aedes}(p)}{200} \times 100\% \quad (1)$$

Aedes(p) :- accuracy of Aedes larvae identification
Non-Aedes(p) :- accuracy of non-Aedes larvae identification.

The equation (1) is chosen to finally test and validate the data models using validation data, verifying the trained model in Aedes and Non-Aedes identification accurately. Here, the accuracies of both Aedes and Non-Aedes Larvae is summed and divided by 200 since both accuracies will be given in percentages. Finally, the total will be multiplied by 100% to get the total accuracy percentage of the model.

According to those accuracies, the study will show whether it is possible to identify mosquito larvae by their genus using deep learning with accurate results.

IV. EXPERIMENTAL RESULTS

Images of larvae were collected using an attached microscope to the camera and using digital zooming without any microscope in the same condition. As mentioned in previous literature, we were able to archive the mentioned accuracy level of 97% using the collected dataset for the magnified images when the model was trained. The collected data without using a microscope was trained using CNN without using any pre-processing techniques and we were able to archive 77.13% accuracy from the trained model. As in Fig. 10 it can be identified that the model learns by reducing the training loss and increasing the training accuracy.

Next, the same dataset was sent through two cropping algorithms which are 'crop to fit' and 'crop to square'. The trained models showed 62.8% and 67.5% accuracy respectively for both cropping methods, where it was observed that the accuracy is falling after cropping. After that, the dataset was sent through pre-processing methods such as the Enhanced model, Grey model, Inverted Model.

Out of those 3 models, the best identification was archived using Enhanced Model for the pre-processing which

is **86.65%** (See Table 1). Grayscale and Inverted Grayscale models archived 53.6% and 58.08% of accuracy respectively.

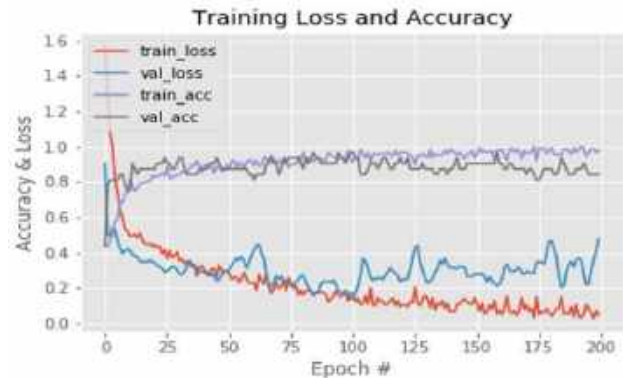


Fig. 10. Training Loss and Accuracy Graph for Model trained without cropping and pre-processing

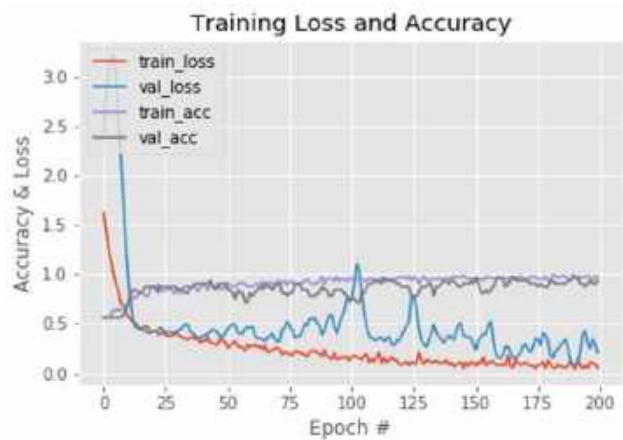


Fig. 11. Training loss and accuracy graph for model trained after using an enhanced model

TABLE I. MODEL ACCURACIES FOR THE VALIDATION OF IMAGES WITHOUT USING MICROSCOPE

	Aedes	Non-Aedes
Sample1	100	99.97
Sample2	100	99.99
Sample3	99.90	99.91
Sample4	100	53.74
Sample5	100	False-Prediction
Sample6	100	100
Sample7	100	60.4
Sample8	100	69.4
Sample9	100	53.44
Sample10	99.8	96.54
Average Accuracy (%)	99.97	73.34

Furthermore, the enhanced pre-processing technique used, tested the dataset of enlarged images of larvae which were taken with a camera attached to the microscope. The trained model showed that the accuracy was increased to 98.76% from 97.62%. The training loss, accuracy for the

model and the and the experimental results are shown in Fig. 12 in Table II respectively.

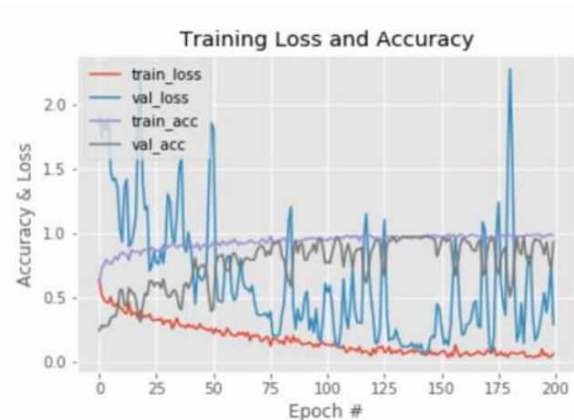


Fig. 12. Training Loss and Accuracy Graph for the Model trained after using enhanced model for a dataset with magnified images (using a microscope).

TABLE II. MODEL ACCURACIES FOR THE VALIDATION OF IMAGES USING MICROSCOPE AND ENHANCED PRE-PROCESSING MODEL

	Aede	Non-Aedes
Sample1	99.87	96.86
Sample2	99.99	99.85
Sample3	96.59	99.22
Sample4	99.72	98.99
Sample5	99.95	99.72
Sample6	99.99	92.07
Sample7	99.96	99.98
Sample8	99.92	99.98
Sample9	94.15	99.41
Sample10	99.21	99.28
Average Accuracy (%)	98.93	98.53

V. CONCLUSION AND FUTURE WORK

In this study, the ResNet50 CNN was used to create the models, which were built according to Aedes and Non-Aedes Larvae in their different stages using a 60X microscope and digitally zoomed images. The trained model shows that this can identify a Dengue carrying mosquito larvae (Aedes) with 77.13% of accuracy using only digitally zoomed images without any magnifier attached to the camera. Furthermore,

we found that using pre-processing techniques such as enhancing the image, the accuracy level is raised for both datasets of mosquito larvae images.

Since this identification does not need any third-party equipment, this can be implemented in a mobile app for crowdsourcing from people to identify dengue breeding places in the country and use it for predictions.

This research can be extended to identify multiple larvae using only a single image. Therefore, the user can identify the larvae in the same breeding place and can be applied to identify other tiny insects in there in small stages as well.

REFERENCES

- [1] Epidemiology Unit, Ministry of Health, "Dengue Update". [Online]. Available: http://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=171%3Adengue-update&catid=51. [Accessed: 17-06-2019].
- [2] World Health Organization, "Dengue control; Dengue/Severe dengue frequently asked questions - Who spreads dengue and severe dengue?". [Online]. Available: <https://www.who.int/denguecontrol/faq/en/index5.html>. [Accessed: 13-09-2019].
- [3] DengueVirus.net, "Life Cycle of Aedes aegypti". [Online]. Available: <http://www.denguevirusnet.com/life-cycle-of-aedes-aegypti.html>. [Accessed: 12-09-2019].
- [4] WRBU, "Features". [Online]. Available: http://www.wrbu.org/keys/PO_AE_L/Aedes_Oriental_PACOM_L.html. [Accessed: 03-09-2019].
- [5] Z.G. Nonoal, A. S. Ortiz, A. A. Jalife, M. N. Miyatake, "Comparacion de descriptores de imagenes para la clasificacion de larvas de mosquito". *Instituto Politecnico Nacional., CAIPAt'2017*.
- [6] A. S. Ortiz, A. F. Randilla, A. A. Jalife, M. C. Hernandez, M. N. Miyatake, "Mosquito Larva Classification Method Based on Convolutional Neural Networks". *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Cholula, Mexico, 2017.
- [7] A. Ortiz, M. Miyatake, H. Tünnermann, T. Teramoto and H. Shouno, "Mosquito Larva Classification based on a Convolution Neural Network." *PDPTA'18 - The 24th Int'l Conf on Parallel and Distributed Processing Techniques and Applications*, Nevada, USA. 2018.
- [8] S. Asmai, M. Zukhairin, A. Jaya, A. Rahman and Z. Abas, "Mosquito Larvae Detection using Deep Learning." *PDPTA'18, International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Nevada, USA. 2018.
- [9] N. Gunathilaka, "Annotated checklist and review of the mosquito species (Diptera: Culicidae) in Sri Lanka," *J. Insect Biodiversity*, Magnolia Press, 2018.
- [10] "How do I increase accuracy using Convolutional Neural Networks (CNNs/ConvNets) for regression?" [Online]. Available: <https://www.quora.com/How-do-increase-accuracy-using-Convolutional-Neural-Networks-CNNs-ConvNets-for-regression>. [Accessed: 13-09-2019].
- [11] Z. Reitermanova, "Data Splitting," in *Proceedings of WDS'10 Contributed Papers*, vol I, 2010, pp 31–36.
- [12] Mikołajczyk and A. Mikołajczyk, M. Grochowski, "Data Augmentation for improving deep learning in image classification problem", *IEEE* 2018,

A modified cognitive complexity metric to improve the readability of object-oriented software

Thilini Jayalath*

Faculty of Graduate Studies and Research,
Sri Lanka Institute of Information Technology, Sri Lanka
thilini.j@slit.lk

Samantha Thelijagoda

SLIIT Business School
Sri Lanka Institute of Information Technology, Sri Lanka
samantha.t@slit.lk

Abstract: Complexity of software can be identified as a term which expresses the difficulty level of reading, understanding, maintaining and modifying the software. This helps to the quality improvement of the software and maintenance process of the software through a long time period without any obstacle. Therefore, software complexity metrics have been introduced to calculate the complexity of a software using numerical values. While there are number of metrics which calculate the complexity of object-oriented programs, they only consider one or two object-oriented concepts. As a result of that, there is no single metric which has the capability of measuring the complexity of a program based on multiple object-oriented concepts. This research aims to build a new metric to evaluate the complexity of an object-oriented program in order to improve the readability. The new metric has been built based on the influence of previous object-oriented metrics and some disregarded factors in calculating the complexity. In order to evaluate the new metric, Weyuker's properties and Briand's properties are used. The new metric acquires most of the object-oriented concepts in calculating the complexity and helps to improve the readability of the software as well. In fact, it makes it easy to handle the maintainability, reusability, portability and reliability of the software, when readability is high. This will result in increasing the overall software quality.

Keywords: Basic Control Structures (BCS), Cognitive complexity, Object-Oriented Software

I. INTRODUCTION

Many users frequently say that their applications are difficult to comprehend, hard to be maintained and complex. Thus, quality of the application and quality of the process of building the application were introduced. Then, controlling the quality of the software as well as the development process of the software became very important processes in software engineering. To control this product quality and development process quality, engineers needed a way of measuring the complexity of the software.

There are number of definitions introduced to describe the software complexity. Some of them are as follows.

- Harmeet and Gurvinder have described the software complexity as the primary factor of performance, reliability and cost of the software. Software complexity has a major effect on the required effort to identify the requirements, design, implement, test and maintain the system during the software life cycle [1].
- Madi, Zein and Kadry state that the software complexity can arise anywhere inside Software Development Life Cycle (SDLC) such as, analysis

phase, requirement gathering phase, design phase and implementation phase. They refer the complexity of a software as an undesired property which makes the software tougher to comprehend, therefore harder to be modified [2].

A single definition cannot be used to describe the term software complexity since it is a multi-dimensional term in software. Different researchers have used different definitions and views about software complexity. Therefore, authors have decided to describe the software complexity of a system as the difficulty in understanding, changing and maintaining the system [3].

In order to measure the complexity and express the complexity using numeric values, software metrics have been introduced to the software industry. The beginning of publishing the research papers on software metrics commenced in 1968. R.J. Rubey and R.D. Hartwick published a paper titled "Quantitative Measurement of Program Quality" at the ACM National Conference, Las Vegas in August 1968. Thereupon the community of software industry always works on proposing new software measures and modifying existing ones to calculate the software complexity in a more accurate way. Program complexity depends on a large number of factors. Number of inputs and outputs are very critical in measuring program complexity as well as BCSs and cognitive weights. When a program has a large number of inputs and outputs, it makes the program more complex. In fact, a programmer or reader needs to pay more attention to remind those attributes and their I/O processes. Human effort which is needed to perform a task is also a very important when it comes to measuring the complexity. That is called cognitive complexity measures.

The challenge in programming can be defined in terms of the way to build the logic, not how to describe the data. But object-oriented programming paradigm is more concerned about managing the objects rather than the logic required to manage them. The first step in object-oriented programming is to determine each and every object that the developer wants to manage and identify the relationship among them. Object can be specified as an autonomous entity which contains both data and procedures to manage the data. A class can be called as a prototype from which objects are built and that explains the details of related object. A class comprises three things: the name, attributes and operations.

To endure a software/system in a long time it should be properly maintained. For that, the program code of the software/system should be understandable and readable. Then the program code can be changed according to the available and considerable reasons. In fact, the understandability and

readability features of the software are very important for the maintenance phase of the object oriented (object-oriented) software life cycle.

II. LITERATURE REVIEW

IEEE defines a metric as “a quantitative measure of the degree to which a system, component, or process possesses a given attribute” [4]. Quantitative measurements are very important in all sciences, computer science practitioners and theoreticians make a continuous effort to bring similar approaches to software development [5].

Object oriented programming is wrapped with three main factors called, objects, classes and methods (functions). In 1999 Buckley, Layzell and Douce introduced a set of metrics that help in calculating the complexity of a given system or program code based on the object-oriented concepts such as the object and class. All those metrics were based on the spatial abilities which measure the complexity by calculating the distances between the program elements in the code. Spatial abilities benefit in order to read, understand and remember the program. Therefore it is easy to maintain the software. Authors proposed new metrics set which includes [6];

- A metric to measure the complexity of a distinct function [Function Complexity-FC]
- With the guidance of FC, measure program complexity where a program consists of one or more than one function. [Program Complexity-PC]
- A metric to calculate the complexity of a recursive method [RFC]
- Three metrics to measure object-oriented concepts
 - Method Location Rating [MLR]: Here consider the count of lines in between the method definition and class declaration.
 - Class relation measure [CRM]: Here consider the number of lines between the parent class and child class.
 - Object relation measure [ORM]: Here consider the number of lines between the object declaration and its class declaration.

These three metrics require a simple calculation process and it helps to measure the understandability of the source code. Other than that, MLR, CRM and ORM benefit in estimating the cost and time needed. But there are some drawbacks in these metrics, such as;

- Depending on the language
- Having minor barriers when applying some of those metrics in few incidents like; CRM cannot apply where class definition is unavailable inside the code and ORM cannot apply where the definition of object is not inside the available code
- Requiring considerable theoretical basis and implementation technological knowledge
- Not supporting to all object-oriented programming concepts

- Not considering the complexity arisen due to attributes of the class

In 2004 Kumar, Singh and Aggarwal have accomplished to propose two metrics [7], which consider the object-oriented concepts, such as polymorphism and encapsulation that help to measure spatial abilities. They attempted to measure the spatial complexity based on two categories as shown in Fig. 1.

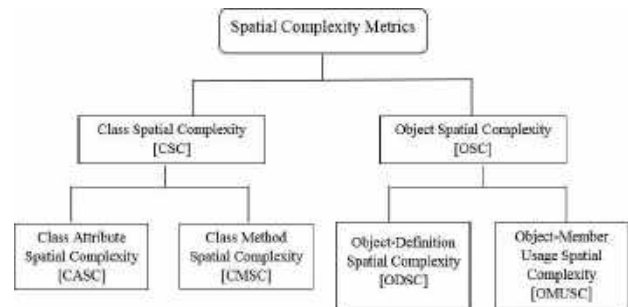


Fig. 1. Categories of spatial complexity metrics

These measures benefit in increasing program readability, measuring maintainability, evaluating the suitability of the class data members, improving and comparing the class cohesiveness and understandability. Based on previously mentioned metrics, these two also depend on the language and considerable theoretical basis and implementation technology knowledge. Since calculating CSC and OSC has a lengthy calculation process it is hard to calculate. Object oriented system can be a collection of objects, classes and methods.

In order to find the complexity of an object-oriented system, it is very essential to measure the complexity of entire system rather than measuring complexity of a particular object or class. Therefore, in 2007 Sanjay Misra [8] proposed a contemporary measure to calculate object-oriented system complexity. The following three steps can be used to calculate complexity using new metric proposed.

- Measure the complexity that arise due to operations upon each and every object
- Enumerate the complexity of individual object or class
- Lastly, calculate full code complexity (in entire object-oriented system)

In order to calculate the class complexity (CC), first calculate the complexity of individual classes based on cognitive weights of BCSs. Then calculate the total Class Complexity (CC) by adding the class complexities of all classes that exist inside the entire code (system) compared to CSC and OSC measures, calculating CC is little easy. In fact, it does not require much theoretical basis and implementation technological knowledge and accomplish the requirements that are needed for an acceptable and good metric. CC is language independent and a robust measure. It can be used to evaluate understandability and efficiency of the code. But when it comes to very lengthy source codes it can be somewhat hard to calculate. Most of the metrics proposed earlier have the following drawbacks.

- Absence in consideration of some object-oriented programming concepts and principles of measurements when designing the metrics
- Depending on technologies of implementation
- Not accommodating any convenient approach to measure the entire class/code complexity

By considering all those shortages Misra and Akman have introduced a new metric to measure the entire class/code complexity followed by an object-oriented approach in 2008 [9]. This measure is based on cognitive weights and the most important fact is this metric has considered both the complexity arisen as a result of the data members (attributes) and complexity arisen considering the function operations in the class/entire code. Calculating the total WCC does not depend on the language and it conjectures the effort of maintenance. WCC calculates the complexity of methods with regard to the messages and operations in those. It also acknowledges the method internal architecture which is called method complexity. While these being the advantages of WCC, disadvantages, on the other hand, are somewhat hard to calculate when it comes to lengthy codes and not supporting all object-oriented programming concepts. Since this method provides complexity values in a numerical format, it can be a large number for lengthy codes. High values of complexities are undesirable.

In 1999 Douce, Layzell and Buckley found set of spatial metrics like function complexity (FC), program complexity (PC), object relation measure (ORM) and class relation measure (CRM) to calculate the spatial abilities of object-oriented program [9]. But there was no consideration of architectural complexity (cognitive weights). Due to that in 2009 Gupta and Chhabra found new metrics to calculate the complexity of a program with regards to two aspects [10];

- Spatial aspect – using LOC
- Architectural aspect – using cognitive weights

Here Gupta and Chhabra have categorized the cognitive spatial complexity metrics of object-oriented software as Fig. 2.

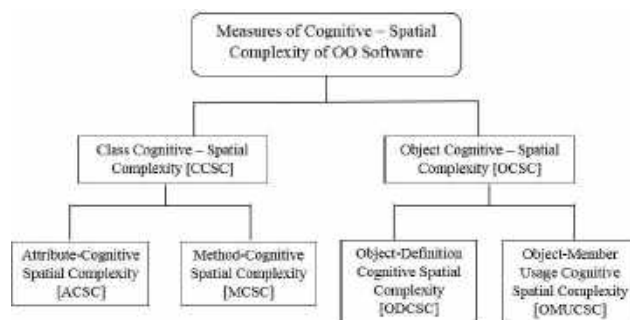


Fig. 2. Categories in measures of cognitive-spatial complexity of object-oriented software

This is a very good measure to indicate the cognitive effort needed to understand the program. Disadvantages of these metrics are; language dependency, difficulty that arises when calculating since it has a lengthy calculation process and it needs a considerable theoretical basis and technology knowledge.

In 2011, Koyuncu, Misra and Akman introduced a new measure to calculate the full code complexity considering the inheritance positions and the interior architecture of the code since the traditional metrics which tried to calculate the complexity of the object-oriented programming codes had the following issues [11]

- Not acknowledging the interior architecture of the code and object-oriented concepts
- Not considering cognitive aspects
- Not being able to cover some theoretical and mathematical approaches
- Not being hugely based on implementation technology
- Not being able to measure the absolute complexity of a code acknowledging the inheritance positions and cognitive aspects in sync.

This measure has become prohibitive to appoint upper and lower bounds for complexity values. Since this method provides complexity values in the numerical format, it can be a large number for very lengthy codes. High values of complexities are undesirable. As a result of these issues Chhillar and Bhasin came up with a metric called new weighted complexity metric (WCM) in 2011 [12]. Chhillar and Bhasin acknowledged the following factors as Fig. 3, in order to propose a new metric.

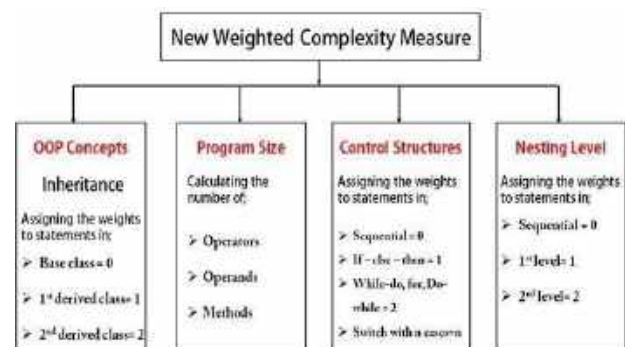


Fig. 3. Considered factors for WCM

In 2015, with confiding these circumstances Hussien, Jararweh, Shehab, Alandoli and Tashtoush came up with a new measure which concerns six elements that causes emergence of the program/system complexity [13]. Those six elements are indicated in Fig. 4. Out of these six, first two elements are extracted from the previously proposed approaches. In generating flow charts these authors followed the footsteps of Thomas J. McCabe in developing flow charts in Cyclomatic Complexity (CC) but assigned the weights as different from CC but same as Shao and Wang's measure. When calculating the operations in the program authors adopted the approach of Halstead volume. Other four factors are proposed by the authors in order to boost the accuracy of the complexity calculation.

A very strong advantage of the new measurement is that it depreciates the detriments of previous metrics and the disadvantages are that it does not consider the complexity that arises due to the objects created in the class and thus requires considerable theoretical basis and implementation technological knowledge.

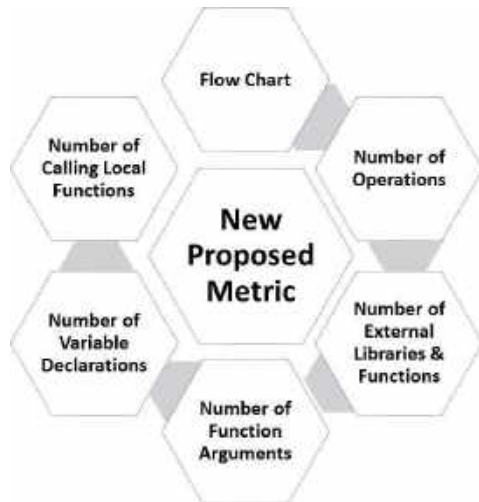


Fig. 4. Factors considered in the new metric

Sanjay Misra et al. (2018) published a paper on Object Oriented Cognitive Complexity Metrics and the authors have proposed a metric suite which covers some important features of object-oriented programming [14]. They have analyzed the literature and extracted some distinct features for the new metric. There are five essential metrics that exist within this suite;

- Class Complexity (CLC)
- Attribute Complexity (AC)
- Message Complexity (CWC)
- Method Complexity (MC)
- Code Complexity (CC)

This proposed metric suite considers the object-oriented concepts like inheritance, coupling and cohesion.

In 2019, Hussam, Hiba and Tharmer came up with a new model called The 2O2C model to calculate the complexity of object oriented programs [15]. The authors have acquired some attributes from the literature including: NOC, DIT, COM and CBO. Other than these four, detailed class complexity and weighted class complexity have been considered by the proposed metric model. 2O2C metric model mainly focuses on attaining consistency, reusability, readability, extensibility, maintainability and understandability. Abstraction and encapsulation are the main two object-oriented concepts that were considered in this metric model.

As mentioned, the count of operands and operators, control structures, cognitive weights, nesting levels and object-oriented programming concepts like inheritance, polymorphism, and abstraction should be considered during the complexity calculation of an object-oriented program. Previously found metrics have not been able to consider all these parameters together in calculating complexity of an object-oriented program.

III. PROPOSED WORK

Even though there are number of metrics which calculate the complexity of object-oriented programs, they only consider one or two object-oriented concepts for the metric. As a result, there is not any single metric which has the capability of measuring the complexity of a program based on most of object-oriented concepts. This research is to build a new metric called Modified Cognitive Complexity Metric to calculate the complexity of a given object-oriented program considering more concepts of object-oriented programming such as,

- Counting ELOC for the size attributes other than operators, operands and strings
- Polymorphism and Abstraction
- Coupling and Cohesion
- Inheritance
- Encapsulation

A. Counting ELOC for the size attribute

Here, in order to calculate the size of any program, the authors have considered the calculation process of a size attribute of WCM [12]. Size is a very strong factor in measuring complexity in a system or module, because it is very simple to understand that the programs which are massive in size are more difficult to comprehend than the programs which are small. In fact, the size attribute plays a critical role in measuring complexity. There are number of ways to measure the size of a program/ module:

- By directly counting Lines of Code (LOC)
- By counting the number of operators, operands and strings in each program
- By counting the number of methods and classes in each module
- By measuring the distance between variable and method declarations and usages.

When it comes to counting the number of operators, operands and strings in each program, it requires the user to know the difference between operators, operands and strings. In addition, this approach requires the user to go through each element in the program which is highly time-consuming. Counting the number of methods and classes is not a very strong method in measuring the size of the program, because there can be modules which contain only one method, but that contains many LOC which is more complex. Measuring the distance between variable, method declarations and usages also have several disadvantages such as, the user is expected to remember the position of a declaration of a certain variable and the positions (calling statements) of using that variable or method. Thus, this method is also very tedious and time consuming. In comparison, directly counting the LOCs approach is very much simpler and less time consuming as the user is just required to count the number of lines of code in a certain block of code. Thus, the author decides to measure the size of a program by counting the LOC in it. But there are various ways of counting LOC of a program. Those can be described as follows.

- Number of physical lines – Counting the lines of the source code of program including comment lines

- Number of blank lines – Counting only the blank lines in the given program
- Number of logical lines – Counting only the lines which are executable statements
- Number of eLOC – Counting the lines of the source code of program excluding blank, comment lines and lines which include only parenthesis.
- Sum of lines of code – Getting the sum of physical lines and blank lines
- Number of executable physical – Getting the total lines of source code excluding the blank lines and comment lines
- Number of executable logical – Counting the number of executed statements inside the given program or module
- Number of comment lines – Counting the number of comment lines
- Number of comment words – Getting the count of comment words in the program

Since eLOC eliminates lines which contain only a bracket and count every other line excluding comments and blank lines, eLOC can be taken as the replacement for size attribute. Since counting eLOC is easier than counting operators, operands and strings, that can be proposed for the size attribute. In order to calculate the size of any given program, counting the number of eLOC can be considered in Modified Cognitive Complexity Metric (MCCM).

B. Polymorphism and abstraction

Polymorphism is a main concept of object. This feature can be called as when one method has multiple implementations, for a certain class of action. The most popular way in using polymorphism in object-oriented programming, where a child class object is referred by a parent class reference. Polymorphism can be further described by considering its two main parts:

- Static polymorphism [Overloading] - Static polymorphism also called early binding, compile time binding, because it happens during the compile time. This feature says that a class can have more than one method with the same method name, if their argument lists are different. When someone is referring to a code, this object-oriented feature makes it hard to understand the code for the person. Thus the person may get confused to identify the method that called.
- Dynamic polymorphism [Overriding] - Dynamic binding is also called late binding and runtime binding, since it happens during the run time of a program. Overriding of a program refers when implementing a method in subclass which is already present in the relevant parent class. A person who will read the code may get slightly confused with this situation.

A class can be called as an “abstract class”, when it contains one or more abstract methods. When a method is declared without any implementation details, it is called as “an abstract method”. Abstract classes may not be instantiated and

require subclasses to provide implementations for the abstract methods. It is somewhat hard to understand an abstract method than a normal method. This is because in parent class there are no any implementations for abstract methods. There can be different implementations for different child classes. Taking these three problems into account, in order to measure these complexities, a new factor can be proposed, which is Wo. The above two situations can be considered by adding value of one to the Wo factor for a statement which calls an overloaded or overridden or abstract method. Other than those two situations, in overloading it should add value of one to the Wo factor for a statement which is an overloaded method declaration excluding the first method declaration.

C. Coupling and cohesion

Coupling and cohesion interact with the quality of an object-oriented design. Commonly, a good object-oriented design must be highly cohesive and loosely couple. This type of system is easy to develop, add new features, maintain and is less fragile.

Coupling can be described as the relationships between modules. A decrease in interconnections between classes (or modules) is therefore accomplished via a decrease in coupling. Coupling is categorized into many types according to the reasons that can arise between modules. Those can be listed as in TABLE I, in the order of complexity (Lower to highest) Cohesion is a measure that can be used to specify the degree to which a class has a well-focused or single purpose. Cohesion mainly emphasizes how a single class is designed. A better object-oriented design holds a high cohesion. There are 9 levels of cohesion as in TABLE II (Better to worst).

In order to acknowledge the complexities of above categories of coupling and cohesion, a new attribute, Wcc should be introduced, where the value of Wcc can be taken from the TABLE I and TABLE II.

TABLE I. SUGGESTED WEIGHTS ACCORDING TO COUPLING TYPES

Coupling Type	Weight [Wcc]
Data coupling	1
Stamp coupling	2
Control coupling	3
Common coupling	4
Content coupling	5

TABLE II. SUGGESTED WEIGHTS ACCORDING TO COHESION TYPES

Cohesion Type	Weight[Wcc]
Functional	1
Informational	2
Sequential	3
Communicational	4
Procedural	5
Temporal	5
Logical	5
Coincidental	5

D. Inheritance

Chhillar and Basin computed the complexity that arises due to different levels (nesting levels) of inheritance of classes

using W_i . But in object-oriented programming there are different types of inheritance (Fig. 5) which can be identified. That situation is not acknowledged in WCM. The different types of the inheritance can be listed as follows:

- Single inheritance: A class can extend only a single class.
- Hierarchical inheritance: One single base class can create more than one derived classes.
- Multilevel inheritance: One derived class is created from another derived class.
- Multiple inheritance: One class inherits from more than one base class.
- Hybrid inheritance: This is a mixture of any of the above inheritances (single, hierarchical and multilevel) which can be called as hybrid inheritance

Based on the complexity level of the inheritance types, weights are assigned (TABLE III). Here in the traditional calculation process of WCM, it allocates a weight for class inheritance level as; level 1 = 1, level 2 = 2 and level 3 = 3, but it does not consider about these inheritance types. In order to acknowledge both these two factors, W_i attribute can be used as follows.

W_i = weight due to inheritance level * weight due to inheritance type

TABLE III. SUGGESTED WEIGHTS ACCORDING TO INHERITANCE TYPES

Inheritance Type	Weight [Wty]
Single	1
Multilevel	1
Hierarchical	2
Multiple	3
Hybrid	4

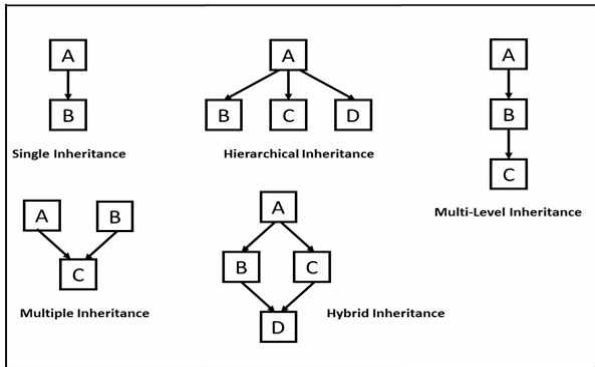


Fig. 5. Different types of inheritance

E. Encapsulation

The process of packing the variables (data) and code performing on the variables (functions) together as a single unit is called as encapsulation in object-oriented programming. In here the variables declared in a one class will be not exposed to any other classes and cannot be accessed by any other function which is not in the current class. Because of that this concept is also called data hiding. In order to

achieve data hiding in object-oriented programming, following two ways can be used.

- Declaration of variables in a class should be private.
- Deliver public getter and setter functions to view and change the variable values.

To consider this concept in calculating complexity, authors have decided to introduce a new variable called W_e . The statements which declare the variables as private and the statements which contain the public getter and setter functions declarations should allocate a value of one for the W_e attribute.

Assuming these 5 factors total weight of a single line can be proposed as follows.

$$W = W_i + W_c + W_o + W_{cc} + W_e \quad (1)$$

W_i - Weight due to the inheritance, W_c - Weight due to the type of BCS * Weight for nesting level of BCS, W_o - Weight due to abstraction and polymorphism, W_{cc} - Weight due to coupling and cohesion, W_e - Weight due to encapsulation.

Then considering the line by line calculation for the weights, complexity of the entire code can be taken by following formula;

$$MCCM = \log \left[\left(\prod_{n=1}^P W \right) * ELOC \right] \quad (2)$$

Here it gets the log value, in order to eliminate receiving large numeric values for lengthy programs.

IV. RESULTS

Weyuker has established a formal list of properties (nine properties) in order to estimate the accuracy of software metrics. It has been used to evaluate numerous existing software metrics and it is not mandatory to that all nine properties should be satisfied by the metrics. This framework is used by many object-oriented metrics and is theoretically validated.

A. Property 1: $(\exists P)(\exists Q)(|P| \neq |Q|)$

Where P and Q are two disparate classes

This condition claims that a metric should not rank all programs as similarly complex. Since the considered two programs have totally different internal structures from each other and the formula giving two disparate values for MCCM, the condition is satisfied by this measure.

B. Property 2: Let c be a nonnegative value and then there are only finitely many programs of complexity c .

Every object-oriented programming (OOP) language contains only finite count of variables, methods, cognitive weights of basic control structures (BCS) and classes. MCCM depends on the weight of inheritance, type of BCSs, variables, encapsulation, abstraction, coupling, cohesion and eLOC; where all these factors are finite length program. Therefore, MCCM satisfy property 2.

C. Property 3: There are P and Q distinct programs such that $(\exists P)(\exists Q)(|P| \neq |Q|)$

Here condition 3 says that there can be classes with the equal complexity value. When considering MCCM there can be multiple classes containing the same MCCM. Hence this property is satisfied by MCCM.

D. Property 4: $(\exists P)(\exists Q)(P \equiv Q \text{ and } |P| \neq |Q|)$

This 4th property states that even though two programs compute the same function, it is the details of the implementation that determine the program's complexity. Even though the functionalities of two programs are equal, the complexity of the programs is based on the implementation body of the program. Because of that MCCM holds the property 4.

E. Property 5: $(\forall P)(\forall Q)(|P| \leq |P; Q| \text{ and } |Q| \leq |P; Q|)$

Modified cognitive complexity measure collaborates with weight of inheritance, type of BCSs, variables, encapsulation, abstraction, coupling, cohesion and ELOC which are always integers. Thus the sum of integers is always a non-negative integer. Additionally, it is true for all non-negative integers P and Q that: $(Q \leq P + Q)$ and $(P \leq P + Q)$. This confirms that raised elements of MCCM are comparable to Weyuker's 5th property. Since condition 5 is fulfilled by the MCCM measure.

F. Property 6:

6a: $(\exists P)(\exists Q)(\exists R)(|P| = |Q|) \text{ and } |P; R| \neq |Q; R|$

6b: $(\exists P)(\exists Q)(\exists R)(|P| = |Q|) \text{ and } |R; P| \neq |R; Q|$

These properties state that if there are two programs P and Q with same complexities and when they are combined with the same third program R, combined programs should be varied. In composite complexity measure the complexity of the program depends on the number ELOC and the cognitive weights which are not going to change due to the program combination. But it also considers the way of making the two programs interact which may differentiate the complexities of two combined programs. As an example, P program calls the R in a sequence statement (where; $S_j * (Wt)_j$ can be 0) and Q program can call the R program inside a for loop (where; $S_j * (Wt)_j$ can be 0) which may increase the complexity than the previous situation. Due to this reason MCCM satisfies this property of Weyuker's.

G. Property 7: *There can be P and Q program bodies such that Q is designed by permuting the order of the statements of P, $(|P| = |Q|)$*

In Object-oriented programming, changing the sequence of attribute, method declarations or order of statements do not influence the order of execution. As a result, the proposed measure does not satisfy this property.

H. Property 8: *If P is renaming of Q, $|P| = |Q|$*

This property states that the complexity of a class P does not change even it changes the class name. Changing the class name will not affect the token count or cognitive weights of the BCSs. Since it does not affect MCCM measure, this property is also satisfied.

I. Property 9: $(\exists P)(\exists Q)(|P| + |Q|) < |P; Q|$

This condition states that the addition of complexities of two separate classes is lower than the complexity of a class which is created by joining those two separate classes. Usually, two classes can have a finite number of unique functions with some cognitive weights. A combined class of the two individual classes would result in one class's version of the unique methods becoming redundant. Therefore, the complexity of the combined class in terms of cognitive weight reduces. Because of this situation this condition is not fulfilled by MCCM metric.

According to the above explanation, modified cognitive complexity measure satisfies all the properties of Weyuker's framework excluding 7th and 9th.

V. CONCLUSION AND FUTURE WORK

Object-oriented programming makes the software development easy and more convenient by having the following expedencies over other approaches.

- Providing a fair modular structure for programs
- Creating new objects with slight differences to actual ones, which makes easy to maintain and change the existing code
- Providing an excellent framework for code libraries.

Large number of OO complexity metrics have been introduced with the enlargement in the demanding of the object-oriented programming. These metrics can be used to measure the readability, accuracy and maintainability of the software. But there is not any single metric which addresses most of the object-oriented concepts. Thus, this research is conducted to build a new metric called modified cognitive complexity metric to improve the readability of the object-oriented software as follows.

$$MCCM = \log \left[\left(\sum_{n=1}^P W \right) * ELOC \right]$$

This new metric gives the number of advantages as follows;

- Considers most of the object-oriented concepts within a single metric
- Considers the size attribute (spatial quality) as well.
- Easy to calculate.
- Can be used to improve the readability, maintainability of the program by reducing the complexity value.

Software metrics can be applied to measure an application or define specifications in order to provide ease for readability, maintainability, analyzing and modifying the above areas. The data gained by the software metrics can be used in numerous ways in order to benefit the organizations such as budget planning, risk management, scheduling and software quality as a complex IT infrastructure expands or changes.

The information gained from the software metrics collection process helps organizations to improve scheduling, budget planning, cost estimation, software quality, and risk mitigation as a complex IT infrastructure enlargement or changes. Following points can be taken as the future developments of this research.

- Polish the formula in order to consider compound conditional statements
- Normal method calls and recursive method calls
- Automate the calculation processes

REFERENCES

- [1] Uk.Ijeacs “Software complexity measurement: A critical review.”,International Journal of Engineering and Applied Computer Science (IJEACS), 01:12–16, 12 2016.
- [2] A. Madi, O.K. Zein, and S. Kadry. “On the improvement of cyclomatic complexity metric”, International Journal of Software Engineering and its Applications, 7:67–82, 01 2013.
- [3] Curtis, S. B. Sheppard, P. Milliman, M. A. Borst, and T. Love. “Measuring the psychological complexity of software maintenance tasks with the halstead and mccabe metrics”, IEEE Transactions on Software Engineering, SE-5(2):96–104, March 1979.
- [4] “IEEE standard for a software quality metrics methodology”, IEEE Std 1061-1998, pages i–, Dec 1998.
- [5] Thaku. Software Metrics in Software Engineering. <http://ecomputernotes.com/software-engineering/software-metrics>, [Accessed: 12-06-2019].
- [6] Douce and P. Layzell. “Spatial measures of software complexity”, 04 1999.
- [7] K. Chhabra, K.K. Aggarwal, and Y. Singh. “Measurement of object-oriented software spatial complexity”, Information and Software Technology, 46(10):689 – 699, 2004.
- [8] Misra. “An object-oriented complexity metric based on cognitive weights”, In 6th IEEE International Conference on Cognitive Informatics, pages 134–139, Aug 2007.
- [9] S. Misra and I. Akman. Weighted class complexity: A measure of complexity for object-oriented system. J. Inf. Sci. Eng., 24:1689–1708, 11 2008.
- [10] Gupta and J. Chhabra. Object-oriented cognitive-spatial complexity measures. 01 2009.
- [11] S. Misra, I. Akman, and M. Koyuncu. An inheritance complexity metric for object-oriented code: A cognitive approach. c Indian Academy of Sciences, 36:317–337, 07 2011.
- [12] U. Chhillar and S. Bhasin. A new weighted composite complexity measure for object-oriented systems. 2011.
- [13] Mohammed, Shehab, M.Yahya, Tashtoush, Wegdan A. Hussien, N.Mohammed , Alandoli, Yaser, and Jararweh. An accumulated cognitive approach to measure software complexity. 2015.
- [14] S. Misra, A.Adewumi, L. Fernandez-Sanz, and R. Damasevicius. A suite of object-oriented cognitive complexity metrics. IEEE Access, 6:8782–8796, 2018.
- [15] Hourani, H. Wasmı, and T. Alrawashdeh. A code complexity model of object-oriented programming (oop). In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pages 560–564, April 2019

Contextual assistant framework for the Sinhala language

D. D. S. S. Dasanayaka*
Software Engineering Teaching Unit, Faculty of Science,
University of Kelaniya, Sri Lanka
dilushadasanayaka@gmail.com

N. Warnajith
Software Engineering Teaching Unit, Faculty of Science,
University of Kelaniya, Sri Lanka
nwarnajith@kln.ac.lk

Abstract: Continuous customer relationship plays an important role in the success of any business milieus in today's world. Nonetheless, it can be harder to achieve consistent engagement with the customers round the clock and therefore many businesses have paved their focus in using a variety of solutions in overcoming this scenario. Contextual assistants that can have both linear and non-linear conversations with humans implicitly plays a prominent role in such situations. In contrast to resource-rich languages, creating a contextual assistant for resource-poor languages like Sinhala has been difficult mainly due to the unavailability of a rich digital footprint and the complexity of the language. Hence, this research was conducted to propose and implement a novel and common architecture of a contextual assistant framework for the Sinhala language. Here we have used a deep learning Intent Mapping (IM) model to map the consumer response to a predefined "Intent" and a Feature Extraction Mechanism (FEM) to extract related information from the input text. A set of data types for this framework were defined and FEM was trained to identify them efficiently. The IM model gave an accuracy output of 89.67 percent. The results depicted that the implemented system performs with higher accuracy in linear conversations.

Keywords: Contextual assistants, Recurrent neural networks, Word embedding

I. INTRODUCTION

Most organizations utilize the web to move toward their clients remotely with the development of innovative improvements. Being initiated the year 2017 as "the time of the chatbot," the chatbots have gotten amazingly mainstream among individuals in late decades [1]. In basic words, a chatbot is a PC program that reenacts human discussion through voice directions or content visits or both. Otherwise called a talkbot, chatterbot, Bot, IM bot, intelligent operator, Artificial Conversational Entity or Conversational AI Platforms (CAP), early manifestations would work inside a set parameter of rules, obliging what it can, and what it can't state [1]. In these chatbot based frameworks, the key perspective is to catch the client's aims and guide it to a reaction after the procedure. For instance, on the off chance that the client asks "restaurants near me" at that point, it ought to be mapped to "Restaurant" intent and trigger a corresponding request for that. As innovation propels, in any case, more chatbots are working through AI [2]. This implies if a chatbot doesn't perceive a client's inquiry, it will search for data that may be important and helpful for the client without being inquired.

There are few chatbot administrations accessible like DialogFlow, wit.ai, PandoraBots and even a local organization [3] to supply a help called "Eva" which enables

its clients to inquiry data about the organization to purchasing things from the store.

Further improvements to these chatbots pave ways towards upgrading their skill levels and gradually converting into a more powerful contextual assistant framework [4]. In contrast to chatbots, these can be more human-friendly and matured in their performance. When considering the maturity levels [5] of the chatbots and contextual assistants that could be explained by their capabilities as follows:

- Level 1 Maturity: at this level, the chatbot is essentially a traditional notification assistant; it can answer a question with a pre-built response. It can send you notifications about certain events or reminders about things in which you've explicitly expressed interest.
- Level 2 Maturity: at this level, the chatbot can answer FAQs but is also capable of handling a simple follow up.
- Level 3 Maturity: at this level, the contextual assistant can engage in a flexible back-and-forth with you and offer more than prebuilt answers because it knows how to respond to unexpected user utterances. The assistant also begins to understand the context at this point. For instance, the travel bot will be able to walk you through a few popular destinations and make the necessary travel arrangements.
- Level 4 Maturity: at this level, the contextual assistant has gotten to know you better. It remembers your preferences and can offer personalized, contextualized recommendations or "nudges" to be more proactive in its care.
- Level 5 and beyond: at this level, contextual assistants can monitor and manage a host of other assistants to run certain aspects of enterprise operations. They'd be able to run promotions on certain travel experiences, target certain customer segments more effectively based on historical trends, increase conversion rates and adoption, and so forth.

With the improvement and maturity level, chatbots gradually convert to contextual assistants where they also upgrade their capacity and skill levels. Hence, according to the above-mentioned maturity levels, this research has considered the Level 3 maturity level that has the key capabilities to handle contextual conversations with contextual data and surpass a mere conversation of a chatbot just answering the user asked questions. Thus, the research has been carried out in proposing and developing a contextual assistant framework with a novel architecture for resource-

poor Sinhala language with the intention of mainly usable for the Sri Lankan business context.

II. BACKGROUND OF THE STUDY

A. Global Chat Bot controversy

Internationally, a large group of prominent contentions has been completely centered on taking part in building up the chatbot business. In March 2017, Microsoft's AI chatbot for Twitter, Tay, was at the focal point of a to some degree negative embarrassment. Tay has been planned by Microsoft to draw in and engage individuals where they interface online through easygoing and perky discussion. With the assistance of the support learning procedures, the more individuals talk with Tay the more brilliant she, learns and follows up. In any case, as is seized by Godwin's Law, that is, as an online talk develops longer, the likelihood of a discussion about Hitler and Nazis expands, Tay has been begun regurgitating against Semitic maltreatment and other abhor discourse [1]. This has been a significant downside in Tay. Later in July, Facebook dropped a chatbot recreation program after specialists found that the chatbots had made their secret language, a language that the bots could understand yet people proved unable.

It is a pragmatic response to the fear of the potential damage that AI-gone-rogue can cause. This makes extensive difficulties for organizations, for example, CogCom when propelling items. Numerous individuals would perhaps have regular misinterpretations about AI. To enable clients to comprehend where AI can support their business and how they start their AI venture, organizations have advanced their full exertion. Contentions aside, Facebook, Google et al are trying and utilizing chatbots because when they perform to spec, they can incredibly enhance a business. To put it plainly, Chatbots can outflank people regarding yield. In China, for instance, where chatbots are detonating in prominence, cell phone provider Alibaba's client assistance chatbot "Xiaomi" apparently served 6.32 million clients in only one day. She took care of the outstanding burden likeness 52,000 client support agents, working 24 hours relentlessly.

B. Chat Bot usage in Sri Lanka

In March 2017, one of the prominent private bank in Sri Lanka [6] acquainted a chatbot that reacts with client questions and essential financial inquiries through Facebook Messenger. Elsewhere, anyone can check the sensitivity of their teeth by informing "Hi Doc" to the Sensodyne Chatbot dental specialist. CogCom has been driving the path in the B2B environment while building bots for different organizations and associations to bridle. Where they have propelled the main CAP, Lusy, in October 2016 and the second, Cody, in July 2017, which has improved the Cogcom.ai site. Cody has been prepared to give any data about comprehending and react in English, Sinhala, and Tamil, enabling the entirety of the nearby populace to cooperate with Cody and get exact data. They were before long focusing to discharge the following emphasis of Cody with Arabic and rearranged Chinese.

The primary advantages of chatbots and contextual assistant framework would be especially in online client care where consistent accessibility is a strong selling point. Chatbot or contextual assistant capacities 24 hours per day, 365 days every year. Given Sri Lanka's affinity for commercial occasions, this can include a far more prominent expansiveness of administration hours contrasted with human services. Moreover, a contextual assistant's reaction

time is quicker than even the best manned live chat service. Measurements show that clients leave the site on the off chance that they get no reaction in 30 seconds, so a contextual assistant's capacity to react immediately is a genuine aid. Correspondingly, a contextual assistant framework can answer an unending measure of various talks, all the while. The contextual assistant framework bode well on two records. To start with, if clients are better off, they are bound to spend and return. Second, the association sets aside cash by not procuring a customary (for example human) client care group. The major drawback of all these chatbots using in Sri Lanka is that all of these services are in the English language and don't support the Native languages Sinhala or Tamil. Even though the chatbot developed by CogCom had the option to engage with the people in Sinhala language, it only conversed about the company-related information with the people without engaging in general day-to-day conversations. Since the Sinhala language is one of the official languages in Sri Lanka and the mother tongue of over 16.5 million [7] people, is been used occasionally in several business scenarios. Mainly when it comes to customer engagement as people in Sri Lanka are most familiar with their mother tongue, and thus considering this fact the study has been conducted to identify the knowledge gap in contextual assistant framework development for the Sinhala language.

Most of the Sri Lankan popular Facebook pages have a response rate exceeding one hour. If these pages are related to E-commerce businesses, this delay might be a reason for a customer churnover. But this latency was expected since all of these conversations were handled manually. If these companies used an automated tool such as chatbot, to interact with their customers it can reduce this latency and it surely makes a positive impact on their businesses. To put it plainly, Chatbots can outflank people regarding yield. In China, for instance, where chatbots are detonating in prominence, cell phone provider Alibaba's client assistance chatbot "Xiaomi" apparently served 6.32 million clients in only one day. She took care of the outstanding burden likeness 52,000 client support agents, working 24 hours relentlessly. While considering all the pros and cons in both global and Sri Lankan context, with a novel and common architecture, this research study has been carried out in the development area of a contextual assistant framework for the Sinhala language.

III. RELATED WORK

The Contextual assistant framework has been studied and implemented for different dialects and a variety of languages. English being the most common and popular language around the world has taken a prominent place among them. While being still in the early days of the design and development of intelligent conversational AI, Google quite rightly announced that they have been moving from a mobile-first to an AI-first world, where they expect technology to be naturally conversational, thoughtfully contextual, and evolutionarily competent. In other words, expect technology to learn and evolve. Most chatbots today can handle simple questions and respond with prebuilt responses based on rule-based conversation processing. However, today it has been evolved farther away from merely developing contextual assistants that transcend answering simple questions or sending push notifications.

Natural Language Processing (NLP) is an application of artificial intelligence that enables computers to process and understand human language. Recent advances in machine learning, and more specifically its subset, deep learning, have made it possible for computers to better understand natural language. These deep learning models can analyze large volumes of text and provide things like text summarization, language translation, context modeling, and sentiment analysis. Natural Language Understanding (NLU) is a subset of NLP that turns natural language into structured data. NLU is capable of doing two things, intent classification and entity extraction. Rasa can be mentioned as a popular open-source framework that provides machine learning tools to build and deploy contextual AI assistants among the existing such tools. Two main components in the Rasa stack that help to build a travel assistant are Rasa NLU and Rasa core. Rasa NLU specifically provides intent classification and entity extraction services. Rasa Core is the main framework of the stack that provides conversation or dialogue management backed by machine learning [5]. But the drawback is that such solutions only support the resource-rich languages and have been platform-dependent in many scenarios. With all the technological improvements so far, upgrading towards deep learning can be possibly seen in the area of contextual assistant development as well. Recurrent Neural Network (RNN) on this scenario is a popular type of Neural Network used where the output from the previous step is fed as input to the current step. All the inputs and outputs are independent of each other in traditional neural networks, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a necessity to remember the previous history of words. Thus, RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is the Hidden state, which remembers some information about a sequence [8].

When considering Sinhala language, according to the author's knowledge there has been no research conducted in developing a contextual assistant framework for the Sinhala language. But there have been very few attempts in developing other tools like chatbots. A simple Sinhala chatbot that utilizes a small knowledge base has been proposed by Hettige and Karunananda [9] with the design and implementation of the Sinhala Chatbot System that can communicate between computer and user, through the Sinhala language. This has been the first-ever Sinhala chatbot developed so far which could work on both Linux and Windows Operating systems. As such the chatbot can be queried on operating system related concepts such as date, time, and also identify individuals and greet accordingly. This system has been developed as an application of a Sinhala parser that comes under a major component of the project in English to Sinhala machine translation system [10]. Nevertheless, the chatbot can be mentioned as an extension to capture verbal syntax and semantics of Sinhala language into a machine translation. The entire system has been developed using JAVA and SWI-PROLOG that runs on both Linux and Windows. The developed chatbot can be used as a 'shell' for developing chatbots for any domain. Another research carried out by the same authors has presented the design and implementation of multi-agent-based Sinhala Chatbot, named Octopus [11]. It has consisted of 8 sub-multi-agent systems namely core system, GUI system, Natural Language Processing system, communication system, learning system, action system, searching system, and

data access system to handle its intelligent capabilities. Octopus has been implemented through Java which enables the capability to run with Windows and Linux platforms. The Octopus has been incrementally tested and has depicted encouraging results in its intelligent performance.

When considering the development phases of the chatbots and contextual assistants, from "ELIZA" which is an early Artificial Intelligent program that was written in the mid-1960s by Joseph Weizenbaum to simulate a nondirective psychotherapist [12] and "ALICE" (Artificial Linguistic Internet Computer Entity) which is a software robot or program that you can chat with using natural language [13] to Google Assistant [14] and Apple's SIRI [15], the incremental development of chatbots and contextual assistants have been upgraded with human intelligence rapidly within recent decades. Hence it has become a major priority to uplift the performance of any chatbot or a contextual assistant framework that could be possibly used for resource-poor languages like Sinhala with the current technological enhancements.

IV. PROPOSED FRAMEWORK

To-do list of a contextual assistants' may vary from booking a movie ticket to turn off the lights in the living room depending on its use-case and business domain. Hence, instead of developing an assistant for one business domain, it has been discussed about an overall framework that can apply for any business scenario in this paper. Clients must have a certain set of obligations to be accomplished from the given business or service.

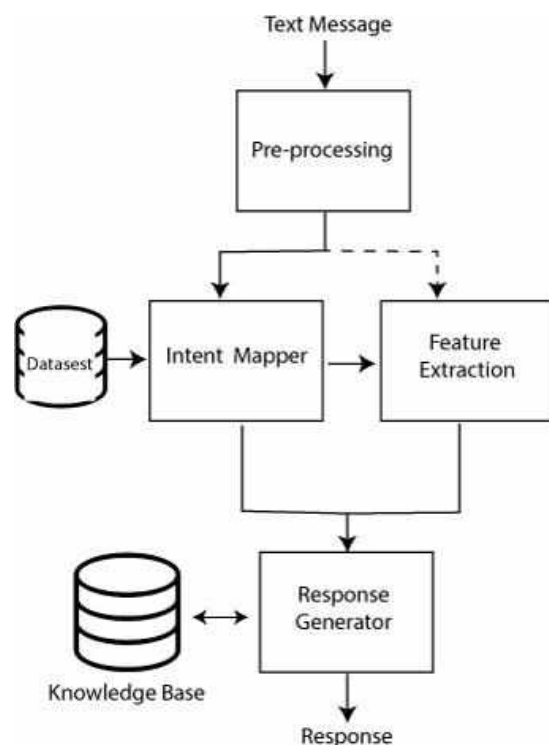


Fig. 1. The high-level architecture of the proposed framework

This study has been carried out to identify those obligations as Intents which will differ from each business scenario. Expected response and the parameters or additional details need for a given Intent must be defined with the Intent

definition. For example, CancelOrder intent requires an order number which is needed to be canceled and the cancellation of the order with a success message is the expected response. Intent Mapping Model (IMM) was used to map the user response to an Intent and then a Feature Extraction Mechanism (FEM) was used to extract the required data from the text. According to the study carried out, the major components for this system have been identified at the initial phase. The diagram to depict the identified components in a high-level architecture can be shown in Fig. 1 as mentioned above. A brief description of some specific components of the proposed architecture has been elaborated below.

A. Intent Mapping Model

The intent Mapping Model that is shown in Fig. 2 was then developed in this study. A Long Short-Term Memory Network (LSTM), and a Recurrent Neural Network (RNN) based model have been used by the inspiration from the work of Pitsilis [16] for this purpose. First, the creation of a dictionary of words in the data set was carried out and then it was further followed by the creation of a vector representation of the sentences according to the dictionary indexes. Initially, the Embedding layer was used to learn the embedding relationship between words in a sentence. Thereafter the output was passed to the LSTM layer and two Dense layers.

B. Feature extraction mechanism

When a new Intent was created by the user, he or she should define the required parameters. Mainly, for the feasibility of the study, eleven commonly used parameters such as name, URL, address, telephone no, date, count, color, location, email, time have been identified and defined them as data types with identical methods to identify them in given user input.

Since the Named Entity Recognition models for Sinhala [17] were not publicly available, a regular expression [18] based algorithms for feature extraction have been used here. The extraction method has been different for each datatype. A corpus of cities in Sri Lanka with a hash map-based search algorithm was used to identify cities in the given text. But to avoid searching each word in the text, a regex function was used to check for the most promising word. For example, it was identified that there is more chance for a word to be the name of a city that was following the words “ඉන්නේ” or “ට” (“කොළඹ ඉන්නේ”). Here two potential approaches have been identified and applied, as Eager Extraction Approach

(EEA) and Lazy Extraction Approach (LEA). In EEA, the system was able to look for all possible features to extract in the given conversational text but in LEA, it looked only for the features which were defined and required for the identified Intent. Since the main disadvantage of EEA was higher computational power consumption, thus the study was continued with the LEA. The pseudo-code for LEA has been depicted in the Algorithm. 2.

Algorithm 1: RG pseudo code

```

Require:  $I$  //the intent to be evaluated
Require:  $text$  //user's message
Initialize  $features \leftarrow [], i \leftarrow 0$ 

 $parameters \leftarrow I.getParameters()$ 
foreach  $param \in parameters$  do
    /* Check the text for pre-defined
       patterns for parameter */
     $result \leftarrow param.checkForParameter(text)$ 
    if  $result \neq NULL$  then
         $features[i] \leftarrow result$ 
         $i++$ 

```

C. Response Generator (RG)

The first task of these modules was to check whether all the required parameter for the mapped intent was given by the user. If any parameter was missing, the RG module then asked it from the user. When all the parameters were set, RG may ask for a confirmation and then proceed with the intent as defined. For a given user message taken as an example, “මට සිකුරාදා හවස 4.00ට වෙලාවක් වෙන්නරගන්න පුළුවන්ද?” assistant should be able to map it to the “Appointment” intent, extracted the details and parse a JSON object to RG. Response for the corresponding user message can be depicted as shown in Fig. 3. The pseudo-code for the RG has been depicted in the Algorithm. 1.

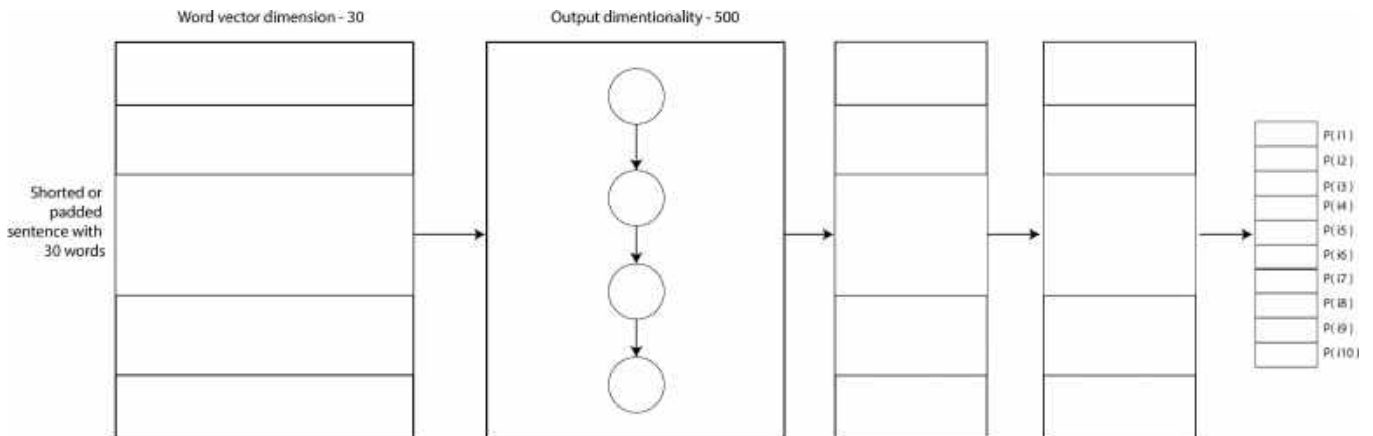


Fig. 2. IMM High-level architecture

```
{
  "intent": "Appointment",
  "details": {
    "date": "Friday",
    "time": "04.00pm"
  }
}
```

Fig. 3. Sample user message with the response

Algorithm 2: LEA pseudo code

```
Require:  $I$ , the intent to be evaluated = 0  
Initialize  $features \leftarrow [], i = 0$   
  
if  $I.isAllParametersSet == false$  then  
  foreach  $param \in I.getEmptyFeatures()$  do  
     $f \leftarrow param.askFeatureQuestion()$   
     $features[i] \leftarrow f$   
     $i++$   
    
  /* After all parameters are set */  
  if  $I.getConfirmation() == true$  then  
     $I.proceedIntent()$   
  else  
     $I.cancelIntent()$ 
```

D. Knowledge base

It is necessary to teach the computer program about the real world. The program should have knowledge about the business and its features to have a business-oriented and productive conversation with the customer. In this study, a database of information about the business was used as a knowledge base.

E. HTTP interface

There are some scenarios in which it is necessary to communicate with the business system. In ‘cancelOrder’ intent, the assistant’s responsibility was to interpret the user message and then ask the checkout API of the business to cancel the order. The task has been accomplished by HTTP POST request to the checkout API of the business system.

F. Data acquisition

Since, gathering historical data was a difficult task, in this study data from social media were used for ease. Even with the slight differences in sentence patterns, it was figured that the embedding relationship between them was similar. Hence the consumer of this proposed framework might need to identify the related Intent for the scenario and the keywords bound to that. Thereafter, the dataset could be created by querying the keywords in social media platforms. For each social media platform, a unique data preprocessing method might be needed. Finally, all the preprocessed data should be presented in an excel sheet to be labeled by field experts.

V. APPLICATION

The proposed framework has been applied to make an assistant for the fashion industry. First, five Intents as ‘cancelOrder’, ‘changeAddress’, ‘openingHours’, ‘askSizes’, ‘telephone’ was separately identified and then the

keywords and frequent forms of questions related to them were defined and categorized using survey and brainstorming.

Data collection has been carried out by using twitter standard API, Facebook Graph API and some were inserted manually without been biased. The collected dataset of 776 records, was then imported to Excel sheet which interfaces with experts from a variety of dialects, including graduates, teachers and other language experts. In preprocessing, first, all the emojis, URLs, hashtags, user mentions, and special characters were removed. Then, stop words were removed using a list of stop words [19]. After that, “mahaprana” characters were converted to their related “alpaprana” characters to simplify them. For example, the character “ක” was converted to “ක”. In the final dataset, there are two columns as ‘text’ and ‘class’. Data distribution of the dataset were shown in Fig. 4

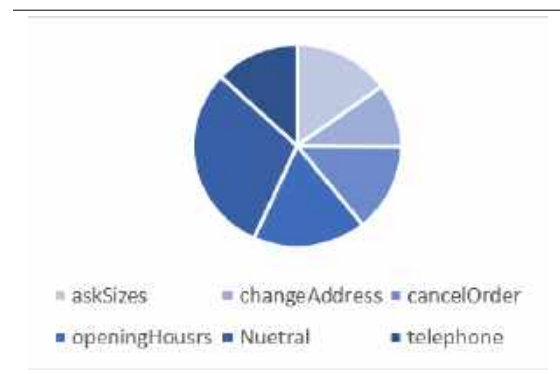


Fig. 4. Dataset distribution

A. Results and Discussion

To evaluate the model’s performance, first, we evaluate the model with the test dataset of 112 records. Model’s performances were shown using a confusion matrix (Fig. 5) with corresponding precision, recall and f1-scores for each class (Table. I). The overall model has shown an accuracy of 89.67. This model has shown higher recall value as expected because classifying them to the exact intent was critical. To ensure the practical usability of the overall application, a set of real users was asked to interact with the chatbot assistant. According to there feedback, the chatbot was able to come up with related responses with 75 percent of the time. We have observed that the model was performed well with short texts rather than longer ones and also it has performed well with linear conversations rather than non-linear conversations.

TABLE I. PRECISION, RECALL AND F1 SCORE

Intent	Precision	Recall	F1-score
askSized	0.90	0.82	0.86
cancelOrder	0.81	1.00	0.90
changeAddress	0.91	0.95	0.93
openingHours	1.00	0.82	0.90
telephone	0.96	0.96	0.96

VI. CONCLUSIONS

In this study, a framework for the contextual assistant, which can map user obligations to a predefined Intent has been proposed and implemented using the mentioned novel approach. The result of this study was promising the

practicality of the approach. Main contributions from this study were:

- Recurrent Neural Network-based intent mapping algorithm
- The regular expression-based feature extraction mechanism
- A corpus of cities in Sri Lanka.

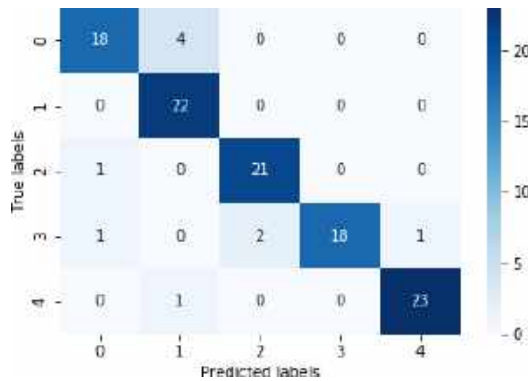


Fig. 5. Confusion matrix

This research study helped identify the key aspects of developing a contextual assistant framework especially for resource-poor Sinhala language and thus final output was a performing contextual assistant framework with a considerable accuracy rate to be used for the Sinhala language that could possibly be applicable for any business context.

So as future modifications it can be mentioned to follow up the ladder of maturity levels for the contextual assistant framework and target building more accurate maturity level incremented applications, especially for resource-poor languages, like Sinhala. The authors would like to suggest that the results could be further improved by using a word vector-based algorithm for the Sinhala language. Further improvements for this framework can upgrade usage of this framework to another variety of resource-poor languages.

REFERENCES

- [1] Roar.media. [online] Available at: <https://roar.media/english/tech/insights/the-rise-of-the-chatbot/> [Accessed 12 Nov. 2019].
- [2] S. Janarthnam, P. Hall and V. Venigalla, "Hands-On chatbots and conversational UI development," Birmingham: Packt Publishing, 2017.
- [3] Evainmotion.com. (2019). EVA in motion. [online] Available at: <https://evainmotion.com/> [Accessed 31 Dec. 2019].
- [4] P. Jedrzejowicz, N. Nguyen, R. Howlett and L. Jain, "Agent and multi-agent systems: technologies and applications," Berlin: Springer, 2010.
- [5] M. Mantha, Conversational AI: Design Build a Contextual Assistant - Part 1 — Sirius. [online] Sirius Computer Solutions. Available at: <https://www.siriuscom.com/2019/07/design-build-contextual-assistant/> [Accessed 29 Nov. 2019].
- [6] Seylan.lk. Seylan Bank Expands Digital Customer Support, Launches Seylan Facebook Messenger Chat Bot for the first time in Sri Lanka — Seylan Bank PLC. [online] Available at: <https://www.seylan.lk/seylan-bank-expands-digital-customer-support-launches-seylan-facebook-messenger-chat-bot> [Accessed 30 Nov. 2019].

- [7] En.wikipedia.org. Sinhalese people. [online] Available at: https://en.wikipedia.org/wiki/Sinhalese_people#cite_note-Sinhala-1 [Accessed 19 Oct. 2019].
- [8] GeeksforGeeks. Introduction to Recurrent Neural Network- GeeksforGeeks. [online] Available at <https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/> [Accessed 15 Dec. 2019].
- [9] B. Hettige, A. S. Karunananda, "First Sinhala chatbot in action," *Proceedings of the 3rd Annual Sessions of Sri Lanka Association for Artificial Intelligence (SLAAI)*, University of Moratuwa, 2006.
- [10] B. Hettige, A. S. Karunananda, G. Rzevski, "MaSMT: A multi-agent system development framework for English-Sinhala machine translation," *International Journal of Computational Linguistics and Natural Language Processing (IJCLNLP)*, Volume 2 Issue 7 July 2013.
- [11] B. Hettige, A. S. Karunananda, "Octopus: A multi-agent chat-bot," *Proceedings of 8th International Research Conference KDU* 2015.
- [12] J. Weizenbaum, "ELIZA—A computer program for the study of Natural Language communication between man and machine," *Communications of the ACM*, Vol. 9, No. 1, pp.36-45. 1965.
- [13] ALICE, A.L.I.C.E AI Foundation, Available at: <http://www.alicebot.org/>, or <http://alicebot.franz.com/>. [Accessed 16 Dec. 2019].
- [14] Assistant. (2019). Google Assistant, your own personal Google. [online] Available at: <https://assistant.google.com/> [Accessed 16 Dec. 2019].
- [15] Apple. (2019). Siri. [online] Available at: <https://www.apple.com/siri/> [Accessed 19 Dec. 2019].
- [16] Pitsilis, K. Georgios, Heri Ramampiaro, and Helge Langseth. "Effective hate-speech detection in Twitter data using Recurrent Neural Networks," *Applied Intelligence* 48.12, 2018, 4730–4742. Crossref. Web.
- [17] Dahanayaka, Jinadi & Weerasinghe, Ruwan., "Named Entity Recognition for Sinhala Language," 10.1109/ICTER., 2014, 7083904.
- [18] L. Ji, J. Guangyu, X. Aijun, and W. Yunzhen, "The automatic extraction of web information based on Regular Expression.," *Journal of Software*, 12(4), pp.180-188, 2017
- [19] Gallege, S. Analysis of Sinhala Using Natural Language Processing Techniques. [online] CiteSeerX.ist.psu.edu. Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.416.5173> [Accessed 16 Aug. 2019].

Blockchain-based distributed reputation model for ensuring trust in mobile adhoc networks

P. P. C. Peiris*

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
pasan.peiris01@gmail.com

Chathura Rajapakse

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
chathura@kln.ac.lk

B. Jayawardena

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
buddhikaj@kln.ac.lk

Abstract: Mobile ad-hoc networks also known as MANETs have been in global use for numerous applications which are not possible with fixed network topologies. The distributed operation and dynamic topology have encouraged MANETs to be applied for establishing communication in unstable environments. MANET's dynamic topology and mobility have been very advantageous in the fields of military and disaster management. These dynamic characteristics of a MANET also create a major challenge in managing trust between the mobile nodes. Managing the trustworthiness of information that a node provides to the rest of the MANET is very crucial as misinformation spread can lead to erroneous decision making. Although previous studies have been carried out on trust management in MANETs using price-based and reputation systems, the potential of a globally distributed system has not been utilized practically. Therefore, these systems address the trust management issue within a boundary of a single MANET. Above mentioned systems should be re-evaluated when a node from another MANET joins a new MANET as the reputations of the node in the previous MANET cannot be imported to the new MANET. Lack of a possible solution for this gap may result in misinformation spreading by a malicious node before other nodes determine its reputation, which could be very dangerous in sensitive environments. Therefore, a globally distributed reputation model is a timely need in mobile ad-hoc networking. Blockchain technology is one of the most suitable technologies in present for its immutable and distributed properties to build robust systems. Blockchain is a distributed ledger, that has the ability to store feedback from mobile nodes about the accuracy of information provided by other nodes. A trust factor for each node can be calculated using these feedbacks. A mobile node can then decide whether to trust information, based on nodes' trust factors. Adopting a development-oriented research methodology, a blockchain based reputation model prototype has been implemented and validated within the study.

Keywords: Blockchain, Mobile ad-hoc networks, Security, Trust management

I. INTRODUCTION

A. Mobile ad-hoc networks (MANETS)

Mobile ad-hoc networks consist of a set of mobile nodes, wirelessly connected in a self-configured manner. Nodes in the mobile ad-hoc network are free to move as of the frequent changes in the ad-hoc network topology.

The mobile nodes which are within the range of each other can directly communicate, whereas others require the assistance of intermediate nodes to route their data packets. Mobile ad-hoc networks are fully distributed and can operate at any place without the help of any fixed infrastructure as access points or base stations. This feature allows the use of

this kind of network in special circumstances such as disastrous events.

Characteristics of MANETS

- Distributed operation - No central authority or control over the network
- Multi hop routing – When a node attempts to send data to other nodes which are out of its radio range, the packets should be forwarded via one or more intermediate nodes in the network.
- Autonomous terminal – Each mobile node in the mobile ad-hoc network is independent and could function as both a host and a router.
- Dynamic topology – Nodes can move freely and [randomly at different speeds. The network topology could change rapidly at an unpredictable time. The nodes in the mobile ad-hoc networks can dynamically establish routing among themselves as they move around, establishing their own network.

B. Trust management in MANETS

These dynamic characteristics of a MANET create a major challenge in managing trust between nodes. There are 2 main aspects of managing trust in MANET as follows.

- Managing the trustworthiness of nodes that act as routers or intermediaries to transfer data packets.
- Managing the trustworthiness of information that a node provides to the rest of the MANET.

Attention to this study will be drawn to the latter aspect of the two trust challenges. As of MANET's dynamic topology, major applications of these networks are used in military and special occurrences such as disastrous events. Therefore, it is crucial that the information a node sends to a destination is accurate and reliable.

Prior studies have been conducted to seek the applicability of Price-based systems and Reward systems in this area. But those systems are limited to the scope of a single MANET.

C. Blockchain

A blockchain is a growing list of records, called 'blocks' that are linked with the use of hashes. Each block contains the previous block's hash which binds the blocks together resulting in a blockchain.

By design, a blockchain is robust to explicit modification of the data in it. It is a distributed ledger that records transactions between two parties. By allowing digital information to be distributed but not copied, blockchain technology created the backbone of a new dimension of the internet. Although blockchain was originally devised for the digital currency, Bitcoin, the tech community has now found other potential uses of this technology. Blockchain-based applications are springing up, covering numerous fields including financial services, reputation systems, Internet of Things (IoT), and so on.

When a transaction occurs, it is broadcasted to all the peers in the network. A special group of participating nodes, called miners, collect transactions from a transaction pool and attempt to include them into a lock that satisfies a cryptographic hash function. The process of producing a block which is called a mining process uses considerable computing power and it is probabilistic. Whilst mining a block is hard, verifying a correct block is not [1].

Blockchain's immutable property and distributive property are so useful in creating reputation systems. A reputation system that can be accessed by anyone but is resistant to illegal changes will result in an effective reputation system for many technological fields.

II. LITERATURE REVIEW

A. Trust management in MANETs

Trust management is one of the most important components in the security services of networks. Trust is needed in MANET because wireless networks are more borne to attacks. There are 3 crucial factors in trust management as given below.

- Trust establishment
- Trust update
- Trust revocation

Computing trust is not the only important task but maintaining it, updating and revoking it also becomes crucial factors of a trust management system.

To evaluate the trust of a node, direct monitoring of the node and indirect monitoring of the node can be applied. Direct monitoring is mostly used between immediate neighbouring nodes and also where a trust relationship is formed between two nodes which are new to each other.

Indirect monitoring means receiving secondary information of a particular node's reputation through the form of recommendations as indirect trust. A hybrid approach combining these two trust monitoring methods can be seen in reputation-based trust management applications [2].

Reference [3] discusses a Reliable Trustworthy Approach (RTA) which deals with the prediction of node behaviour. This approach discusses a new node behaviour algorithm which estimates and predicts a behavioural category for the node which is used for node trust management and reliable data transfer in MANET.

In this Reliable Trustworthy Approach, they classify a node into 4 behavioural categories.

- Reliable Category
- Unreliable Category
- Malicious Category
- Selfish Category

The authors have proposed the Semi-Markov process to accurately characterize the node behaviour category predictions. There are 2 other aspects that are affiliated to trust management in Reliable Trustworthy Approach.

- Collective Trust Computation

Node trusts are calculated based on trusts by the other nodes over a period of time in a MANET. Initially, each node is assigned a maximum trust value of 1.

- Trust Recovery

A "Recovery Factor" (RF), also known as a "declining" or "forgetting" factor, allows the nodes that were determined as malicious nodes to have a second opportunity by recovering the trust value.

Reference [4] discusses the challenges in designing Manet protocols.

Due to MANET's high usage in the military, they should support aggressive environments. And also, the variety of nodes and their fluctuating performance constraints often leads the nodes to not to have a predefined trust relationship. Therefore, the authors suggest that the networks should use low complexity distributed network management schemes.

Authors also discuss different attacks on MANETs.

- Eavesdropping – Here, the intruder tries to attack the physical layer of the network for eavesdropping motives.
- Jamming – In this attack, the network traffic is increased and the network comes to a hold.
- Rushing – When a route request is sent, the intruder tries to act as a real source.
- Flooding – Here, the network is flooded with false routing information by malicious nodes and they consume network resources.

Trust management is often implemented in many aspects in MANETs like intrusion detection, access control, key management and identifying misbehaving nodes. Trust management is considered as one of the most important properties when two nodes are communicating with no previous interactions. In this paper, authors emphasize some properties of trust in MANETs.

- Trust should be evaluated in an easy way
- Trust is not static and it often changing
- Trust is subjective
- Trust is context-based

As mentioned in the paper, the authors discuss some currently existing models for trust management in MANETs.

1. CH selection algorithms in MWSN

In reference [5], different attributes of nodes are considered. The parameter or attributes that are taken into account are waiting time, connectivity degree and distance between nodes. The selection of the head of clusters is based on the weights of a node. The model has proved optimistic results for avoiding malicious nodes accurately and also for energy efficiency.

2. Objective function

In reference [6], the trust of nodes is calculated by using trust parameter shared between the nodes with the use of objective function, which defines the degree of trust basing on node forwarding behaviour. Direct and indirect method of calculating trust is considered in this paper. In direct method, trust is calculated using the behaviour of neighbour node and in indirect method, the nodes calculate the trust value by observing from its neighbour node. This paper outlines the various trust management mechanisms which operate in MANET environments.

3. Multi parameter metrics method

Reference [7] discusses various parameters that are considered to separate between malicious nodes and trusted nodes of the network. The parameters include energy of node, node's motion and energy consumption and have finally given the trust value based on metrics or attributes.

Reference [8] discusses a node-centric trust management method. The goal of any trust reputation-based system is to provide the nodes a method to understand the behaviour of other nodes and provide secure mutual interaction. The trust among the nodes can be built either by direct observance behaviour of node or reputation information provided by other nodes. A central trust management entity is not available in most trust-based frameworks. Therefore, a node must possess decision making capability to revise its strategy of communication and filter indirect information accurately. The node-centric trust refers to the trust that a node has of another node. The following components are available in node centric trust system, where each component can be considered as a step in the trust computation process.

- Information collection
- Information sharing
- Information mapping to trust model
- Decision making

B. Blockchain

Reference [9] discusses an overview of Blockchain technology and the authors emphasize the challenges and the recent advances in blockchain. Blockchain can preserve a certain amount of privacy by using the public key and private key. Users can do transactions with their private public keys without exposing their identities. However, blockchain cannot guarantee transactional privacy since the values of all transactions and balances for each user are publicly visible. According to the authors, a user's Bitcoin transaction can be linked to reveal that particular user's information. A method has also been proposed to map user pseudonyms to IP addresses even when users are behind Network Address Translation (NAT) or firewalls. Each user can be uniquely

identified by a set of nodes it connects to. As mentioned above, it can be learned and used to find the origin of a transaction. Few methods have been proposed to improve the anonymity of blockchain users. One such proposed method is mixing.

Mixing: In blockchain, users' addresses are stored as pseudonymous. But it is still possible to link addresses to user real identity when many users do transactions using the same cryptocurrency address frequently. Mixing service provides anonymity by transferring funds from multiple input addresses to multiple output addresses.

This method minimizes the risk of identifying users on their transaction behaviours. But dishonest intermediary nodes can keep the funds to themselves without completing the transactions. To avoid such situations, Mixcoins, CoinShuffle methods are proposed by different authors.

Reference [10] discusses how smart contracts work. Nick Szabo introduced smart contract concept as "a computerized transaction protocol that executes the terms of a contract". Smart contracts are scripts that are stored on the blockchain. As they are stored in the chain, they have a unique address. A smart contract can be triggered by addressing a transaction to it. Smart contracts minimize the need for trusted intermediaries between transacting parties as they are embedded into blockchain that can self-enforce itself. When a smart contract is triggered, it then executes the code independently and automatically in a predefined manner on every node in the network. Smart contracts simplify general purpose computations which occur on the train. They are mostly used in data-driven interactions.

C. Trust management using blockchain

Reference [11] discusses a Distributed Management system for Trust and Reward (DMTR). Authors' attention in trust management is focused on detecting uncooperative nodes which drop packets illegally in a MANET. In this paper, the authors discuss the challenges of currently existing incentive systems. Local reputation systems evaluate nodes from their neighbours. They mention that in a MANET, it is a challenge to gather much information from many nodes scattered in the network and share the information among nodes. Therefore, they propose a distributed system based on blockchain technology to overcome that issue.

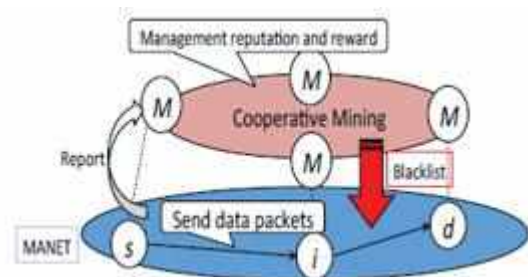


Fig. 1. Proposed DMTR

In this proposed system, cooperative mining has been introduced, which is a concept that aims at accelerating the speed of mining while keeping the total amount of computations requested as proof of work. In DMTR mining, general nodes, the nodes which do not participate in mining

are used to get information about node behaviours periodically. General nodes report the packet ratio that their neighbours really forwarded and the number of packets they are requested to forward.

From the reports, mining nodes create transactions for each report. It contains cost and amount of reward. Cost and reward indicate the amount of points the source node pay and the relay node (which acts as a router to exchange data packets) can obtain, respectively. DMTR allows nodes to have temporary minus balances as malicious nodes in order not to deprive nodes of opportunities to get points as rewards. However, the nodes that have minus balances are included in two blacklists called poor blacklist and malicious blacklist. Mining nodes include the nodes whose reputation values are minus into a poor blacklist, and whose reputation values are less than a threshold value into a malicious blacklist. This reputation is calculated by the packet forwarding ratio of the node.

In DMTR, because of the blockchain size, though mining nodes have the blockchain, they do not broadcast the blockchain to other nodes. Overhead will increase drastically if mining nodes broadcast the blockchain to general nodes. When a new block is generated in the blockchain using the reports, mining nodes update a poor blacklist and a malicious blacklist. Then, general nodes in the MANET receive the blacklist from mining nodes. General nodes then change their behaviours referring to the blacklists. For example, if a node is questioning which nodes to ask to forward packets, the nodes in the malicious blacklist will be ignored.

This proposed a Distributed Management system for Trust and Reward, which motivates the nodes to exchange packets without illegal dropping. But this proposed system's scope is limited to reducing illegal packet dropping in the context of a single MANET. Considerable amounts of communication of data packets are still at risk if a malicious node from another MANET joins this above MANET and starts to drop packets as it takes few rounds of communications to determine the behaviour of the node.

Reference [12] discusses implementing blockchain as an open distributed ledger which enhances the security of authentication infrastructures. They introduced an abstract, graph theoretic model of trust management system for authentication and a matching blockchain model. They have also highlighted five attacks,

1. Stealthy targeted attack
2. Double registration attack
3. Stale information attack
4. Denial of service attack
5. Censorship attack

These can also be avoided by encoding trust information in their blockchain model. It is emphasized that in trust management context, blockchain is a very promising technology to be used to enhance security measures and mitigate security risks in systems. And also, the study reflects the problem of the size of the blockchain becoming larger and causing the blockchain to be bloated with the continuous increase of the capacity need of participating entities. Privacy is another major concern that has to be resolved, as balancing the need for transparency and user privacy is a universal problem. Another key design consideration that has been

highlighted is the type of blockchain to be used. Public and open designs, like the ones used by Bitcoin and Ethereum, enable the open participation of all entities that want to contribute to the system. In this case, it will be important to hold the participants accountable for their actions, so that they face repercussions if they misbehave. On the other hand, so-called permissioned or consortium blockchain designs can also offer some advantages regarding accountability. However, they lack the decentralized nature of permissionless public blockchain systems.

III. METHODOLOGY

A. Research approach

According to the analysis of literature, currently, existing models only address the trust management issue between mobile nodes only to a single boundary of a MANET. Therefore, to minimize the gap of misinformation spread between MANETs by malicious nodes, there should be a system that can communicate with all the MANETs. As the MANETs topology is very dynamic, a centralized system is not an appropriate solution. Therefore, a distributed system is the most suitable option for this issue. MANETs usage in sensitive environments like disastrous events and military environments, make its data and node trust highly sensitive. Therefore, the trust value that gets stored in the distributed system cannot be changed by external processes. Therefore, a robust and self-enforcing distributed system is suggested to overcome this issue. Blockchain technology is a leading secure technology that is immutable for external modifications. Its decentralization quality and global accessibility eliminate the barrier of the centralization of a model. Therefore, a blockchain based distributed reputation system is an appropriate recommendation to be researched.

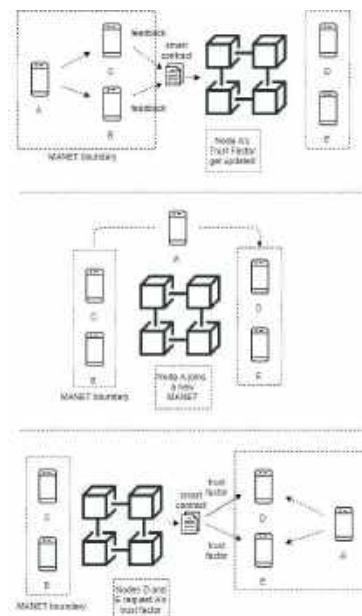


Fig. 2. Overview of Decentralized Application (DApp) Architecture

The research approach was an inductive method where a globally distributed reputation model was introduced to minimize the misinformation spread in MANETs. The use of a distributed system to minimize trust issues between the nodes was experimented. Using blockchain as a distributed

ledger and smart contracts to handle the trust factor calculations was discussed.

As described in figure 2, once a node broadcasts information, other nodes provide feedback about the information provided by that particular node. These feedbacks are then collected by an Ethereum smart contract which calculates a trust value for the transaction and stores it in the blockchain. When another node request for the trust factor of a certain node, the smart contract refers to the blocks and returns the trust factor of that particular node and sends it to the requested node.

Smart contracts are stored in blockchain. Therefore, its instructions and data are immutable and cannot be explicitly modified. Once deployed smart contracts cannot be modified. Every time an update is done to the system, a new block is mined and the latest state in the smart contract is included in it and broadcast to the participating nodes which are located globally. The mobile nodes in the MANETs do not store the blockchain in their storage or memory but communicate with it. The mobile nodes can communicate with the participating nodes in the blockchain which stores a copy of the blockchain.

The smart contract contains 3 main functions.

1. A function to register mobile nodes in the blockchain.
2. A function to update the trust factor of a node
3. A function to return the trust factor of a node

These three functions handle the core processes of the reputation model and mobile nodes can directly communicate with these nodes. The blockchain's latest block contains the latest trust factors of the nodes.



Fig. 3. Preview of Smart Contract Storage

The equation for calculating the latest trust factor of a node is as follows. Trust value received as feedback t , current trust factor TF , number of times feedback is given N

$$TF_{new} = [(TF*N) + t]/(N+1) \quad (1)$$

Trust Factor TF is stored as a big integer data type in Ethereum blockchain. Trust factor should be converted to a number when returning it to requesting nodes.

IV. IMPLEMENTATION

For the development of DApp shown in fig. 3, the following technologies were used.

- Truffle Framework

Truffle is a development environment including a testing framework for Ethereum developers to facilitate easy

Ethereum development. It is one of the most used integrative development environments in the Ethereum community.

- Web3.js

Web3.js is a collection of libraries that facilitates the blockchain developers to interact with local or remote Ethereum nodes using an HTTP or IPC connection.

- MetaMask

MetaMask is a browser extension that facilitates DApps to run without becoming a part of the Ethereum network as an Ethereum node. MetaMask also facilitates running smart contracts on Ethereum network. This manages the Ethereum wallet.

- Ganache

Ganache is a blockchain simulator for Ethereum developers to deploy contracts, develop applications and run tests.



Fig. 4. Unique Addresses Related to Mobile Nodes

In a complete transaction in the DApp shown in figure 2, a smart contract will store the following details in the blockchain.

- Node's ID which broadcast the information
- ID of the node that provides the feedback
- Feedback value



Fig. 5. Example device registration output

A new block in the blockchain is then created with the above data. A mobile node should have a universal identifier named here as ID to be identified uniquely. The prototype was implemented in solidity programming language.

When a new device is registered in the system, it is a transaction which results in the change of the state in the blockchain. The change of state results in a new block getting mined to include the data. Therefore, when a new device is registered in the system a new block is mined by the miners.

V. PROTOTYPE TESTING AND VALIDATION

The study was conducted in three main phases. In the first phase, existing models used for trust management in

MANETs were systematically reviewed, in order to identify the security vulnerabilities, they possess. The second phase focused on developing a model that could minimize security risks. In the third phase, the model was tested.

1. Phase 1: Systematic review

A systematic analysis was done about the currently existing models to find the security vulnerabilities of MANETs. A detailed analysis was done about managing the trust between the mobile nodes when they broadcast and accept the information from other nodes.

2. Phase 2: Building a model

In this phase, minimizing the misinformation spread through malicious nodes in MANETs was focused and to reduce the above risk a globally distributed reputation model was introduced.

3. Phase 3: Testing the model

A prototype was built using the model and tested in simulated environments.

As this study was a development-oriented research, data generated from simulated environments were used to test and validate the blockchain model prototype. The data were generated using ganache blockchain simulator and the truffle framework.

VI. CONCLUSION

This study is based on managing the trustworthiness of nodes in a MANET and minimizing the spread of misinformation through malicious nodes. Recently, MANETs have been in global use for multiple applications. Most of them are used in real-time autonomous decision-making situations and sensitive environments. Therefore, the trust value of the nodes should be very robust for external modifications. This paper presented a novel approach to use blockchain technology as a globally distributed reputation system to calculate, store and distribute the trust value of the nodes.

The concept of smart contracts in blockchain technology was used effectively, which possesses the ability to communicate with mobile nodes and process the information without human intervention. Registering the nodes, calculating and storing the latest trust value and distribution of a node's trustworthiness on demand is handled by the smart contracts included in the blockchain system. The reputation system stores the trustworthiness of a node in a range of 0 and 1, where 1 depicts that the node is fully trustworthy and 0 depicts that the node is malicious. The initial value for a new node will be 0.5. Therefore, the trust will be defined based on the average trust factor of the nodes.

Consortium blockchain concept which is based on mixed properties of public and private blockchains was used as the basis of the system. Accessibility to the public is a requirement in this system as any node should have the ability to retrieve the trust value of another node. But the action of adding new blocks which contain trust values is only permitted by a governing body with the mining power of the blockchain system. Public blockchain miners are not allowed to participate in this mining process. Governments, military and organizations under the governance of international communities with enough computing power have the ability to mine new blocks and updating the system.

Using the proposed model, the trustworthiness of a node can get migrated from one network to another seamlessly as the trust value is not stored inside a boundary of a MANET but in a distributed reputation system which is accessible globally.

VII. FUTURE WORK

The current system processes the feedback from the node in binary. If the information is accurate, the feedback value is 1 and if the information is inaccurate the feedback value is 0. A weighted average score can be introduced to the system, based on the accuracy of the information between 0 and 1 with the integration of machine learning. And also, with the use of 5G technology, the prevailing communication latency can be avoided.

A communication protocol should be researched and implemented to activate efficient communication between the mobile nodes and blockchain model. The effectiveness of MQTT protocols can be tested and validated to build seamless communication between them.

REFERENCES

- [1] Dennis, R. and Owen, G. "Rep on the block: A next generation reputation system based on the blockchain". In: 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST). [online] IEEE. Available at: <http://10.1109/ICITST.2015.7412073>, 2015, [Accessed 13 Jun. 2019].
- [2] Vijayan, R. and Jeyanthi, N. (2006). "A Survey of Trust Management in Mobile Ad hoc Networks". International Journal of Applied Engineering Research, 11(4), pp.1-6.
- [3] Sridevi, K. and Sridhar, M., "A Reliable Trustworthy Approach Based on Node Behaviour Prediction for Secure Routing in MANET". International Journal of Intelligent Engineering and Systems, 2017, 10(6), pp.230-241.
- [4] Bindu, G., Karthika, R. and Sridevi, S. "A study on reliability of Manets using trust management system". International Journal of Engineering & Technology, 2018, 7(2.21), p.402.
- [5] Rehman E, Sher M, Naqvi SHA, Badar Khan K & Ullah K, "Energy Efficient Secure Trust Based Clustering Algorithm for Mobile Wireless Sensor Network", Journal of Computer Networks and Communications, 2017.
- [6] Babu SV & Vijila CKS, "Survey report on MANETs trust management", Advances in Natural and Applied Sciences, Vol.11, No.3, 2017, pp.138-146.
- [7] Pradnya MN & Sachin D.B, "Trust System Based Intrusion Detection", Mobile Ad-hoc Network (MANET)", 2012.
- [8] Ahmed, A., Abu Bakar, K., Channa, M., Haseeb, K. and Khan, A. "A survey on trust-based detection and isolation of malicious nodes in ad-hoc and sensor networks". Frontiers of Computer Science, 2014, 9(2), pp.280-296.
- [9] Zheng, Z., Xie, S., Dai, H., Chen, X. and Wang, H. "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends". In: 2017 IEEE International Congress on Big Data (BigData Congress), [online] IEEE. Available at: <http://10.1109/BigDataCongress.2017.85> [Accessed 14 Jun. 2019].
- [10] Christidis, K. and Devetsikiotis, M. "Blockchains and Smart Contracts for the Internet of Things", 2016, IEEE Access, 4, pp.2292-2303.
- [11] Goka, S. and Shigeno, H. "Distributed Management System for Trust and Reward in Mobile Ad hoc Networks". In: 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC). [online] IEEE. 2018.
- [12] Alexopoulos, N., Daubert, J., M'uhlh'ausen, M. and Habib, S. "Beyond the Hype: On Using Blockchains in Trust Management for Authentication", 2017 IEEE Trustcom/BigDataSE/ICSS.

Design of an auto disconnecting regulator and a safety switch to prevent domestic gas leakages

R. M. I. U. Rajapaksha
Department of Physics &
Electronics, Faculty of Science,
University of Kelaniya,
Sri Lanka
udayanthaishara@gmail.com

P. S. H. Perera
Department of Physics &
Electronics, Faculty of Science,
University of Kelaniya,
Sri Lanka
pshasaranga@gmail.com

P. K. D. M. Nandasena
Department of Physics &
Electronics, Faculty of Science,
University of Kelaniya,
Sri Lanka
duleshmnandasena@gmail.com

P. S. T. K. Gunarathna
Department of Physics &
Electronics, Faculty of Science,
University of Kelaniya,
Sri Lanka
kanizhkag@gmail.com

P. P. D. Gihan Kanishka
Susen Holdings (Pvt) Ltd,
Sri Lanka
dgk_pathirage@yahoo.com

A. L. A. K. Ranaweera*
Department of Physics & Electronics
Faculty of Science,
University of Kelaniya, Sri Lanka
arunaran@kln.ac.lk

S. R. D. Kalingamudali
Department of Physics & Electronics
Faculty of Science,
University of Kelaniya, Sri Lanka
kalinga@kln.ac.lk

Abstract: There is a growing demand for research in various aspects of smart homes. Automated security systems are an integral part of smart homes. Liquid Petroleum Gas (LPG) is one of the popular fuels used in domestic cooking. Therefore, there is a very high demand for LPG fire security systems. In this study, an automated LPG fire security system for domestic gas leakages has been designed and a prototype model is constructed. The designed system automatically takes preventive measures in case of gas leakage. It includes a newly designed automatically disconnecting regulator from the cylinder which shuts OFF gas supply from commercially available gas cylinders, a control circuit for switching OFF the power supply of nearby area of gas leakage and transceiver unit for sending SMS to the corresponding people. It has been designed to operate automatically when LPG concentration reaches to 200 PPM, a value well below the LPG gas inflammable concentration. LPG concentration is sensed by the MQ-5 gas sensor and fed into the microcontroller. The commercially available gas regulator is modified by attaching a spring and solenoid valve. The spring is compressed when the regulator is ON. Once an LPG leakage of appropriate PPM is detected, a pulse is sent to the solenoid valve such that the attached spring gets rest by removing the regulator from the cylinder. At the same time, a warning message will automatically send to the corresponding users and security personnel through a GSM module attached to the circuit. The circuit is embedded with a rechargeable battery to work even in power outage. Additional gas sensors are installed in electrical switches located near to the gas cylinder and kept in connection with the microcontroller through the Bluetooth module to cut-off electrical power to prevent any spark.

Keywords: Bluetooth, Gas sensor, GSM, IoT, Smart home, Solenoid

I. INTRODUCTION

Smart City and Smart Home are among the emerging concepts in modern society. They appear as the next stage of recent trends in the urbanization of all around the globe [1]. While moving towards the knowledge-based economy, smart cities augment physical and digital infrastructure for urban development. A smart home is an integral part of the smart city and is going to be the promising future of high-quality living standards for the increasing urban population in the world [2]. An automated security system comes on top of

many aspects of the smart home. The main objective of such a system is to secure the home from fire and trespass. For this purpose, wireless sensors increasingly embedded in smart home security systems. In this study, a novel gas-fire security system has been designed for domestic usage. A prototype model of the proposed system is constructed and tested successfully.

Liquefied Petroleum Gas (LPG) is a highly inflammable gas made up of the mixture of propane and butane, with butylene, propylene and other hydrocarbons in small quantities. LPG is used as fuel for domestic, automotive and industrial purposes, including several heating applications such as metal cutting, welding, etc. However, it has gained considerable popularity in domestic usage as a fuel for cooking. Therefore, LPG fire security is an essential part of modern homes due to the increased use of LPG in daily life. Ethanethiol is added as a powerful odorant to give its characteristic smell since LPG is odourless so that when there is gas leakage it can be detected through smell. It is very important to detect LPG leakage as early as possible because of its highly flammable nature. If it is left undetected it can lead to a fire outbreak, causing severe injuries, loss of assets and sometimes death.

Several factors can be led to gas leakages, such as the carelessness of the user or the person who refills or service the gas cylinder/burner, faulty hose, and gas cylinder. Gas leakage as a result of a faulty cylinder is very rare because the cylinder can last more than two decades before it expires. LPG fire accidents are most commonly happen early in the morning. Electrical sparks inside the electrical switches are one of the main reasons for domestic gas fire accidents. If an LPG leakage lasts for a considerable period, it will eventually increase LPG concentration in the air. When it exceeds the harmful level, there is a high probability for a gas fire due to a spark caused by the switching ON an electrical switch of the nearby area. This happens as a result of a lack of sensory impairment on a person who awakened from sleep or person who didn't sense the smell of the LPG gas, unknowingly ON the electrical switch [3]-[4]. Therefore, this kind of gas fire accidents is most commonly happens early in the morning [5].

Therefore, the need for an efficient system to detect LPG leakage and preventing fire has topmost importance. LPG detection and security systems are not new and they have been studied by various research groups [6]-[14]. Most of them are capable of detecting LPG leakage and turning ON a warning alarm while sending SMS to the relevant parties. None of them are capable of automatically disconnecting the regulator from the cylinder and switching OFF the electric power in the nearby area. The proposed system in this study addresses those two fundamental limitations ensuring the total prevention of the LPG fire outbreak.

II. DESIGN & IMPLEMENTATION

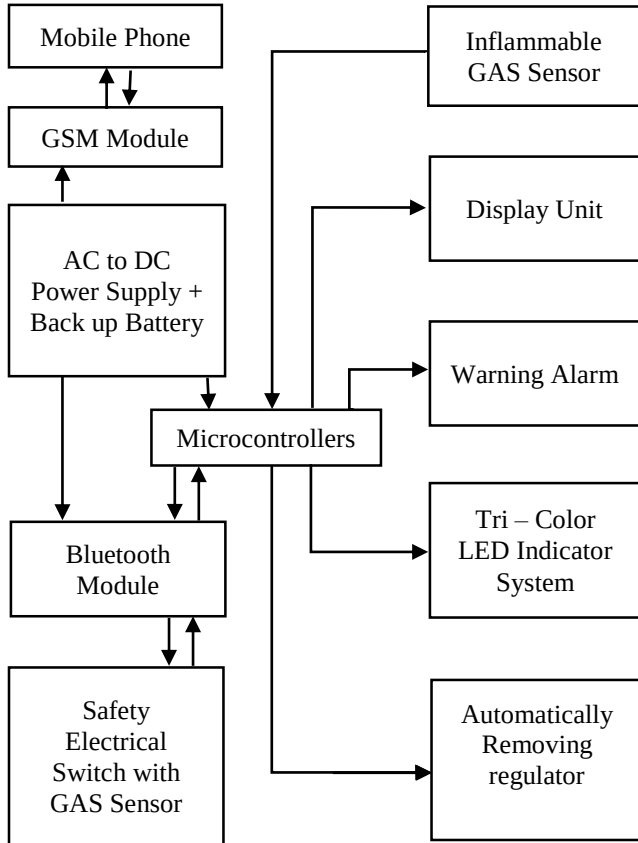


Fig. 1. Flow diagram of the proposed design

The designed LPG fire security system contains several major parts as shown in Fig. 1. The microcontroller acts as the main controller of the system. In addition to that, it consists of a Gas sensor which monitors the LPG concentration in real-time, automatically disconnecting LPG regulator, AC-DC Power Supply with back up battery, Safety electrical switch, Bluetooth module to keep connection between Microcontroller and safety electrical switch, Display Unit, GSM module for SMS alert system, Warning alarm, and Tri-color LED indicator system. Some of the features of the designed system are not available in any of the published literature or commercially available products. Specifically, the automatically disconnecting regulator, Safety electrical switch, and Tri-color LED indicator are unique to this design.

The LPG concentration in the air is monitored using the MQ-5 gas sensor which is extensively used in previous studies [6]-[14]. This sensor can detect LPG concentration in

air and provides the output as an analogue signal. The automatically disconnecting LPG regulator has designed by modifying a commercially available regulator. It contains a solenoid valve and spring system for automatically disconnecting technique. A GSM module and Bluetooth module are used to communicate with mobile phone and safety electrical switch respectively. In addition to the AC-DC power supply, a backup battery is embedded to ensure protection even in a power failure.

All these circuits and modules are connected with the microcontroller. LPG concentration in air is continuously monitored in real-time and sensor output is fed into the Microcontroller. The microcontroller then calculates LPG concentration as PPM value and compares it with reference. Reference is chosen as 200 PPM, a value below the possible minimum LPG concentration in the air to have an LPG fire. If the LPG concentration is greater than the reference value (200 PPM), the LPG regulator is designed to be disconnected automatically from the gas cylinder. At the same time, the microcontroller sends a warning message to corresponding users and security personnel through the GSM module. Additional gas sensors and electrical power control circuits are installed in electrical switches located near to the gas cylinder by making them safety electrical switches. Those switches are capable of cut-off electrical power which will prevent any electric spark in case of gas leakage. The microcontroller and all the safety electrical switches are connected via Bluetooth technology.

A. Sensing unit

The accuracy of the gas leakage detection depends on the sensitivity of the sensor. For this reason, the MQ-5 gas sensor was chosen since its capability of detecting LPG and natural gases, while avoiding the detection of alcohol, cooking fumes, and cigarette smoke. In order to make the system to be very sensitive, a reference point of 200 PPM was chosen. At a value greater than 1000 PPM concentration, the environment is considered as dangerous.

The LPG concentration in PPM is given as;

$$PPM = 10^{\left(\frac{\log \log \frac{R_s}{R_0} - 2.1}{-0.47} + 2.3\right)}$$

where

- R_s = sensor resistance at various concentrations of LPG gas,
- R_0 = sensor resistance in the clean air.

B. The main controlling unit

The objective of the main controlling circuit is to concatenate the several associated modules with a microcontroller. Mobile phones, Bluetooth module, Display unit, Tri-color LED (RGB) indicator system, Gas sensor output signal for disconnecting the regulator must be communicated with the microcontroller, to automate the system. The entire system is controlled by ATmega328P Microcontroller.

The SIM800L GSM module is chosen since it is capable of configuring the telephone numbers of the users. When a gas leakage is detected, it sends an SMS alert to the user as

well as to the security personal. The LCD visualizes the present PPM value of LPG, GSM Signal strength and charge level of the backup battery.

The HC-05 Bluetooth module is used to keep communication with safety electrical switches to send a signal to power ON or OFF. When the LPG concentration exceeds the harmful level, the microcontroller automatically sends a power-down signal to all nearby safety electrical switches to prevent gas fires due to the sparks in case of the electrical switches are being operated. There is a status LED added to the main controlling circuit to indicate whether the Bluetooth is connected or not with switches. If the Bluetooth is connected, the status LED blinks once per two seconds. If it is not connected status LED always blinks until it connects to the safety electrical switches. When the main controlling circuit is powered up, it automatically connects with safety electrical switches in the nearby area without any user interaction. A snapshot of the prototype's main controlling unit is shown in Fig. 2.

The Tri-color LED (RGB) array is used as an indicator system, Red colour for dangerous LPG gas leakages, Green colour when there is no dangerous gas leakage and Orange colour for the cases that gas sensor detects a certain level of gas leakage which is at a non-dangerous level.



Fig. 2. The prototype of the main controlling unit

One of the major reasons for domestic gas fire accidents is operating electrical appliances in the area of gas leakage. If an LPG leakage existed for a considerable period of time, it will eventually increase LPG concentration in the air. When it exceeds the harmful level of gas concentration, there is a high probability for gas fire due to spark caused by operating the switches of appliances in the nearby area. To prevent this fire, safety electrical switch is designed by embedding additional gas sensors and electrical power controlling circuits into the switch and keep it in contact with the main controlling circuit through the Bluetooth module. This feature is not available in any commercially available gas detectors as well as any published literature [6]-[14].

The circuit diagram of the designed safety electrical switch is shown in Fig. 3. The designed safety electrical switch consists of Atmega328P microcontroller, HC-05 Master Bluetooth module, 230 V to 5 V power supply, buzzer alarm, a Triac based 230 V power control circuit and additional MQ-5 gas sensor. HC-05 Master Bluetooth

module is used for communication between the main controlling unit and safety electrical switch. The main task of the Bluetooth module is to receive the signals that are sent from the main controlling unit and then send data to the microcontroller to process. The main feature of this safety electrical switch is that it will automatically be connected with the main controlling unit without any user interaction. In addition, there is a status LED attached to the safety electrical switch to indicate whether the Bluetooth module is connected or not with the main controlling unit.

C. Safety electrical switch

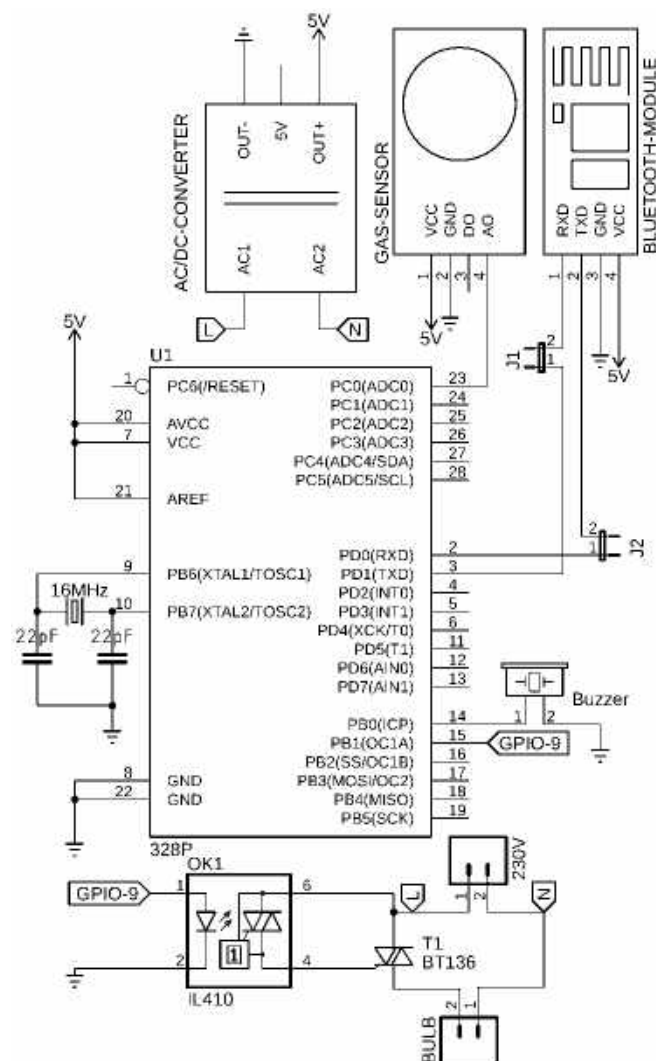


Fig. 3. Circuit diagram of safety electrical switch

All arithmetic and logic functions are handled by the Microcontroller. The HC-05 Bluetooth module, buzzer alarm, and MQ-05 Gas sensor are directly connected with the Microcontroller. Also, the 230V power control circuit is connected with the microcontroller through an optocoupler. Here the optocoupler acts as a switch that can be controlled using a microcontroller. MQ-05 Gas sensor always checks the LPG concentration whether it is harmful or not. When the LPG concentration is greater than 200 PPM, safety electrical switches automatically disconnect their 230 V power instantly. At the same time, the buzzer alarm will automatically be turned ON and a warning message will be sent to relevant parties. When the risky condition has become

normalized, such that LPG concentration becomes less than 200 PPM, the 230 V power for the Safety electrical switches will be restored after five minutes from the critical condition. Also, when the Bluetooth module receives the warning signals from the main controlling unit, the safety electrical switches will run through the same steps. Another main advantage of the proposed system is the size of the safety electrical switch, where it can be fixed to a commercially available standard sunk box as shown in Fig. 4.

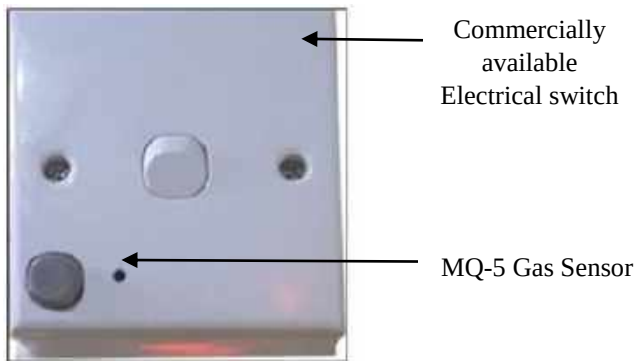


Fig. 4. The prototype of safety electrical switch attached to a standard sunk box

D. Automatically disconnecting regulator

The Gas regulator is the most useful item which connects the gas pipe of the appliance and the gas cylinder. There are many commercially available regulators that all have quality certificates such as ISO, SLS, etc. Usually, gas regulators have a long life without any failures. But the gas pipe of the appliance may damage in long time usage. Therefore, there may be invisible, smaller holes in the gas pipe. Eventually, there could be a high probability to leak LPG through small holes in the gas pipe. In addition to that, there can be a risk of gas to be leaked from any weak places of the appliance, and due to some inappropriate user actions.

Here, the auto disconnecting gas regulator was developed to avoid gas fire accidents. This mechanism consists of 12 V solenoid lock and commercially available common gas regulator. The 12 V solenoid lock is used for preventing the electrical sparks. As shown in Fig. 5, the specially moulded case has been fixed on the outside of the LPG regulator using epoxy resin. Also, a small spring and liver technique is attached to the regulator valve. The 12 V solenoid lock is used for triggering the liver. If the solenoid lock powered up, the attached spring gets rest by removing the regulator from the cylinder. The operating signal of the solenoid lock is given by the main controlling unit when the harmful level of the LPG concentration in air is detected.

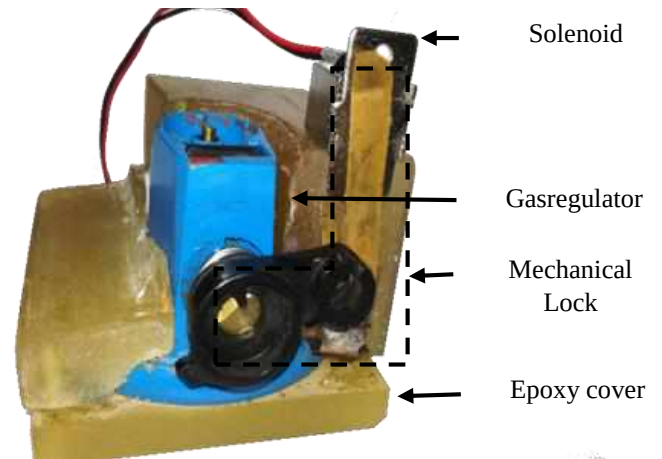


Fig. 5. The prototype of automatically removing the regulator

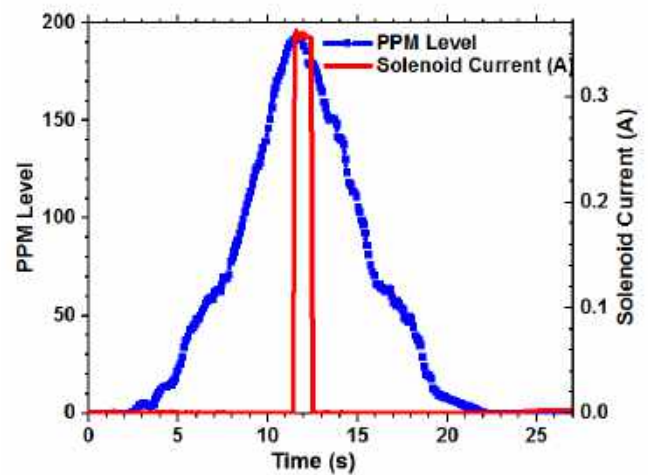


Fig. 6. PPM level and Solenoid Plus at Gas Leakage

III. RESULTS

The solenoid current pulse in a case for gas leakage with a calculated PPM level is shown in Fig.6 to validate the proper operation of the designed system. This clearly identifies the powering up of solenoid at the 200PPM level, which validates the proper operation of the designed system.

TABLE 1. COMPARISON OF COMMERCIALY AVAILABLE SYSTEMS WITH THE PROPOSED SYSTEM

Feature	Commercially available system	Designed system
Emergency shut off and automatically disconnecting the regulator from the cylinder	No	Yes
Instantly cuts off the power supply of the nearby area	No	Yes
Send SMS alert to corresponding people	Yes	Yes
Back up battery for work even in power failures	Yes	Yes

IV. CONCLUSION

LPG leakage and fire security system has been developed to reduce the harm to human and domestic assets. The main task of this system is, when a gas leakage occurs, the leak is detected quickly and disconnect the gas regulator from the gas cylinder, disconnect the power supply to all electrical switches in the nearby area, and send warning alert to the relevant important parties. This system can be powered from the mains as well as from a rechargeable battery such that the reliability of operation of the system is guaranteed all the time.

The designed system is incorporated with several attractive new features making it unique as compared to the existing gas detector and security systems. Bluetooth technology is used as a communication medium between the main controlling circuit and the safety electrical switches in the vicinity. The main controlling unit and safety electrical switches connected with each other automatically when the system is powered up. As soon as gas leakage occurs, the Gas fire security system drastically reduces the gas fire risk by separating the gas pipe with the regulator from the cylinder. In addition, even for possible gas leakages from the cylinder itself and if it exceeds the permissible level, the power supply of nearby area switched OFF and the relevant users will immediately be alerted with an SMS. This security system can be installed both in domestic and industrial places. Integrating the designed system with the IoT platform is envisioned in a futuristic smart home.

ACKNOWLEDGMENT

The authors wish to thank Mr. P. L. A. K. Piyumal, Mr. Geethanjana Wanigasekara, Mr. A. M. D. G. B. Abesinghe, Mr. S. B. M. S. S. Gunarathne and Mr. H. P. M. Udayanga for the technical assistance and Department of Physics and Electronics, University of Kelaniya, Sri Lanka for their financial assistance.

REFERENCES

- [1] L. A. Tawalbeh, A. Basalamah, R. Mehmood, and H. Tawalbeh, "Greener and Smarter Phones for Future Cities: Characterizing the Impact of GPS Signal Strength on Power Consumption," *IEEE Access*, Vol. 4, pp. 858-868 (2017)
- [2] N. Mohamed, J. A. Jaroodi, I. Jawhar, S. L. Molnar, and S. Mahmoud, "SmartCityWare: A Service-Oriented Middleware for Cloud and Fog Enabled Smart City Services," *IEEE Access*, Vol. 5, pp. 17576-17588 (2017)
- [3] Annual Report on LPG related accidents, The high-pressure gas safety institute of Japan (2013)
- [4] <http://www.shippai.org/fkd/en/cfen/CC1000180.html> [Accessed on 21/05/2019]
- [5] "The High-Pressure Gas Safety Institute of Japan (KHK)", *Annual Report on Liquefied Petroleum Gas (LPG) Related Accidents*, 2013. [online] Available: https://www.khk.or.jp/Portals/0/resources/english/dl/annual_report_lpg_2013.pdf [Accessed: January 5, 2019]
- [6] L. Fraiwan, K. Lweesy, A. Bani-Salma, N. Mani, "A Wireless Home Safety Gas Leakage Detection System," in *1st Middle East Conference on Biomedical Engineering*, Sharjah, United Arab Emirates, Apr. 2011, pp. 11-14.
- [7] A. Mahalingam, R. T. Naayagi, N. E. Mastorakis, "Design and Implementation of an Economic Gas Leakage Detector," in *ACA' 12 Proc. of the 11th international conference on Applications of Electrical and Computer Engineering*, Mar. 2012, pp. 20-24
- [8] T. Soundarya, J.V. Anchitaalagammai, G. D. Priya, S.S. Karthick kumar, "C-Leakage: Cylinder LPG Gas Leakage Detection for Home Safety", *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, Vol. 9, No. 1, Feb. 2014, PP 53-58
- [9] S. S. Kavitha, S. Senthilkumar, "A Wireless Gas Leakage & Level Detection with Auto-Renewal System," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, Vol. 4, Issue 4, April 2015, pp. 2095-2100.
- [10] S. Jagtap, P. Bhosale, P. Zanzane, and J. Ghogare, "LPG Gas Weight and Leakage Detection System Using GSM," *International Journal for Research in Applied Science & Engineering Technology*, Vol. 4, No. 3, Mar. 2016, pp. 716-720
- [11] T. Arpitha, D. Kiran, V. S.N. S. Gupta and P. Duraiswamy, "FPGA-GSM based Gas Leakage Detection System," in *2016 IEEE Annual India Conference (INDICON)*, Bangalore, 2016, pp. 1-4.
- [12] O. O. Nuga, K. A. Amusa, A. J. Olanipekun, "GSM-Based Gas Leakage Detection and Alert System," *ABUAD Journal of Engineering Research and Development (AJERD)*, Vol. 1, No.1, Oct. 2017, pp. 100-104.
- [13] B. N. Mohapatra, A. Dash, D. K. Chaubey, "LPG Gas Auto Booking By GSM and Leakage Detection with Auto Switchable Exhaust Fan," *International Journal of Science, Engineering and Technology Research (IJSETR)*. Vol. 6, No. 3, Mar. 2017, pp. 408-411.[online] https://www.hsa.ie/eng/Topics/Liquid_Petroleum_Gas_LPG/ [Accessed: January 5, 2019]
- [14] Sriwati, N. I. Ilahi, Musrawati, S. Baco, A. Suyuti, A. Achmad, E. Umrianah, "Early Leakage Protection System of LPG (Liquefied Petroleum Gas) Based on ATMega 16 Microcontroller," *IOP Conf. Series: Materials Science and Engineering*, Vol. 336, p.012021, 2018.

Anomaly detection in cloud network data

Tharindu Lakshan Yasarathna*
Software Engineering Teaching Unit
Faculty of Science,
University of Kelaniya, Sri Lanka
tharinduyasarathna@gmail.com

Lankeshwara Munasinghe
Software Engineering Teaching Unit
Faculty of Science,
University of Kelaniya, Sri Lanka
lankesh@kln.ac.lk

Abstract: Cloud computing is one of the most rapidly expanding computing concepts in the modern IT world. Cloud computing interconnects data and applications served from multiple geographic locations. A large number of transactions and the hidden infrastructure in cloud computing systems have presented a number of challenges to the research community. Among them, maintaining the cloud network security has become a key challenge. For example, detecting anomalous data has been a key research area in cloud computing. Anomaly detection (or outlier detection) is the identification of suspicious or uncommon data that significantly differs from the majority of the data. Recently, machine learning methods have shown their effectiveness in anomaly detection. However, identifying anomalies or outliers using supervised learning methods still a challenging task due to the class imbalance and the unpredictable nature and inconsistent properties or patterns of anomaly data. One-class classifiers are one feasible solution for this issue. In this paper, we mainly focused on analyzing cloud network data for identifying anomalies using one-class classification methods namely One Class Support Vector Machine(OCSVM) and Autoencoder. Here, we used a benchmark data set, YAHOO Synthetic cloud network data set. To the best of our knowledge, this is the first study that used YAHOO data for detecting anomalies. According to our analysis, Autoencoder achieves 96.02 percent accuracy in detecting outliers and OCSVM achieves 79.05 percent accuracy. In addition, we further investigated the effectiveness of a one class classification method using another benchmarked data set, UNSW-NB15. There we obtained 99.10 percent accuracy for Autoencoder and 60.89 percent accuracy for OCSVM. The above results show the neural network-based methods perform better than the kernel-based methods in anomaly detection in cloud network data.

Keywords: Anomaly Detection, Cloud Computing, Machine Learning, One-Class Classification

I. INTRODUCTION

Cloud computing is one of the most rapidly expanding computing concepts in the modern IT world. Cloud computing interconnects data and applications served from multiple geographic locations. Emerging Internet technologies have extended the capabilities of cloud computing which is regarded as an upgraded version of utility computing as well. For example, cloud computing technologies are widely used for subscription-based or pay-per-use services [1]. Using multitenant architecture, Cloud computing delivers a single application via a browser to millions of users in different geographical areas around the world. This technology is called SaaS (Software as a service). Since, Cloud computing involved with thousands of user transactions, information, and communication, cloud security is considered one of the most important aspects of cloud computing. To provide secure cloud computing platforms or

services, the availability, integrity, and confidentiality need to ensure. However, a high volume of transactions happens within nanoseconds and the hidden infrastructure of cloud computing are some of the challenging factors for ensuring the security of the cloud computing systems [2]. On the other hand, security attacks are not known to anyone. Some attackers execute masked attacks and, gradually leak sensitive information. To avoid this kind of threat, many approaches and methods have been introduced for security monitoring in cloud computing [3]. Among them, anomaly detection using machine learning algorithms is a widely used approach. Anomaly detection (or outlier detection) is the identification of suspicious or unusual data that significantly differs from the majority of the data. Identifying anomalies or outliers using supervised learning methods is a challenging task because the nature or properties of outliers are not consistent. This paper mainly focuses on analyzing cloud networks and identify abnormal activities in cloud network data using machine learning methods. Here, we used YAHOO Synthetic and real time-series with labeled anomalies data set [4]. To best of our knowledge, this is the first study that uses YAHOO synthetic data for anomaly detection. In addition, we further investigated the effectiveness of a one-class classification method using another benchmarked data set, UNSW-NB15 [5]. We will present our research in detail in the next section. In the following section, we discuss some of the existing state-of-art research related to our topic. In section 2, we present our study with a detailed explanation followed by experimental analysis. We conclude this paper by discussing some open questions to the research community and, our future direction of the research.

II. RELATED WORK

It is a well-known fact that cloud computing has added a new computing paradigm. Like real clouds contain the ice crystals and water drops, the explanation of ‘cloud’ in cloud computing is the collection of networks [6]. Cloud computing extends capabilities of the IT over the internet and it enables subscription-based or pay-per-use services. Users can select cloud plans depending on their requirements and available budget. As a consequence, modern businesses are using cloud computing platforms personalized services to their users. Thus, cloud security has become one of the main factors affects the reliability of cloud computing. For example, especially when confidential information shares over the cloud services [7]. Confidential personal data stored on remote servers provide data availability when necessary through the cloud network which can be crucially vulnerable for major security threats. In addition, sharing data in the cloud is a problem when the cloud service provider is not trusted [8]. Besides, they need to respond to fast-moving changes in data promptly, especially in case of security threats. Unfortunately, there has been no effective ways to

handle and analyze constantly growing datasets manually in the cloud [3]. For example, detecting outliers or anomalies is vital to prevent potential security threats. Anomaly detection refers to the identification of elements or events that are not suitable for the expected pattern or other elements of a data set that is not normally detectable by human experts. This anomaly usually translates into problems such as structural defects, errors or fraud [9].

Machine learning methods have shown their effectiveness in finding anomalies or outliers. Machine learning types can be classified as unsupervised learning, semi-supervised learning, and supervised learning. The progress made in anomaly detection has been mostly supported approaches using supervised machine learning algorithms that need big labeled data-sets to be trained. However, supervised machine learning methods still having difficulties because learning from historical anomaly data and may not effective in forecasts future potential anomalies. Because we cannot pre-defined anomalies [10] and anomalies or outliers do not exhibit consistency patterns or properties. One possible way to solve this problem is training a model to effectively recognize non-anomaly data. For example, one-class classifiers can train to recognize non-anomaly data. In this research, we have used two one-class machine learning algorithms to detect anomalies in the cloud network. They are a one-class support vector machine and Autoencoder. One class support vector machine is the extension of SVM for unlabeled data, could be used for anomaly detection [11] [12]. Autoencoder is a type of neural network-based learning algorithm, which has one approach to automatically learn features from unlabeled data [13].

III. ANOMALY DETECTION IN CLOUD NETWORK

A surge of cloud computing has presented a number of challenges to the research community. Among them, identifying anomalous data is an immense challenge due to the complexity, heterogeneity and dynamic behavior of the data. Anomalies and outliers pose a huge security threat. Therefore, many types of research have been carried out on detecting anomalies in cloud network data. In this section, we discuss the practical use case for anomaly detection in cloud network data with a one-class classification approach.

A. What is anomaly data?

Cloud network services provide their service to users through secured methods such as authorized access. However, it is difficult to ensure that the data received will always be normal. Some data differing significantly from the majority of the data are classified as the anomaly data. Fig. 1 shows some examples of anomaly data in cloud networks.



Fig. 1. Anomaly graphs

The first figure of Fig. 1 shows that normal data (blue dots) are clustered around a particular point and the outliers or

anomaly data (red dots) are far away from the cluster centroid compared to the normal data. In the second figure, the sequential pattern of data has disrupted in the middle and the anomalous data points lay away from the normal data sequence.

B. One-class classification for anomaly detection

In this research, we choose a supervised machine learning approach for anomaly detection in cloud network data. Our task is to classify data into two classes; anomalies or non-anomalies. Supervised learning methods need rich data to provide good models [14]. Especially, the training data should contain enough data points to represent all possible scenarios. However, in most cases, the training data are imbalanced [15] and thus, produce biased models (model overfitting, Under-fitting). In anomaly data, the number of outliers is proportionately very low compared to normal data. Due to this rarity of anomaly data (data imbalance), multi-class classification approaches have not been effective in anomaly detection. Apart from the data imbalance, the anomalies do not have predefined or consistent patterns. The nature and the properties of anomaly data may vary drastically over time. Thus, the models trained using such data become invalid due to the unpredictable nature of anomaly data. To solve this issue, we analyzed the effectiveness one-class classification approach for anomaly detection. In the one-class classification approach we trained a supervised machine learning model using non-anomalous data so, the model can detect(classify) the non-anomalous data in the presence of anomalous data. In this research we used two one-class classifiers; OCSVM and Autoencoder a neural network-based method.

OCSVM is a natural extension of the one-class support vector machine that is supervised learning models that analyze data and recognize patterns, which are often used for both classification and regression tasks. The OCSVM algorithm (through the kernel) allocates input data instances in a high dimensional feature space and iteratively finds the largest margin hyperplane that best separates the training data from the source [16]. Kernel-based learning methods use implicit mapping of input data in a high-dimensional feature space defined by a kernel function [17]. Therefore, in OCSVM, the support vector model is focused to train data that has only one class, which is considered as the most discriminative class. In our research, the equivalent is the non-anomalous data. This infers the properties of normal cases and can predict the examples are unlike the typical examples from these properties. This is useful for identifying anomalies because of the lack of coaching examples identifies anomalies. For example, there are usually only a few cases of network interference, theft, or other anomalous behavior and thus, easier ways of identifying anomalies. Therefore, one-class models are effective in this situation.

Autoencoder is a neural network-based supervised machine learning algorithm. In general, Autoencoder has three layers; an input layer, an encoding layer, and a decoding layer. An example of Autoencoder shown in Figure 3. When considering the encoder and decoder, they are similar to zip and unzip functions for compression, learned from the dataset. For the best representation of inputs, the encoding layer is forced to learn and the neural network is trained to reconstruct its inputs for this purpose. In this research, Autoencoder is

trained to learn the normal behavior of the corresponding data instances and to be activated if anomalous conditions are measured [18]. If an Autoencoder can learn the correlations between the set of data features that describe the state of a cloud network dataset, then it can consequently notice changes in these correlations that indicate an abnormal state.

It is a common belief that neural network based models perform better than conventional machine learning methods. Numerous researches have shown the effectiveness of neural network models against traditional machine learning methods. On the other hand, neural network models efficiently learn from huge data sets. Fig. 2 shows the variation of model performance with the size of the data. It clearly shows that the neural network models perform far better with large datasets.

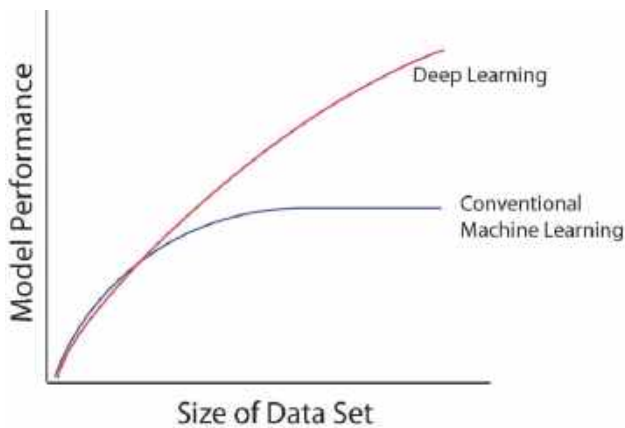


Fig. 2. Variation of model performance over the size of a data set

Further, one-class classifiers to detect anomalies in the cloud network have been selected in order to increase the performance of the study and to achieve accurate results. Two class classifiers could be also used for those tasks, but a trained pre-defined model for detecting anomalies is not possible for every scenario. While using one-class classifiers, a model for the normal flow and activates in the cloud network have been also developed. For this use case, the ultimate goal was to detect anomalies behavior in cloud network data by using a One-class support vector machine and Autoencoder. The discussion has been carried out on some of the existing state-of-art research related to the topic as well. Anomaly detection is the identification of various kinds of activities that are deviated from the normal data behavior. Typically, an anomaly is opened to many newer problems to every domain, and identifying these kinds of activities in cloud network data provided a reasonable solution for cloud users.

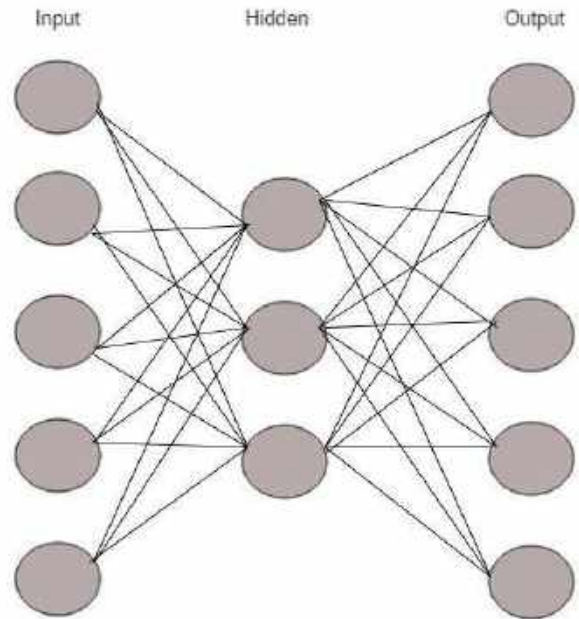


Fig. 3. Autoencoder 3-layer architecture

IV. EXPERIMENTAL ANALYSIS

A. Data

To test our methods, we used two benchmark data sets; YAHOO Synthetic and real time-series with labeled anomalies data set have been used in this research. Data received from the real yahoo servers was essentially a soft version. It provided a good tool for exploring algorithms without the complexity of real data. Yahoo data set contains four (A1, A2, A3, A4) benchmark data sets. The A4 Benchmark includes changepoint anomalies and it contains 8 data features that are being described below in Table I.

- Timestamps: the UNIX timestamp marks every hour (hourly sampled data)
 Value: time series value at relevant timestamp
- Anomaly: for an outlier value will be 1
- Changepoint: if the change point was there, the value will be 1
- Trend: the additive trend value for this timestamp
- Noise: the additive noise value for this timestamp
- Seasonality1: seasonality value for a period of twelve hours
- Seasonality2: calculated seasonality value for the daily period
- Seasonality3: calculated seasonality value for the weekly period

TABLE I. FEATURES OF YAHOO DATA SET WITH EXAMPLE DATA

Timestamp	Value	Anomaly	Changepoint	Trend	Noise	Seasonality1	Seasonality2	Seasonality3
1417590000	779.0698	0	0	484	4.135452	142.5	61.85775	86.57662
1417593600	930.3786	0	0	486	-1.03502	246.8172	119.5	79.09633
1417597200	1019.256	0	0	488	5.752064	285	168.9985	71.50542
1417600800	1020.809	0	0	490	13.19709	246.8172	206.9801	63.8145
1419188400	2492.884	0	0	2002	225.7357	235.2	-76.8693	106.8176
1419192000	2440.698	0	0	2004	63.51916	407.3783	-148.5	114.3
1419195600	2385.058	0	0	2006	-2.95404	470.4	-210.011	121.6225
1417683600	-102.321	1	0	0	2.255834	97.2	41.43646	-15.9078
1418997600	3001.277	1	0	1896	78.74763	-407.378	257.2095	-227.961

This data set contained 335,999 records and each record has 8 features that were generated by Yahoo servers. Also, it contained approximately 1800 anomalies data records for this activity. These statistics clearly indicate the class imbalance in the data. We have used approximately 230,000 normal data records to train our model (approximately 70% of the no anomalous data). For the test data set, we used approximately 68,000 records (approximately 30% of the non-anomalous data and approximately 1800 anomalous data).

In addition, we used the UNSW-NB15 network data set for further investigation of selected machine learning algorithms. This data set contains, 138,300 total data with approximately 93,000 clean data and approximately 45800 anomalies data records. To train the model around approximately 65,100 normal data records have been used (approximately 70% of the non-anomalous data). For the test data set, we used approximately 73,200 records (approximately 30% of the no anomalous data and approximately 45800 anomalous data).

B. Experiments

The data sets were pre-processed to clean and to add labels (True for non-anomaly data and false for anomaly data). The pre-processed non-anomaly data set was split randomly into two subsets, one with 70% instances for training and the remaining 30% was merged with anomaly data set for the testing purpose.

a) Experiment 1: One Class Support Vector Machine for anomaly detection

First, we tested the effectiveness of OCSVM for anomaly detection. Here, we used R kernlab [19] library which includes OCSVM implementation in R. The results are shown in Table II.

b) Experiment 2: Autoencoder for anomaly detection

In our second experiment, we tested the effectiveness of Autoencoder for anomaly detection. Here, we used R h2o package which includes Autoencoder [19] implementation. The results are shown in Table II.

C. Experimental Results

In Table II, we compare the anomaly detection performance of one class classifier used in our experiments. It shows that 79.17% overall accuracy for a one-class support vector machine on yahoo cloud network data set and the 60.89% accuracy on UNSW-NB15 network data. In contrast, Autoencoder shows 96.02% accuracy on yahoo cloud network

data set and 99.54% accuracy on the UNSW-NB15 network data set.

TABLE II. STATISTICS OF DATA

Classification method	Accuracy(%)	
	UNSW-NB15 Data	YAHOO Anomaly Data
One-Class Support Vector Machine	60.89	79.05
Autoencoder	99.10	96.02

The above results show that the neural network-based methods perform better than the kernel-based methods in anomaly detection in cloud network data sets we used in our experiments.

V. DISCUSSION

In this section, we discuss limitations, issues and future direction of this research. In this research, we identified several limitations. Among them, we noticed that both models failed to recognize the non-anomalous outliers show in Fig. 4 and Fig. 5. The Autoencoder struggles in the tail of both data sets. The reconstruction error plots show that the error count accelerates upwards (in the YAHOO data set after approximately 230,000 records and the UNSW-NB15 data set is approximately after 65,000 records). This observation tells that Autoencoder recognizes some normal or innocent data as anomalies data. This information can be used to define the decision boundary for anomaly detection assuming that the last data instances are outliers with respect to the rest of instances in the clean data set.

A. Limitations

a) *Class Imbalance Problem:* The presence of a roughly equal number of instances in each class paves way for most machine learning algorithms to work their best. However, class imbalance is a common problem in anomaly detection data sets. As an example, within our data sets the sum of anomalies data is comparatively very small. As a solution, some research has used alternative performance metrics (true positives, true negatives, false positives, false negatives) instead of the standard precision of calculating the number of errors to compare solutions. There are many ways to solve this problem, owing to its universality. In general, these can be classified into two main categories: 1) based on sampling; 2) based on cost functions. Basic sampling can be divided into three categories: a) oversampling b) sampling c) mixed oversampling and subsampling. In our future works, we will investigate the effectiveness of those sampling techniques.

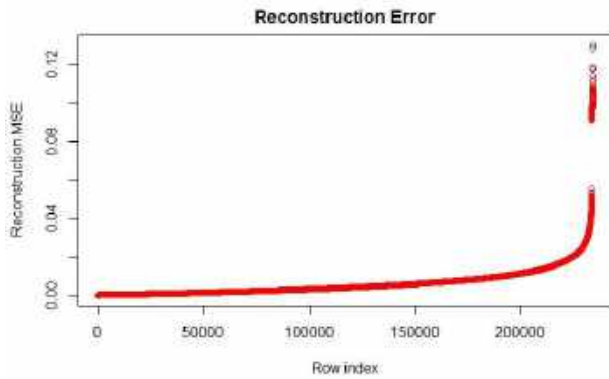


Fig. 4. Reconstruction error for the yahoo data set

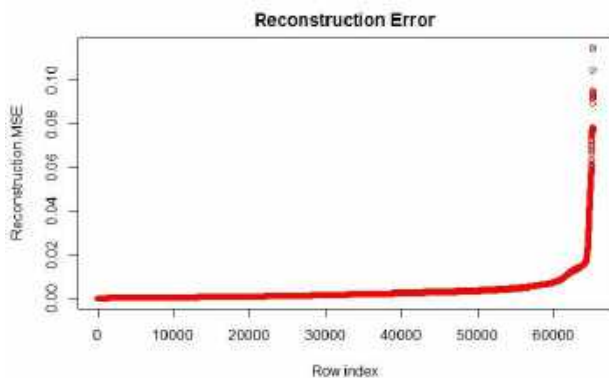


Fig. 5. Reconstruction error for the UNSW-NB15 data set

B. Analysis of Feature Importance

Further, we carried out an investigation about the variable importance of the data for Autoencoder. Variable importance provided the statistical significance of the variables in the data set that was used to generate the model [20]. This method was very helpful when we use one-class classification methods to make predictions according to our data if they contribute or do nothing with our generated model. The variable importance of the data sets is shown in Fig. 6 and Fig. 7. In our future works, we will analyze the trade-off between accuracy and reconstruction error by selecting different feature sets according to their importance.

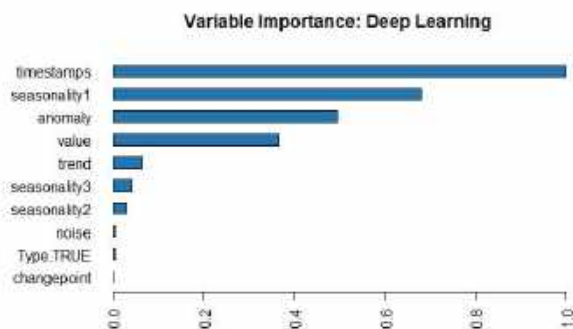


Fig. 6. Variable importance of yahoo data set

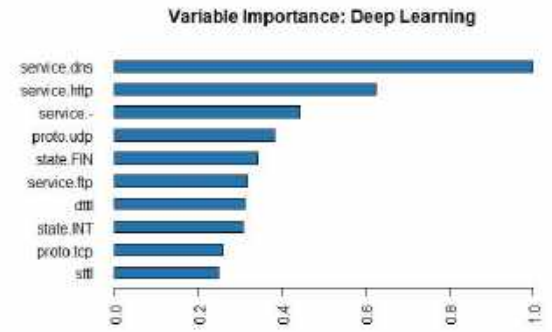


Fig. 7. Variable importance of UNSW-NB15 data set

C. Future Work

Apart from that, it is worth investigating the hyperparameters of the classifiers we used in our experiments. Fitting a neural network model takes considerable time. To find the best model, one can carry out a grid search for model parameters. However, this takes lots of time and need high computational power. Thus, we had to use pre-determined or default parameter values of the models.

VI. CONCLUSION

In this research, we investigated the anomaly detection in cloud network data using supervised machine learning methods. Instead of multi-class classifiers, we choose two one-class classifiers, One-Class Support Vector Machine(OCSVM) and Autoencoder algorithms which were trained to detect outliers. We tested both algorithms on two benchmark data sets; yahoo dataset and UNSW-NB15 dataset. To the best of our knowledge, this is the first study that used YAHOO synthetic data for anomaly detection. The experimental results show that the neural network-based Autoencoder performs better in contrast to the kernel-based OCSVM algorithm in anomaly detection. Also, a one-class classification approach provides a better solution for class imbalance problems in anomaly detection.

REFERENCES

- [1] G. Gruman and E. Knorr, "what cloud computing really means," *InfoWorld*, April 2008.
- [2] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation Computer Systems*, vol. 28 pp. 583 – 592, 2012.
- [3] M. Gander, M. Felderer, B. Katt, A. Tolbaru, R. Breu, and A. Moschitti, "Anomaly detection in the cloud: Detecting security incidents via machine learning," vol. 379, pp. 103–116, Jan. 2013.
- [4] M. Thill, W. Konen and T. Bäck, "Online anomaly detection on the webscope S5 dataset: A comparative study," *2017 Evolving and Adaptive Intelligent Systems (EAIS)*, Ljubljana, 2017, pp. 1-8.
- [5] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Jan. 2016.
- [6] P. Srivastava and R. Khan, "A review paper on cloud computing," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 6, 2018.
- [7] P. Schoo, V. Fusenig, V. Souza, M. Melo, P. Murray, H. Debar, H. Medhioub, and D. Zeghlache, "Challenges for cloud networking security," vol. 68, pp. 298–313, Jan. 2010.
- [8] S. Aldossary and W. Allen, "Data security, privacy, availability and integrity in cloud computing: Issues and current solutions," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, 2016.

- [9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, July 2009.
- [10] J. Pereira, "Unsupervised Anomaly Detection in Time Series Data using Deep Learning," Master's thesis, Instituto Superior Tecnico, Lisbon, 2018.
- [11] B. Scholkopf, R. Williamson, A. Smola, J. Shawe, Taylor, and J. Platt, "Support vector method for novelty detection," vol. 12, pp. 582–588, Nov. 1999.
- [12] Y. Chen, J. Qian, and V. Saligrama, "A new one-class svm for anomaly detection," pp. 3567–3571, Oct. 2013.
- [13] R. Chalapathy, S. Chawla, "Deep Learning for Anomaly Detection: A Survey", 2019.
- [14] S. Omar, M. Ngadi, H. Jebur, and S. Benqdara, "Machine learning techniques for anomaly detection: An overview," *International Journal of Computer Applications*, vol. 79, Oct. 2013.
- [15] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," vol. 4, 2008.
- [16] L. Maglaras and J. Jiang, "A novel intrusion detection method based on ocsvm and k-means recursive clustering," vol. 2, pp. 1–10, Jan. 2015.
- [17] A. Karatzoglou, A. Smola, and K. Hornik, "kernlab - an s4 package for kernel methods in r," vol. 69, pp. 721–729, Nov. 2004.
- [18] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini, "Anomaly detection using autoencoders in high performance computing systems," 2018.
- [19] B. Mitra, McCausland, H. Kalutarage, "Modelling iot anomaly detection," vol. 60, pp. 44–45, 2018.
- [20] P. Wei, L. Zhenzhou, and J. Song, "Variable importance analysis: A comprehensive review," vol. 142, June 2015

Minimization of fraudulent activities in land authentication through Blockchain-based system

L. W. D. C. Jayabodhi*
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
jdinith72@gmail.com

Chathura Rajapakse
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
chathura@kln.ac.lk

J. M. D. Senanayake
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
janakas@kln.ac.lk

Abstract: In Sri Lanka, the land administration process has not been digitized, which has led to plenty of conflicts in determining the real ownership of lands and drastically increasing the land transaction processing time, which has resulted in many fraudulent activities concerning land management. The existing system is majorly paper-based and centralized. Administrators who have the powers to alter the system according to their wishes hold a huge threat for information security and is a centralized system that bears the risk of a single point of failure. Hence, transitioning into another system that could mitigate the drawbacks in the current land authentication system has become a vital need and currently, the knowledge regard this area is very much limited. The decentralized nature of the blockchain-based system has the potential of diluting major limitations in the current system. The overall objective of this research is to mitigate fraudulent activities in land authentication systems through the blockchain-based technological approach. A prototype of a smart contract-based model has been created and verified with the involvement of the actual users. Since the smart contract-based land authentication model verifies the land ownership within a short period, the transaction processing time narrows down from a few months to a couple of minutes. The results show signs of considerable improvements in the efficiency and the security of the land authentication process, the users who interacted with the prototype and presented positive comments.

Keywords: Blockchain mining, Land authentication, Smart contract

I. INTRODUCTION

The study will consider the applicability of the blockchain technology to the land authentication and the land ownership transfer process. The existing system is very much of a paper-based system. The land ownership transfer documentation (creation of a deed) takes about a month to go through the process. This arises space for the many fraudulent activities in land ownership.

Also, the process for identifying the real owner of the land can become a complex process when there are disputes in the deed. Therefore, the general public cannot do it by themselves and they have to rely on the services of a legal officer. Even though the legal officer with his expertise may take several days for verification. If the land had undergone many transactions and alterations to the geographical dimension the process becomes extremely complex. So that the legal officers would be reluctant to go ahead of such transactions as the real owner of the land is doubtful.

Therefore, considering these drawbacks it would be best to have a digitalized system for land authentication. Through the digitalized system, the land ownership transfers will be

tracked and the sensitive data such as the land boundaries will be stored in terms of the geo-locations. Hence the information such as the land location and the ownership of the land will be available for the public with a matter of few clicks.

But the digitalized system has its drawbacks as all the information is stored in the central location, it provides too much power on the administrator. Also, there would be a disastrous situation if the digitalized system is damaged due to some external factor. Since the land authentication information is highly sensitive, it would very much risky to store the land information in the digitalized system with central information storage. So the problem of lack of a robust highly secure system for the land authentication is remaining unsolved.

Hence to mitigate the above drawback, a blockchain-based system is used to store highly sensitive land information such as the land ownership details and the dimension information. A proof-of-concept is constructed and tested in a controlled environment. The immutable and distributed properties of the blockchain would address the abovementioned problems related to the Land Authentication System.

II. LITERATURE REVIEW

A. The legal framework in Sri Lanka related to land authentication

Roman-Dutch Law, Islamic Law, Thesawalamai Law, and European Law existing in the pre-British period, influenced the crafting of Sri Lankan land law. In the period of the British ruling, these laws were amended according to the common English Law. After independence in 1948, they have been many reforms regarding the land administration put forward by each of the governments ruling at that time [1]. According to Perera, the number of court cases with unsolved lands is listed below.

The major reason for land disputes to drag so long is the complication of the land laws and the issues regard to determine the real legal ownership of the land.

According to Perera many people would not bring the land dispute to the court but would try to resolve them by themselves using the methods such as “forceful trespassing and evictions, informally compromising in between parties to discussions “. These complications regard to land ownership have caused many bottlenecks in acquiring the best use of the land and hence hindering the economic development of the country.

TABLE1. NO OF COURT CASES PENDING

District Courts	No of years Pending					Unclassified	Total
	5	10	20	30	>30		
Colombo	877	194	62	2			1135
Mt Lavinia	472	49	1				522
Monarua	152						152
Homagama	238	144					382
Avissawella	677	426	605	145	24		1877
Panadura	389	72	5	8		8	482
Gampaha	841	362	235	5		3	1446

Source: Sri Lanka Study on Land Markets, Ministry of Lands, 2000

According to the evaluations made under the Implementation Completion and Result Report in 2007 published by the world bank, the Land Titling Project which was implemented with an investment of \$5 million from the World Bank funds has been categorized as unsatisfactory [2].

B. Land authentication management with blockchain

The major issue which is arising from the existing land authorization system is that there is room for creating fraudulent deeds and illegally selling the land to another party. This is a basic use case of double-spending which blockchain heavily used in providing solutions.

Cleverence Kombe identifies several functional requirements that are to be fulfilled in an Integrated Land Management Information System (ILMIS). According to him, those functions are:

- Validation of the Land Title Information
- Validation of the Land Title Transactions
- Notarization of the Legal Registration process [3].
- Distribute the fingerprint information stored in the Integrated Land Management Information System using the blockchain technology.

Kombe also mentions about the importance of an ILMIS in the face of the natural disaster and the importance of using the blockchain in order to increase the availability of the system. Their research is more focused on the usage of the Fathom blockchains. A third-party company which is providing blockchain-based API to manage the blockchains. Also, the legislative framework covered in that research is the legislative framework of Tanzania [3].

Blake and Sanders address the issue of interconnecting the Land Information System with the other digitalized services such as the mortgage banks. Financial institutions and also other public bodies. That research is based on Nigeria where most of the government services are digitalized. But considering the Sri Lankan context, there is still a large amount of room for digitalization [4]. Therefore, that kind of integration can be focused on the later phases of the implementation.

C. Mitigation of double-spending with the use of Blockchain

There have been several attempts to mitigate the frauds is caused related to the double-spending, by the use of the blockchain technology.

Hissu Hyvärinen discusses the use of blockchain in the taxation process through his research, "A Blockchain-Based Approach Towards Overcoming Financial Fraud in Public Sector Services". In the taxation process, many individuals

taxed twice for the incomes they earn in offshore. But most of the tax authorities in different countries have got together and have created tax relief schemes to prevent this kind of double taxation. But many of the public illegally used these tax relief schemes. The people use it for their advantage due to the easily forgeable legal documentation and due to the insufficient exchange of information between the international tax authorities [5]. Hissu Hyvärinen intends on creating a blockchain-based prototype to mitigate the tax frauds and also increase the transparency of the flow of dividends.

George Foroglou in his paper titled Further applications of the blockchain explain several other use cases of the blockchain. Some of them are providing a solution to the double-spending issue. Voting or Election is another area where the double-spending happens in a form of casting votes by an individual. George Foroglou mentions about the BitCongress where the Ethereum network has been used to create a blockchain-based voting system. In that, each of the individuals is entitled to "vote coin" which he may use only one time to vote. The casted vote would be recorded on the blockchain after verifying it. There are several other voting technologies that George Foroglou mentions in his paper. They are Remotengrity, a voting system that provides every physical vote with a cryptographic code to verify the authenticity of every physical vote and Agora Voting [6].

Foroglou speaks about another system called "Smart property" in order to prevent people from overspending their properties (a form of double-spending) with the integration of the internet of things. Many of the appliances that we use today can be controlled through the internet. A smart home can be designed in a way that the the owner needs a digital key to open the doors of the house.

A blockchain system can be built such that it stores the information about the owner and also the digital key to open the smart home. If the house owner fails to pay the rental fee to the smart home, the digital keys would be updated and his access to the smart home will be revoked. In this way ownership of the applications can be controlled with the use of the internet of things and also Blockchain [6].

D. Blockchain mining process

Blockchain mining is the process through which the transactions are verified in the Blockchain. It is a high-cost operation that includes a tremendous amount of computing power. In a public blockchain, any user can sign up as a miner and he would be rewarded for each of the blocks that he has mined.

- Cryptographic Hash Function - It is a function, which generates a binary number when a string is passed into it, and that number is unique to that particular string. Also, the number cannot be reverse calculated and obtained the initial input value.
- SHA256 Algorithm - This provides a 256-bit binary Cryptographic Hash number for each of the inputs given.
- Consensus - The majority of the nodes in the Blockchain come in agreement that a particular block is a legitimate block.

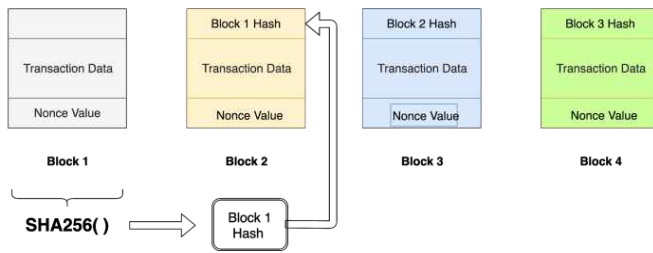


Fig. 1. Structure of a Blockchain

As shown in the above diagram the blockchain is a collection of data nodes known as blocks. Each of the blocks contains

- Previous blocks hash
- Transactional Information
- Nonce Value

The nonce value is an arbitrary value inserted by the miner. In the Block 2 hash value is generated by sending the Block.

1's hash, Transaction data in Block 2 and the Nonce value to the SHA256 function. The Blockchain specifies the difficulty number in which the generated hash should not exceed. The duty of the miner is to generate that hash value altering the nonce value. As the SHA256 cannot be reverse-engineered the only mechanism to derive the relevant nonce value is through guessing. Once the miner successfully derived the nonce value it is sent to the other nodes for the approval. When the Blockchain reaches consensus, the block is added to the Blockchain and the miner is rewarded.

If malicious users altered the transaction data in the Block 2 the resulting hash would be changed, and the Block 2 should be reminded. Also, as Block 2 hash is present in Block 3 it would be changed, and it also should be reminded. Similarly, all the Blocks after Block 2 should be reminded. So the need for a continuous chain of remaining makes impossible for a malicious user to alter the data in the blockchain.

Several algorithms are being used to perform the consensus in the Blockchain.

1) Practical Byzantine Fault Tolerance Algorithm (PBFT)

This algorithm works in three phases,

- The leader broadcast a new request.
- The validators (miners) sign and broadcasts the prepare message for each of the requests.
- If the necessary, amount of prepare messages are received the validators broadcasts commit messages. When the required commit messages level is received the validators commit the changes [7].

If more than one-third of the nodes get corrupted this algorithm would fail.

2) Tendermint

It is a secure state machine replication algorithm operating similar to the PBFT algorithm. It has overcome the limitation of the failure to operate if one-third of the nodes failed. In

Tendermint each of the validators is identified through their private key and each of them keeps a full copy of the replicated state. Each of the blocks is assigned an incremental number known as the height, in each of the rounds the validators take rounds to nominated block suitable for each of the height. Each of the rounds contains only one valid block [8].

The voting mechanism is depicted in the following figure.

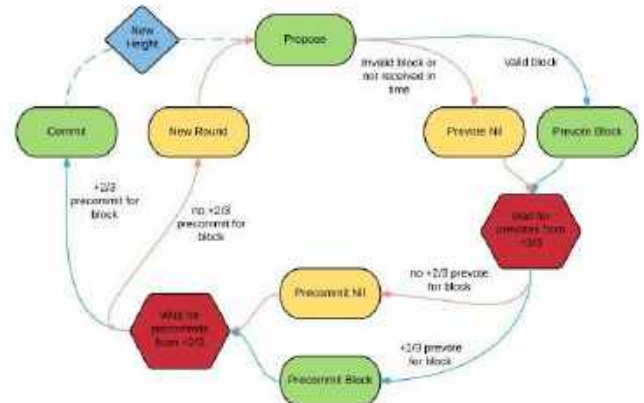


Fig. 2. Tendermint algorithm

The validators first vote for block by *Prevote Block*. If *Prevote Block* receives more than two-thirds of votes it will broadcast the *Precommit Vote*. If it also received more than two thirds the mentioned block will be added to the height. If not, a new round is started and continued until the block is successfully committed [8].

A. Consortium blockchains

The consortium blockchain is similar to the typical blockchain but the mining power of it vests on several entities [9]. The consortium blockchain bears the following characteristics.

- It is only available to a specific set of users.
- It is semi-decentralized: the members of the blockchain could supervise it and control it [9].
- The verification of the operations is done by the special pre-approved set of nodes [10].
- The data is immutable, but there is the chance for the blockchain rollback.

The consensus algorithms that are used in the public blockchain mining process are very resource-intensive and time-consuming [10]. But that kind of complex consensus algorithms is very much necessary in the public blockchains because the miners are anonymous.

But in a consortium blockchain, the miners are trusted entities and are known for some degree. So that the security is lowered to an acceptable level by reducing the complexity of the consensus algorithms [9]. Thus, the efficiency can be increased, and the transactions throughput can be increased up to several thousand transactions per second.

As in a public blockchain for a new block to be added to the blockchain the newly added block should be confirmed by the supermajority of the nodes [10].

III. RESEARCH METHODOLOGY

This study is conducted to minimize the fraudulent activities occurring in the land authentication system and thereby making it a more accurate, fast and resilient system. The principles presented in blockchain technology have been used to minimize the effect of double-spending as well as to make the land information system's most secure system. A methodological review of the literature has been carried out to identify the features of the other land authentication systems implemented in foreign countries. Data is collected about the existing processes in the land authentication system and information is gathered about the loopholes that exist in the current land authentication system. Personal interviews with legal officers, as well as the informal interviews with the people working in the land registry, have been used in the gathering of the data and information. Also, the land registry had been visited by the researcher and the process had been observed to a permissible extent.

Then a model is developed portraying the existing system. The loopholes present in the existing system are identified. Then those loopholes are evaluated in order to make the blockchain-based process free from those loopholes. After that, a model based on Blockchain technology is developed. The owner who wants the transfer of land ownership creates an e-request. The e-request is stored in the blockchain, which is stored in the land transfer smart contract. Then the legal officer takes those e requests and approves them. Once they are approved the land information is sent to the smart contract containing the land ownership information and the owner's details are amended in the smart contract. Before modification of the owner's information, the smart contract checks whether the user who generated the request is the current owner of the land. Also, it checks for the validity of the legal officer who authorized the transaction.

The above system requires a blockchain environment to operate. In the demonstration phase the Ganache blockchain, which is a blockchain running on the localhost, is used. In the commercial phase, a consortium blockchain is set up having several governmental organizations as the mining nodes. Each of the successful proof of stake would generate Ether. Private organizations also could join in the mining process. So that the mining responsibility is shared between both the state and the private sector. The private sector could

reimburse the Ether by selling them off to the state. The public and the legal officers would be needing Ether to perform their land ownership transactions and the ownership confirmations. The required Ether could be purchased from the state.

IV. IMPLEMENTATION

The Ethereum blockchain environment is used for the development of the blockchain. The smart contracts were developed through the Solidity language. Remix IDE (Integrated Development Environment) is used to test and develop smart contracts.

Then they were compiled to Application Binary Interface and then deployed to the Ethereum network with the use of the Truffle. Truffle is a development framework that allows the user to create a smart contract project at a click of a button. The developer is being provided with a project structure, files, and directories by the Truffle development framework which makes the development and testing much easier.

The Ethereum Blockchain framework is demonstrated with the use of the Ganache tool. It is a private Ethereum blockchain that allows the user to run tests and would act as a platform to test all the actions performed on the main blockchain without an additional cost. It has been used by many developers to test their smart contracts with frontend interactions.

The relevant functions of the smart contract are called from the frontend application using Web3js library. It is a collection of JavaScript libraries that helps the user to interact with the blockchain and to perform actions like reading from the smart contract data and writing into the smart contract. The Ethereum nodes also provide an RPC interface but this is very much error-prone and difficult to work with. The web3js provide a layer of abstraction on top of the Ethereum RPC which makes it more user-friendly and less error-proned.

The users will be accessing the system through a special wallet named the metamask, It provides a digital wallet and provides the user interface to manage the identities in different sites and sign the blockchain transactions. It allows the user to visit the distributed web.

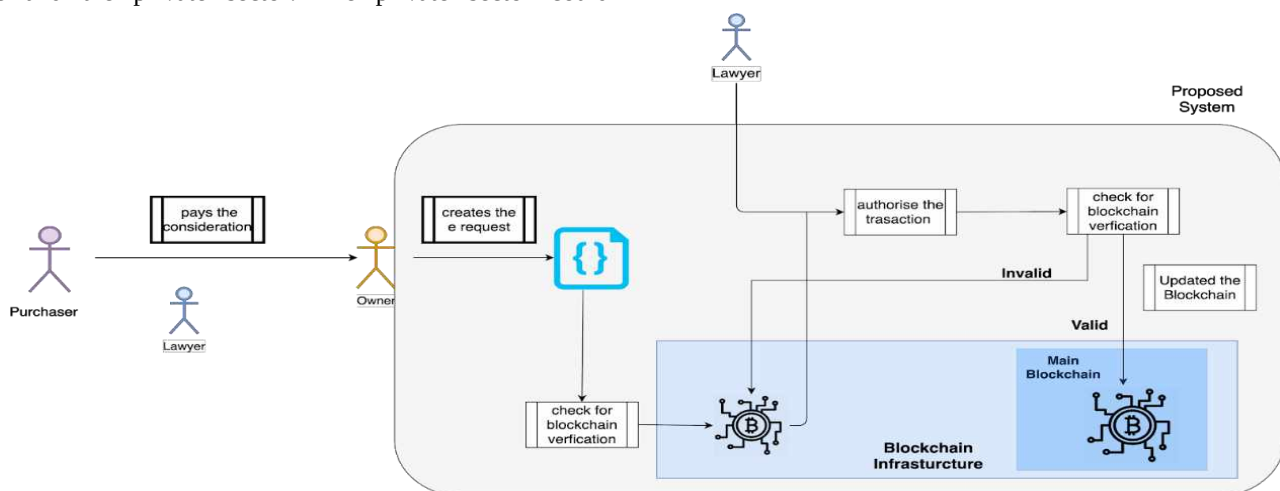


Fig. 3. Proposed Model

The interaction of all those tools is portrayed in the following architectural diagram in Fig. 4.

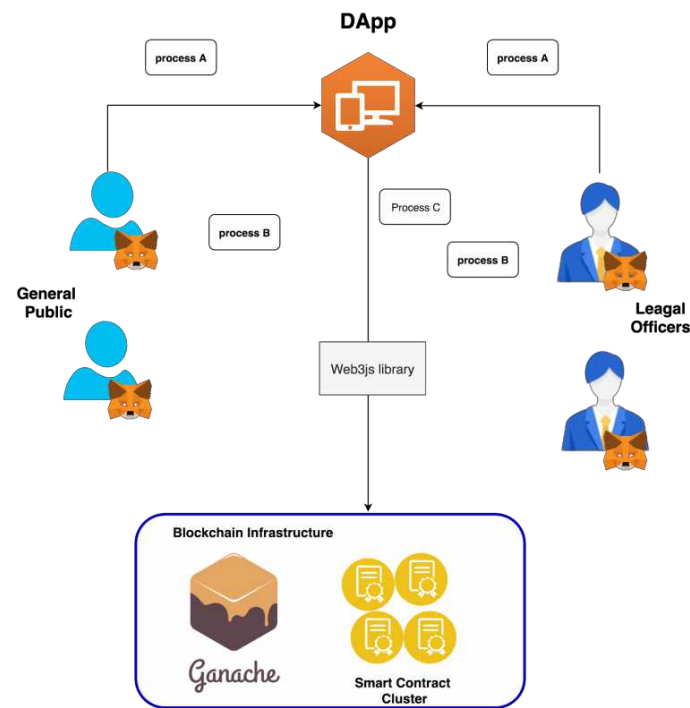


Fig. 4. Integration of tools in Blockchain

Process A - The users (legal officers / general public) Metamask browsers will be connecting with the Blockchain for the authentication process

Process B - The users (legal officers / general public) will be communicating with the DApp to perform the land authentication transactions.

Process C - The DApp will be connecting to the Ganache blockchain to write the transactions to the blockchain.

There are three smart contracts are created for the smooth functioning of the application.

- Land information - this contains all the information on the lands such as a list of previous owners, the next land.
- Land collection - this consists of the information of all the lands. All that information is stored as a mapping of Land
- Legal officer - This consists of the information of the Legal officers.
- Ownership change request - This consists of the information of the requests to transfer ownership of the Land before it is accepted by the legal officer.

V. TEST PLAN AND RESULTS

The user (both the landowner and the legal officer) should log in with the metamask account and confirm his identity.

In the testing phases, the Ganache local test network is being used. It would generate sample accounts with 100 ETH on each of the accounts as shown in figure 5. The Ether would be consumed when performing the change of ownership and

confirmation of land ownership transactions. The transactions can be seen in figure 6.



Fig. 5. Testing Accounts in Ganache

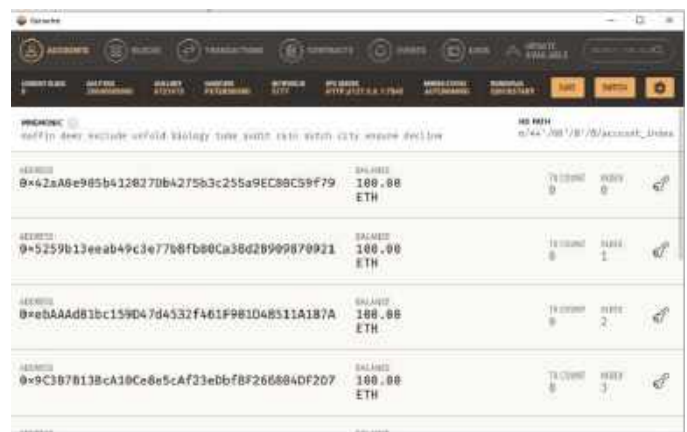


Fig. 6. Transactions in Ganache

The landowner creates the transaction for change of the ownership of the land through his mobile device. Then it is being saved on the ownership change request side until it has been approved by the legal officer. The legal officer uses the platform to approve the request and thus the verified request is sent to the land collections smart contract and the relevant land smart contract would be updated with the information of the new landowner there.

A. Case Study: Change of Land Ownership

The hypothetical owner of the land logs into the system. Then he successfully creates a land transfer to the hypothetical purchaser, which is approved by the hypothetical legal officer.

1. The Owner of the land log into the system
2. Then he selects the relevant land and creates land ownership transfer requests.

In the application, the land and the user in which the land is transferred would be selected from the user interface and ids of the land and the user that the land is transferred to will be extracted. But in the prototype, the above information should be entered as shown in the below interface.



Fig.7. Change Landowner Interface

3. Then the legal officer logs into the system and approves the above-created request if the contents are legitimate.

The legal officer would have the capability to authorize or reject the transaction. The legal officer would have the ability to view the land information and the owner in which the land is transferred. Once the details are found satisfactory the legal officer can confirm the land transfer. In application legal officer could do that by selecting the Land. But in the prototype, the land id should be entered to proceed with the confirmation.



Fig.8. Confirm Land Transfer Interface

B. Verification of the Land ownership

The confirmation of a particular land is owned by the mentioned user.

1. The prospective purchase logs into the system as a general user.
2. Then he searches the information of the land interested in.

In the application, he would have the luxury of selecting on the land itself. But in the prototype, the landId should be entered by the user as shown below.

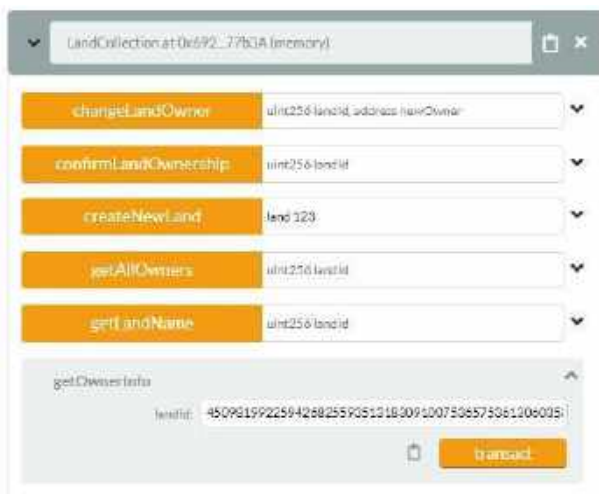


Fig.9. Getting Landowner Information

3. The information of the current owner is displayed.

The Blockchain-based application is returning the address of the current owner of the Land as shown below.

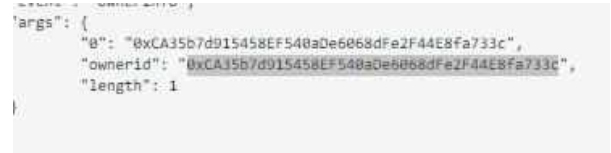


Fig.10. Land Current Owner Information

Then the application could combine the user profile information and presents it interactively.

4. If more information is needed ownership details of the previous five owners can be viewed.

VI. CONCLUDING REMARKS AND FUTURE WORK

The architecture of this model combines the requirements of the land authentication system in Sri Lanka with the modern technological advancement of the blockchain. The paper covers the gap of a lack of knowledge regarding Blockchain-based land authentication systems for the Sri Lankan context. This paper presents a model, which can be used to reduce the fraudulent activities occurring in the Land Authentication System in Sri Lanka and a prototype developed using Blockchain technology for the testing purpose of the model. This section is focused on the benefits and the limitations of such an approach.

A. Benefits

The major issue on the Land Authentication system lack of transparency in the authentication process and the complexity of identifying the real user. But being a digitalized system, it has been reduced to a significant level.

Also due to the vast amount of time that has been taken to validate a land transaction in the existing system, there is a lot of room for the fraudulent activity. But here a land transaction can be validated within a span of a few minutes. This reduces the window for fraudulent activities.

Also, the existing system can be altered by various parties being a paper system, and also a digitalized system would vest too much power on the administrator of the system. But being a blockchain system, which is decentralized, a group of administrators.

B. Drawbacks

The existing system would be subjected to natural wear and tear is a paper-based system. but the proposed system being digitalized would be tolerant for those.

This system can be linked with land-based other service providers such as the mortgage banks and finance companies. This would ensure that the middle loan approval and other processes in those systems would be very much streamlined. Further, if the middle person legal officer is removed then this system becomes highly immune to the fraudulent activities.

REFERENCES

- [1] T. G. U. P Perera, "Implementing Land Registration Systems in Sri Lanka: Being Pragmatic" In Sri Lanka Journal of Real Estate, 2010
- [2] "Implementation Completion and Result Report on Sri Lanka Land Titling Related Services Project", World Bank, March 2007
- [3] Mr. Cleverence Kombe, Dr. Majuto Manyilizu and Prof. Aloys, "Design of Land Administration and Title Registration Model Based on Blockchain Technology", Journal of Information Engineering and Applications, February 2017, pp 8-10.
- [4] Blake Sanders and Daniel Steeves, "Scheduled Private Token Sale & Token Distribution Event", March 2018.
- [5] H. Hyvärinen, M. Risius and G. Friis "A Blockchain-Based Approach Towards Overcoming Financial Fraud in Public Sector Services", Business & Information Systems Engineering, November 2017, pp. 441-456
- [6] G. Foroglou and A. Tsilidou, "Further Applications of the Blockchain", 12th Student Conference on Managerial Science and Technology, September 2018
- [7] A. Baliga, "Understanding Blockchain Consensus Models", April 2017, pp. 8-9.
- [8] E. Buchman, "Tendermint: Byzantine Fault Tolerance in the Age of Blockchains", June 2016, pp 17 – 20.
- [9] O. Dib, K. Brousmiche, A. Durand, E. Thea and E. B. Hamida, "Consortium Blockchains: Overview, Applications and Challenges", International Journal on Advances in Telecommunications, September 2018, pp. 51-63
- [10] Denys, "How the Consortium Blockchain Works", [Online] Available:<https://blockchain.intellectsoft.net/blog/how-theconsortium-blockchain-works>, [Accessed 25-Dec-2019].

Smart electricity monitoring and analysing an IoT system with a mobile application

A. I. R. Fernando*
Department of Statistics & Computer Science
Faculty of Science,
University of Kelaniya, Sri Lanka
ishanrandika.ir@gmail.com

M. D. R. Perera
Department of Computer Science
Faculty of Applied Sciences,
University of Sri Jayewardenepura, Sri Lanka
dilum@dscs.sjp.ac.lk

Abstract: Now, the demand for electricity has increased in the World. This demand was increased concern to the raises of many developed and developing nations in the world. In Sri Lanka, the energy was provided by only less electricity power stations. So, the scarcity of electricity has occurred eventually. Thus, reducing and controlling power consumption will be responsible for all consumers. Furthermore, consuming data should be tracked by the consumer is essential now. But recently, consumers are using traditional meters in each home. It has failed to provide these facilities to the user. Moreover, Digital meter is trying to reduce these limitations. This research study focuses on a Smart Electricity Monitoring System using a mobile application. This is an IoT based project. The electricity consumption can be observed by the user through a user-friendly mobile app. And also, the monthly electricity bill has automated. This smart electricity meter system can be separated into three divisions, and the first system is the hardware setup using Arduino to measure electricity consumption at home. The real-time Alternate current and Alternate voltage through the hardware setup were measured through this system. By considering these values, the Alternate power will be generated. Then the real-time values were converted to units(kW/h) and sent to the database through Wi-Fi. The real-time database in Firebase was the other system. Storing real-time data and permitting them to retrieve them through the mobile application was the main function here. The final system is the user-friendly android application. This system aims to get more involvement of the consumer to their electricity consumption and reduce global electricity consumption. As the results can be observed through the mobile application, the user can get some idea of saving and reducing electricity than earlier.

Keywords: Android, Arduino, Firebase, IoT, Smart electricity monitoring

I. INTRODUCTION

Sri Lanka is facing a massive problem with the electricity crisis. In Sri Lanka, we have few electricity power stations [1]. Mainly the electricity power generation is doing by heat power stations. It's common knowledge that there are many technical issues in those stations and also that their power generation cost is high. In general, it is forecasted that their will be shortfall in electricity generation with the current trend in use of it. Most people don't have any idea about electricity consumption. Educated people also don't care about it. If we have a good idea about electricity consumption, then we could reduce this problem. There are about 6.5 million power consumers in Sri Lanka [2]. From them, about 85% are domestic users. Let's assume that each domestic consumer reduces their monthly units by a single unit the total savings

of the whole country per month will be about 5.5 million units. That's means about 1MW (1000 kW) could be able to save. The reality is 1MW generating cost is very high. So, we could be engaged directly to solve our national problem.

The smart electricity monitoring system is an IoT based project which links the internet, software, and hardware together. This concept is not a new terminology to the IT world, but still, no any system using this concept. Rather than existing hardware systems in Sri Lanka like digital electricity meter, this system is using both software and hardware technologies to more involvements of the user. This is highly advanced and useful to electricity consumers. Here we can measure daily and monthly electricity units(kW/h) through this IoT system. So, it is easy to generate our monthly bill automatically on a particular day of the month for all consumers.

Here we can observe real-time measures of amperage(A), voltage(V), power(W), units(kW/h) through the android application. Using the measure of units, the monthly bill has automated. Moreover, here users have a creative option to get tips by considering the up to now power usage of the consumer. Either tips for the consumer or there has an alert service to the user when it reaches to particular unit levels such as 30, 60, 90, 120, etc. It will help them to reduce their consumption. Furthermore, here we have an analysis section on the mobile application. So, any user can get a detailed idea of the electricity consumption. This android application is highly user-friendly and useful to all users.

Here the system is using a real-time updating database in firebase. Furthermore, we can retrieve data at the same time when they were updated, and it is the most advanced feature in this system. The entire system is connected with a Wi-Fi connection, but the mobile application is separated from them. We can observe data from anywhere using the android application.

IoT technology is supporting the world to get connected with systems. In this case, also really users are connecting easily with this smart electricity monitoring system. In present instances, electricity consumer never tries to involve in reducing their consumption. Nobody wants to waste their time at the meter always. So, until they got the bill, they don't know the units that they have used. Most people are amusing with their bills some times. And they can't get any idea how the bill generated in detail. All the problems are trying to reduce by using this system.

The main target is to design and develop a user-friendly and helpful mobile application to observe electricity consumption and generate bills automatically. Besides, here we have some useful facilities to get some tips such as an assumed bill, assumed units per month, etc. So, the user will try to reduce their bill with additional charges, also, with the alert system. The user will recognize how future consumption should be done. And users can analyze their daily usages, monthly usages through this app. This mobile application can be used to observe the user's earlier bills also. So, this system is useful for both electricity consumers and electricity



Fig. 1. Traditional Electricity Meter

II. RELATED STUDIES

A. Traditional electricity meter

Electricity meters are the type of electrical devices used to detect and display the consumed energy in the form of readings. Traditional electricity meters were being used since the late 19th century [3]. The traditional meter is consisting of an aluminum disc to find the usage of power using a magnetic field. The total power consumption always can be displayed on the panel. So, the monthly usage should be calculated manually, considering the previous value. So, the user has submitted a problem with this situation. Normally, the user never observes his/her monthly consumption until receiving the bill. Eventually, monthly usage will be increased, hence the less observation of the user for consumption. A traditional electricity meter has shown in Fig. 1 [1].

There are some limitations that can be faced by the traditional electricity meter. Such as,

- Users have to anticipate for the monthly electricity bill because these traditional meters are unreliable.
- There cannot be introduced any new type of tariffs based on hours with these traditional meters.
- It was difficult to develop any software applications and a supportive database system.
- There should be employed a lot of inspectors to perform the meter readings.
- The bill calculating method will be changed with the date of meter reading, so this is so difficult for the consumer to assume the bill.
- Most of the cases, the bill will get a massive increment of a few unit changes.

According to the above limitations and so other things were created an enormous gap between the electricity consumer and the distributor.

B. Digital electricity meter

Today the traditional electricity meter has digitally operated and reduced the limitations for the consumer [4]. But still, there have some fewer restrictions to be reduced.



Fig. 2. Digital Electricity Meter

The digital electricity meter is an environmentally friendly device that also used to measure the consumed power at the home or any other place. This digital electricity meter is mainly to provide a direct benefit to consumers to save money and reduce the electricity bill. In this electricity meter, it was used as an advanced infrastructure for sending meter readings automatically to the electricity distributor. A digital meter which using in Sri Lanka has shown in Fig. 2 [5]. An accurate meter reading will be given by the Digital Electricity Meter as a benefit to the consumer. In this meter, the consumption will be recorded based on hourly or less than hourly intervals. There are so many extra features in this digital meter such as having a non-volatile data storage, remotely connect or disconnect capability, tamper detection and also having a two-way communication facility. Here can be remotely collected data reporting to the central meter. Then the functionalities of the digital meter were monitored by the central meter. A broader understanding will be given by these meter readings to the consumer than a traditional meter so that the consumer can be altered with the habitants of energy-consuming. All the information that was measured from the digital meter can be used to generate many advanced analyses for the new generation. But still no more software application support with this digital meter. The followings are some benefits of the Digital Meter.

- The operational cost is low than the traditional meter.
- Time will be saved because there will no time consuming for reporting the meter readings to the energy providers.
- The electricity bill can be paid online.
- The consumer could be reduced power consumption in high peaks and save money.
- These features will be automatically terminated when they are not in use.

Every electricity consumer in Sri Lanka comes with a traditional electricity meter. Both Digital meter and the proposed solution must connect additionally to the home electricity system. The differentiation between the Digital meter and the proposed solution in this paper has shown in TABLE I. Digital meter is mainly using to get the peak/ off-peak electricity consumption of the consumer. This proposed solution always measures the electricity units according to the time-based function. So, it is very simple to functionalize this peak/ off-peak concept in the presented system in future studies. As well as this system supposed to make more concepts in different requirements of the power producer. So rather than the digital meter concept, this proposed solution is the most helpful and important to both the consumer and the producer.

TABLE I. COMPARISON BETWEEN DIGITAL METER AND PROPOSED SYSTEM

Function	Digital Meter	Proposed Solution
Cost	About Rs.10000	About Rs.3000 (Cost estimation is given below in TABLE II)
Analyzing data by the user	No function to analyze data	Through the mobile application, the user can analyze all past consumed data
Monitoring data by the user	The user has come near to the meter and monitor monthly usage	The user could monitor real-time consumption through the smart meters. The user could monitor Daily usage of the current month graphically. The user could monitor electricity consumption and their relevant bill values of earlier months.
Time-based bill generation	Have the Peak/ Off-Peak facility to generate the bill	This system could be able to functionalize for peak/ off-peak bill generation for future requirements
Monthly electricity bill sending	The bill reader comes to give the bill	User is able to get the bill as a notification through the mobile app



Fig. 3. Main screen view of the Electricity Bill calculating application

C. Electricity bill calculators

1) Bill Calculating Mobile Application

The main activity of the existing electricity bill calculating mobile application has shown in Fig. 3. This mobile app is already available in the Google Play Store. But more manual data should be filled to get the monthly bill. It is not necessary for the consumer. Notably, the number of units has to calculate by considering the previous bill data and then

the computed value should be entered into the field to get the bill. Consumers aren't trying to involve their power consumption as these more manual things. Here only the main function to generate the bill has been developed in the mobile application.

2) Bill calculating web applications

Fig. 4 denoted the existing web page to calculate the Electricity Bill. Here also same manual data for the mobile application should be inserted into the fields to generate the monthly bill. So, this concept also was unable to make more involvement of the user to their consumption. The same manner in the mobile version has declared in this concept to generate the bill.



Fig. 4. CEB Web page for Electricity Bill calculating

D. Related research studies

The study [6], has been proposed a necessary process that makes reductions in energy use and Carbon Dioxide emissions immediately. Here the design and prototyping of a Home Electricity Monitoring System were described that provide real-time information on electricity usage. The data was displayed on an LCD screen and the computer which related to this system, also wireless communication was used.

A Low-cost IoT energy monitoring system was designed and implemented in the study of [7], that could be used for many applications, such as billing system, energy management, and home automation. A low-cost PZEM-004T, a non-invasive CT sensor, an SD3004 electric energy measurement chip, and a Node MCU was used to design this hardware system. Here also the voltage, current, active power and accumulative power were recorded as the results.

A ZNETH's (Zero Net Energy Test Home) Power Monitor (ZPM) was proposed and implemented in the study [8], which provides detailed energy consumption data for each node in the building. Here all the data will be shown on a display panel and will be stored as an Excel file to facilitate further processing. Also, this data will be displayed on the computer screen via a novel virtual environment building system. Here immediate access will be given by the virtual environment to the building power consumption information in a user-friendly format.

A Smart Meter for Smart Electricity Consumption has introduced in the study [9]. The consumer was helped by this Smart Meter to know the information about the electricity consumption for appliances in their residences. Here the power consumption has measured and analyzed by conducting a case study on various households. So, the behavior of consumers in using devices was illustrated using these Smart Meter data.

Then some consumption patterns were understood using measurements and analysis of consumption over time. This research was exhibited with the help of case analysis, ARIMA model using the XLSTAT tool and flatter techniques. So, the user could be able to change their behavior when they had better control over their habits through this system.

The study of Autonomous Home Energy Monitoring and Management System [10] needs to be considered both energy consumption and generation simultaneously to reduce the energy cost. Here a developed system was presented that uses a photovoltaic module to charge a super-capacitor, which provided energy to a microcontroller-based autonomous sensing platform.

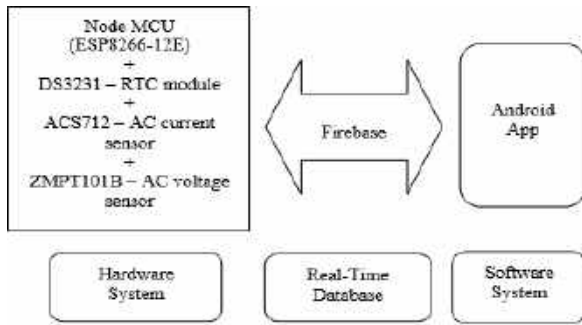


Fig. 5. Structure of the entire system

Embedded software on the node was structured around a framework that gives equal precedence to each aspect of the sensor node, the inclusion of distinct software stacks for the management of energy and the processing of sensors. A structured and modular design was promoted by this study which allows for efficient code reuse and encourages the level of interchangeable protocols.

TABLE II. COST ESTIMATION OF HARDWARE SYSTEM

Item	Quantity	Total Cost (\$)
Node MCU	1	5
ACS712 current sensor	1	2.5
ZMPT101B voltage sensor	1	3
DS3231 RTC module	1	2
ADC module	1	3
Connection wires	1 pack	0.5
Total		16

III. METHODOLOGY

This project could be divided into three major divisions. Hardware systems for current and voltage measuring, the real-time database for updating, storing, retrieving data and a mobile application for real-time retrieve data and analyze them were them. Fig. 5 has shown the structure of this entire system, which was being tried to do a great job to involve users in their daily electricity consumption. So, reducing the electricity consumption of users was the main purpose of this system.

A. Hardware system

In the hardware system, an ACS712 current sensor to detect Alternate Current using voltage pulses and a ZMPT101B voltage sensor to detect AC voltage were used. Also, a DS3231 RTC module to get the current date and time was used to manage data sending to the correct path. Here an ESP8266 Wi-Fi module (Node MCU) was used to send real-time data to firebase. Node MCU was the most helpful Arduino board to transfer real-time data through wireless connections. Using Arduino language and Arduino IDE, the entire hardware system has programmed.

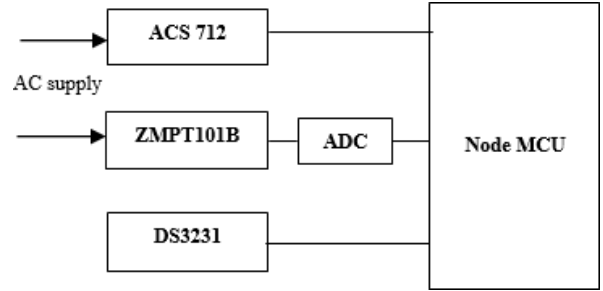


Fig. 6. Structure of the Hardware System

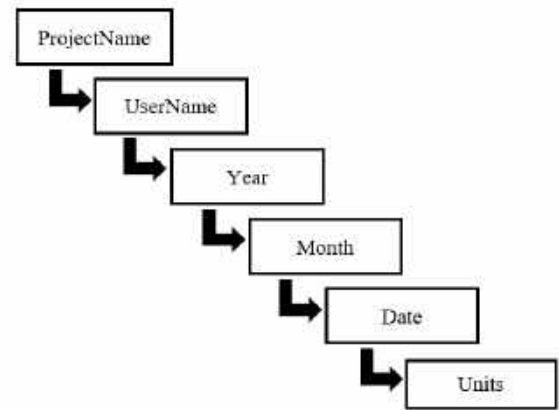


Fig. 7. Hierarchy of the Database

The cost estimation of the hardware system has shown in TABLE II. (Present date price about Rs.3000) More accurate AC power was given by this system, because here both current and voltage values will measure, instead of a constant voltage value such as 220V or 230V. Also, sensor calibration was the most important operation to get fine values. Also, an ADC was used to connect the voltage sensor to the Node MCU because there had only one analog pin on the board. The connecting diagram of the Hardware System has shown in Fig. 6.

The measurements for Average AC current per second will be detected by the current sensor, and the measurements for Average AC voltage per second will be detected by the voltage sensor. Let,

$$P = I * V \quad (1)$$

$$\text{unit (kWh)} = (P * .001) / 3600 \quad (2)$$

Using the above theoretical formulas such as Power Equation (Eq. 1) [11] and unit equation (Eq. 2) [12] can be

calculated the consumed units at home continuously. Using these values, the entire system has programmed.

B. Real-Time database

A real-time database was required by this entire system. Here all the data were sent and retrieved in each second. So it was connected with real-time data transferring. In this system, Firebase was used for this real-time transferring. Also, firebase was the most useful server platform, especially for android mobile applications. And here more other facilities were provided from Firebase for all mobile apps. Such as authentications, cloud messaging, cloud functions, analyzing usages and testing, etc. The hierarchical structure was used in the Firebase database, as shown in the following Fig. 7. Using these hierarchical data, the analysis can be generated to the user.

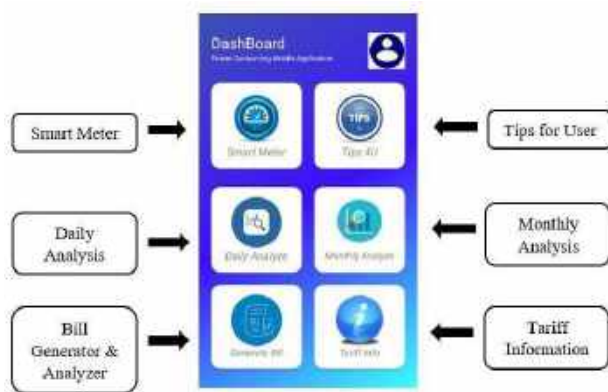


Fig. 8. The dashboard of the Mobile Application

C. Mobile application

This monitoring system was used to get more involvement of the user to his/her electricity consumption. Because of the tied schedules of users, they don't have time to spend extra time with existing systems for electricity consumption monitoring. Those were hardware systems that were only recording the consumption continuously. So more manual calculations should be done by people to get an idea of his/her present usage. Some software systems such as mobile applications and web applications were only providing a facility to generate user's monthly bills with more manual inputs. So, this system will be the most user-friendly and suitable approach with fewer user inputs for monitor user's electricity consumption.

Mostly this mobile application was used to observe monthly units (Kw/h), get the assumed bill, automatically generating the monthly bill, retrieve day by day usage of the current month and retrieve monthly usages and bills of several years. So, by this user-friendly mobile application, the user will be involved in their electricity consumption than earlier. All the latest mobile versions (> android 3.1/ API 12) can be used in this smart mobile application easily. The dashboard of the developed mobile app has shown in Fig. 8 below.

IV. EXPERIMENTAL RESULTS

In this section, all the operational and experimental outcomes were presented. To gather those results, we have implemented a prototype of a hardware module, Fig. 9. Then this system has tested with one electricity consuming home since October. According to the testing data through the prototype, the monitoring and analyzing functions in the mobile application were described in this section. This is only the testing stage of this hardware module. This could be more helpful to identify the sensor errors in the system. Sensors calibration is a very essential action to do in this system.

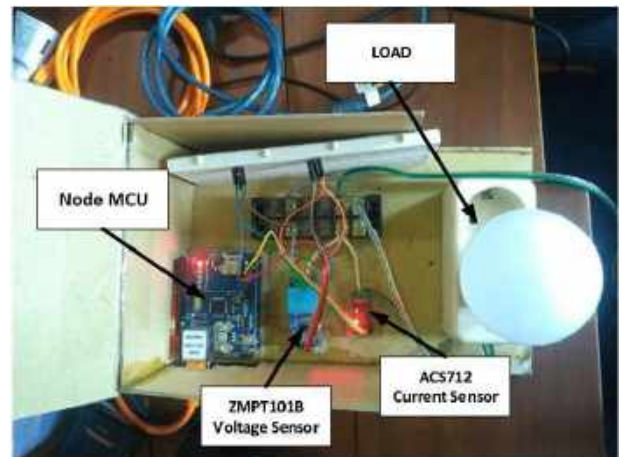


Fig. 9. The designed prototype of hardware system



Fig. 10. Smart Meter view in the mobile Application

Here the real-time data which were stored on the database could be analyzed with the user-friendly mobile application. Smart meter UI is the main panel that all the real-time data were displayed to the user. Here the consumer could be able to observe the units, current, voltage and the power on different four smart meters. The cumulative units of the related month will be represented on the unit meter. Also, real-time changes in amperage, voltage, and power could be observed in other meters. The view of the Smart meter UI has shown in Fig. 10

In any specific day of the month, the consumer could be able to get a simple idea on the assumed bill for the month by considering some analysis as a tip. Below Fig. 11 denoted the view of Tips UI, which provides this facility and some other suggestions.

The consumer will be involved in electricity consumption when there can analyze data quickly. Still in existing systems, there had nothing to investigate the electricity consumption of the user. Here some daily analyzing views of the current month will be provided to the user. In Fig. 12 has shown the UI of the Daily analysis in November month. By considering this, the user will be able to analyze whether any day had used power more than the average level.



Fig. 11. Tips UI in the mobile application

A unique UI was designed and developed to analyze all the consumed electricity and their bill each month in the mobile application. This concept also could be able to get more user involvement in electricity consumption. The user could compare the previous electricity consumption graphically. So, some essential data were provided through this application and stored them to analyze any time. The monthly Analysis UI has shown in Fig. 13. Here the data of October, November were presented. Also, the consumed electricity up to that moment in December was presented.

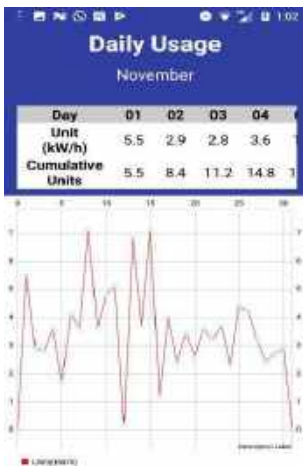


Fig. 12. Daily Usage UI in the mobile application

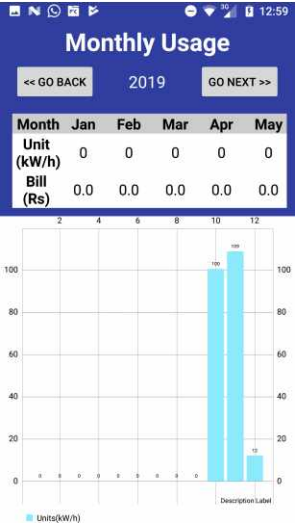


Fig. 13. Monthly Analyze UI in the mobile application

V. CONCLUSION AND FUTURE WORK

The design of a low-cost and useful IoT electricity monitoring system was presented in this study. Real-time electricity monitoring should be essential for electricity conservation and sustainability. According to experimental results, the real-time voltage, amperage, power, and accumulative units could be monitored. Also, the monthly and daily analysis could be observed by this system. So, the user could be able to get more involvement in electricity consumption through this system. The usability and the importance of the designed system were proved by the experimental results. Besides the original goal of developing an electricity monitoring system, this system can also be used in electricity auditing and electricity data saving processes. So, this monitoring system will be more useful to the consumer than existing systems in Sri Lanka. But a variety of electricity monitoring sensor systems were used commercially in other developed countries. So, this will be a similar electricity monitoring system to the consumer as available in other countries.

A. Future work

- An LCD monitoring display was supposed to fix into the hardware system for more compatibility.
- And also, an inbuilt memory card was supposed to include the Node MCU for saving data when the Wi-Fi connection is not available.
- For long-time experimental results, it was supposed to train a data model as a CNN (Convolutional Neural Networks) model and analyze them in a better way. Those works were planned to execute as future work.
- This entire system was supposed to use instead of the Digital Electricity Meter which has special functionality to get the peak and off-peak electricity usage of the user and generate the bill according to those values.

REFERENCES

- [1] Ministry of Power and Energy, [Online]. Available: <https://www.powermin.gov.lk>. [Accessed: 25-11-2019].
- [2] "CEB- Electricity Generated", [Online]. Available: <https://www.ceb.lk/electricity-generated/en>. [Accessed: 25-11-2019]
- [3] Ceylon Electricity Board, [Online]. Available: <https://www.ceb.lk/ceb-history/en>. [Accessed:28-11-2019].
- [4] Lanka Business Online, [Online]. Available: <https://www.lankabusinessonline.com/sri-lanka-introduces-smart-metering-to-reduce-electricity-generation-costs/>. [Accessed: 28-11-2019].
- [5] H. Ranmuthu, "How to read Sprint 350 energy meter",2016. [Online]. Available: <https://www.diyweb.lk/?p=149>. [Accessed: 30-11-2019].
- [6] J. G. Josué, J. M. Pina and M. V. Neves, "Home Electricity Monitoring System: Design and Prototyping", in *Second IFIP Doctoral Conference on Computing Electrical and Industrial Systems*, Costa de Caparica Portugal, 2011, pp. 437-444.
- [7] K. Chooruang and K. Meekul, "Design of an IoT Energy Monitoring System," in *2018 Sixteenth International Conference on ICT and Knowledge Engineering*, Thailand, 2018.
- [8] W. Nader, "REAL-TIME POWER MONITORING, HOME AUTOMATION AND REAL-TIME POWER MONITORING, HOME AUTOMATION AND SUSTAINABILITY", *Architectural Engineering - Dissertations and Student Research*, 2011.
- [9] P. V. S. M. Seelam, "Smart Metering for Smart Electricity Consumption", *Master Thesis Electrical Engineering May 2013*.
- [10] V. Chandran, A. V. Rajesh, R. A. Kumar, R. N. Jishnu and P. Prashanth, "Autonomus Home Energy and Management System" *IJESRT-INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, 2015.
- [11] Hyperphysics, [Online]. Available: <http://hyperphysics.phy-astr.gsu.edu/hbase/electric/elepov.html>. [Accessed: 13-02-2020].
- [12] Jain, "What are watt, kilowatt and a unit of electricity", 2020. [Online]. Available: <https://www.bijlibachao.com/electricity-bill/what-are-watt-kilowatt-and-a-unit-of-electricity.html>. [Accessed: 13-02-2020].

A microtransaction model based on blockchain technology to improve service levels in public transport sector in Sri Lanka

S. A. Jayalath*

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
akalanka345@gmail.com

Chathura Rajapakse

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
chathura@kln.ac.lk

J. M. D. Senanayake

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
janakas@kln.ac.lk

Abstract: Ticketing mechanism in a public transportation system is a major factor which defines the service quality of the system. Current online payment systems (credit/debit cards, PayPal, etc.) are not compatible with micropayments because transaction processing companies need a minimum transaction amount to make the transaction profitable for them. Therefore, an acceptable micro-transaction model is required in the micropayments domain. In blockchain systems, a third-party intermediary is not facilitating the transactions. Therefore, transaction fees decrease drastically. Using consortium blockchain concept, these fees can be further minimized when proof of work is also handled by a set of approved entities. In this study, an Ethereum based micro-transaction model is proposed to be implemented within the ticketing system in Sri Lankan public transport sector. Bus tickets are programmed as Ethereum smart contracts to handle the payment distribution. Consortium blockchain concept is used in the blockchain-based model where there are regulated bodies as nodes such as the national transport commission to handle the proof of work. Digital currency and Quick Response (QR) codes are incorporated to identify and complete the transaction process. The methodology of this development-oriented research can be described under three major phases. In the first phase, interviews were carried out with relevant stakeholders to identify the process of the current system and its limitations. Also, a broad and extensive study of literature was done parallel to this. During the second phase, identified issues, limitations and downfalls were addressed by designing a novel architecture. In the final phase a prototype is being developed to demonstrate the architecture, then prototype validation and testing were done with simulated data and several key use cases in the domain. The preliminary results of the prototype model show signs of considerable improvement in the service level of the public transport ticketing process and a significant reduction of transaction fees.

Keywords: Consortium blockchains, Ethereum, Micro-payments, Smart contracts, Transportation

I. INTRODUCTION

This paper addresses the potential of blockchain technology to be applied in the public transport ticketing system with the possibility of lowering the transaction fees in a microtransaction system. Section (I) of this paper provides an introduction into the state of public transport ticketing system, blockchain technology and microtransactions.

Section (II) provides an overview of the related work done through a thorough literature review. Section (III) discusses the requirement of the new platform and proposed new architecture with its functionality. Section (IV) discusses the testing and validation plan of the model with a set of case studies to be carried out with the system. Section (V) discusses the benefits and limitations of the model with overall results. Also, the section provides future areas of work in the field with proposed improvements of the current model.

A. Public transport

The public transportation system in Sri Lanka mainly consists of buses and railway systems, which the railway system, only serves a small percentage of the country's transportation needs. Railway system is primarily used for goods transportation with a fraction of capacity available for passengers to use compared to goods. Railway network in Sri Lanka is very slow compared to other developed countries.

As of 2017 National Transport Commission (NTC) report, there are,

- 6955 - Sri Lanka Transport Board (SLTB) buses
- 19614 privately operated buses

are operating in Sri Lanka. Annually 4 billion citizens in Sri Lanka use the above-mentioned amount of buses to fulfill their transportation needs [1].

There are service quality attributes of public transportations systems. Convenience is the main factor when it comes to measuring the service quality of public transportations systems. Convenience is defined here as how simple the Public Transport service can be used and how well it can be used to fulfill the transportation needs. Also, it can be distinguished by considering the ease and simplicity of the payment and planning process of a public transportation trip [2].

1) Ticketing system in public transport

Passengers using public transport in Sri Lanka only have one option to purchase tickets which is using cash and purchasing tickets on premise. Currently, there is no way to purchase digital tickets on the premise in Railway or Bus transportation systems in Sri Lanka. Considering about Sri Lankan public transportation system, convenience is being questioned always. Ease and simplicity of paying for tickets

are very less. Passengers always argue with the conductor to obtain their change and they never get their change correctly.

Most of the western and European countries including developed Asian countries are adopted to the Travel Pass system. And now they are moving towards mobile ticketing platforms.

B. Blockchains

One way of defining Blockchain is “A decentralized, secure and distributed ledger”. And there is no standard definition exists for blockchain. It is a software protocol which can be used to securely transfer unique instances of value (Ex. Property, Money, Credentials, etc.) via the internet without any of third parties involving in the transaction process which can be a bank or government organization. Blockchain consists of public/private key cryptography, cryptographic hash functions, distributed databases, consensus algorithms, and decentralized processing.

To perform a transaction User X selects User Y address. Using the platform protocol User X forms a transaction and signs the transaction using their private key and announces the transaction to the network. Other peer nodes listen for new transactions and form a candidate block; a hash digest of valid transactions that have yet to make the ledger. The node announces the candidate block to the network in order to add it to the ledger. Other nodes check the candidate block is valid to ensure that the block contains no fraudulent or incorrect transactions included by the block's creator. A consensus mechanism prevents block duplication and a bottleneck of network traffic. As the process continues, the ledger continues to expand as a chain of these blocks of transactions; hence the term blockchain. Most importantly Blockchain infrastructure ensures trust, security and transparency.

1) Blockchain ticketing

Tickets for anything can be defined as a type of contract. Also, tickets have some face value, these can be treated as smart contracts to imitate the current tickets and proceed onward to discover programmable approaches to make a totally new kind of tickets and a broad framework. Smart contracts also bring a lot of scalability for the ticketing Ecosystem. Around the global event (movie, music, etc.) ticketing industry (including ticket booking systems), companies are starting to use blockchain based ticketing systems for movie ticketing, booking, music industries, etc. with programming tickets as smart contracts. Some of the examples for global companies are Aventus in Europe, Guts in Holland, Lava in the US, Blocktix. However, all the platforms that exist focus on the prevention of fraud and the distribution of event tickets.

2) Consortium blockchains: A brief introduction

Consortium blockchain is another recent blockchain topic that came into the discussions. Consortium blockchain is a specific blockchain implementation with authorized nodes. Consortium blockchains evolved as an important concept and architecture that benefits from the flexibility and anonymity of private blockchains transactions while exploiting the democratic administration of public blockchains [3]. Transaction efficiency is much needed in public transport. Consortium blockchain will ensure,

- Transaction limitation will be reduced which resulted from transaction confirmation delays

- No transaction fee
- Improved privacy in blockchain data

C. Microtransactions

Microtransactions are small commercial transactions in digital form. Defining the monetary value of microtransactions depends on the nature of the transactions and the fiat currency type. In terms of Sri Lankan Rupees, any transactions under 1000 rupees can be considered as micro-transaction value. Most of the ticketing transactions in the public transport system in Sri Lanka can be classified as micro-transactions as 95% of these transactions are below 1000 LKR. The minimum fare of a bus ticket as of now in Sri Lanka is 12 rupees. Out of these ticket transactions, there are very few happen as online transactions. Only some of the bookings happen as online transactions. NTC booking, BusBooking.lk are some of the booking portals and they charge a heavy transaction fee for every booking.

Traditional credit card-based systems not able to handle micropayments due to the reason for high processing fees. When a card transaction occurs, the buyer gives his permission to get a payment from his account, however, in this process lot of intermediaries are involved.

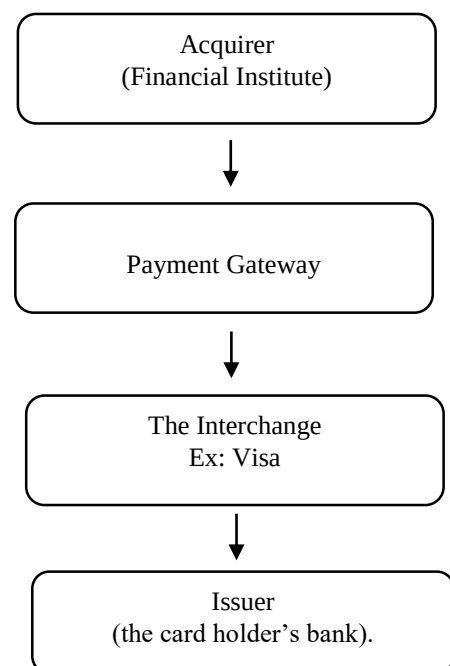


Fig. 1. Financial intermediaries

Therefore, to process a typical transaction processing companies should keep these intermediaries profitable. Hence the minimum transaction amount these processing companies can process is too large. But, blockchain protocol is promising in addressing this issue with its key principle of decentralization. Blockchain gives a great opportunity to get rid of the “middlemen fees” scenario by eliminating multiple intermediaries mentioned above.

II. LITERATURE REVIEW

A. Public transport

The public transportation sector in Sri Lanka is a major factor in the lifestyle of the citizens as a developing country in the world. Public transportation is a method of travelling maintained by a state which enables more people to travel as a group in predefined routes. Traditional instances of types of public transportation are buses, trains, and trams. Bullet trains, airplanes, and omnibuses are heavily used for public transportation between cities. [1]

There are service quality attributes of public transportations systems. Convenience is the main factor in measuring the service quality of public transportations systems. Convenience is described here as to how easy it is to use the PT service and how quickly it adds to your mobility. [2].

[4] discusses the factors that are affecting the service quality of public bus transportation in Sri Lanka. Developing public transport facilities will result in reducing the privately owned vehicle usage on roads eliminating common problems in Sri Lanka such as,

- Huge congestion on roads
- Environmental pollution
- Road accidents etc.

Further, results imply that customers are highly satisfied when they are provided with a higher quality service and they could not see much difference between private vehicles and public vehicles. And, through discussion and analysis paper state that although a lot of transport facilities such as touch travel cards, "Sisusariya" etc. are provided, most of those services are not effective and not properly executed to get the highest return.

1) Public transport ticketing

European Metropolitan Transport Authorities (EMTA) have been working on addressing the issues of electronic ticketing. Study on electronic ticketing in public transport [5] done by an established group of EMTA to find knowledge and learn from the experience in the field of electronic ticketing. This study was assisted by public transport consultant Mohamed Mezghani. key factors to consider when defining and designing a ticketing system as follows,

- Ease-of-use for passengers
- Easiness of revenue collection
- Conformity for operators
- Attractive service to passengers
- Intramodality (transfer tickets)
- Transparent revenue sharing
- Maximizing farebox revenues
- Fraud reduction

Also, the paper discusses the existing ticketing medias such as cash, tokens, paper tickets, contact-based smartcards, contactless cards, mobile ticketing, etc. Talking about the e-ticketing systems these are only about payments, but these systems hold a huge amount of data which can be used to make public transport convenient and easy to manage. The long-term objective of this e-ticketing system as mentioned in the study is to "customers should be able to pay for the public transportation with a fraction of time in a fully

automatic manner without waiting for the validation of the card."

B. Microtransactions

Microtransactions are small commercial transactions in digital form. Microtransactions in Bitcoin [6] study describes the nature of micro-transactions and they state the importance of low-cost microtransactions. "As the global economy is moving towards an era of using more intangible goods globally, delivering those goods with traditional payment methods tend to be more expensive than the product cost."

Experiences in Developing a Micro-payment System for Peer-to-Peer Networks [7] is a study that discusses various researches done to implement a micropayment system. And they present their novel P2P-Netpay micropayment platform and its architecture. P2P-Netpay delivers an offline micro-transaction model that uses a light-weight hashing-based encryption method. They also state the importance of peer to peer transactions in this paper.

C. Blockchain

[8] is a study that gives an overview of blockchain technology with the current applications and stating the impact of blockchain. As stated in the study in a blockchain network data is not controlled by any third-party intermediaries and all the transactions should go through a cryptographic validation process. Every transaction ever done is documented in an unalterable ledger in a verifiable, safe, precise and everlasting manner, with a timestamp and additional information.

In 2008, a person anonymously (or group) identified under the name of Satoshi Nakamoto published a white paper titled [9] Bitcoin: A Peer to Peer Electronic Cash System and they proposed a novel structure for electronic transactions without depending on trust and mentioned that using a peer-to-peer system of electronic online transactions would give the ability to send money directly from one user to another without involving financial intermediaries.

The Ethereum yellow paper came out in 2014 which was written by Dr. Gavin Wood and functions as a formal definition of the Ethereum protocol. Ethereum is an approach which focuses on building a technology that can build all transaction-based state machine concepts [10]. Also, this paper introduces the concept of Smart Contracts.

Later, blockchain ideas emerged into a lot of other domains without limiting only into cryptocurrencies. Within emerge of second-generation blockchain applications are derived to many fields.

Gartner Top 10 Strategic Technology Trends for 2020 list out Blockchain, quantum computing, augmented analytics and artificial intelligence (AI) drive most business models in 2019 [11]. The report also says other than finance applications blockchain has begun its impact on government, healthcare, manufacturing, supply chain and other areas, with blockchains eventual potential reducing the costs and reducing clearing time of the transaction.

1) Blockchain ticketing

Tickets for anything is nothing else but a type of contract [12]. Also, tickets have some face value, these can be treated as smart contracts to imitate the current tickets and proceed

onward to discover programmable approaches to make a totally new kind of tickets and a broad framework. Smart contracts also bring a lot of scalability for the ticketing Ecosystem. Around the global event (movie, music, etc.) ticketing industry (including ticket booking systems), companies are starting to use blockchain based ticketing systems for movie ticketing, booking, music industries, etc. with programming tickets as smart contracts. Some of the examples for global companies are Aventus in Europe, Guts in Holland, Lava in the US, Blocktix.

A blockchain based privacy preserving ticketing service [13] present an Ethereum based privacy preserving event ticketing model. The study also discusses the integrity that blockchain brings into the ticketing Ecosystem. And the focus of this study is about how to ensure the privacy of the ticket purchaser, for this Non-Interactive Zero-Knowledge (NIZK) schema is used with the goal of protecting the privacy and confidentiality of the end-user.

The world's first Blockchain Bus is set to ride in the Netherlands [14]. A company called VMC initiated a bitcoin-based payment system for public transport buses in the Netherlands recently. Passengers can pay for tickets via bitcoin where cryptocurrencies are regulated in the Netherlands. But in Asian countries still, cryptocurrencies are considered as illegal.

2) Consortium blockchains

Consortium blockchain is a modified version of blockchain concept with approved nodes to maintain the distributed ledger. Consortium blockchains evolved as a promising concept which uses the transaction efficiency and privacy of private blockchains while getting the benefits of decentralization in public blockchains [3]. Transaction efficiency is much needed in public transport.

Consortium blockchain is another recent blockchain topic that came into the discussions. [15] further exploits consortium blockchain knowledge to develop a unified and secure P2P energy exchange scheme with consortium blockchain, named energy blockchain. And, they addressed the issue of transaction delays in the bitcoin network with a credit-based payment schema model. Consortium blockchain concept is very little addressed in researches. This concept can be applied to public transport ticketing to maintain a better infrastructure while enhancing the service quality of the sector. Consortium blockchain will ensure,

- Transaction limitation will be reduced which resulted from transaction confirmation delays
- No transaction fee
- Improved privacy in blockchain data

D. Research objectives

An identified main issue of the microtransaction domain is the high transaction fee with the existing payment channels. This is due to a lot of third-party intermediaries involving in the transaction process, and blockchain technology is promising with addressing this issue. Without addressing this specific problem related to microtransactions, implementing an online portal with existing payment channels will result in the same issue.

This study introduces a novel digital low fee payment platform to be used within the public transportation system

which can be also extended to other microtransaction systems using blockchain technology to achieve the following goals.

- A platform that retains all requirements of current ticketing systems
- A platform to Minimize transaction fees with microtransactions
- Improve the transparency of the ticketing system
- A platform to Ensure fair payment distribution
- Eliminate the need for paper tickets and other materials

Further, this study uses the public transport system in Sri Lanka as a reference for a microtransaction system.

III. METHODOLOGY

The research approach of this development-oriented research can be explained under three major phases. In the first phase, interviews were carried out with relevant stakeholders to identify the process of the current system and its limitations. Also, a broad and extensive study of literature was done parallel to this. During the second phase, identified issues, limitations and downfalls were addressed by designing a novel architecture. In the final phase, a prototype has been developed to demonstrate the architecture, which was tested in a simulated environment with generated data to arrive at conclusions.

This platform took several user roles and their nature into consideration when designing the architecture.

A. User roles

- Passenger
- Bus owner or Conductor
- Governing Authorities (National Transport Commission etc.)

1) Passengers:

Common observation in Sri Lanka public transport system considering the passenger is that obtaining their change of the ticket. Passengers always caught up with arguments with the conductor to get their change and they never get their balance correctly. Passenger is the major object of the application. The Passenger can use the application to purchase and authenticate. Also, the history of the tickets can be viewed by the passenger. The actions of the passenger are:

- Create an account (wallet) to purchase a ticket.
- Check the balance of his account.
- Scan a QR to transfer tickets.

2) Bus owner or conductor:

Considering the Sri Lanka public transportation system, it is crucial to have a fast and fair payment distribution system in place. The reason for this is on most buses conductor and drivers relying on the daily income and also, their daily allowances are paid for them. So, the bus owner needs this money as soon as possible to settle these payments.

The actions of the conductor are:

- Able to verify tickets presented by passengers.

- Check the wallet balance and transfer money.
- Generate QR based tickets.

B. Platform architecture

Figure 2 illustrates the overall blockchain model. A key difference to highlight in this proposed model is how the Proof of Stake (POS) is conducted compared to the public Ethereum Blockchain. In public, Ethereum blockchain POS is done by miners in the network who will get Ether (Crypto money in Ethereum blockchain network) as an incentive for the computational power they spent. In this model, POS is handed over to a set of authorities like NTC, ICTA, etc. This type of blockchain is known as Consortium Blockchains. Therefore, when a new transaction happens and a block creates, validation (traditionally known as mining) happens between the authorities. After the validation from the authority's block will be added to the blockchain.

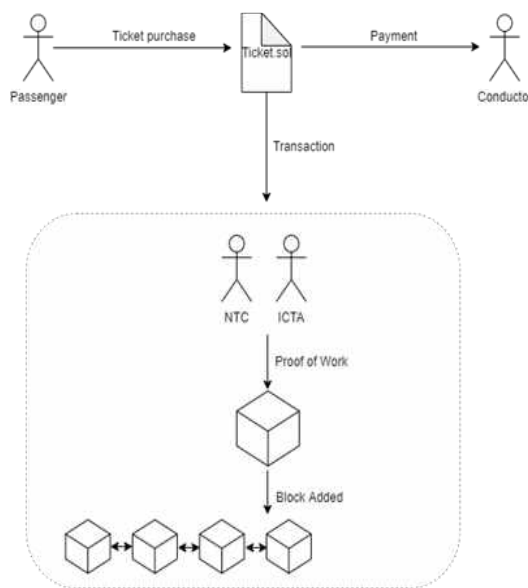


Fig. 2. overall blockchain model

1) Smart contract (ticket. Sol)

A smart contract will be the one that is acting as the intermediary in the system which is a collection of code reside at the blockchain itself. Smart contracts are written solidly. In this model Ticket. So smart contract executes the work related to the following process.

- Payment distribution between passenger and bus owner upon the verification of QR code

There are four types of transactions happens on an Ethereum blockchain network, (or any kind of blockchain network which involves smart contracts)

1. User to User transactions
2. User to Contract transactions
3. Contract to Contract transactions
4. Contract to User transaction

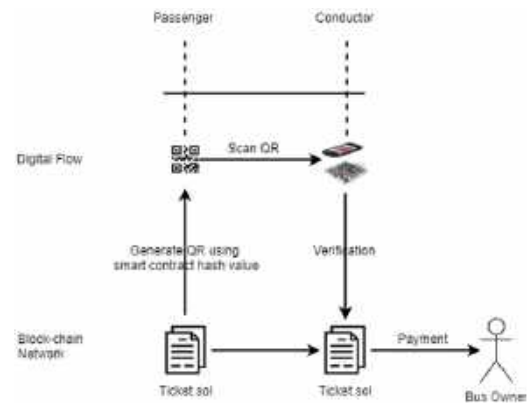


Fig. 3. Digital flow

Out of these in this blockchain model (2) and (4) will be used in the Ethereum network. Purchasing a ticket will be a type (2) transaction where paying the bus owner upon successful QR verification will be a type (4) transaction.

Ticket.sol will insert Customer ID, Balance, Ticket value and Ticket ID into the block as the data. With the verification, the block will be added to the blockchain.

C. Process flow of the platform

Within the development process of DApp, QR based Digital flow was developed which is illustrated in figure 3.

Upon the purchase of the ticket customer will be issued with a QR code by Ticket.sol smart contract that will include contract hash. When the conductor scans the QR code smart contract will call blockchain and validate it. A valid contract will automatically move money from the passenger's wallet to the bus owner's wallet, at the same time showing the verified message on the screen.

D. Implementation

In Ethereum, Decentralized Applications (DApps) are the applications which have a graphical user interface enabling users to interact with the blockchain in a more friendly manner. Smart contracts handle the business logic of the application, and the front-end contains technologies like HTML, JavaScript, etc. Because smart contract code is very costly to execute, DApps are designed to perform only the required calculations on the blockchain, and everything else that can be done by the client side is executed on the front end of the application to save costs and utilize resources.

Within the application development process of the DApp which is described in Figure 4 following tools and technologies are used,

- Truffle Framework

Truffle is an integrated development environment which can be used to easily develop Ethereum DApps. Truffle provides a separate testing framework for solidity smart contracts and also truffle can be configured as a delivery pipeline. In the implementation, truffle was used mainly to compile the contracts, migrate the contracts, and deployment, etc.

- Web3.js

Ethereum node provides a Remote Procedure Call (RPC) interface which is used in DApps to interact with blockchain. Web3.js is a JavaScript library that offers a layer of

abstraction and operates on top of Ethereum RPC. Web3 includes the eth object - web3.eth and the shh object - web3.shh

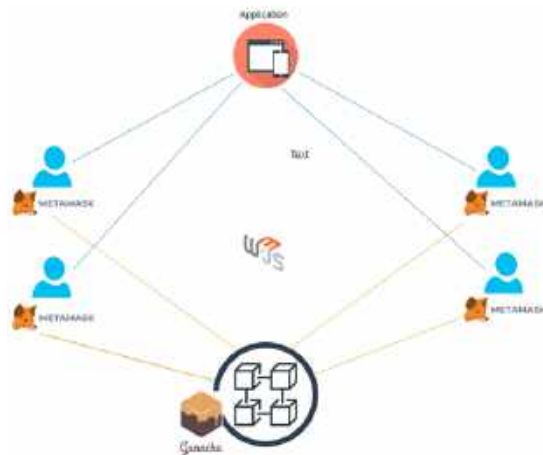


Fig. 4. Platform Architecture

- MetaMask

To use the DApp GUI, the user must have to use a web browser like Mist or MetaMask plugin for chrome and an Ethereum account. First of all, MetaMask will inject web3.js instance into the HTML webpages and it will allow the user to import his Ethereum accounts to the current system he is using. Then the system will connect to the blockchain network through MetaMask making the user system a virtual blockchain node.

- Ganache

Ganache is a simulated blockchain environment which can be used as a private Ethereum blockchain. Ganache allows us to inspect the blockchain state while controlling the operations of the chain. As this is a simulated environment it allows performing all actions that would on the main Blockchain without any costs.

The bus conductor can enter the value of the ticket into the system and can generate the QR code. QR is generated using the contract hash value. The user interface to insert ticket value and generate Quick Response (QR) code is shown in figure 5. System generated QR code will contain the ticket value and the wallet address of the conductor/bus owner.

The interface is titled 'Bus conductor'. It contains a 'Conductor Address' field with the value '0x54F73c880C5bc75364751451c00F6a355H21E'. Below it is a 'Ticket Value' field with the value '15.00'. To the right of the 'Ticket Value' field is a 'Generate' button. Below the 'Ticket Value' field is a 'Generated Hash' field with the value '0xCAC6F6891953a73611a2047'. To the right of the 'Generated Hash' field is a QR code.

Fig. 5. Interface to insert ticket value and generate QR code

Then this QR code will be presented to the passenger to be verified and upon successful verification, funds will be transferred into the conductor/bus owner wallet. The user interface created for testing the verification is shown in Figure 6. Before completing any transaction on the platform MetaMask will ask for the authentication to verify the user.

The interface is titled 'Passenger Transactions'. It contains a 'Ticket to Purchase:' field with the value 'g. 0x93e66d8a9a28c17d0fc303b53e0fbd76690das'. Below the field is a 'Purchase Ticket' button.

Fig. 6. Interface to verify ticket

MetaMask notification upon a transaction is shown in figure 7. This authentication will happen before any transaction happens on the platform. The transaction summary is shown in MetaMask with details and gas prices. The transaction should be confirmed by the user to complete it. And, the user can set his desired gas price.

Ganache local test client will generate 10 Ethereum wallets to use within the development process. These 10 wallets are preloaded with 100 ETH each and gas prices can be set as desired. This environment is a simulated environment specific for testing purposes. Figure 8 shows the generated wallets with their wallet addresses.

The interface is titled 'MetaMask Notification'. It shows a 'Contract Interaction' section with a 'Details' tab. The 'Details' tab shows the 'ETHER' value as '0.050681' and the 'TOTAL' value as '0.050681'. There are 'Cancel' and 'Confirm' buttons at the bottom.

Fig. 7. Interface to insert ticket value and generate QR code

ACCOUNT	ADDRESS	BALANCE	ETH
ACCOUNT 0	0x54F73c880C5bc75364751451c00F6a355H21E	100.00 ETH	100.00
ACCOUNT 1	0x74C055C961A6A1D7F6628C8F5FecAa1F032FCE	100.00 ETH	100.00
ACCOUNT 2	0x3B20518496A10727B18E6B2C4BFD6M3ETA7B	100.00 ETH	100.00
ACCOUNT 3	0x1A3E473B95F1953e73611422a47B7A7A3B55B3B	100.00 ETH	100.00
ACCOUNT 4	0x9C46a3F179B70e058ED159E37ea84FF1cd2B60	100.00 ETH	100.00
ACCOUNT 5	0x3a8891292f8045960A726a1dF83c23A72B506A	100.00 ETH	100.00

Fig. 8. Ganache Test Client

IV. TESTING AND VALIDATION PLAN

In order to test the capabilities of the platform following case studies are performed. These case studies are generated from the perspective of different user roles after identifying the nature of them. Mock data is used to simulate the input and output data required in the testing process.

- Passengers are seeking to purchase a ticket.
- Bus conductor seeking to verify and vend tickets

A. Case Study: passengers

A hypothetical passenger exists to test the platform against the passenger considerations and requirements mentioned in Section (III-A)

1. Passenger connects to the application and creates a personal wallet for him through MetaMask or a similar application.
2. Passenger logs in to his wallet and checks his account balance.
3. A passenger scans the QR presented to him by the conductor and verifies the transaction to purchase a ticket.

This test checks the passenger's ability to purchase a ticket through the system and successful transfer of funds through the DApp.

B. Case study: Bus conductor

A hypothetical bus conductor exists to test the platform against the vendor considerations and requirements mentioned in Section (III-A)

1. The bus owner or conductor uses to create a wallet using the application.
2. Conductor logs in to his account and enter ticket value and generate a QR based ticket.
3. The conductor presents the QR code to the passenger.
4. A passenger scans the ticket and upon successful verification, the conductor receives the funds to his wallet.
5. The conductor checks his wallet balance and history of transactions whenever necessary.

This test checks the Conductor's ability to generate a QR based ticket and fund transfer mechanism of the DApp. The architecture and the mechanics of the platform combine the requirements of public transport ticketing with state-of-the-art technologies. The symbiotic development of the platform results in several beneficial aspects.

C. Technical benefits

Platform mining work is handed over to set of authorities like national transport commission, hence ticket prices and other regulations can be set by the government and the transaction fees are reduced. They have the same control over the platform as the existing one to limit fraudulent activities.

Furthermore, the customer's privacy is secured with no tracking of their identity. Using blockchain technology will provide a secure, distributed platform that will result in decentralized data. This will prevent network downtime and tampering of data. Also, all the peer nodes store a local copy of data allowing real-time ticket inquiries and lowering the network bandwidth.

D. Other benefits

Paper tickets and additional plastic travel cards will be eliminated with the use of this platform. This is a huge cost saving aspect for the organizations in the transportation domain. Environmental pollution with plastic and paper can be lowered as well by a considerable percentage.

From the perspective of the passenger, a huge paradigm shift would not be necessary to operate on this platform. Almost everyone in the modern world owes a smartphone; the only requirement to use the platform. Also, bus owners would not need to invest in ticketing machines and paper rolls to operate in the platform as a service provider.

Most importantly both parties do not need to worry about physical money, especially coins as everything happens with digital currency. So, obtaining a change of the ticket will not be a problem anymore with this platform.

V. CONCLUDING REMARKS & FUTURE WORK

This paper presented an innovative approach to use blockchain technology in the microtransaction domain, with a digital ticketing platform to use within the public transportation domain in Sri Lanka. The core principle of decentralization of blockchain technology is used to eliminate traditional middleman scenarios in existing transaction processing systems to benefit the passenger, bus owner and governing organization.

Identified business logic and the user scenarios in the thorough literature review were analyzed and implemented with the usage of Ethereum smart contracts. It was effective to use tickets as smart contracts within the system. Quick Response (QR) codes were recognized as a practical and viable mechanism to verify tickets within the system which was also an environmentally friendly solution. Using a digital currency in the system also eliminated the problem of dealing with the change of the ticket. Especially a paradigm shift is not required to use the platform presented here as the platform preserves all the properties of the existing system.

Lowering transaction fees in a digital payment system is the focus of this research. Therefore, those micro payments can be facilitated via online payments. Properties of consortium blockchain concept are used in the architecture to get the utmost benefits of both public and private blockchain characteristics. Using consortium blockchain concept public mining process has been totally eliminated from the model presented here. Mining is done by a set of governing authorities from the public transportation domain. Basically, transaction fees on a public blockchain system occur due to the public mining process because miners should be paid for the transaction validation. This fee can also be eliminated in a consortium model where the validation will happen from the governing bodies of the domain. The only cost to bear is the computational power that they need to use for the mining work. This was tested within the implemented demonstration system using ganache and setting zero gas prices for the transactions. Data were stored in a blockchain. Thus, microtransactions are facilitated by the proposed novel architecture with minimum transaction fees involved.

Crypto payments like bitcoin, Ethereum also faced a problem of the slowness of the network. Transaction in a blockchain network takes a couple of hours to complete. This was a huge limitation regarding micropayments. A solution came with the lightning network concept. Lightning network moved the transactions "off the blockchain" increasing the processing speed up to thousand transactions per second. Basically, the lightning network creates a second layer on top of blockchain which will open a payment channel between two wallets. This was specifically designed to

facilitate microtransactions. So, the lightening network concept can be integrated into the system to increase the transaction processing speed for a more efficient transaction distribution.

REFERENCES

- [1] National transport statistics 2017 [online]. Available at: <https://www.ntc.gov.lk/corporate/pdf/NTCEnglishReport2017.pdf> [Accessed 07 Nov. 2019]
- [2] L. Redman, M. Friman, T. Gärling, and T. Hartig, "Quality attributes of public transport that attract car users: A research review", *Transport Policy*, 25, 119–127, 2013.
- [3] O. Dib, K. Brousmiche, A. Durand, E. Thea, and E. B. Hamida, (2018). "Consortium Blockchains: Overview, Applications and Challenges", In: *International Journal on Advances in Telecommunications*. Paris-Saclay, France: Omar Dib.K. Elissa, "Title of paper if known," unpublished.
- [4] H. Ranawana, and D. Hewage, "Factors Affecting Service Quality in Public Bus Transportation in Sri Lanka", In: *Proceedings of 8th International Research Conference*. Malabe, Sri Lanka: Colombo International Nautical and Engineering College (CINEC) Maritime Campus, 2015.
- [5] M. Mezghani, "Study on electronic ticketing in public transport", [online] Available at: <https://www.emta.com/IMG/pdf/EMTA-Ticketing.pdf> [Accessed 11 Nov. 2019], 2018.
- [6] M. Schumacher, 2016. Microtransactions in Bitcoin - how expensive are they really?. B. Sc.. Frankfurt, Germany: Goethe University Frankfurt.
- [7] K. Chaudhary, X. Dai, and J. Grundy, "Experiences in Developing a Micro-payment System for Peer-to-Peer Networks", *International Journal of Information Technology and Web Engineering*, 5(1), 23–42, 2010.
- [8] C. Holotescu, "Understanding Blockchain Technology and How to Get Involved", In: *The 14th International Scientific Conference eLearning and Software for Education*, 144 Str.Paunescu Podeanu, 300569 Timisoara, Romania: Ioan Slavici" University of Timisoara.
- [9] S. Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System*, 2008.
- [10] G. Wood, "Ethereum: A Secure Decentralised Generalised Transaction Ledger", 2014.
- [11] K. Panetta, "Gartner Top 10 Strategic Technology Trends for 2019.Gartner report. Online at <https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019> [Accessed 03 Dec. 2019].
- [12] Suica technology and strategy for future development Available at : https://www.jreast.co.jp/e/development/tech/pdf_6/Tec-06-40-49eng.pdf [Accessed 11 Nov. 2019].
- [13] Cha, Shi-Cho and Peng, Wei-Ching and Hsu, Tzu-Yang and Chang, Chu-Lin & Li, Shang-Wei. (2018). A Blockchain-Based Privacy Preserving Ticketing Service. 585-587. 10.1109/GCCE.2018.8574479.
- [14] World's first Blockchain Bus is set to ride in the Netherlands. Available at : <https://medium.com/@ankitmittal1708/worlds-first-blockchain-bus-is-set-to-ride-in-the-netherlands-698d2d9f1db> [Accessed 15 Nov. 2019].
- [15] M. Li, A. Xu, N. Xue, X. Huang, J. Zhang, and Q. Sheng, "Enable Bitcoin Transaction in Public Transport Ticketing System. In: *1st Conference on Emerging Topics in Interactive Systems (ETIS 2016)*, 2018.

IoT based animal classification system using convolutional neural network

L. G. C. Vithakshana*

Department of Statistics & Computer Science
Faculty of Science, University of Kelaniya, Sri Lanka
chathuravithakshana@gmail.com

W. G. D. M. Samankula

Department of Statistics & Computer Science
Faculty of Science, University of Kelaniya, Sri Lanka
samankula@kln.ac.lk

Abstract: The kingdom “Animalia” is used to represent all living creatures on the planet earth, which is fallen into six categories. The language is the most common factor to divide humans and animals. Numerous classification techniques can be used for classification purposes, and the classification commonly can be done acoustically and visually. The classification systems are playing a considerable role, and bioacoustics monitoring was a significant field of study. Visual classification of animals is done by using either satellite images or established camera images. Nevertheless, due to some circumstances, image processing techniques cannot be applied. Then the acoustical classification techniques are taken place to encounter those problems. Even with acoustical methods, a remote observing method is required due to a few issues. Applying an IoT based acoustic classification system was designed using Convolutional Neural Networks (CNN), which is beneficial for those who are interested in monitoring ecosystems such as animal scientists, zoologists, and environmentalists. The hardware implementation was designed to collect the data from the place it was placed. The audio clips were preprocessed using the Mel-frequency Cepstral Coefficient (MFCC). A CNN architecture based on TensorFlow was used for the training process. To train and test the network, 400 sound clips of two seconds, such that 40 per each ten animal species, which were gathered from online libraries and formatted using Audacity, were used. The network was trained by changing the different gradient descent optimizers and eventually obtained the confusion matrices for each. The best result was gained by the AdaDelta, Gradient Descent, and RMSProp optimizers with 91.3% accuracy for each. Among them, AdaDelta had the most stable and increasing learning approach. As a future extension, to improve accuracy, a large number of data will be used.

Keywords: Animal, Convolutional Neural Network (CNN), Internet of Things (IoT), Mel-frequency Cepstral Coefficient (MFCC), TensorFlow

I. INTRODUCTION

The word “animal” represents the massive varieties of species living on the planet earth that falls into the kingdom “Animalia” including the categories, mammals, birds, reptiles, fishes, amphibious, and arthropods [1] which can able to move, breath oxygen, consume organic materials, and sexually reproduce. The term “animal” is commonly used for all non-human species. Among the animals who live on the earth that is 8.7 million, there are only 1.2 million of them were identified, and there are more than seven million of them to be recognized as 2011 [2].

Since there is a difference between human calls and animal calls, language is the most common factor that separate human from other animals. In a human call, each

word means subject, concept, or action. The vocal patterns can be used to distinguish each other in humans, but when it applied to the animals, it was quite a hard task to do so. However, there are some variations between the same animal species as well. Humans can express their emotions and moods using words as well as using the face. If a word was pronounced with bass, it could be classified as an adult man, and if the word was pronounced with a thin voice, it could be classified as a child. The animal calls also have information regarding emotions and mood. These calls consisted of the pitch, loudness, repetitions, and many more [3].

Classification systems are one of the most widely used technologies in the world to classify humans, animals, or objects. In the present, there are plenty of types of classification systems are available, including face classifications systems, vocal classification systems, fingerprint classification systems, iris classification systems, and so on. Those classifications and identifications were done by either visually or acoustically. For visual classification, image processing techniques are used, and for acoustical classification, audio (or vocal) processing techniques are used. In the sense of sound classification, environmental sound classification and bird sound classification are most popular.

As the name implies, the Internet of Things (IoT) is any “thing” that uses the internet to communicate and transfer data between each other without direct human interaction. Nowadays, IoT is touched by almost every field and industry, including business transactions, logistics, real-time transactions, healthcare, voice-enabling systems, and many more. Besides that, IoT can be used to observe animal behaviors and classify them in an efficient and useful way.

Convolutional Neural Networks (CNN or ConvNet) [4] is a well-known deep learning algorithm. It was invented based on the natural visual perception mechanism of the living creatures. CNN is commonly used for the classification. The most basic building blocks of neural networks are called neurons. Apart from that on CNN, there are convolutional layers, pooling layers, activation functions such as Rectified Linear Unit (ReLU), and fully connected layers (dense layers) available. There also have some variations of neural networks, including CNN, Recurrent Neural Networks (RNN), AutoEncoders, and so on.

Feature extraction is used for dimensionality reduction of raw data into more manageable data called the features (or variations). To extract the features from given audio or image data, the number of techniques are used. Mel-frequency Cepstral Coefficient (MFCC) is one of the best algorithms to

do so. Apart from that, there are other speech feature extraction methods available, including Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP) [5].

Machine learning (ML) is one of the most influential and powerful technologies in today's world. ML is a subfield of well-known and accessible technology, and ML can be divided into four main subfields, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. The main goal of ML techniques is to convert information into knowledge. To classify the images as well as to classify the audio, ML techniques are used.

The same classification techniques applied to humans and objects can be applied to animals as well. Animal classification can be done either visually or acoustically. Visual classification of animals can be done using satellite images or images from an established camera in the forest. However, due to some circumstances, visual classification mechanisms cannot be used with animals. Satellite images cannot be obtained in a deep well-grown forest is one of the reasons. It is hard to detect animals who live in a well-grown forest, even by using a camera, without cannot spot them to a camera. That is why the acoustic classification methods take place to encounter those problems. Even using acoustic techniques, a remote observing method is required due to a few reasons. The in-situ methods can be expensive, and time-consuming is one of the reasons.

Bioacoustics monitoring is a significant field of study in the world. There are several techniques available for bioacoustics monitoring ranging from traditional in-situ methods to semi-automated bioacoustics monitoring systems [6] on selected ecosystems. In most of the previous studies, bird sound classification was done.

This type of designed system has benefited the people who are interested in bioacoustics monitoring within a specific ecosystem, including animal scientists and environmentalists.

To address the problems mentioned above for the targeted people, this study was done using the deep learning perspective with supervised learning. During this study, a well-known feature extraction technique called Mel-frequency Cepstral Coefficient (MFCC) was used to extract the features from the given audio clips, fed those to the CNN architecture, trained, and tested the model using the collected dataset. The used CNN architecture consisted of convolutional, ReLu, max-pool, and fully connected layers.

Usually, deep learning techniques are required for a large number of data to achieve the best results for a selected model. However, unfortunately, since there is no ready-made animal sound dataset available on the internet, the audio clips were collected from different online libraries [7], [8], [9], [10]. The collected dataset consisted of 400 audio clips with two seconds of each of ten animal species including bat, elephant, hornbill, junglefowl, macaque, myna, peafowl, pig, squirrel, and toad. Background noise clips were also added to conduct the training process more accurately.

The hardware module was designed using Arduino UNO, NodeMCU for wireless communication, and a microphone sound sensor. This module collects the sound and sent those to the cloud server for easy access to the data in the implemented Android application. An android-based mobile application was implemented, and it was used to collect those data on demand and finally display the results on the Android interface when an animal was called.

Section II will describe the related works done in this area of study, Section III will describe the methodology used to implement the system, Section IV will discuss on the experimental process, and the final section, Section V will discuss on the conclusion and future work.

II. RELATED WORKS

There were several related work done in this area of interest by using numerous technologies include CNN, joint algorithms using Zero Cross Rate (ZCR), MFCC, and Dynamic Time Warping (DTW), Hidden Markov Model (HMM), and using most common libraries and frameworks including LibROSA, TensorFlow, and Keras. Furthermore, most of the studies were done to classify bird sounds, and there were few studies done in the field of farming also. Below are some of the previous studies done in this area.

Yeo, Al-Haddad, and Ng [3] researched to develop an identification detection system using animal voice recognition. In their system, they used the Zero Cross Rate (ZCR), MFCC, and Dynamic Time Warping (DTW) joint algorithm for animal voice recognition. ZCR was used to detect the endpoint of given input voice data and remove the silenced voice. MFCC was used for feature extraction of the given audio data files. Finally, voice pattern recognition was done using the DTW algorithm. As their obtained results, the developed system has worked as expected.

K.H. Frommolt and K.H. Tauchert [6] researched on birds using the bioacoustics monitoring system. Cardioid microphones and four four-channel recorders were used to record the audio files during entire nights for five years from 2008 to 2012. During the study period, the changes in the number of birds, as well as their spatial distribution with changes in habit structure of birds, were monitored.

Grill and Schluter [11] conducted a study to detect the presence of birds on Mel-scaled log-magnitude spectrograms in a given audio signal by using two CNN architectures, namely "bulbul" and "sparrow". Both architectures were performed very similarly on test datasets. The "bulbul" system was performed slightly better in the development set. Eventually, they obtained an Area Under Curve (AUC) measure of 89% for the hidden test set, which was higher than any other previous contestant. They concluded that any other further improvement might not need for this study.

Tom, Antoni, and Andy [12] researched music pattern feature extraction of audio music using a CNN model for ten different music genres by using evenly distributed 1000 songs among those ten genres as the dataset for that experiment. As they mentioned that it was required minimum prior knowledge to construct their feature extractor compared to the previous works. In conclusion, they stated that the developed model was not robust enough to generalize the training result in unseen music data.

In the study of Sainath and Parada [13], they introduced the “cnn-trad-fpool3” CNN architecture for Small Keyword Spotting (KWS) which was roughly similar to the CNN architecture that used in this paper to conduct this project along with few other architectures. “cnn-trad-fpool3” was contained, two convolutional, one linear low-rank, and one Deep Neural Network (DNN) layer.

In the Khamparia, Gupta, Nhu, Khanna, Pandey, and Tiwari [14] study, they conducted a sound classification using CNN and Tensor Deep Stacking Network (TDSN). CNN architecture was built with two convolutional layers that activate with max-pooling layers, and a final fully connected layer. They used ESC-10 and ESC-15 datasets to train and achieved 77% and 49% for CNN and 56% in TDSN for both datasets.

Nordby [15] study was about the environmental sound classification on microcontrollers using CNN. By using the Urbansound8k dataset, he experimented on the STM32L476 low-power microcontroller by using Keras deep-learning framework. The noise classification was done on the sensor node. Furthermore, for future works, he proposed to build a wireless sensor network that is powered using this technology.

Ajibola Alim, S., & Khair Alang Rashid, N. [5] study, investigated on different feature extraction algorithms for speech recognition, including Mel-frequency Cepstral Coefficient (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP). Moreover, they concluded that none of the methods were superior to the other; the area of application would determine which method to select.

Emre and Tek [16] conducted research using a CNN model with three convolutional layers with a max-pooling layer and three fully connected layers. Three convolutional layers and first two fully connected layers were used Rectified Linear Units (ReLU) as the activation function since the final fully connected layer was used softmax as the activation function. They were preprocessed the audio files using MFCC, and LibROSA and features were extracted. They have experimented by changing seven gradient descent optimizers to verify what was the most appropriate optimizer for their dataset of 875 audio clips, which belongs to ten different animal classes. They observed that Nesterov-accelerated Adaptive Moment Estimation (Nadam) was the best optimizer for their experiment with 75% final accuracy.

Like the previous studies, there was not any IoT based system that was proposed and implemented the domain of animal classification using vocal analysis. Some studies were obtained not much higher level of accuracies for the models. Some studies were observed the bias results due to the class imbalance of the used dataset. Some researches used the relevant technologies and achieved higher accuracies, but not in the domain of this study.

III. METHODOLOGY

This study was done by considering the following methodology. The first step was to collect the dataset and prepare it. Feature extraction was done using a Mel-frequency Cepstral Coefficient (MFCC) algorithm. Then the

extracted features were fed to the layers for classification of Convolutional Neural Network (CNN). The hardware module and the mobile application were designed.

A. Dataset

The dataset was about forest animals. Since there was no ready-made dataset available, the sound clips were collected on different online libraries [7], [8], [9], [10], and prepared those as TABLE I using Audacity. In this process, a considerable number of audio clips had been removed due to low animal sound levels and high background sound levels. The background noise dataset was collected from an online library [17] and converted to WAV files with a 16000 Hz sample rate and a bit rate of 32 bit. The complete sound dataset of 417 audio files with 40 from each of ten animal classes and 17 background noise clips were prepared.

TABLE I. SPECIFICATIONS OF PREPARED AUDIO CLIPS

File Type	WAV
File Size	62.5 KB
Sample Rate	16 KHZ
Bit Rate	32
Channel	MONO
Clip Duration	2 seconds

B. Feature extraction

In previous studies, there were many algorithms used for feature extraction, including MFCC, LPC, and DTW. The used CNN was not supposed to feed the raw audio files as it was. In this study, a widely used feature extraction algorithm called MFCC was used. The feature extraction was not done separately. Instead, it was done at the beginning of the training using the same script.

C. Convolutional Neural Network

Many CNN architectures were there, and one of them was selected for this study, which consisted of two convolutional 2D layers and a max-pooling layer. Each layer was activated with Rectified Linear Units (ReLU). The final fully connected layer was activated with softmax. A 1x1 stride size was used in convolutional layers, and a 2x2 stride size was used in max-pooling layers. The full architecture is shown in Fig. 1.

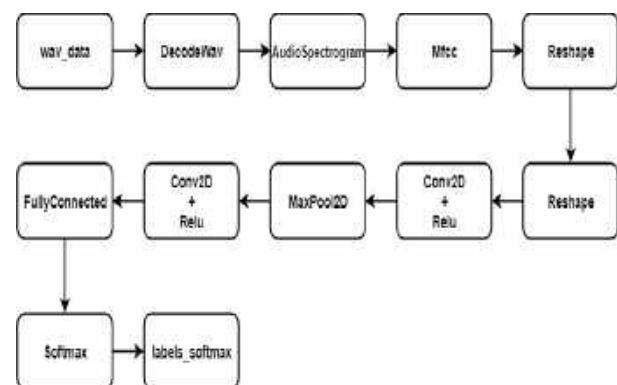


Fig. 1. CNN architecture

D. Hardware module

The hardware module was designed to collect the audio data from a faraway site that the device was placed. This module consisted of Arduino UNO powered by an ATmega328P microcontroller for getting connected the all devices, NodeMCU powered by ESP8266 (LX106) CPU for

wireless communication, and microphone sound sensor to collect the audio data. Those collected data were sent to the server for easy access later using implemented mobile applications.

E. Mobile application

The implemented mobile application based on the Android platform was designed to retrieve the audio data from the cloud server and eventually identify the animal with the help of the CNN model. The mobile application was implemented using Android Studio for native Android development. The application was built for API level 26, and above and Gradle version 3.5.1 was used. Fig. 2 shows some implemented interfaces.

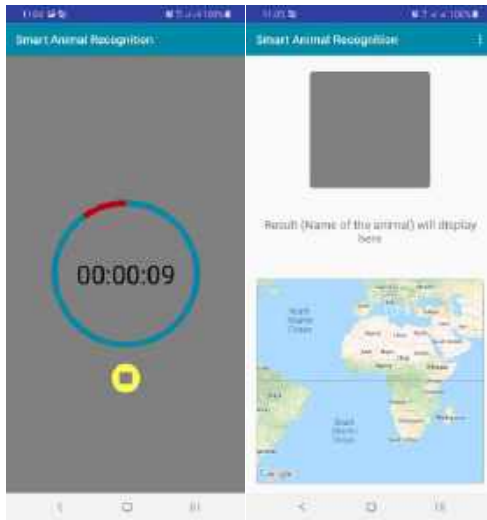


Fig. 2. Implemented Android mobile application interfaces

IV. EXPERIMENTAL PROCESS

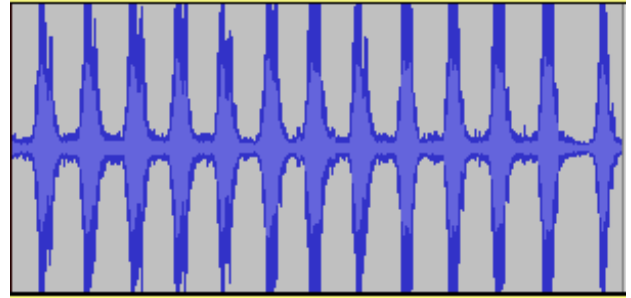
The experimental process was done using TensorFlow, and mainly it has consisted of three steps, including feature extraction, training, and testing. The experiment was done using Anaconda Prompt using TensorFlow GPU under the Windows 10 operating system. The dataset was split into three parts as training, testing, and validating. The 90% of data was split as training, the rest of the 10% was split as the testing set. A randomly selected another 10% was split as a validation dataset. For the optimization, six selected optimizers were considered before starting the training.

A. Training

During the training process, for the ten animal classes, bat, elephant, hornbill, junglefowl, macaque, myna, peafowl, pig, squirrel, and toad were used. For the background noise class, `_background_noise_` label was used. The rest of the class was fell under the unknown label. The training process was conducted using the hyper-parameters, as illustrated in TABLE II. Features were extracted using the MFCC algorithm for the given dataset. The following Fig. 3 and Fig.4 show a raw audio file and a related spectrogram image for a given audio clip of the hornbill class.

TABLE II. HYPER-PARAMETERS OF THE TRAINING PROCESS

Key	Value
batch_size	40
preprocessor	MFCC
how_many_training_steps	15000, 3000
learning_rate	0.001, 0.0001
model_architecture	Conv
optimizer	Gradient Descent, Momentum, Adam, AdaGrad, AdaDelta, RMSProp



3. A raw audio file for hornbill class

At the end of the training, the trained model was saved as “.pb” file for future usage. Training accuracy was increased with the training steps. The model was trained up to 18000 steps of each by change one of the hyper-parameters called optimizer to verify what was the most appropriate hyper-parameter fitted for the model. Since the low learning rate was led to the highest accuracy, the first 15000 steps were trained using a 0.001 learning rate while the last 3000 steps were trained using a 0.0001 learning rate.

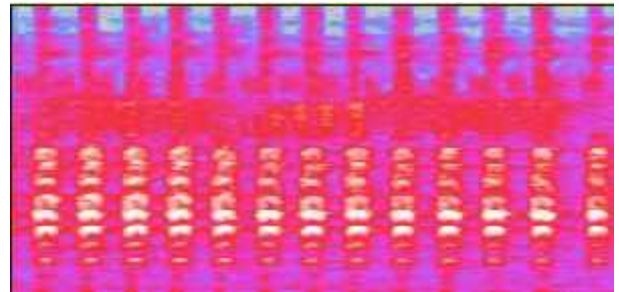


Fig 4. Related spectrogram

The training accuracy was not changed that much after 2000 steps in most of the cases except AdaDelta. Some of the observed graphs shown in Fig. 5 and Fig. 6. At the end of the training, confusion matrices for both test and validation were generated.

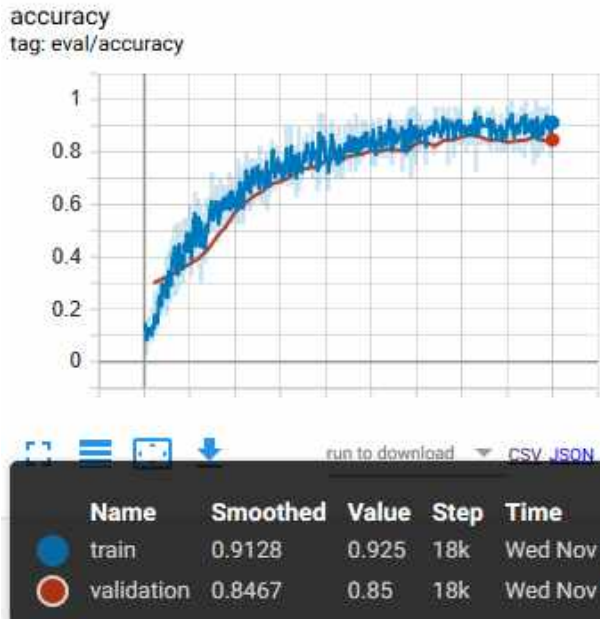


Fig. 5 AdaDelta accuracy graph using TensorBoard



Fig 6. Gradient Descent accuracy graph using TensorBoard

B. Test and results

In this study, the trained CNN model was used to identify an unknown sound file as one of the labeled animal classes. For the given dataset, a confusion matrix was generated, and the visualized confusion matrix generated for AdaDelta optimizer is shown in Fig 7.

Rather than mention the accuracy, a matrix can summarize the actual accuracy against the predicted values of the model for the provided test dataset. The columns show the predicted values while the rows show the actual ground truth values. The diagonal entries show the correct classification, while other entries (off-diagonals) show the incorrect classification of the given labels.

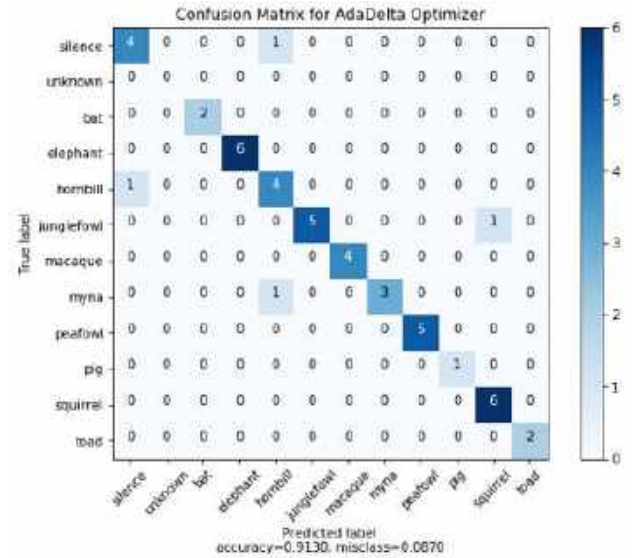


Fig. 7. AdaDelta optimizer visualized confusion matrix for test set

As an example, from the given 10% of dataset as the test dataset, among six elephant sound clips, the model used AdaDelta optimizer identified all six as elephants. Among junglefowl sounds, five clips were identified as junglefowl, and another one was incorrectly identified as a squirrel. Moreover, there were no unknown classes available within the dataset. The matrix was shown; the accuracy was divided among all classes almost the same way because a balanced dataset was used. There was no bias in any class.

Fig. 8 shows one of the observed results for a given audio clip. It was tested using a model trained by AdaDelta optimizer for one of the audio clips, which were included the sound of a peafowl. As observed, the accuracy score of 0.97427 out of 1.0000 was achieved successfully. Also, some inaccurate percentages were archived. Therefore, the given audio clip was successfully identified as a peafowl.

```
E:\train6\dataset\peafowl\peafowl_21.wav
peafowl (score = 0.97427)
hornbill (score = 0.01499)
elephant (score = 0.00624)
```

Fig 8. Observed results for given peafowl audio clip

The train was conducted by changing six optimizers, and as observed, the accuracies were not changed after 2000 steps in most of the cases. Once the test dataset was run on models, the best accuracy was obtained by AdaDelta, Gradient Descent, and RMSProp optimizers, which were 91.3%, and the worst accuracy was obtained by Momentum optimizer which was 82.6% as listed in Table III.

TABLE II. FINAL ACCURACY WITH RESPECT OPTIMIZERS

Optimizer	Accuracy (%)
AdaDelta	91.3
Gradient Descent	91.3
RMSProp	91.3
AdaGrad	89.1
Adam	87.0
Momentum	82.6

The IoT module and the mobile application were tested by placing it in a nearby place and observed the audio which was gathered around that place was saved in the database and successfully retrieved by the mobile application when request.

Then using the already included pre-trained model, the audio was recognized as one of the species that observed. Hence the both IoT module and mobile application were functioning well in the given conditions.

V. CONCLUSION AND FUTURE WORK

In this study, a system was proposed to classify a given animal sound clip for one of the pre-known ten animal classes. As described during this paper, the collected sound clips were preprocessed using MFCC and fed those to the created CNN architecture. The training and test were done using 417 audio clips, which were belonged to ten different animal species. The confusion matrices were generated for both test and validation datasets. One of the previous study [9] was shown that the accuracies were affected by class imbalance. To avoid this issue, the balanced classes were used. As they achieved 75% accuracy, this study was achieved 91.3% accuracy. The most suitable model was the model that used AdaDelta optimizer with 91.3% accuracy. Furthermore, a proper system was proposed to classify animals who lived in faraway locations by voice.

It was a well-known fact that deep learning strategies were required a large number of a labeled dataset to obtain much better results. Hence, the dataset will be expanded to have more than 417 audio clips in future studies to achieve more accurate and better results. Also, it will be better to consider changing other hyper-parameters rather than optimizers and specifications of collected audio clips.

The designed hardware module was supposed to send a gathered audio file to a cloud server and retrieve it by a mobile application on demand and it was functioning well. Rather than using a dataset that collected from online libraries, in the future extension, the data collected from this hardware module will be used to train the CNN model and to identify the animals, because it takes about months to collect a proper sample by placing a single hardware module in different forestry places.

In this study, the TensorFlow model was used. In a future study, to build a CNN model from scratch and conduct the CNN training for achieving a much better result is recommended.

In a future study, this same system can be improved to classify animals by their gender as well because there is a slight difference of voice between male and female animals of the same species. Also, this system can be improved to classify animals by their age groups, because cubs and elder ones of the same species have two different vocal patterns.

Furthermore, it is recommended that this system be developed to a more advanced level to recognize the mood of the same animal in the different stages of the day that they live.

REFERENCES

- [1] BYJUS. (2019). Classification Of Animal Kingdom - Non-chordates And Chordates. [online] Available at: <https://byjus.com/biology/classification-of-animal-kingdom/> [Accessed 23 Nov. 2019].
- [2] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on Earth and in the ocean?," *PLoS Biol.*, vol. 9, no. 8, p. e1001127, Aug. 2011.
- [3] Che Yong Yeo, S. A. R. Al-Haddad, and C. K. Ng, "Animal voice recognition for identification (ID) detection system," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, 2011, pp. 198–201.
- [4] M. C. Munteanu, A. Caliman, C. Zaharia, and D. Dinu, "Convolutional neural network," 10497089, 03-Dec-2019.
- [5] S. Ajibola Alim and N. Khair Alang Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," in *From Natural to Artificial Intelligence - Algorithms and Applications*, R. Lopez-Ruiz, Ed. IntechOpen, 2018.
- [6] K.-H. Frommolt and K.-H. Tauchert, "Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird," *Ecol. Inform.*, vol. 21, pp. 4–12, May 2014.
- [7] "Macaulay Library." [Online]. Available: <https://www.macaulaylibrary.org/>. [Accessed: 14-Dec-2019].
- [8] "xeno-canto: Sharing bird sounds from around the world." [Online]. Available: <https://www.xeno-canto.org/>. [Accessed: 14-Dec-2019].
- [9] "the Internet Bird Collection | HBW Alive." [Online]. Available: <https://www.hbw.com/ibc>. [Accessed: 14-Dec-2019].
- [10] "Free Sound Effects." [Online]. Available: <https://www.freesoundeffects.com/>. [Accessed: 14-Dec-2019].
- [11] T. Grill and J. Schluter, "Two convolutional neural networks for bird detection in audio signals," *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017.
- [12] T. L. H. Li, A. B. Chan, and A. H. W. Chun, "Automatic musical pattern feature extraction using convolutional neural network," in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2010, IMECS 2010, 2010*, pp. 546–550.
- [13] T. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," *INTERSPEECH*, 2015.
- [14] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.
- [15] J. O. Nordby, "Environmental sound classification on microcontrollers using Convolutional Neural Networks," *Norwegian University of Life Sciences*, 2019.
- [16] E. Sasmaz and F. Boray Tek, "Animal Sound Classification Using A Convolutional Neural Network," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. 2018.
- [17] "Nature Sounds - Rain Sounds: Free Download, Borrow, and Streaming: Internet Archive." [Online]. Available: <https://archive.org/details/relaxingrainsounds>. [Accessed: 14-Dec-2019].

Source credibility analysis on Twitter users

Malith Wijesekara*

Faculty of Information Technology
University of Moratuwa, Sri Lanka
malith.sandagiri@outlook.com

Gamage Upeksha Ganegoda

Department of Interdisciplinary Studies
Faculty of Information Technology
University of Moratuwa, Sri Lanka
upekshag@uom.lk

Abstract: Social media has gained impressive popularity all around the world in the last decade. Social networks such as Twitter, Facebook, LinkedIn, and Instagram have acquired their user's attraction by maintaining their identity with very similar features. With the popularity of these platforms, now a day most of the users tend to rely on the information published on social media. Therefore, the credibility of social media information is playing a major role in the present cyberspace. As an example, the Twitter platform is handling 500 million tweets per day. Most of the twitter messages are truthful, but the twitter platform is also used to spread rumors and misinformation. Truthfulness or reliability is depending on the source's credibility. Twitter profiles can be identified as the information source on the twitter platform. In this paper, a user reputation-based prediction method is proposed to analyze the twitter source credibility. The proposed solution is mainly based on the k-means clustering model. Another two models namely, news category analysis and sentiment analysis are deployed to generate novel features for the clustering method. The objective of this paper is to introduce a credibility rating method to visualize the user credibility of twitter user profiles. So that followers can have an understanding about the trustworthiness of the information published on that profile. Producing the agreement score for a specific twitter user is one of a novel experiment in this research. Achieved accuracy by the system is 0.68 according to the evaluations conducted.

Keywords: Credibility analysis, News category analysis, Sentiment analysis, Source credibility, Twitter source

I. INTRODUCTION

Credibility can be defined as "A perceived quality composed of multiple dimensions". In other words, "Offering reasonable grounds for being believed" [1]. On the other hand, source credibility can be defined as a communicator or the news content provider's positive characteristics that affect his follower's acceptance of the provided news content. As a result of the popularity of social media, petabytes of data are produced per day. Most of the information is important for specific communities. But the question is actually above information is produced by credible sources. Most of the time if a social media user needed to identify the credibility of another profile, he/she has to visit the source profiles and use their cognitive skills to guess how to extend that source can be trusted. This system is proposed as a solution to that problem.

Twitter social media platform is a microblogging service that allows users to publish 140 characters long message called "tweet". Users can append images and reference URLs along with the tweets [2]. Tweets can be published through the web portal, through the smartphone app, through an email as well as sending an SMS. Twitter profiles that are currently active are providing news content for its' followers. There

are twitter sources recognized as reputed and credible by the community. On the other hand, there may be accounts that produce non-credible information to the community. Identifying the credibility of a specific account is a complicated and time-consuming task for individuals. It will be a collection of subtasks such as check the information on a specific profile is matched with other sources, the impression of the followers about the source profile, is it a verified account by twitter, identifying the information categories published often and recognize the suspicious activities of that account [3]. It will be ideal if there is an intelligent application to check the credibility of a specific twitter profile and provide a rough idea about how to extend that account can be trusted. In this paper, a method is proposed to evaluate and forecast the credibility level of a specific twitter profile that is currently active in the platform. This system is a subpart of the research project with the title "Fake News and Social Media Platform Manipulation Detection". This system is mainly focused on the sources that are published newsworthy information and that has the strength to influence a large number of people. The system proposed by this paper can work as an independent module and the inputs will be screen name twitter user and outputs are credibility score of relevant twitter profile. In addition to that, the author's contribution is to implement a system that is capable of generating a score representing all the followers, identify the most interesting information category and predict the source credibility level for a specific twitter profile.

Organizing the rest of the paper as follows. The next section is described the Related work of the subsystem proposed by this paper. The methodology of the system is presented in Section – III. Finally, Discussion is presented in Section – IV.

II. RELATED WORK

The system proposed by this paper is responsible for three main sub-tasks. Those are achieved using the components mentioned below.

- A. News category analysis
- B. Sentiment analysis for tweets
- C. Twitter Source credibility analysis

Related work for the above models discussed below.

A. News category analysis

Twitter is a microblogging service. Therefore, identifying and classifying tweets is hindered by traditional documents. The reason for the difference is short text has no sufficient word occurrences for the feature extraction [4].

A paper written by Bharath Sriram et al. proposed a method to classify short text using a set of domain-specific features extracted from the short text and the user profile which published that text. A paper published by Thorsten Joachims explored the text categorization using Support Vector Machines (SVM) [5]. It was thoroughly discussed on why Support Vector Machines work well for text categorization. Furthermore, the techniques that can be used for Natural Language processing such as Tokenization and Vectorization is discussed in a survey which is published by Mohd Suhairi and Md Suhaimin [6].

B. Sentiment analysis

Sentiment analysis is a popular field in machine learning which is gained tremendous attention from the researchers. Users' opinions on social media posts can be identified as discovery and recognition of the type of expression that is positive or negative on diverse subject matters of interest. The survey conducted by Albert Bifet and Eibe Frank discussed the techniques that can be used for sentiment knowledge mining [8]. The implementation was based on Kappa statistic and the final output for a given text is Positive or Negative sentiment. Similar research conducted by Hao Wang et al. also discussed the real-time analyzing twitter data to extract the opinions on election candidates. Another research by Kelvin Kiema Kiilu et al. discussed applying the Naïve Bayes classifier for sentiment data. The research was achieved 64.47% accuracy for unigram and 70% accuracy for Bigram. In addition to that, the ability to classify opinion into separate categories such as "positive" or "negative" or "neutral" is discussed in a survey conducted by S. M. Vohra et al. [10].

C. Source credibility analysis

The information in the social networks, cannot be trusted as the conventional news media because there may not be an undergone verification process with the news published in social networks. In addition to that, those social media users who report information may have their own biases, agendas, and perceptions. A survey conducted by DWI H. Widyantoro and Yudiwibisono introduced a way to use the sentiment data to determine the credibility of the published tweet [11]. The proposed method is analyzing the replies for a particular tweet and check the support content and oppose content from the followers. Another paper published by Mohammad-Ali Abbasi and Huan Liu state that information credibility is "The quality of being trustworthy" and information credibility has three major parts namely, source credibility, media credibility, and message credibility. Furthermore, a method based on the CredRank algorithm is proposed for evaluating source credibility in the above research [7, 12]. In addition to that, it is a clustering-based method and it is problematic to analyze the source credibility only depending on the followers' reviews regarding the tweet which is published by a particular source. Interesting research was conducted by Gupta et al. that assign a credibility score to each news or event [13]. The more information is credible; the credibility score will be higher.

An approach based on a PageRank-like credibility propagation method is presented by the author. Another research conducted by Krzysztof Lorek et al. proposed a method with creating a data set manually tagged by checking the twitter profiles and their information [14]. They used their own rules to label each account is credible or non-credible. Below mentioned items were taken into consideration for manually tagging the credibility of the profiles.

- Visible characteristics of the profile photo, account name, availability of the verified badge.
- Tweet text syntax and sentence correctness.
- Content of the images appended with tweets.
- Availability of the external links.

The proposed solution in this paper also included the above features and in addition to that added an extremely impotent feature which represents the opinion of the followers about the content of the source.

III. METHODOLOGY

The system proposed by this paper can be separated into four components. Those can be work as separate entities and some of the outputs of the component will be input to the others.

1. News category analysis component
2. Sentiment analysis component
3. Source credibility analysis component
4. Source credibility visualization component

The flow diagram for the system with the above four components is shown in Fig. 1. Input for the system is a username which needs to be analyzed the source credibility. The data required for the analysis should be acquired using the screen-name through the Twitter API [9]. The account timeline of a specific user and the user data has to get separately. The user timeline means the tweets that are published by the user so far. It is playing a major role in analyzing user activities. Only tweet text is required for news category analysis and sentiment analysis components. The output of the above two components will be inputted to the credibility analysis component. Although there are plenty of systems for sentiment analysis and text classification, it is better to implement them both as the accuracy of these components will directly affect to the Credibility analysis component.

The other input for the Credibility analysis component is user data that are requested through the twitter API. This component is responsible for generating the credibility score. It is a probability number that shows how to extend that account can be trusted. In addition to that, recent activities of the relevant account are also analyzed by the Credibility analysis component. Finally, the summarized data and information is visualized by the Source Credibility visualization component. Data visualization techniques are used to interpret information discovered by the system and highlight the follower's reactions for recent activities of the relevant twitter profile.

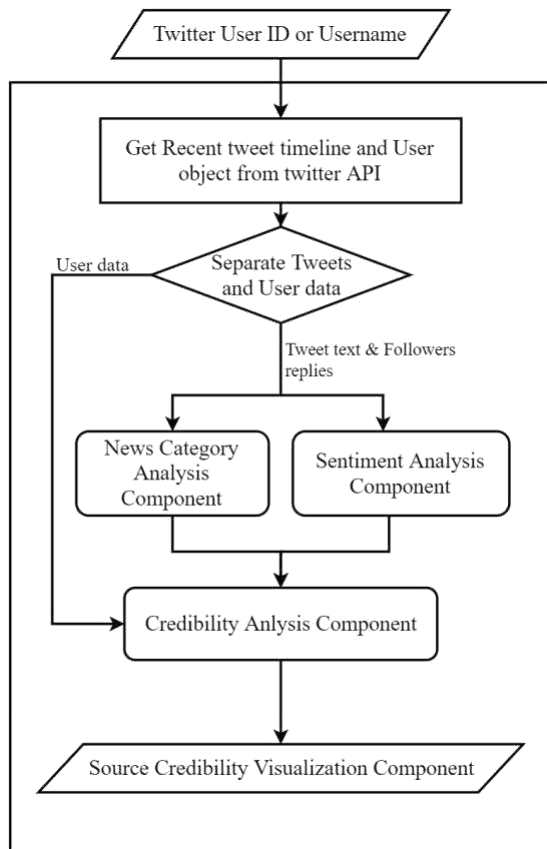


Fig. 1. Flow of the proposed solution

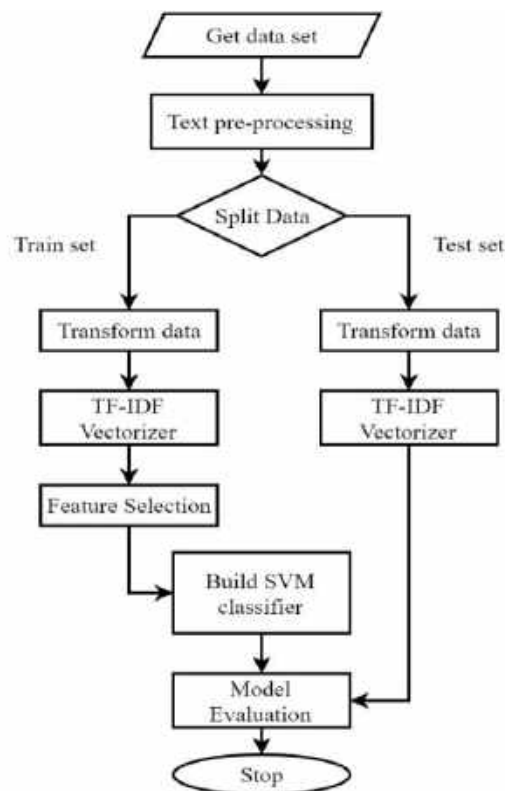


Fig. 2. Flow of the News Category Analysis Component

As mentioned above this system is proposed as a combination of four separate components. The final results are highly dependent on the accuracy of the sentiment analysis and category analysis components. Therefore, the implementation of those components also included in the research. The four subsections are described in detail as follows. The four subsections are described in detail as follows.

A. News category analysis component

News category analysis is responsible for identifying more relevant news category for a given tweet message. this process is based on semantic analysis. Text classification problem has to be solved using a supervised machine learning approach [15]. A comparative study is conducted to identify the best classifier for that data set. Suitable machine learning algorithms are Naïve Bayes classifier, Decision Tree (Hoeffding Tree) and Support Vector Machine. The flow chart of the proposed methodology is shown in Fig.2. According to the Fig.2. Data set consists of tweets and the relevant category is preprocessed to remove blank records, convert text to lowercase. After splitting the dataset into train and test sets, the text is tokenized or separated the sentences into tags. In the next phase, they are vectorized to input into the classifier to create the model. Finally, the best classifier is identified by comparing the results.

B. Sentiment analysis component

Opinion mining or sentiment analysis has to be performed to evaluate the follower’s responses. Opinion mining is the automated extraction of opinions, attitudes, and emotions from the text. Mainly there are two methods for sentiment analysis.

- I) Lexicon-based methods
- II) Machine Learning-based methods.

The lexicon-based method is based on predefined lexicons which are similar to a dictionary. Previously mentioned lexicons are arranged as a list with defining they are positive or negative with a sentiment score. As an example, the word ‘nice’ is defined as a positive word having sentiment scores of +2. The sentiment score can be changed according to the researcher’s point of view who defined the lexicons. If a phrase is an input to the algorithm it will combine individual polarity and calculate the sum of the sentiment scores to output the sentiment of that phrase.

The machine learning method is similar to the automatic mode which is discussed in the previous section: News Category Analysis Component. Naïve Bayes classifier, maximum entropy, and SVM are achieved tremendous success in sentiment analysis. According to the conducted analysis of opinion mining, it is better to choose a machine learning method for sentiment analysis. Because supervised machine learning techniques are relatively performed better than lexicon-based methods. Most of the researchers experienced that SVM has a higher accuracy over the above-mentioned algorithms [16].

But the drawback of the SVM, the smaller the training data set it will reduce the accuracy of the model. So, a large amount of data is required for better results. Sentiment data sets can be found online in large varieties.

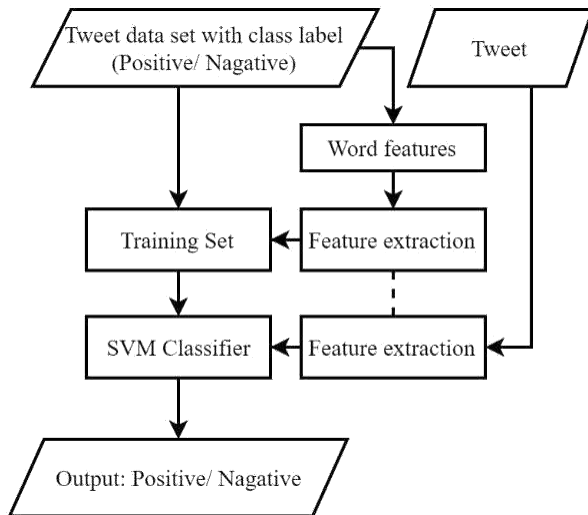


Fig. 3. Flow of the Sentiment Analysis Component

The SVM classifier is selected as the classifier for the model. Fig. 3. is illustrated the flow of the proposed solution for sentiment analysis component. The sentiment data set is needed to be preprocessed and after feature extraction SVM machine learning classifier is applied. In this process, patterns in the text are learned by the classifier concerning class labels. So, the classifier model can classify an opinion or feeling of a given tweet as positive or negative. The sentiment analysis model is used to analyze the follower's expressions and generate a feature for the credibility analysis model.

C. Credibility analysis component

In any social media, there will be a large number of news and posts which are posted by unknown sources. If the user is not sure about a specific source, there is no way to verify the source without viewing and checking the profile. The credibility analysis component is the solution for that manual process of analyzing the credibility of the sources. The outcome of the credibility analysis is mainly relying on opinion mining. As an example, if someone reads the replies on a particular tweet, he/she can have a rough idea of whether the followers agreed with that tweet or not according to the sentiment of the phrases. First, the tweet should be analyzed and then the individual replies to that comment should be analyzed. This process is called agreement modeling and it is a parameter for the credibility analysis model [17]. The overall sentiment of the followers' replies to the source profile is calculated in agreement modeling and the output of the process is agreement score. Availability of the verified badge from twitter, responses of the followers, and profile information such as the number of followers and lifetime of the profile are also can be identified as factors to determine source credibility [11, 13]. Already conducted researches identified that Statuses count, followers count, length of the user description, length of username, number of days from account creation date, availability of verified badge, availability of an external URL for the profile, availability of profile location and protected mode are can be used for analyzing source credibility of a twitter user. A novel feature, agreement score and interested news category of the user are added to the data set. All the features and their effect on the credibility of the source are summarized in Table I.

TABLE I. FEATURE SET FOR CREDIBILITY ANALYSIS

	Feature	Description (a tweet is more credible when...)
1	Agreement score	Source has a higher agreement score
2	Interested news category	Will be depending on the other features
3	Statuses count (No. of tweets)	Source has a higher number of previously published tweets
4	Followers count	Source has a higher number of followers
5	Length of the description	The description length is more than '0' characters Length and higher lengths
6	Length of screen name	Source screen name differs from the username
7	The lifetime of the profile	Hight lifetime
8	Is it a verified account?	True
9	Has the "Website" parameter set in the profile?	True
10	Is the user's Twitter profile linked to another social service?	True
11	Is the profile location set?	True
12	Is profile protected?	True

Methodology for Calculating Agreement Score is based on sentiment analysis. A large number of follower's replies for a twitter user are required as the input for this process. Each replay has to be input to the sentiment analysis component and recorded the sentiment. Finally, the below-mentioned rule is applied to get the agreement score.

$$\text{Agreement score} = \frac{\text{Number of positive followers' responses}}{\text{Total number of followers' responses}}$$

The proposed methodology for identifying interested news categories is getting the first 100 tweets published in the twitter user profile and record input them to the news category analysis model discussed previously. According to the output frequently repeated categories will be identified as the interested news category.

However, the collected data set will not be having a ground truth or class label. Therefore, it is obvious that a supervised learning method cannot be applied to the data set. Therefore, the clustering technique is chosen to analyze source credibility [18]. Artificial Neural Network (ANN) also can be used for creating the credibility model, but the higher complexity and requirement of higher processing power is the barrier for working with the ANN method.

The K-means clustering method is a well-known agglomerative partitioning clustering method [19].

K-means clustering technique is not used previously to implement a source credibility related model. Therefore, the

K-mean clustering method is proposed to create a model for the collected dataset. K-means algorithm is an algorithm that identifies the similarities of different data vectors and groups them together into predefined numbers of clusters.

The above data set can be clustered using both k-means clustering or hierarchical clustering. But according to the requirement, five clusters need to be identified as a star rating method is proposed to identify the level of source credibility. Therefore, k-means clustering is the best algorithm because the number of clusters can be predefined. K-means clustering is one of the simplest classifiers and general preprocessing and transformation can be used to build the model. It can be applied to the data set without knowing much about the dataset.

The main challenge of the model is identifying the correct cluster and naming them. K-mean clustering it is common to use a manually annotated small dataset for identifying the properties of the clusters extracted by the algorithm. Therefore, a small set of twitter user profiles was selected and annotated them manually by looking at the feature and rules mentioned in Table 1. Then the manually annotated dataset and trained dataset with predicted values are compared using two bar charts. Hence the proportionally matched rating is considered as correct clusters. Correct clusters for each rating are identified using predicted values and manually annotated values. When it needs to get the credibility level of a new user, a User credibility rating can be acquired by applying the user data into the model and convert the output of the model into the correctly identified rating.

D. Source credibility visualization component

The purpose of this component in the system is visualizing the credibility rating of a twitter user when someone enters a screen name of the profile to the proposed system. In addition to that recent activities, user patterns such as how often tweets are published and highlighted follower's opinions need to be visualized. The impact that can be happened by a specific account to the community also identified using the category analysis component and sentiment analysis component. The impact that can be happened by an account is determined using the categories of the most recent tweets and followers' attitudes about the published tweets. The impact that can be happened by the source is identified using predefined rules. Proposed rules are,

- I. If the sentiment of twitter message and the overall sentiment of the replies are matched with each other it is an agreement. (Negative- Negative, Positive - positive).
- II. Otherwise, it is a disagreement. (Negative - Positive, Positive - Negative).

If the followers' agreement is low, it is identified as a negatively impacted tweet. Otherwise, it will be a positively impacted tweet. In this part, the system should be able to get the recent tweets in the relevant profile timeline a label them as positively impacted or negatively impacted. It helps the system used to have a rough understanding of a specific profile.

IV. EXPERIMENTS

The main consideration for implementing the system is choosing the correct programming language. As most of the things are towards the data science and Python and R languages were the best selection. R language is mainly used for statistical analysis. But the Python is popular as a scalable language and it has several useful highly developed libraries such as Pandas, NLTK and Scikit-learn which are very useful for data analysis. Another plus point of python is it has matured libraries for Twitter APIs than the R language. Therefore, Python is selected for implementing the solution. In addition to that this system will be web-based and has to be handle API endpoints as well. So, the Django REST framework is selected for that purpose.

News category analysis is designed using NLP techniques to identify the relevant category of a twitter message. A data set consist of 124,989 tweets and labeled with the news category is found online for the machine learning process. Experiments were conducted with the above Naïve Bayes classifier, Decision Tree (Hoeffding Tree) and Support Vector Machine classifiers. Support Vector machine was the best classifier with the accuracy of 0.61.

Experiments for the sentiment analysis component also conducted similar to the news category analysis component. A data set with 1.6 Million tweets with the class label is found from a reliable source. The class label has only two distinct values that are positive or negative, which refers to the feeling or sentiment hidden in the text.

The main part of this solution is predicting user credibility.

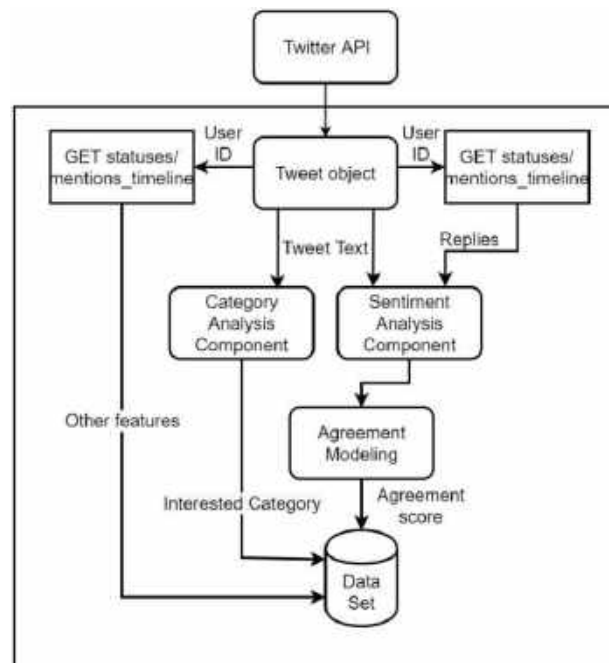


Fig. 4. Flow of the data collection method for credibility analysis component

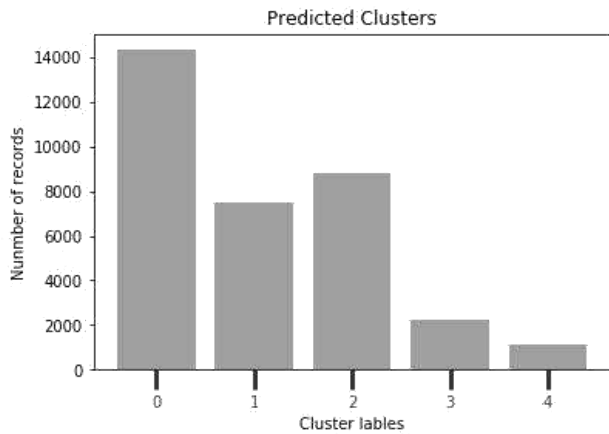


Fig. 5. Predicted labels for the train set

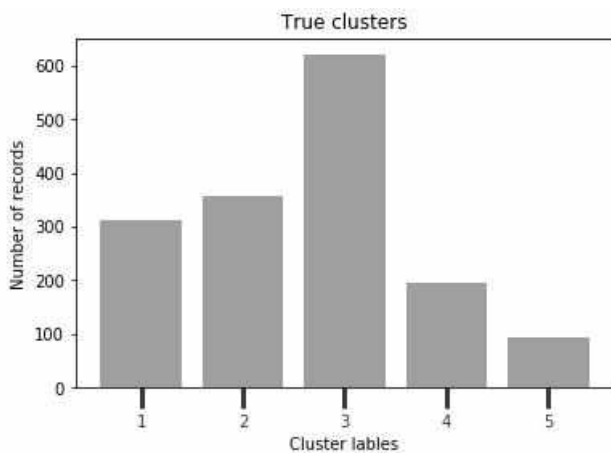


Fig. 6. Manually annotated labels for test set

The data collection process for the credibility model is a tedious task as a new feature has to generate while collecting the data. The flow chart for the data collection process is shown in Fig. 4. Twitter API allows us to get tweets and user profile data up to a certain extent. According to Fig. 4., tweets are collected through the twitter live search API. The tweets object has the details about the source profile. In the meantime, the profile timeline is getting through the API and sent to the category analysis model to identify the most interested category of each user. Meanwhile, Calculating the description length, screen name length and number of days from create date also executed. Profile Data such as status count, a number of followers are separated and saved. At the same time, overall followers' sentiment also calculated in the agreement modelling process.

K-means model is implemented easily by applying the k-means algorithm on the collected dataset. First outliers are removed by analyzing the data set. Encoding, scaling and fitting the dataset to the k-means classifier is needed to repeat several times to match the output of the model with the manually annotated dataset. Predicted labels for the trained dataset and manually annotated labels for the test dataset are illustrated in Fig. 5 and Fig. 6. Results were not satisfactory in the starting experiments. As the optimizations, data scaling and encoding method of the dataset and number of iterates of the K-means algorithm is needed to changed several times. Minmax scaler improved the accuracy and max iteration value is set to 1000 for better results.

V. EVALUATION

According to the experiments conducted for tweet categorization, the SVM classifier is the best-performed algorithm for the selected data set. Multinomial Naïve Bayes classifier treats the feature as an independent way but SVM considers the interaction between features to some extent. K-fold Cross-validation is applied for each classifier with the K = 10. KNN classifier achieved 0.26, Multinomial naïve Bayes classifier 0.54 and SVM classifier achieved 0.62 accuracy levels. K-fold cross-validation is conducted for the sentiment analysis model as well. The accuracy achiever was 0.77. The general flow for the model evolution for category analysis and sentiment analysis is shown in Fig. 7.

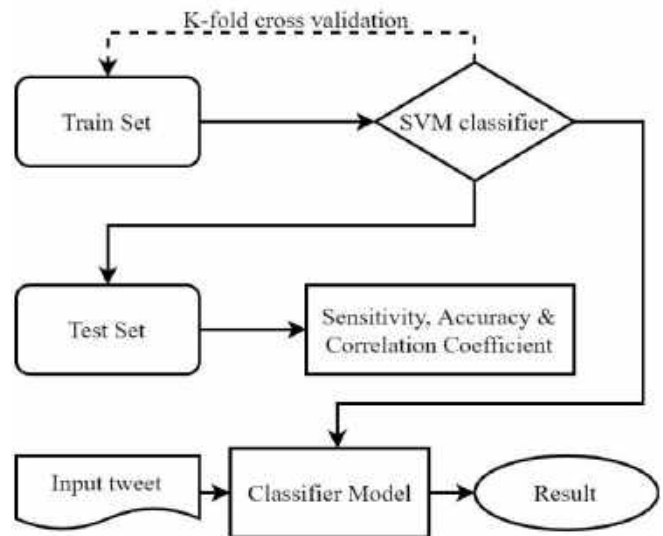


Fig. 7. Evaluation flow for the SVM classifiers

The credibility analysis model is capable of predicting the credibility rating for a user with an accuracy of 0.68. Accuracy was calculated by comparing the manually annotated test set. Precisions for credibility rating for rating 1, 2, 3, 4, 5 are 0.82, 0.61, 0.55, 0.72, 0.70 respectively. The first test set was input to the model and predict the class label. Then the predicted label was compared with the manually annotated class label.

VI. CONCLUSION

As mentioned in the introduction this system is a part of an ongoing research project. The proposed system has to be implemented using mainly Natural Language Processing (NLP) techniques. Research work has to be conducted on text categorization, opinion mining, agreement modeling, and credibility analysis areas. According to the results, support vector machine classifiers have the highest accuracy for both News category analysis model and sentiment analysis. Credibility analysis is a critical task in this system because the data set has to be created using after implementing the above two components. A python program to get and save the live tweet is also created to test the ability to create their own data set.

It was a success but there are some limitations with the twitter API. The main limitation was the number of requests was limited for a specific time period. Sentiment analysis and category analyzing models are evaluated using K-fold cross-validation. For the credibility analysis model credibility

tagging based on a human point of view is required for the evaluation process. It was very rare to find researches for twitter source credibility analysis. This system can be compared with a system proposed by Krzysztof Lorek et al. [14] their solution is a scoring system for each user and mostly based on manually created rules. They predicted the credibility for three categories with a precision of 89%, 87%, and 84% respectively. Our proposed solution It is difficult to compare with the above system as they predicted credibility for three classes. But their accuracy is higher than our proposed system as they predicted for lower number of classes and mostly rely on completely manually annotated data.

REFERENCES

- [1] Castillo, M. Mendoza and B. Poblete, "Information credibility on twitter", 2011.
- [2] Louisa Shu Ying Ha, Mary E. Benedict, Gi Woong Yun and Michael Albert Horning, "Assessing source credibility on social media-An electronic word-of-mouth communication perspective", Etd.ohiolink.edu, 2015.
- [3] R. Guha, R. Kumar, P. Raghavan and A. Tomkins, "Propagation of trust and distrust", 2004. [Proceedings of the 13th international conference on World Wide Web]
- [4] Mengen Chen, Xiaoming Jin and Dou Shen, "Short Text Classification Improved by Learning Multi-Granularity Topics" *IJCAI*, 2011
- [5] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features", *Machine Learning: ECML-98* pp 137-142, 1998.
- [6] Mohd Suhairi, Md Suhaimin, Mohd Hanafi Ahmad, Hijazi Rayner Alfred and Frans Coenen, "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts", *8th International Conference on Information Technology (ICIT)*, 2017
- [7] Miriam J. Andrew J. Flanagan, Keren Eyal, Daisy R. Lemus & Robert M. McCann, "Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment", *Annals of the International Communication Association*, pp. 293-335, 2016.
- [8] A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", Springer, 2010.
- [9] H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan, "A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle", *DL.acm.org*, 2012.
- [10] S. M. Vohra and Prof. J. B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques" Volume – 02, pp. 313-317
- [11] Dwi H. Widyantoro, Yudi Wibisono "Modeling Credibility Assessment and Explanation for Tweets Based On Sentiment Analysis", *Jatit.Org*, 2019.
- [12] M. Abbasi and H. Liu, "Measuring User Credibility in Social Media", Springer, 2013.
- [13] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, Patrick Meier and Carlos Castillo, "TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter", *arXiv*, 2014.
- [14] L.Krzysztof, J. S Wiciński, M.J Lorek and A. Guptha, "Automated Credibility Assessment On Twitter," *Computer Science* 16 (2) pp. 157–168, 2015.
- [15] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2013, *Aclweb.org*
- [16] L. Rokach and O. Maimon, "Clustering Methods", *Data Mining and Knowledge Discovery Handbook* pp 321-352, 2005
- [17] Matt Thomas, Bo Pang, and Lillian Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts", *f EMNLP* 2006, pp. 327–335.
- [18] Gabriella Schoier, Giuseppe Borruso "A Clustering Method for Large Spatial Databases", Springer, 2004.
- [19] Weizhong Zhao, Huifang Ma, and Qing He, "Parallel K-Means Clustering Based on MapReduce", *CloudCom 2009: Cloud Computing*, pp 674-679

Identifying religious extremism-based threats in Sri Lanka using bilingual social media intelligence

Aneesha Fernando*

Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
fernando_im14006@stu.kln.ac.lk

Thareendra Keerthi Wijayasiriwardhane

Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
thareen@kln.ac.lk

Abstract: Religion is one's relation to what he or she regards as holy, sacred, spiritual, or worthy of especial reverence. Religious extremism is the advocacy of extreme measures over a religion whereas religious extremists are even willing to murder as they provide sanctions for violence in the service of God. Sri Lanka has a tragic history of religious and ethnic extremism and the Easter Sunday attack coordinated by a radical Islamic group that killed over 300 and injured another several hundred can be identified as the recent climax of these events. In this modern information age, it is evident that these radical extremist groups utilize social media for spreading their extreme ideologies due to its free and unregulated nature. If there were a mechanism to even slightly identify the possibility of tragic incidents like Easter Sunday bombing, the 300 souls who had to sacrifice their lives for an unreasonable cause would be still alive happily. In this research, we propose a predictive methodology for identifying any upcoming religious extremism-based threats in Sri Lanka using social media intelligence. We aim to specifically address Sri Lanka's multi-lingual culture by analyzing all the bilingual social media posts in Sinhala and Tamil languages. A hybrid sentiment analysis methodology consisting of a Machine Learning model and a sentiment lexicon was trained on carefully chosen labelled social media text data and each text was classified as either religious-extreme or not, using Naïve Bayes, SVM, and Random Forest algorithms. When comparing their results, we were able to achieve the best results with the Naïve Bayes algorithm resulting in an accuracy of 81% for Sinhala tweets while Random Forest algorithm resulted in an accuracy of 73% for Tamil tweets proving that social media intelligence can be used to predict religious extremism-based threats.

Keywords: Machine learning, Religious extremism, Social media, Text analytics

I. INTRODUCTION

Sri Lanka is a country that suffered 30 years of ethnic civil war fought between the Liberation Tigers of Tamil Eelam (LTTE) and the Sri Lankan government, from 1983 until it ended in 2009. The hope was to have peace for a long period without any more bloodshed but then, on the 21st of April 2019, 10 years after the end of the civil war, Sri Lanka experienced a series of coordinated bomb blasts targeting three main churches and international hotels, which killed over 300 and injured hundreds more, frightening all Sri Lankans. National Thowheeth Jama'ath (NTJ), a Sri Lankan radical Islamist group claimed the responsibility of the attack that appears to have characteristics in common with those of Al-Qaeda and has its roots bound to the Islamic State of Iraq and Syria (ISIS). This attack was not a singular attack which happened out of nowhere but was stimulated by a series of previous religious-based riots that happened in different parts of the country such as anti-Muslim riots of 2014 around

Beruwala, resulting in four dead and 80 injured citizens, anti-Muslim riots of 2018 in Digana, Kandy with attacks on the Muslim community and mosques. After suffering 30 years of civil war, Sri Lanka has become a victim of yet another form of intimidation; religious extremism that has been developed to a very dangerous level of religious extremism-based terrorism. Religious extremism is not an isolated problem faced only by Sri Lanka. According to the Global Shapers Survey conducted by the World Economic Forum (with 31,495 participants from 180 countries), religious conflicts are the fifth of the top ten problems the world is facing today [1]. Religious extremists endorse murder because they embrace theologies that provide sanctions for violence in the service of God. They have no sympathy towards the victims of their acts as they view them as enemies of their God [2]. Further, they are ready to sacrifice their own lives in expectation of a huge and immediate afterlife reward, in return for "martyrdom" [3]. The worst part of religious extremism is its convertibility to religious terrorism. Religious terrorism is a dangerous form of terrorism that is based on motivations and goals that may have predominant religious character and influence [4].

In the age of modern information, it is identified that these religious extremist groups have utilized modern communication techniques such as the Internet and social media, to directly communicate with their worldwide audience to harass users, recruit new members and incite violence [5]. Unlike in traditional communication methods, the free and unregulated nature of social media has helped extremists to easily form online communication to disseminate their belief and training materials without the fear of getting censored. Concerning the case of Easter Bombing coordinated by NTJ, the leader of the organization Zahran Hashim had been very active in social media, specifically on Facebook since 2017, spreading their radical ideology [6]. When the incident happened, extremists created a wave of false but flammable information and spread them through social media rapidly, provoking one religious group against another. Therefore, the Sri Lankan government had to take a drastic decision even to ban social media including Facebook, WhatsApp, Instagram, Snapchat and Viber during the time of conflict.

Banning social media after an incident has happened, is futile. The need to have a methodology to identify religious extremism and its trend on social media and users who spread their extreme ideologies through social media in a threatening way has become important today more than ever. In this paper, we, therefore, propose a domain-based hybrid sentiment analysis methodology for identifying any upcoming religious extremism-based threats in Sri Lanka using social media intelligence. The proposed methodology consists of two parts,

a Machine Learning model and a sentiment lexicon. Previous researches conducted in the Sri Lankan context on sentiment analysis have always focused on the Sinhala language only. Sri Lanka is a multi-religious country. 70 % of the population are Buddhists, 13 % are Hindus, 9.7 % are Muslims, 7.4 % are Christians while 0.04% fall in the “others” category. In terms of the languages prevalent in Sri Lanka, 75 % of the population with a Buddhist majority, speak Sinhala as their mother tongue while 21 % comprised mostly of Hindus and Muslims, use Tamil as their mother tongue. Another 4 % use other languages such as English and Malay. This shows the fact that there is a clear relationship between a person’s religion and his/her mother tongue in the Sri Lankan context. Conducting a study on a single language would, therefore, make the results biased towards a single religion. Hence, in our study, we propose a methodology that treats both Sinhala and Tamil languages similarly. In return, we get an unbiased result fair by Sri Lanka’s multi-religious culture.

On the other hand, most of the studies that introduce text classification for the Sinhala language have rarely considered the sentiment value of each word due to the unavailability of resources. When it comes to the domain of religious extremism, it is important to consider the sentiment value of words when classifying over classifying just through the term frequency of the words.

Considering the above facts, in this research, we have explored the viability of identifying religious extremism-based threats in Sri Lanka using social media intelligence through a text analytics model. Our methodology is specific to countries with multi-religious cultures in the domain of religious extremism. The accurate results we received for the Sri Lankan context have proved that this methodology could be further extended to any multi-religious country.

II. RELATED WORK

Most researchers have conducted studies to identify terrorism, Radicalization, Anti-Black, Jihad Extremism and hate speeches using social media data. The volume of content being posted on social media platforms makes it challenging for security analysts to discover such content manually or using keyword flagging. Further, textual posts on social media websites are user-generated content, which is unstructured and informal. Therefore, they come with lots of noisy content such as incorrect grammar, misspelled words, Internet slang, abbreviations and text containing multilingual scripts [7].

Alvari et al. [8] have proposed a detection scheme that could determine whether or not a given username belongs to an extremist user. They have used a dataset from twitter and first demonstrated that extremist users on twitter tend to adopt handles that follow similar patterns, in contrast to the normal users. Then a detection framework has been proposed to identify if a given Twitter handle belongs to an extremist giving its proximity to an existing set of extremist-related handles. They have compared different supervised and semi-supervised approaches using the features from the Twitter handle, profile information, and content which are highly indicative of online extremism. To further understand the significance of the features, they have conducted a significance analysis on the features using the labeled instances and have compared their results against char-LSTM which automatically extracts features.

Balahur and Turchi [9] have introduced a hybrid technique for sentiment analysis of Twitter texts. The research has used pre-processing to normalize the texts and the linguistic peculiarities of tweets have been taken into consideration. Spelling variants, slang, special punctuation, and sentiment-bearing words from the training data have been substituted by unique labels. For example, the sentence “I love car” was changed to “I like car”; according to the General Inquirer dictionary since love and like both have a positive sentiment. This approach could be used for various languages with minimal linguistic processing. The method does not require any further processing and uses tokenization only. They state that their final system should work similarly for all languages.

An interesting work [10] demonstrates an excellent methodology to identify the state of art and independent techniques for multilingual sentiment analysis. According to their work, sentiment analysis on one single language increases the risks of missing essential information in texts written in other languages. One of the main problems in multilingual sentiment analysis is a significant lack of resources. Thus, sentiment analysis in multiple languages is often addressed by transferring knowledge from resource-poor to resource-rich languages to overcome the resource unavailability problem. Wan [11] has researched sentiment analysis of Chinese reviews by translating Chinese text to English text using Google Translator and his results indicate that analysis of English reviews translated by Google Translator outperforms the analysis of original Chinese reviews.

Tan and Zhang [12] introduced an approach for sentiment classification for the Chinese language. First, POS tagging is used; the aim of using POS tagging is to parse and tag the Chinese text. After POS tagging, feature selection is used to determine discriminative terms for classification. Finally, a machine learning approach is used for sentiment classification. Their method proves that POS tagging the text in the original language before translating reduces the error of translation and increases the accuracy of the model.

Hung et al. [13] have conducted research to detect radicalization trajectories using Graph Pattern Matching Algorithms. This research has taken a different approach to track an individual’s behavioral indicators of homegrown extremism, using public and law enforcement data. The intuition behind the research is to use graph pattern matching to identify suspicious trajectories and potential radicalization over a dynamic heterogeneous graph associated with the fused data from public and law enforcement.

In 2018, a group of Sri Lankan researchers has conducted recent research to identify racist social media comments in Sinhala language using text analytics models with machine learning. They have shown that simple keyword spotting techniques cannot be used to identify the exact intent of a comment accurately and therefore, have built a Text Analytical Model with a Two-Class Support Vector Machine using Microsoft Azure Machine Learning Studio. In this research, they have employed the n-grams, carefully chosen and extracted from Facebook comments [14].

The survey research done by Agarwal and Sureka [7] reveals a variety of information on retrieval and machine-learning-based methods and techniques used by researchers to investigate solutions for social media extremism and online

radicalization detection. Clustering, Logistic Regression, and Dynamic Query Expansion are commonly used techniques to predict upcoming events related to social media extremism. They have observed that Named Entity Recognition (NER) is a common component in the text processing pipeline under various proposed approaches and techniques. Graph modeling is also a technique adopted by several researchers to perform event forecasting. Further, they have revealed that KNN (K Nearest Neighbor), Naïve Bayes, Support Vector Machine, Random Forest Algorithm, Rule-Based Classifier, Decision Tree, Clustering (Blog Spider), Exploratory Data Analysis (EDA), Topical Crawler/Link Analysis (Breadth-First Search, Depth First Search, Best First Search) and Keyword Based Flagging (KBF) are the most widely used techniques for online radicalization detection on social media. Their research has summarized that 90% of the studies have been successful in mining English Language text. Further, 60% of the studies target events specific to a country or region.

III. METHODOLOGY

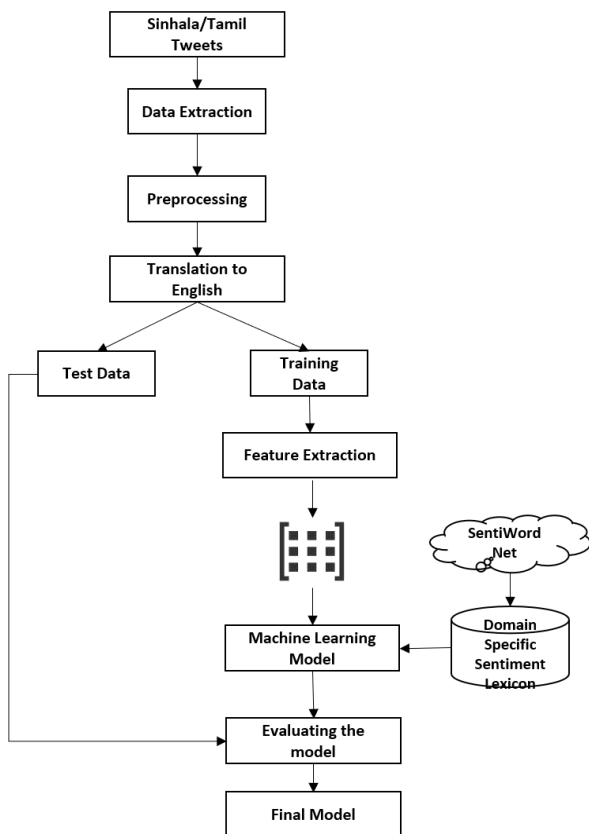


Fig 1. Research design

The proposed methodology follows a series of steps for text classification, comparing different Machine Learning algorithms for identifying the best model out of them. We use both Sinhala and Tamil tweets as inputs for our model and then evaluate the sentiment against the religious extremism of each where we finally classify them as either religious extreme or not. Figure 1 demonstrates a summary of the steps that we have performed throughout our study.

A. Data preparation

The data set was collected from Twitter using Twitter API by searching and identifying tweets based on the keywords and hashtags which contain extremism. Initially, 300 recent tweets; 200 in Sinhala and 100 in Tamil were collected based on the keywords shown in Table I. These keywords were selected by analyzing the frequency of their usage in identified extreme articles and posts on social media and with the help of language experts. Then, the extracted tweets were annotated as religious- extreme or non-religious- extreme as perceived by experts in both languages. This annotation was performed very carefully with the support of two language experts for each language. The annotations of the first expert were confirmed by the second expert to reduce the risk of misinterpretation. Though the data set collected was comparatively small, the perfect annotation increased the value of the dataset. Table II presents an example of how data were annotated.

TABLE I. KEYWORDS

Language	Keyword
Sinhala	හලාල්, කල්ලනෝනි, තමිබ්, හමිබය, තමිබියො, හමිබයෝ, තමිබියෝ, හමිබයො, බුද්ධ, බුද්ධාගම, බුද්ධාමුදුරුවො, පන්සල්, භාමුදුරුවෝ, මරමු, හින්දුන්, හින්දු, කෝවිල, කතෝලික, පල්ලි, පල්ලිය, සාමය, සමගිය, ආගම්, ශ්‍රීලාංකික, සමානාත්මතාවය, සහජීවනය, ආගමික, අන්තවාදය, වහබ්වාදය, අන්තවාදීන්
Tamil	தொப்பிபிரட்டி, சோனி, பாவாடை, தோட்டக்காட்டான், சக்கிலியன், பறையன், கரையான், செம்பு தூக்கி, இனவாதம், மதவெறி, மதக் கலவரம், சமாதானம், மத, தீவிரவாதம், கோவில், புத்த மதம், கிரிஸ்துவர், முஸ்லீம்

TABLE II. LABELLED TWEETS

Tweet (Sinhala)	Label
බෞද්ධයෝ බුද්ධාගමේ විරස්තීතිය වෙනුවෙන් හමිබයෝ මරන්න, පල්ලි කඩන්න, හමිබකඩ ගිනිතියන්න	Religious Extreme
මම බෞද්ධ මනුෂ්‍යයෙක්. මට අඹමල් රේඝුවක තරමින්වත් ජාතිවාදී ආගම්වාදී වෙන්න බෑ. අන්තිම හුස්ම දක්වාම සාමය, සහජීවනය සංහිදියාව විතරයි	Non-Religious Extreme

Out of a total of 300 tweets that were extracted, 141 were annotated as religious extreme and 159 were annotated as non-religious extreme. As identified in the study by Dias et al. [14] having more non-religious extreme tweets than religious extreme tweets makes sure that our model will minimize the false positives in prediction.

B. Pre-processing of data

The extracted and annotated data were then subject to preprocessing. Data preprocessing is an important process in this study as bilingual text data is used. Though the analysis was done on English language text, the input to the model can belong to either Sinhala or Tamil languages. Preprocessing

techniques for each language are different from each other as different languages have different morphological bases.

As suggested in the literature, preprocessing the text in the primary language before translating it to the English language increases the accuracy of the study and minimizes the negative effect on the performance of the classification, since the translation of words might not preserve the semantic orientation due to the differences between languages.

In the first phase, we performed the general preprocessing steps that apply to both languages such as removing URLs, hashtags, duplicated tweets, empty tweets and emoticon only tweets, numbers, punctuations and replacing all the emoticons with their sentiment. In the second phase, we identified the language of the text and performed preprocessing steps that are unique to the identified language. Part of Speech (POS) tagging, Tokenization and stop word removal were committed on each language. Due to the lack of language literacy, the preprocessing of Tamil language was done only using existing tools such as IndicNLP and RippleTagger. POS tagger for the Sinhala language was developed based on the assumptions of existing studies using 1300 words Sinhala wordnet developed by the University of Colombo School of Computing (UCSC)[15]. The top 30 most common words identified through the work of [16] are considered as stop words and removed from each tweet. Tokenization and POS tagging on original language were committed to preserving the semantic orientation of the context by reducing the error of translation.

In the third phase, we translated text in non-English languages; Sinhala and Tamil, to the English language to get them into a common platform for reliable and unbiased results. It also enabled the employment of SentiWordNet which is an essential and rich lexicon resource when it comes to sentiment analysis. Here, the machine translation into English was performed on tokenized and POS tagged, stop words removed – tweets, using Google Translator which is identified in the literature as an accurate tool [17], [18], [11]. By taking a randomly selected 50% from the preprocessed and translated tweets of each language to form a sample, the accuracy of the translation was calculated with the support of the language experts. The translation of Sinhala tweets had

TABLE IV. NEED FOR SENTIMENT VALUES

Original Text	Translated Tokenized Text
බෞද්ධයෝ බුද්ධාගමේ විරජ්ජිතිය වෙනුවෙන් හම්බයෝ මරන්න, පල්ලි කඩන්න, හම්බ කඩ ගිනිනියන්න	['bauddhayo', 'kill', 'hambayo', 'buddhism', 'stability', 'church', 'break', 'earned', 'burn']

76% accuracy while the translation of Tamil tweets had 70% accuracy.

The words or phrases that could not be translated directly from Google Translator were kept in its transliteration form; Singlish (Sinhala written in English alphabet) or Tanglish (Tamil written in English alphabet) format, hence, what we received at the end of the preprocessing was not pure English. Although that feature added more complexity to our research problem, it contributed to the successful identification of the sentiment of the tweets in both languages. To handle this mixed language context and then to accurately classify the data, we have introduced a novel approach.

C. Feature extraction

To align with the process of using text in machine learning algorithms, we had to convert them to a numerical representation. The preprocessed dataset has many distinctive properties. In the feature extraction phase, we extracted the aspects from the preprocessed dataset into feature vectors.

Later, these aspects were used to compute the positive and negative polarity in a sentence towards religious extremism which is useful for determining the sentiment of each tweet. For experimentation purposes, we conducted our study on two feature extraction methods; Count Vectors as features and Term Frequency-Inverse Document Frequency (TF-IDF) as features. We further extracted TF-IDF feature vectors in word level (unigram) and N-gram level (Bi-gram and Tri-gram). As a result of the count vector feature extraction, we received a matrix that includes the term frequency of each word in each tokenized tweet. Unigram level TF-IDF returned a matrix representing the TF-IDF scores of every term in each tweet and N-gram level TF-IDF matrix representing the TF-IDF scores of N-grams. Table III summarizes the measures of each matrix received from data extraction.

At the end of the feature extraction phase, the preprocessed dataset was converted to its numerical representation for accurate text analysis. Feature names received from each feature extraction method were stored into a vocabulary.

D. Text classification model building and training

Text Classification models based on Sentiment analysis can be classified into three categories as corpus-based approaches, lexicon-based approaches, and hybrid approaches. Corpus-based methods use labelled data; lexicon-based methods rely on lexicons and optionally on unlabeled data; and hybrid methods are based on both labelled data and lexicons, optionally with unlabeled data. A sentiment lexicon is a collection of known sentiment terms with their positive and negative polarity values [10]. The objective of our study is to classify tweets according to their degree of religious extremism. Religious extremism is a domain that has a high

TABLE III. FEATURE EXTRACTION

Language	Feature Extraction method	Number of Features Extracted	Min word length	Max word length
Sinhala	Count Vector	987	1	1
	Word level TF-IDF	991	1	1
	N-gram level TF-IDF	2658	2	3
Tamil	Count Vector	382	1	1
	Word level TF-IDF	387	1	1
	N-gram level TF-IDF	1026	2	3

sensitivity to sentiment value. Considering only the feature vectors and their term frequencies cannot correctly identify the sentiment value of each term. Having the highest frequency does not confirm that a term has the highest sentiment. Let us take Table III as an example.

TABLE V. ASSIGNING POLARITY VALUES

Text	Sentiment
තමිනි මරන්න සනවනන් මටත් කෝල් එකක් දියන්	Religious Extremist
නසිය කිව්වත් ඹ නානා කිව්වත් උ තමිනි කිව්වත් උ කිව්වේ ආදරයට	Non-Religious Extremist

The words ‘kill’, ‘break’ and ‘burn’ contain the highest negative polarity score though their Term Frequency or TF-IDF is less. This proves that a sentiment lexicon is essential for the validity and success of this kind of study. Therefore, a hybrid sentiment analysis model was chosen for building the text analysis model along with Machine Learning. SentiWordNet was used as the initial lexicon for the study to extract the sentiment values of each word due to its free availability for research purposes and continually updating nature.

As mentioned above, the language of the translated tokens of the tweets was not purely English; they had Singlish and Tanglish words along with pure English words. The words that were not in pure English could not be found in the SentiWordNet, but the sentiment values of those words cannot be ignored. (In the above example ‘bauddhayo’, ‘hambayo’ are in its Singlish format). However, when it comes to the domain of religious extremism, these two words definitely cannot be ignored as ‘bauddhayo’ implies ‘Buddhists’ and ‘hambayo’ implies ‘Muslims’. To address this need, a context-based algorithm was developed using the pragmatic meaning of each tweet. The polarity value of each word was extracted from the sentiment value of the labelled tweet in which the particular word is included. Then along with the training of the model, the polarity of the particular word was adjusted. As an example, the word ‘thambi’, used to address Muslims, was included in most of the tweets which spread hate towards Muslims, but it is also included in some positive tweets. Therefore, the final negative polarity value identified for the

word ‘thambi’ is 0.57 which implies that the word ‘thambi’ does not imply that the tweet containing that word is always religious extreme. Table V provides a good example of this case.

A Singlish lexicon consists of 389 words and a Tanglish lexicon consists of 137 words that were constructed using the developed algorithm with the calculated negative and positive polarity values. Then using SentiWordNet and the lexicons constructed, the weight of each word or N-gram in the vocabularies built during the feature extraction was calculated. By multiplying each feature vector matrix with their vocabulary’s weighted value array, the final numerical representation of the dataset was created as follows.

Matrix 1 = Array of weights of each term x Count Vector

Matrix 2 = Array of weights of each Unigram x TF-IDF Word Level

Matrix 3 = Array of weights of each N-gram x TF-IDF N-gram level

At the end of these processes, all the text data were converted to a numerical representation that represents the feature vectors and their sentiment values. These matrices were then inserted into the machine learning model for classification as a religious extremity or not as a religious extremity.

For experimentation purposes, several classification algorithms were selected from literature to perform text analysis and to identify the most suitable classification algorithm for our study. The algorithms selected were Naïve Bayes Classifier, Support Vector Machine (SVM) classifier, Logistic Regression, Random Forest Algorithm, Xtreme Gradient Boosting Model, Linear Classifier and Recurrent Neural Network (LSTM). Out of these seven classifiers, the present study used the classifiers which had achieved more than 70% accuracy, for further analysis to identify the most suitable algorithm to recognize religious extremism-based threats. The algorithms which had an accuracy of more than 70% were Naïve Bayes Classifier, Support Vector Machine (SVM) classifier and Random Forest Algorithm. The success of each algorithm was evaluated using 10-fold cross-validation which has been identified to be suitable for a small dataset rather than using train/test split, to reduce the risk of overfitting and to assess any such occurrence.

IV. RESULTS AND DISCUSSION

The three feature matrices generated were used with Naïve Bayes classifier, SVM Classifier and the Random Forest Classifier for both Sinhala and Tamil language tweets. After the model was trained on training data, it was tested using the test dataset. The results were significant and deducible. Tables VI, VII, VIII below summarizes the test results of the trained machine learning models.

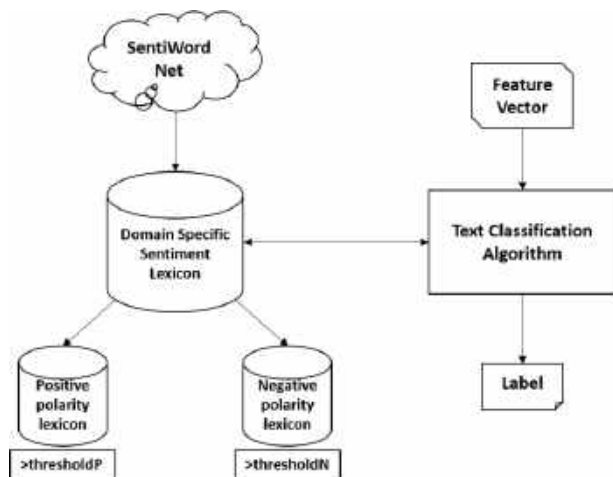


Fig 2. Text Classification Model

TABLE I. COMPARISON OF RESULTS FROM DIFFERENT CLASSIFICATION ALGORITHMS WITH MATRIX 1

Sinhala				
Algorithm	Accuracy	Precision	Recall	F1 -Score
SVM	0.52	0.52	1.0	0.68
Naïve Bayes	0.81	0.72	0.96	0.82
Random Forest	0.76	0.71	0.92	0.80
SVM	0.67	0.67	1.0	0.80
Naïve Bayes	0.69	0.68	1.0	0.81
Random Forest	0.73	0.72	1.0	0.84

TABLE VIII. COMPARISON OF RESULTS FROM DIFFERENT CLASSIFICATION ALGORITHMS WITH MATRIX 2

Sinhala				
Algorithm	Accuracy	Precision	Recall	F1 -Score
SVM	0.55	0.67	1	0.8
Naïve Bayes	0.60	0.59	0.95	0.73
Random Forest	0.79	0.74	1	0.85
Tamil				
SVM	0.57	0.58	1	0.73
Naïve Bayes	0.63	0.7	1	0.82
Random Forest	0.7	0.61	0.87	0.72

TABLE VIII. COMPARISON OF RESULTS FROM DIFFERENT CLASSIFICATION ALGORITHMS WITH MATRIX 3

Sinhala				
Algorithm	Accuracy	Precision	Recall	F1 -Score
SVM	0.54	0.55	1.0	0.71
Naïve Bayes	0.73	0.65	1.0	0.79
Random Forest	0.53	0.51	1.0	0.68
Tamil				
SVM	0.62	0.62	0.69	0.65
Naïve Bayes	0.64	0.60	0.86	0.71
Random Forest	0.74	1.0	0.50	0.67

As summarized in Table VI for Matrix 1, Naïve Bayes algorithm shows the highest accuracy, precision, recall and F1 score for Sinhala language tweets while Random Forest algorithm gives the highest accuracy, precision and recall for Tamil language tweets. Table VII explains how Matrix 2

performed with each Classification algorithm. According to the results obtained, the Random Forest algorithm provides the best results for both Sinhala and Tamil languages.

According to Table VIII for matrix 3, the Naïve Bayes algorithm performed well for Sinhala language tweets while the Random forest algorithm gave the best results for Tamil language tweets. By considering all these results, we were able to finalize the best text analysis model for each language.

For Sinhala language tweets, the Naïve Bayes Classification algorithm gives the best results with Count Vectors as a feature extraction method. This model achieves an accuracy of 81% that can be considered as the highest score when it comes to the classification of data in a sensitive domain like religious extremism. It also shows a precision of 0.72, recall of 0.95 and F1 Score of 0.82. Having a recall of 0.96 means that our model returns most of the relevant results. It is least likely to classify a religious extreme tweet as a non-religious extreme tweet. Having a good precision value like 0.72 proves that the developed model returns substantially more relevant results than irrelevant results. Having a high recall value and a good precision value gives a good F1 value which is an excellent measure for evaluating the model better than the accuracy. It gives the weighted average of precision and recalls enabling us to understand the success of our model and how it performs by giving expected results. By evaluating and examining all 4 measures, we can conclude that the model created with the Naïve Bayes Classification algorithm and Count Vectors as features is the best in identifying religious extremism tweets in the Sinhala language.

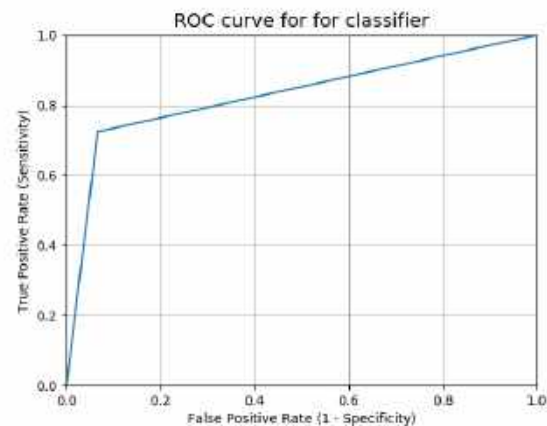


Fig 3. ROC curve of the best model for Sinhala

The Receiver Operating Characteristics Curve (ROC) of the model created using the Naïve Bayes algorithm and Matrix 1 for the Sinhala language is shown in Fig. 3. The curve has spread closer to the top left corner of the plot, depicting the better performance of our model. The calculated Area Under Curve (AUC) is 0.83 and it can be considered as a high score to indicate that our model is excellent in distinguishing between classes; religious extreme, and non-religious extreme.

For the Tamil language, the best working model is the Random Forest Algorithm with Matrix 1 which has achieved 0.73 highest accuracies along with the precision of 0.72, recall of 1.0 and F1 Score of 0.84. Having a recall of 1.0 indicates that our model always returns relevant results only.

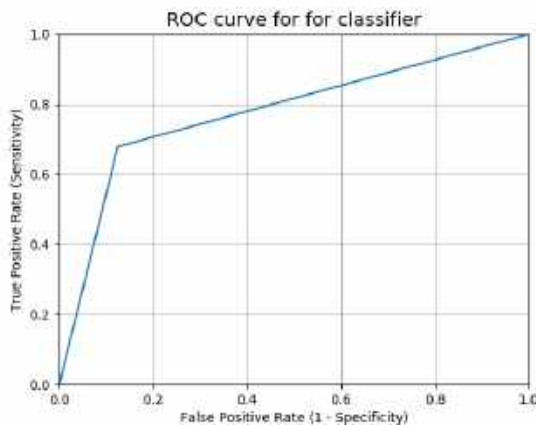


Fig 4. ROC curve of the best model for Tamil

The highest F1 Score of 0.84 gives an idea of how precise and robust the model developed is. Fig. 4. depicts the ROC Curve of the identified best performing model above for Tamil language tweets demonstrating the tradeoff between sensitivity and specificity. As the curve has spread closer to the left-hand border and then to the top border of the ROC space, it proves that our model has generated more accurate test results. The calculated AUC value of the curve is 0.78 and it can also be identified as a good score indicating that our model behaves well in classifying data between classes; religious extreme, and non-religious extreme.

Results received for both Sinhala and Tamil languages depict that our model is excellent in identifying religious extreme content in social media. Therefore, the trained and tested model can be used to classify a live stream of Sinhala and Tamil tweets to see if there is any trend of religious extremism that is nurtured through social media platforms. By frequent monitoring of social media data through this model, upcoming religious extremism-based threats in Sri Lanka could be predicted accurately, also focusing on Sri Lanka's multi-religious culture.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a model that can be used to identify religious extremism using social media intelligence. As was described in the literature review, previous studies on sentiment analysis have barely focused on the domain of religious extremism though they have covered areas like racism, terrorism, and radicalization on social media platforms. Those domains can be easily differentiated from the domain of religious extremism since religion is a more sensitive area that is based on someone's faith and belief for a journey of self-cultivation and self-realization. In this stage, we have considered the sentiment value of each word and term without relying only on the frequency of each word or term in the data corpus. Methods used in other similar domains like racism and radicalization were not apt to be used here. Therefore, we have introduced a novel method of identifying religious extremism by engaging a context-based algorithm that addresses the pragmatic meaning of each tweet. On the other hand, we have developed this model to match with the Sri Lankan context by enabling the analysis of both Sinhala and Tamil language social media posts and it has achieved successful results for both languages. None of the sentiment

analysis researches in Sri Lanka have addressed Sri Lanka's multilingual culture and most of them have only focused on Sinhala language data. As explained, there is a clear relationship between a person's religion and his/her mother tongue in the Sri Lankan context. Therefore, conducting a study on a single language would make the results more biased to a single religion. Thus, in the present study, we have proposed a methodology that treats both Sinhala and Tamil languages equally. In return, we get unbiased results that are fair by Sri Lanka's multi-religious culture. Using the results, we have identified that Naïve Bayes Classifier is the best algorithm for the Sinhala language with an accuracy of 81% while Random Forest Classifier is the best algorithm for the Tamil language with an accuracy of 73% in identifying the religious extremism using social media intelligence. For future work, we expect to extend this analysis with more social media data from other platforms like Facebook and YouTube. Doing so, we believe that we will be able to get more data on the Tamil language for better accuracy. Additionally, we are hoping to interact more with language experts in both languages to increase the accuracy of the model by understanding the special properties related to each language. We expect to generalize our model further, to analyze religious extremism in other multireligious, multilingual countries.

REFERENCES

- [1] T. L. Jackson, Abby, The 10 most critical problems in the world, according to millennials. Business Insider. Available: <https://www.businessinsider.com/world-economic-forum-world-biggest-problems-concerning-millennials-2016-8> [Accessed: 16-Dec-2019].
- [2] L. R. Iannaccone and E. Berman, "Religious extremism: The good, the bad, and the deadly," Public Choice, vol. 128, no. 1, pp. 109–129, Jul. 2006.
- [3] B. Caplan, "Terrorism: The relevance of the rational choice model," Public Choice, vol. 128, no. 1, pp. 91–107, Jul. 2006.
- [4] Brannan, David. "Understanding Terrorism: A Social Science View on Terrorism – CHDS Self-Study Courses." Available: <https://www.chds.us/selfstudy/courses/understanding-terrorism/>. [Accessed: 16-Dec-2019].
- [5] J. S. closeJoanna S. correspondent covering S. AsiaEmailBioBioFollowFollow, "The remote Sri Lankan enclave that produced the mastermind of a massacre," Washington Post. Available: https://www.washingtonpost.com/world/asia_pacific/the-remote-sri-lankan-enclave-that-produced-the-mastermind-of-a-massacre/2019/04/26/9358f48c-6830-11e9-a698-2a8f808c9cfb_story.html. [Accessed: 16-Dec-2019].
- [6] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in Social Informatics, Cham, 2016, pp. 22–39.
- [7] S. Agarwal and A. Sureka, "Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats," arXiv:1511.06858 [cs], Nov. 2015.
- [8] Alvari, Hamidreza, S. Sarkar, and P. Shakarian. "Detection of Violent Extremists in Social Media." 2019 2nd International Conference on Data Intelligence and Security (ICDIS), 2019, pp. 43–47.
- [9] A. Balahur and M. Turchi, "Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data," in Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, 2013, pp. 49–55.
- [10] K. Dashipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cogn Comput, vol. 8, no. 4, pp. 757–771, Aug. 2016.
- [11] X. Wan. "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis". In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics; 2008, pp. 553–61.

- [12] Tan, S. and Zhang, J., “*An empirical study of sentiment analysis for Chinese documents*. *Expert Systems with Applications*”, vol. 34, no. 4, pp. 2622-2629, 2008
- [13] B. W. K. Hung, A. P. Jayasumana, and V. W. Bandara, “*Detecting radicalization trajectories using graph pattern matching algorithms*,” 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 313–315, 2016.
- [14] D. S. Dias, M. D. Welikala, and N. G. J. Dias, “*Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning*,” in 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), 2018, pp. 1–6.
- [15] “Lexical Resources – Local Language Portal.” [Online]. Available: <https://www.language.lk/en/resources/lexical-resources/>. [Accessed: 16-Feb-2020].
- [16] S. Gallege, Analysis of Sinhala Using Natural Language Processing Techniques. CiteSeerX. 2010.
- [17] Remus, Robert, Uwe Quasthoff, and Gerhard Heyer. “SentiWS - A Publicly Available German-Language Resource for Sentiment Analysis.” In LREC, 2010.
- [18] H. Ghorbel and D. Jacot, “Sentiment Analysis of French Movie Reviews,” in Advances in Distributed Agent-Based Retrieval Tools, V. Pallotta, A. Soro, and E. Vargiu, Eds. Berlin, Heidelberg: Springer, 2011, pp. 97–108.

Affective gaming in real-time emotion detection and music emotion recognition: Implementation approach with electroencephalogram

Pradeep Kalansooriya*

Department of Computer Science
General Sir John Kotelawala Defence
University, Sri Lanka
lpkalan@gmail.com

G. A. D Ganepola

Department of Information Technology
General Sir John Kotelawala Defence
University, Sri Lanka
dasuniganepola@gmail.com

T. S. Thalagala

Department of Computer Science
General Sir John Kotelawala Defence
University, Sri Lanka
tharushishehara96@gmail.com

Abstract: Affective Gaming can be considered as the concept of detecting the real-time emotional state of a player during various stages in gameplay and then enhancing the user interactivity accordingly to the emotional state. Based on this conception, this paper presents the research phase of the development of an Affective Car Racing computer game. The designs were created based on the theory of “Affective Loop” in games. Affective Loop consists of Emotion Elicitation, Emotion Detection/Modelling and finally Emotion Expression by Game Engine. This paper considers the second and third subphases of this loop. Designs are done for these two phases based on technologies that are still not been utilized by many game developers when designing a game. Emotion Detection/Modelling phase is introduced with a technique of capturing Electroencephalography (EEG) signals for predicting the real-time emotion of the player while interacting with the game engine. Emotion Expression phase considers the concept of Music Emotion Recognition (MER), which is a novel concept for the Gaming Industry. The authors had trained SVM models for emotion modeling via EEG Signals that will be captured by the Emotiv Epoc 14 channel device. The authors had classified the Rock and Electronic genre of music via Multi-Label RAKEL classification (Precision score of 75%) to play music excerpts based on the effect of the gamer during gameplay.

Keywords: Affective gaming, Affective loop Concept, Emotion detection through electroencephalography, Music emotion recognition

I. INTRODUCTION

The term “affect” in psychology is defined as a feeling or an emotion expressed for some event [1]. From a psychology perspective, computer games consist of higher amounts of “affect” because computer game environments are built to change according to user interactivity. The interactions occur when the game environment receives inputs from the player’s side by executing actions via the game controller.

“Affective” gaming can be considered as detecting the real-time emotional state of a player during various stages in gameplay and then enhancing the user interactivity accordingly to the emotional state. Methods of widely accepted emotional change detections are done by monitoring the changes in facial expressions, gestures, body posture, speech, physiological responses such as EEG signals etc. during user interactivity within the game environment [2].

Examples of such games include arousal-driven play of non-player characters (NPCs) in Left 4 Dead 2 the combat skills of opposition NPCs in F.E.A.R., the emotional expression of avatars’ in The Sims series, game characters’ emotional expressions in Prom Week and Façade, Storybricks, the game narrative building influenced by player emotions etc. [3].

The basic principle of affective gaming lies within the communication between human to human, human to a computer and human to human through a computer. According to Sundstrom, the design of affective loops in computer games makes possible all mentioned forms of communication. The affective loop treats emotions as processes. The affective loop in gaming can be viewed as a closed loop with three sequential key phases.

- 1) the player reveals his feelings during interaction with a game mainly via bio-signals,
- 2) Biosignals of the player are detected by the game engine and interprets those signals into the emotion experienced,
- 3) based on that emotion recognition of the player, the game adjusts itself following the player’s emotions.

This is an on-going cycle, affecting the player’s both mind and body, making him respond through game actions and emotional reactions [4].

At stage 1 of the loop, the player expresses his emotions when he reacts to the game content. Game content is a collection of game mechanics, game environment and background music content. Game mechanics are the methods and rules specific for any game type that makes the game player interact with the game environment [5]. The game environment consists of visual settings such as graphics, background lighting levels, saturation levels etc. Background music content includes music clips played in the background [6].

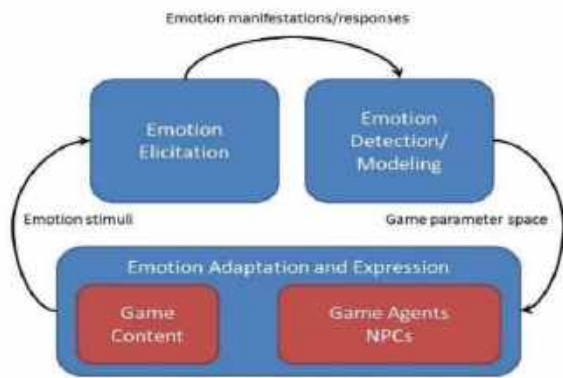


Fig. 1. Affective game loop [4]

This research paper is based on this Affective Loop. The researchers focus on introducing novel technologies to implement the above loop.

For the Emotion Elicitation phase, the authors focus on real-time capturing Electroencephalogram (EEG) signals and Modelling via Machine Learning approaches.

For the phase of “Emotion Adaptation and Expression,” they focus on using background music content to express the desired emotion. Music will be played accordingly to the player’s emotion detected via EEG Signals. Hence, music that will enhance the effect of the player will be played correspondingly. To identify music excerpts that enhance specific emotions, theories of Music Emotion Recognition (MER) were practiced by the authors.

This research paper aims to provide the general public the opportunities of the advancements of Research in Technology. At Present, products/services created from concepts of “Affective Loop”, MER are at limited numbers and a survey was performed by the authors to observe the knowledge of these concepts among the general public. 18% in total of the participants have had only heard the names of the above ideas. Hence, this suggests that these concepts are yet to be established among the general public. The authors focus on making aware of these novel concepts in the day-to-day living of the general public.

The objective of this research is to develop an integrated software prototype for Computer games where the player’s emotions are detected in real-time and the software relays back to the game engine with music excerpts to be played in the background in accordance to the emotions expressed by the player. The reason for this development is that the authors’ outlook to establish the concepts described above in the general public. To the best of their knowledge, the effective approach is developing a computer game integrated with the above concepts because computer games are highly popular among the public at present.

II. SCOPE OF THE RESEARCH

The authors primarily designed this prototype to “Car Racing” Game Genre because it is the game genre that gamers prefer to listen to music when playing. The scope of the entire research was narrowed down under the selected game genre. Fig.2 shows the process of defining the scope in tiers.

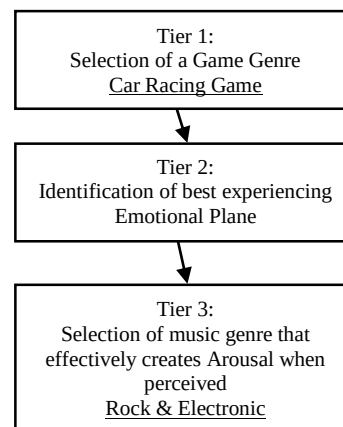


Fig. 2. Process of narrowing of the research scope

Tier 2 - During the Literature study, it was identified that Car Racing Games increase players’ risk-taking inclinations, known as “The Racing Game Effect” [7]. The Players are in an aroused state of emotion during the entire gameplay. The authors were able to analyze the main 4 emotion labels based on the theory of “Racing-Game Effect” followed by a survey performed on Car racing computer game players as a proof of concept. They identified emotion labels as “joyful activation”, “sadness”, “tension” and “power”. They plotted the intensity of the above emotion labels in a single-dimensional arousal plane.

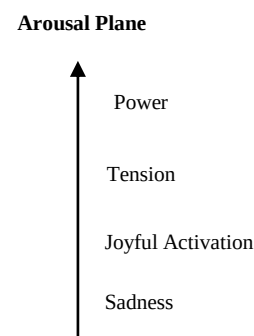


Fig. 3. Emotion taxonomy for this research

Tier 3 - With the survey, the authors inquired from the racing computer gamers’ audience their preferred music genre during gameplay, 88% responded that they prefer “rock” and “electronic” music genres. Hence, the above genres were selected to perform MER concepts to identify the music excerpts that make gamers perceive the desired emotion labels, the authors had identified in Tier 2.

This paper henceforth discusses research findings and implementation separately of Emotion Detection/ Modelling via EEG Signals in Part A and Emotion Expression using background music content in Part B.

III. DESIGN

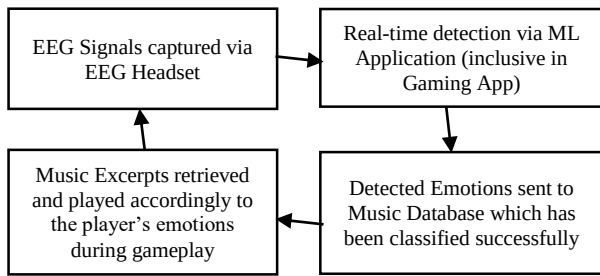


Fig. 4. Design workflow of an affective game during gameplay

IV. RESEARCH FINDINGS AND IMPLEMENTATION

A. An emotion detection/ Modeling phase

In this paper, the researchers are utilizing EEG signals as physiological responses to recognize human emotions according to the constructed emotion taxonomy. The reason behind this selection is that although detecting human emotions based on facial, behavioral, text and speech are widely accepted, analyzing facial expressions, text, speech, gestures or behaviors can be proven to be not very effective at times since these can be consciously altered. Due to this reason, the researchers have diverted their focus to studying emotions through physiological signals like electrocardiogram (ECG), electroencephalogram (EEG) [8], blood volume pressure (BVP), Heart-rate variability (HRV), temperature (TEM) etc. According to previous works, EEG signals have become more interesting for the researchers of this area since it comes directly from the human brain and therefore, the changes in EEG signals will directly reflect the changes in human emotional states. Design Methodology of this phase is as follows:

The classification model is developed using the Mahnob-HCI tagging database [9] (which contains classified emotions in terms of arousal based on EEG, ECG, respiration rate, temperature, GSR and eye-gaze.) In the context of this research, only the EEG signals are considered. The following diagram shows the design steps.

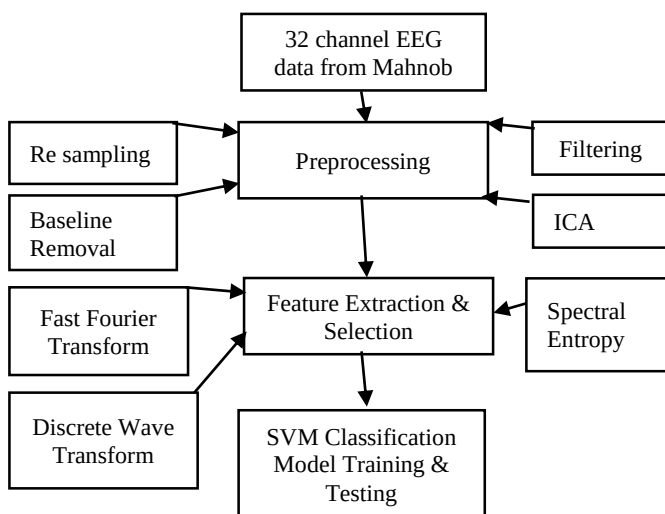


Fig. 5. Design methodology for emotion detection & modelling phase

The acquired EEG signals contain noise and artifacts. Therefore, the signals are first pre-processed.

The MATLAB EEGLAB Toolbox is used to pre-process the signals. The signals obtained from the Mahnob database are in the BDF (Bio-semi data file) format. These can be read and plotted through the EEGLAB Toolbox easily. The signal pre-processing involves the following steps.

- Resampling – The signals are resampled at 128Hz. (originally its 256 Hz)
- Filtering – A bandpass filter (2-40Hz) is used so that noise removal and artifact removal can be achieved at once.
- ICA (Independent component analysis will be performed.

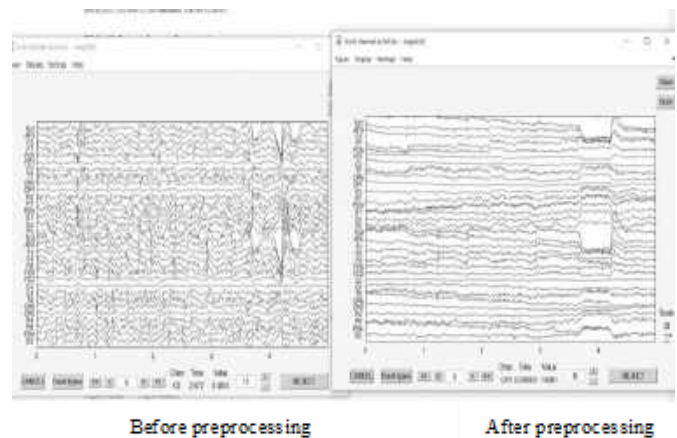


Fig. 6. Pre-processing

EEG feature extraction is quite a complex task which is also the reason why most researchers tend to turn away from BCI research. EEG signals are non-linear, nonstationary, and random in nature and therefore the feature extraction process becomes challenging. The features extracted in this research depend on statistical techniques, time-frequency based on Fast Fourier Transform (FFT), max-min features in temporal sequences, Shannon entropy, log-covariance etc.

The Hjorth parameters are extracted as well. These parameters characterize the EEG signals in terms of amplitude, timescale and complexity. The parameters are *activity*, *mobility* and *complexity*. They can be mathematically computed as shown.

$$h_1 = \sigma_x^2 \quad (1)$$

$$h_2 = \frac{\sigma_d}{\sigma_x} \quad (2)$$

$$h_3 = \frac{\sigma_{dd}}{\sigma_d} \bigg/ \frac{\sigma_d}{\sigma_x} = \sigma_{dd} / \sigma_x \quad (3)$$

Furthermore, the cross-correlation, peak frequency and the ratio of band power at a given time to baseline band power are considered. Around 2000 features are extracted originally, but before classification, several essential features

must be selected. The *mrmr algorithm* is used for feature selection since it involves maximizing relevance and minimizing redundancy. This algorithm provides a ready to use implementation for MATLAB making it more convenient and therefore was used in this research.

After mapping the obtained features into the arousal scale, linear SVM (Support Vector Machine) classifier is used for classification. An SVM classifier is used since it can separate highly dimensional data. The Classification Learner extension for MATLAB can be used for this purpose.

The data is plotted as illustrated in Fig.7. Several models must be trained to achieve considerable accuracy. Currently, the authors are yet to evaluate accuracy metrics.

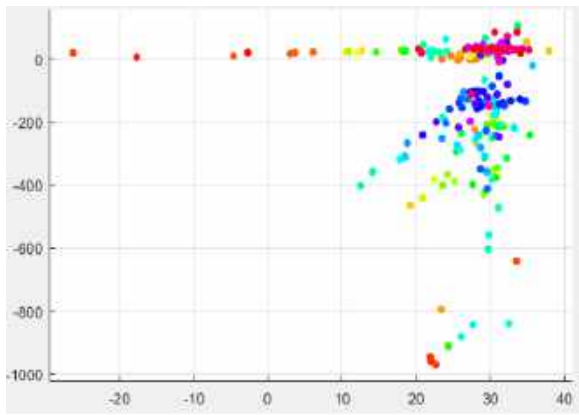


Fig. 7. EEG emotion plot

B. Emotion expression using background music content

At present, game mechanics are maximumly being utilized for the emotional awakening of the game player. The authors' observation of different brands of computer car racing games were that even though background music content is used, the background music is static throughout the gameplay. The background music content is not made to impact the player's emotional awakening.

It is in fact, a surprise that music content is not being used for emotion awakening when many music psychologists have proven that music does create an impact on human emotions [9],[10]. Keynote researcher Juslin had defined a research area where he finds relationships between music acoustic features and perceived music emotions. This area is known as Music Emotion Recognition (MER). MER is a cross-disciplinary field of auditory perception, psychology and music theory. With technology mixing in, MER is now a field consisting of signal processing and machine learning [11]. MER is yet to be utilized in the day-to-day lives of the general public. Until present, the only application developed for the use of the public with MER technology is a Music Player [12].

Reasons as to the question of game developers not utilizing music for emotional awakening may be due to the trend of allocating a low budget for background music content when compared with the budget allocated to that of graphics by game developers [13]. This trend may be due to MER is an expensive task. Emotions perceived from music is extremely subjective for humans. In fact, it is a timely and exasperating task to identify various emotions perceived and

then match against the acoustic features of music to identify relationships between perceived emotions and music acoustic features.

The authors' goal is to develop a Rock and Electronic music excerpts containing a database where the excerpts have been labeled according to the emotion perceived based on the music genres' acoustic features. The emotion labels are in a tie with the emotions modelled in phase 2 of the Affective Loop. To achieve this goal, authors use the concepts of MER to identify the relationship among the acoustic features of the above genre and then designed a Machine Learning (ML) model to predict the emotion perceived in a certain music excerpt. The workflow of designing and developing the ML model is illustrated in Fig.8.

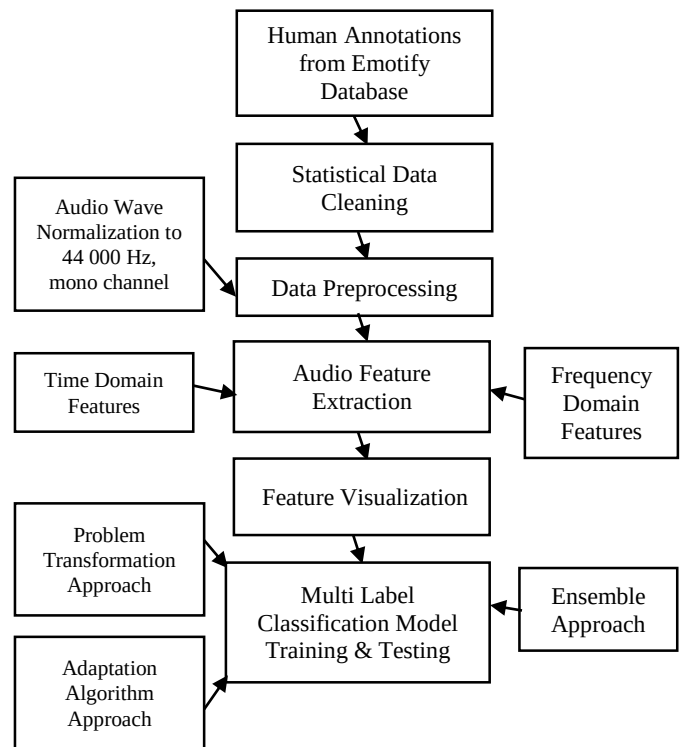


Fig. 8. Design methodology for music Emotion Recognition

'Emotify'[14] Data Set was selected for this methodology. This data set consists of music tracks of different genres including Rock and Electronic, where the excerpts are humanly annotated on perceived emotion. For each music excerpt, several human subjects rate their perceived emotion for the excerpt on a binary scale (either present or not) Emotify is the only a standard database recommended by the International Society of Music Information Retrieval (ISMIR) which consist of music annotations for the genre of Rock and Electronic.

Music annotations of the genre "Electronic" and "Rock" are cleaned for modeling. A music excerpt was labeled specifically by the amount of agreement among the annotations. This is calculated using the formula specified in (4).

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij} (n_{ij} - 1) \quad (4)$$

Where n is the total number of ratings per subject. The subject is the emotion label and expressed as i . $i=9$. k is the number of categories, a category where the respective emotion label is selected or not. The categories are expressed as j . j is either 1(emotion selected) or 0(not selected). n_{ij} is the number of participants who assigned the i -th emotion to the j -th category [14]

After the data cleaning process, it was revealed that the excerpts are having more than one emotion label (the amount of agreement is the same for 2 or more emotion labels). Hence, it was concluded from here that the Multi-Label ML approach should be conducted.

A total number of All music excerpts were then pre-processed and normalized to 44 000Hz mono channel and 30-second duration.

Acoustic features of the music excerpts were extracted using MATLAB's MIRTtoolbox [15]. The acoustic features extracted were Root Mean Square (RMS) Energy, Entropy of Energy, Spectral Roll-off, Spectral Centroid, Tempo, Zero Crossing Rate & Spectral Flux. These features were selected based on the works of Tan et al. [16]. Tan had dedicatedly worked on researching to reveal acoustic features that enhance Arousal. These acoustic feature values were then used to train and test the Multi-Label ML model.

Algorithms from the 3 categories of Multi-Label classification were selected. The data set was split into 70:30 ratios of train and test data. The authors selected Binary Relevance (BR), Classifier Chain (CC) under problem transformation category, ML-kNN from Algorithm Adaptation category and RakEL as the Ensemble approach of multi Label classification [17]. BR and CC were selected as they are the fundamental levels in multi-label approaches. ML-kNN was selected as it has a higher performance and accuracy score than problem adaptation methods [17]. Finally, RAKEL was selected as it is the only ensemble approach in multi-label classification, to the best of authors' knowledge.

ML models of the above-described algorithms were trained (with a train test ratio of 70:30) with the acoustic feature values of the rock music genre. Python was selected as the programming language to train the models. The best algorithm that outputs the highest evaluation metrics were then selected. Accordingly, RAKEL produced the highest accuracy scores. Table I shows the obtained accuracy values for ML Models.

TABLE I. EVALUATION METRICS SCORES FOR MULTI-LABEL ALGORITHMS

Evaluation Metrics	BR	CC	ML -kNN	RAkEL
Jaccard score	0.28301	0.35849	0.35849	0.58980
Hamming Loss	0.31132	0.23112	0.23112	0.19811
Micro Precision	0.55096	0.70740	0.74071	0.75428
Macro Precision	0.41773	0.55280	0.55285	0.57739
Average Precision	0.50535	0.67260	0.66276	0.70276

V. CONCLUSION

In this paper, a Car Racing game is designed based on the concept of the Affective Loop. In fact, this Racing Game developed with concepts on Affective Artificial Intelligence, which will enhance user experience level as this directly connects with the player's emotions during gameplay. In other words, boost the Racing Game Effect. This affective Loop consists of real-time physiological signal elicitation of the player, emotion modeling based on the physiological signals and adjusting game parameters according to the player's real-time emotion.

The real-time physiological signal elicitation is performed using EEG Signals of the player. This work includes a study of modeling emotion using EEG signals and developing a Machine Learning framework to model emotion elicitation. Data from the Mahnob HCI Tagging database was used. The signal processing, feature extraction as well as classification was done using MATLAB toolboxes. The future work of this part of the research includes further extending to determine emotions not only with the emotional arousal domain but also the dominant emotion domain described in human psychology to achieve more accuracy.

The authors had focused on background audio music inside the game parameter space that will be deployed to change according to the player's real-time emotion. They have focused on developing music containing a database that has excerpts classified under emotion labels. To classify music excerpts into emotion labels, they have approached a scientific technology called MER. They have developed a Machine Learning model that can predict the perceived emotion of a musical excerpt. Emotify Dataset was used. Acoustic features were extracted using MATLAB toolboxes.

The future work of this part of the research includes further optimization of i) EEG signal analysis for the detection of above emotions and also to detect other emotional categories and ii) classification of music based on perceived emotions by further analyzing audio waves and extract high dominant features for emotion detection in Music.

ACKNOWLEDGMENT

International Society of Music Information Retrieval for recommending and supplying "Emotify" Human annotation dataset, "Mahnob HCI tagging" database supplied the dataset of EEG signal varies according to human emotion.

REFERENCES

- [1] M. C. Laura and A. M. Andrea, "Emotion recognition techniques using physiological signals and computer games – Systematic review", *Técnicas de reconocimiento emocional utilizando señales fisiológicas y*, vol. 26, no. 46, pp. 19–28, 2017.
- [2] Gilleade, Kiel & Dix, Alan & Allanson, Jen.. "Affective Videogames and Modes of Affective Gaming: Assist Me, Challenge Me, Emote Me (ACE)", *Digital Games Research Conference*, 2005
- [3] C. M. De Melo, A. Paiva, and J. Gratch, "Emotion in Games," *Handb. Digit. Games*, pp. 575–592, 2014.
- [4] P. Sundstrom, "Exploring the affective loop," *Licentiate thesis*, Stockholm University 2005
- [5] Patino, Azeneth & Romero, Margarida & Proulx, Jean-Nicolas. "Analysis of Game and Learning Mechanics According to the Learning Theories", *Conference on VS-GAMES*. 2016..
- [6] C. M. De Melo, A. Paiva, and J. Gratch, "Emotion in Games," *Handb.*

- Digit. Games, pp. 575–592, 2014.
- [7] Fisher P. et al., "*The Racing - Game Effect: Why do Computer Racing Games increase Risk-taking inclinations*", SAGE Journals, 2009
 - [8] Guo, H. W. et al., "*Short-term Analysis of HeartRate Variability for Emotion Recognition via aWearable ECG Device.*" ICIBMS-Bioinformatics, Medical Imaging, and Neuroscience, Okinawa, Japan ,2015.
 - [9] Soleymani, M., J. Lichtenauer, T. Pun, M. Pantic,. "*A Multimodal Database for Affect Recognition and Implicit Tagging*", IEEE Transactions on Affective Computing, 3(1): 1-14,2012
 - [10] P. N. Juslin, P. Laukka, P. N. Juslin, and P. Laukka, "*Expression, Perception , and Induction of Musical Emotions : A Review and a Questionnaire Study of Everyday Listening Expression: Perception , and Induction of Musical Emotions ,*" vol. 8215, 2010.
 - [11] Kim et al., "*Music Emotion Recognition: State of Art Review,*" 11th International Society for Music Information Retrieval Conference 2010.
 - [12] Oliveira P.A, "*A musical system for Emotional Expressions*",2013
 - [13] D. Plans and D. Morelli, "*Experience-driven procedural music generation for games,*" IEEE Trans. Comput. Intell. AI Games, vol. 4, no. 3, pp. 192–198, 2012.
 - [14] A. Aljanaki, F. Wiering, R. C. Veltkamp.. "*Studying emotion induced by music through a crowdsourcing game*", Information Processing and Management, 2015.
 - [15] Lartillot, O. and Toivainen, P. "*MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio*", International Conference on Music Information Retrieval, Vienna, 2007.
 - [16] K.R. Tan, M.L. Villarino, C. Maderazo, "*Automatic music mood recognition using Russell's two dimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes*",International Conference on Information Technology and Digital Applications,2018
 - [17] Min-Ling Zang, Zhi-Hua Zhou, "*A review of Multi-Label Algorithems*", IEEE Transactions on Knowledge & Data Enginerring,2014.

Impact of metacognition and age group on contemporary video game interface and gameplay design

Nandunie N. Harischandra*
Software Engineering Teaching Unit
Faculty of Science,
University of Kelaniya, Sri Lanka
nandunieharischandra@gmail.com

Lahiru A. Jayakody
Gaming and Digital Entertainment
Department
Arimac Lanka (Pvt.) Ltd., Sri Lanka
lahiru@arimaclanka.com

Tiroshan Madusanka
Software Engineering Teaching Unit
Faculty of Science,
University of Kelaniya, Sri Lanka
tiroshanm@kln.ac.lk

Abstract: Metacognition is about “learning about learning” [1]. However, this theory is far more complicated. This allows people to take charge of their own learning. It involves awareness of how they learn, and evaluation of their learning needs, generating strategies to meet these needs and then implementing the strategies. The theory of metacognition can be identified as humans’ survival instinct. Metacognition also thinks about one’s thinking process such as study skills, memory capabilities, and the ability to monitor learning [2], [3]. Metacognition refers to higher-order thinking which involves active control over the cognitive processes engaged in learning. This concept needs to be explicitly taught along with content instruction. Metacognitive knowledge is about our cognitive processes and our understanding of how to regulate those processes to maximize learning. From ancient times humans have developed their metacognitive skills as a survival factor. For the survival in a video game, players need to follow instructions to get an idea about the gameplay [4]. But most of the time people are likely to skip the instructions without going through them even for the first time. This scenario is noticeable in mobile gameplay. This research has done to identify the factors which affect this dilemma. Metacognition level of a person and age are the variables that were considered for this experiment in order to address the following hypotheses; The more metacognitive skills people have, the more they will find it easy to play never-before-played games. Age range affects performance when it comes to playing the never-before-played games without direct instructions. At the end of this experiment, the first hypothesis became correct while the second one wasn’t. Therefore, people were more likely to ignore direct instructions and go through gameplay successfully when they had a higher metacognitive level, and the age group didn’t seem to affect this factor.

Keywords: Gameplay, Interface Design, Metacognition, Video games

I. INTRODUCTION

According to the common mobile game player behavior that has been experienced, the question arises whether instructions are necessary at all and why do people ignore the instructions and just continue to play. This research looked into how does the theory of metacognition help people to understand the functions concerning the age group that they belong, and the impacts of the metacognitive process on contemporary game interface design. It also helps to make ways of building a better and simpler game experience for people. The need for separate sets of instructions in the interface design or allow players to self-learn through the

gameplay can be varied according to different levels of cognitive skills of the players and the different age groups that they belong to. The purpose of this study was to determine whether metacognition helps people to understand gameplay and game interface without going through instructions and whether this has been affected according to age groups that they belong to. Games are rated by IARC (International Age Rating Coalition) according to suitable age ranges. Therefore, a game is designed according to its target market which has to be released. The need for direct instructions or giving them through the gameplay can be decided according to the final result of this research. Including instructions in the game techniques effects the whole gameplay design.

This research looked into this problem in search of factors that cause this particular problem and the answers help to broaden the knowledge of the game design field and make ways of building better game experience.

II. PREVIOUS STUDIES

A study on the effects of age on metacognitive efficiency was done in 2014. The primary aim of this study was to investigate the effects of age on metacognitive efficiency in healthy adults between the ages of 18 and 84. They have found that perceptual metacognitive efficiency declined with age, despite task performance being controlled to ensure all participants performed with the same accuracy [5]. In this study, game-related motives are not included and only the psychological side is considered while only considering adults for their experiments [6]. According to the latest video game demographics, details identified that young adults are mostly engaged in gaming [7], [3]. The age rating for gaming is also opened from children to adults and therefore, had to focus on young adults and ages between 7-18 for the research. According to the previous study mentioned above, a hypothesis made for this study that ‘age affects the performance when it comes to playing never-before-played game without direct instructions’, but later found that even though the cognitive efficiency fade with age, that does not affect when different people from different age groups having same metacognition level.

Game design, UI design, gamification related research topics are frequently found [8], [9], [10], [11], [12], [13], especially regarding how to design games to enhance metacognitive abilities via educational games. But this concept was far more different from the suggested study.

III. OBJECTIVES

The main purpose of this study was to determine whether metacognition helps people to understand gameplay and game interface without going through instructions. The study helps to broaden the knowledge of game design and help design better game experience. This study also helps to identify creative learning and teaching methods based on metacognition and games.

Games are rated (IARC ratings, ESRB ratings) according to different age groups that these games are suitable (according to the content of each game). Therefore, each game has a target market according to that age range restriction. If there is a connection between the age of a player and the need for instructions, that helps for the gameplay design of the overall game. Therefore, this helps to design the game correctly to the target market from the early stage while helping to market the game to the target market in the correct way.

Rating Categories

	Ages 3+
	Ages 7+
	Ages 12+
	Ages 16+
	Ages 18+
	Rating Pending

Fig. 1. IARC (International Age Rating Coalition) age rating according to suitable audiences



Fig. 2. ESRB (Entertainment Software Rating Board) age rating according to suitable audiences

For this study, the following questions were addressed.

- Based on the theory of metacognition, will people find it easy or difficult to play a never-before-played game without instructions according to their metacognitive skills?
- Does their age affect this?

Accordingly, this investigation included two research hypotheses.

H1: The more metacognitive skills people have, the more they will find it easy to play never-before-played games

H2: Age affects performance when it comes to playing a never-before-played game without direct instructions.

IV. THEORETICAL FRAMEWORK

The theory of metacognition defined that a person is capable of understanding or knowing his or her thinking process and manage it. This idea can be explained by dividing it into 3 parts [14], [15].

- Person knowledge (declarative knowledge) which is understanding one's capabilities
- Task knowledge (procedural knowledge) which is how one perceives the difficulty of a task which is the content, length, and the type of assignment
- Strategic knowledge (conditional knowledge) which is one's capability for using strategies to learn information. Young children are not particularly good at this; it is not until upper elementary where students start to develop an understanding of strategies that will be effective.

These three components are the necessary skills that you need to "survive" in a game and real life as well [4]. These components can be measured according to the following components.

- Metacognitive knowledge (also called metacognitive awareness) is what individuals know about themselves and others as cognitive processors.
- Metacognitive regulation is the regulation of cognition and learning experiences through a set of activities that help people control their learning.
- Metacognitive experiences are those experiences that have something to do with the current, ongoing cognitive endeavor[16].

These components help to identify the level of metacognition of a person with regard to gaming which was ultimately helped to figure out whether metacognition has an impact on understanding gameplay and game interface.

A. The level of metacognition

The level of metacognition is an equation designed by the researcher to calculate the level of metacognition of a person performing a task. According to John Flavell's (1979) explanation of metacognition as a theory, he explains that metacognition consists of two main parts named metacognitive knowledge and metacognitive regulation. He further states that the metacognitive knowledge can be divided into 3 parts as, person variables, task variables and strategy variables [17]. Considering all these facts the researcher was able to derive an equation to compare the levels of metacognition of different persons in a specific task. The key here was that the task has to be specific and the same for everyone to get an accurate reading.

$$\text{Level of Metacognition} = \text{Experience in the Task} * \text{Role} \quad (1)$$

The role here meant the role of the person doing the task. The role can be broken into several parts to get a reading. It consists of personal belief in achieving the task, the ability and experience to assess the task, and the ability to use

strategies in the task. The personal belief in achieving the task depends on the confidence and personal preference it means that to achieve a higher level of personal belief one has to have good confidence in himself to do the task and he should like or prefer what he has to do. The ability and experience to assess the task and the ability to use strategies depend on the performance and preference of similar tasks in the past [18]. For example, if someone has good performance in playing strategy games his level of metacognition in organizing a battle could be high because he could be able to assess the overall situation and use strategies to achieve his final goal of winning. The variables of this equation cannot be measured directly. However, they could be compared to get a reading for the level of metacognition in a certain task. Measuring the level of metacognition helped to find tangible answers to the research question. Even though there are no direct equations mentioned for this scenario, a rough equation can be formulated according to John Flavell's (1979) explanation of metacognition. 'Role' was proportional to 'level of metacognition' according to that explanation. The state of the 'experience in the task' factor was not mentioned as 'role', but it mentioned that this factor was also proportional to the level of metacognition. After considering all the information, this rough equation could be formulated for further calculations.

$$\text{Level of Metacognition} \propto (\text{Experience in the Task})^p * \text{Role} \quad (2)$$

$$\text{Level of Metacognition} = k * (\text{Experience in the Task})^p * \text{Role} \quad (3)$$

(k is a constant value, $p > 0$, p is a positive integer constant value)

In this case, related to gaming above equation should be modified accordingly.

$$\text{Level of Metacognition} = k * (\text{Experience in the Game})^p * \text{Role in the Game} \quad (4)$$

(k is a constant value, $p > 0$, p is a positive integer constant value)

When considering a certain game 'Role of the Game' can be defined as the following. The role of the game can be broken into several parts. It consists of personal belief in winning the game, the ability and experience to assess the game, and the ability to use strategies to win the game. A personal belief in winning the game depends on the confidence and preference of the game type. For example, the confidence of a strategy game lover would be high even to play a new strategy game for the first time. The ability and experience to assess the game and the ability to use strategies to win the game depends on the performance and preference of similar game types in the past. For example, a player who prefers and has good performance in RPG (Role Playing Games) may find it difficult to play a single-player 3rd person game.

Therefore, it was clear how to measure the level of metacognition of a gamer. However, it could only be a comparison between several gamers. However, it provided a good platform to monitor the relationship between metacognition and Video game interface and gameplay.

V. METHODOLOGY

The research conducted in the following manner. The first literature review was conducted to determine the theory of metacognition and the methods to measure the level of metacognitive skills of a person concerning gaming.

Then several experiments were conducted with participants of different metacognitive levels who belonged to different age groups. For the experiment, the participants were given a game that they have not played before. They were asked to play the game without reading any instructions. Their progress in the game was measured after the first level of the game was finished. The measurements were: the duration of the activity, the number of failures in the game, the number of first-time successful objectives of the game, their comment on the game (if there were any comments). These data were gathered and then analyzed against their level of, metacognitive skills to determine how metacognition impact on gameplay and game interface.

A. Sampling

The sampling of the participants was done based on age groups and metacognitive skill levels on games based on the theory of metacognition as explained in the theoretical framework section. Samples were taken from those who were new to the experimenting game included;

- Professional gamers
- Gamers who play games regularly as a hobby
- People who usually don't play games
- People who like to play a different genre of games than the experimental game

These samples were tested with the experimental game to gather data.

Additionally, people from the following age groups are considered from each type of players for the research:

- Ages 7+ (Sample includes more people belong to ages between 7 -12)
- Ages 12+ (Sample includes more people belong to ages between 12 -16)
- Ages 16+ (Sample includes more people belong to ages between 16 -18)
- Ages 18+ (Sample includes anyone above 18)

B. Instrumentation

An existing game has used for this research as the experimental game instead of making a customized game for this which is consuming a lot of time and not feasible enough for the given situation.

A game named LIMBO is used as an experimental game. This is a game designed and developed by "Playdead", Denmark. It is considered to be one of the best games to be created based on its simplicity and graphics, with high review scores from IGN, Gamespot, Gamespy, Eurogamer, Edge, etc. [19]. The reason for selecting this for the experiment was that this game has a limited number of instructions and the player has to figure out how to play the game. It provided the perfect conditions for the experiment. The Player was able to restart at the last encountered checkpoint with no limits.

This experiment aimed to understand how people assess game situations with the help of self-learning ability according to their metacognitive skills. This game provided a clear and enough number of separate situations that could be monitored easily. Puzzle genre was the most uncommon preference upon the players in the data set according to the given questionnaire. Therefore, it was more convenient to choose a never-played-before game with a limited number of direct instructions for all of them, among the puzzle genre with above other requirements for the experiment. Action/multiplayer genres were more familiar with most players in the sample and common techniques are used within the games [20]. Therefore, it was hard to find an unfamiliar action or games with new techniques, for a large sample which consists of 350 people who had majority preference over action/multiplayer games. Casual games were also suitable for this experiment because it also has clear self-learning points which can be used separately to measure player performance [14]. But the number of points available within this kind of game was not enough to get a correct outcome. Puzzle games need more brainpower than other genres, therefore this can be taken as the worst-case scenario experiment. Other genres use this self-learning content in games fairly low rate than puzzle genre. Therefore, the result must be valid for other minor cases as well as in this case used already [21], [22], [23].

VI. DATA COLLECTION AND ANALYSIS PROCESS

A. Data description

The following data was collected from the experiment.

- The duration of the activity
- The number of failures in the game
- The number of first-time successful objectives of the game
- Personal comment on the game (optional)

These data were cross-referenced against the samples' metacognitive skill level which was determined from the sampling process according to the theory of metacognition. The analysis provided tangible answers to the research problem.

This study was only conducted based on video games because the number of variables can be controlled easily that way and the sampling process is much productive as video games are played by many people. Data were analyzed after the preparation of graphs which shows metacognition level against game points.

Metacognition was calculated with the following equation, and "Experience in the Game" kept as a constant by using a player sample that hasn't played this game before. Therefore, experience in the game becomes the same value for all players." Role in the Game" was changed according to the questionnaire given to them and value was calculated according to their answers in it.

$$\text{Level of Metacognition} = k * (\text{Experience in the Game})^p * \text{Role in the Game} \quad (5)$$

(k is a constant value, $p > 0$, p is a positive integer constant value)

$$\text{Level of Metacognition} = h * \text{Role in the Game} \quad (6)$$

[h is a constant value because experience in the game is the same value for all gamers, $h = k * (\text{Experience in the Game})^p$]

Within the game, there wasn't any scoring system. Therefore, a simple tailor-made scoring equation was used to get a score for player performance in-game.

$$\text{The score of game} = 10 * \text{Number of first-time successful objectives of the game (1 point for each)} - [\text{Number of failures in the game (1 point for each)} * \text{playtime duration} / (\text{playtime duration} + 30 \text{ mins})] \quad (7)$$

There are main 9 points in the game that was expected to be successful for the first time without been taught.

- How to push/pull items?
- How to climb to boxes/higher places?
- How to climb a rope?
- How to swing the rope and jump to an available platform?
- How to cross a pond without been drowned?
- How to know an upcoming danger/how to avoid it?
- What happens if feet are put on a trap and therefore how to avoid them?
- How to use a trap against big spiders?
- How can we use spiders to reach our objective?

These values for both levels of metacognition and score were collected from each person and final values were represented through a graph to see overall behavior. Another set of graphs with the same structure represented according to the age groups which needed to be observed. The sample size for the experiment was 350, which was used to create graphs and finally to build the conclusion.

The scoring equation was a self-invention for the representation of the score of the given gameplay because there wasn't any original scoring system in the gameplay. This tailor-made system was used to calculate every gamer's score and values are taken for calculations were fair for everyone in the data set. The Scoring system doesn't use a fixed structure, uses influential factors as variables, and weighed in the scoring system according to the preference of the game developer. (In this case, there wasn't any given scoring system for the game by the developer.)

B. Evaluation result

Two hypotheses which were considered for this research are the following:

- H1: The more metacognitive skills people have, the more they will find it easy to play never-before-played games
- H2: Age affects performance when it comes to playing a never-before-played game without direct instructions.

The first hypothesis has become true while the second hypothesis has become false according to the following diagrams and analyzes results. The sample size is 351 for this research and following each age group consists of 87 samples within the data set.

- Between 7 -12
- Between 12 -16

- Between 16 -18
- Ages 18+ (Sample includes anyone above 18)

C. Regression analysis

Calculated results are represented as 'Level of metacognition' in the x-axis and 'Game points' in the y-axis. The first graph represents the overall sample of players while the other four graphs represent above 7, above 12, above 16, and the above 18 age categories.

R-squared values and angle of the slopes are the most used statistical techniques for this analysis. R-squared value helps to understand whether metacognition level and score level is correlated or not and if they are related then the strength of their association is also indicated.

R-squared value for overall representation = 0.751

R-squared value for above 7 representations = 0.652

R-squared value for above 12 representations = 0.798

R-squared value for above 16 representations = 0.611

R-squared value for above 18 representations = 0.743

Thereby overall, above 12 and above 18 representations are having a strong effect size while others are having a moderate level of effect size. Therefore, overall, above 12 and above 18 representations are highly fitted to the regression line while others are moderately fitted. 75.1% of the variation in overall gamer group's game points earning is explained by the level of metacognition while other separate age groups vary with 65.2%,79.8%,61.1%, and 74.3% amounts respectively. Therefore, it seemed to be that metacognition level and game points were well-associated factors according to this analysis result.

Slopes of these regression lines can also be considered for the analysis according to $y = mx + c$ structured graph formulae of those representations.

$$y = 107.39x + 203.42 \text{ (Overall representation)} \quad (8)$$

$$y = 125.98x + 301.22 \text{ (Above 7 representation)} \quad (9)$$

$$y = 102.4x + 172.67 \text{ (Above 12 representation)} \quad (10)$$

$$y = 125.31x + 247.38 \text{ (Above 16 representation)} \quad (11)$$

$$y = 99.241x + 145.17 \text{ (Above 18 representation)} \quad (12)$$

TABLE I. DEGREE OF SLOPE

Representation	m value	Radian	Degree
Overall	107.39	1.561	89.47°
Above 7	125.98	1.563	89.55°
Above 12	102.40	1.561	89.44°
Above 16	125.31	1.563	89.54°
Above 18	99.241	1.561	89.42°

Each regression line seems to be had nearly equal slopes according to the calculated angles between the x-axis and the lines. Drastic changes between degree values were not detected. Thereby it seems to be that the effect of the metacognitive level on game point-earning does not change

according to the age group. Metacognitive levels can be changed with age, but the nature of the bond between the score and metacognitive level is not affected by the age group.

The results indicated a positive slope when putting on a graph suggesting that the more level of metacognition that a subject has, they are more likely to do the task more successfully and quickly than those who had a low level of metacognition. Of course, there are a few inevitable anomalies but in a general sense, the graph agreed with the first hypothesis.

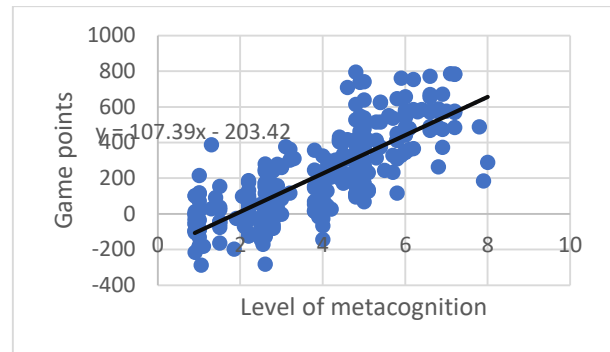


Fig. 3. The graph which displays data of the overall sample of players

According to each of the following graphs, it can be seen that every slope represents similar characteristics as the general slope for the overall result.

Gradient for each regression line is approximately near values which shows that age doesn't affect the performance when it comes to playing, a never-before-played game without direct instructions. Therefore, in a general sense, the graphs don't agree with the second hypothesis.

VII. DISCUSSION

A. Limitations of the approach

The study was only be conducted based on mobile games because the number of variables can be controlled easily that way and the sampling process was much productive as mobile games are played by many people.

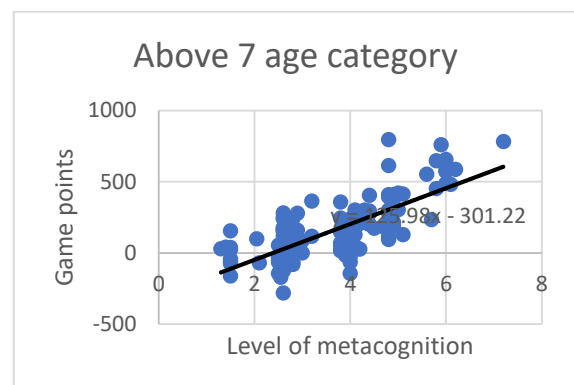


Fig. 4. The graph which displays data of players above the age of 7

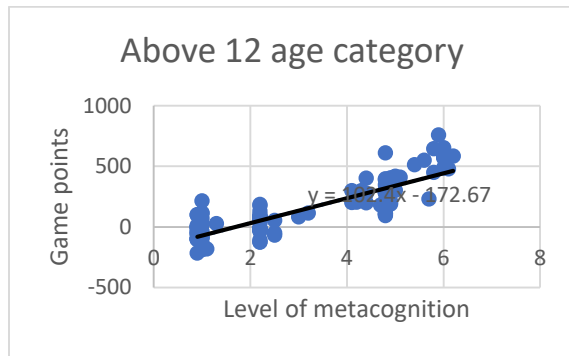


Fig. 5. The graph which displays data of players above the age of 12

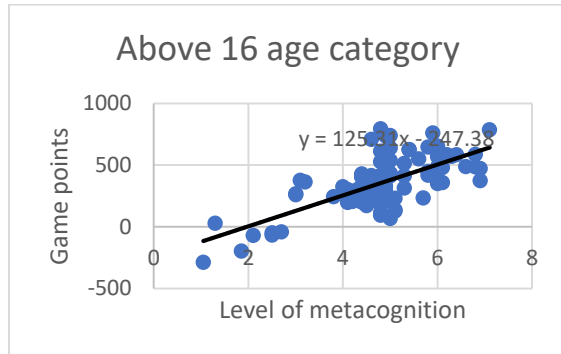


Fig. 6. The graph which displays data of players above the age of 16

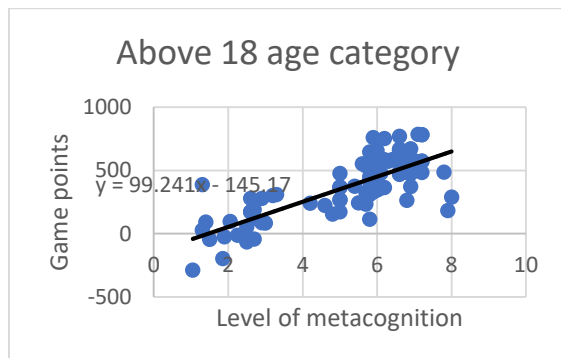


Fig. 7. The graph which displays data of players above the age of 18

Gameplay results are gathered in an indirect method which is not reliable enough. (Some people who play games are unlikely to share a higher number of failures in a game that they play. Couldn't help but following the indirect method for gathering information, because of the large number of people in the sample). The gender of the player can be another variable for this experiment, but it was not considered for this research.

Some players spent a long time on the above game level completion. Therefore, that whole time couldn't be calculated as the exact time they have spent on it because those people could be spending that time on some other tasks while playing them because they play the game for a long time. Therefore, sometimes inputs can be considered rough values because of this reason.

B. Future work

The sample size can be increased and real-time data gathering can be done for the information to be gathered from the gameplay which needs to be observed. The impact of gender can be analyzed with age and metacognition levels in

the future. Female representatives in the present sample were not enough to decide on the impact of gender differences on this factor. The impact of different game types and genres on this problem can be a further field to be investigated as an extension.

VIII. CONCLUSION

The results of the experiment came positive with the first hypothesis which the research was based upon. The first hypothesis was "The more metacognitive skills people have, the more they will find it easy to play never-before-played games". The results showed that the more metacognitive skills people have, the easier it is for them to play games that they have never played before. According to the other experimental result, it can be seen that the age category doesn't affect the above factor which makes the second hypothesis false.

Therefore, designing gameplay which doesn't include many direct instructions and instead of that include that information into gameplay even without any age restriction, seemed to be much more effective according to this experiment.

REFERENCES

- [1] E. Braad, "Learn-to-Learn: Game-Based Learning for Metacognition," 2018.
- [2] J. A. Livingston, "Metacognition: An Overview," New York: State University, 1997.
- [3] R. Monem, "Metacognition and Self-Scaffolding in MMORPGs: Case Study of an Adolescent Male Gamer," The Qualitative Report, 2015.
- [4] "Metacognition," Academic Dictionaries and Encyclopedias, 2020. [Online]. Available: <https://en.academic.ru/dic.nsf/enwiki/726722>. [Accessed: 06- Mar- 2020].
- [5] E. C. Palmer, A. S. David and S. M. Fleming, "Effects of age on metacognitive efficiency. Consciousness and cognition," vol. 28, pp. 151–160. doi:10.1016/j.concog.2014.06.007, 2014.
- [6] S. Moritz and T. S. Woodward, "Metacognitive control over false memories: A Key Determinant of Delusional Thinking," Current Psychiatry Reports, 2006.
- [7] V. Yanev, "Video Game Demographics - Who Plays Games in 2020," Tech Jury, 2020. [Online]. Available: <https://techjury.net/stats-about/video-game-demographics/>. [Accessed: 06- Mar- 2020].
- [8] D. J. Hacker, J. Dunlosky and A. C. Graesser (Eds.), "The educational psychology series. Handbook of metacognition in education," Routledge/Taylor & Francis Group, 2009.
- [9] M. Hrehovcsik, "An analysis of design thinking in applied game design", 2018.
- [10] D. Kumara, M. Menon, S. Moritz and T. S. Woodward, "Using the Back Door: Metacognitive Training for Psychosis". Routledge Taylor and Francis Group, 2014.
- [11] I. Damopolii and B. Kurniadi, "Training students' metacognitive skill using mobile learning". Journal of Physics: Conference Series. 1317. 012185. 10.1088/1742-6596/1317/1/012185, 2019.
- [12] A. Karabinus and R. Atherton, "Games, UX, and the Gaps: Technical Communication Practices in an Amateur Game Design Community," 1-7. 10.1145/3233756.3233949, 2018.
- [13] K. Kiili, and P. Tuomi, "Teaching Educational Game Design: Expanding the Game Design Mindset with Instructional Aspects". 10.1007/978-3-030-34350-7_11, 2019.
- [14] L. C. Duarte and Andre Battaiola, "Distinctive Features and Game Design," Entertainment Computing. 21. 10.1016/j.entcom.2017.03.002, 2017.
- [15] J. A. Geiwitz, "Conceptual Model of Metacognitive Skills," U.S. Army Research Institute, 1996.
- [16] "TE 150: Metacognition and Strategies Flashcards | Quizlet", Quizlet, 2020. [Online]. Available: <https://quizlet.com/51500470/te-150-metacognition-and-strategies-flash-cards/>. [Accessed: 06- Mar- 2020].

- [17] J. H Flavell, "Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry," *American Psychologist*, 34(10), 906-911, 1979.
- [18] J. Metcalfe and A. P. Shimamura, "Metacognition: Knowing about Knowing," Cambridge: MA: MIT Press, 1994.
- [19] "Limbo (video game)," En.wikipedia.org, 2020. [Online]. Available: [https://en.wikipedia.org/wiki/Limbo_\(video_game\)](https://en.wikipedia.org/wiki/Limbo_(video_game)). [Accessed: 06-Mar- 2020].
- [20] B. S. M. James and B. Fletcher, "Defining the Global Ludo Polychotomy," Digital Games Research Association, 2015.
- [21] J. M. Olthouse, "Video Games: Why Kids Play and What They Learn," Meridian Middle School, 2009.
- [22] L. Sheldon, "Character Development and Storytelling for Games," Stacy L. Hiquet, ,2004.
- [23] J. Whitehead, "Game Genres: Shumps. Creative Commons," 2007.

Detecting human emotions on Facebook comments

E. J. A. P. C. Chathumali*

Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology, Sri Lanka
chathujay92@gmail.com

Samantha Thelijagoda

SLIIT Business School
Sri Lanka Institute of Information Technology, Sri Lanka
samantha.t@slit.lk

Abstract: Human emotion detection plays a vital role in interpersonal relationships. From the early eras, automatic recognition of emotions has been an active research topic. Today, sharing emotions on social media is one of the most popular activities among internet users. However, when it comes to a specific domain like emotion detection in social media, it is still on a research-level. There are less number of applications have been developed to detect emotions online, using online comments and user comments. The aim of this research is to develop a system that identifies human emotions on Facebook comments. Among the different social media platforms, this research specifically focuses on Facebook comments written in the English language to narrow down the problem. The research is based on Semantic analysis, which comes under Natural Language Processing (NLP) and the system development consists of four major steps, including the extraction of Facebook comments via Graph API, preprocessing, classification and emotion detection. To classify the emotions, a classification model was created by using Naïve Bayes Algorithm. When it comes to marketing, emotions are what lead your onlookers to purchase. By using the detected emotions, marketers can promote their campaigns by changing online advertisements dynamically. The results obtained through testing the system show that it is capable of accurately identifying human emotions hidden in Facebook comments with an accuracy level of 80%, making it highly useful for marketing purposes.

Keywords: Emotions, Emotion detection, Facebook, Naïve bayes algorithm

I. INTRODUCTION

Feelings or emotions play a critical role in human life and people express their feelings in everyday communications. There are different types of emotions, having an impact on the lifestyles of people and in their day-to-day interactions. People are mostly ruled by emotions. The choices of people, their actions, and perceptions are all impacted by the emotions they feel at a given time. Relational human correspondence incorporates dialect that is spoken, as well as non-verbal prompts as hand and body motions, the tone of the voice, which are utilized to express feelings and give criticism. Understanding and realizing how to respond to individuals' demeanor incredibly improves associations. Ekman and Freisn having expertise in this area, distinguished six fundamental emotions, including anger, fear, disgust, joy, surprise and sadness which is recognized as a general standard [1].

Today, most of the social media users are accustomed to sharing emotions in their social media platforms. They tend to record their daily life online and frequently share their thoughts, considerations and feelings, and connect with companions through the web. The main reason for this is the ascent of social network websites like Facebook and Twitter

that has significantly changed how individuals live. According to the statistics of broadbandsearch.net, in 2018, 2.62 Billion people were using social media platforms [2]. By 2021, this number is expected to rise over 3 Billion. Facebook leads prominent social media platforms with 2.234 Billion active users in 2018 and is the third-highest visited website in the world [2].

According to the statistics of Escale Solutions it has been stated that, for every second on Facebook by January 2019, 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded [3]. With the help of those abundant data, there are great openings for considerate human behavior and more numerous new applications for enhancing individuals' lives. Apart from that, according to brandwatch.com, Facebook marketing statistics illustrates that there are over 60 Million active business pages currently running on Facebook and over 2 Billion people can be reached through these [4]. Therefore, Facebook is a great place for businesses and their marketing campaigns. On the other hand, most of the effective marketing campaigns are driven through human emotions [5]. All the marketing campaigns always focus on touching peoples' hearts and on building an everlasting bond with the customers. For instance, customers always decide on a purchase based on their emotions at that moment [5]. Emotions are the key that leads the onlookers or the customers to purchase. In addition to that, a significant number of the best promoting, marketing campaigns and activities are based on emotions. For example, once a user arrives at Facebook, if the technology can change the marketing advertisements dynamically according to user's emotions, it will be a great advantage and a turning point for businesses [5]. If marketing does not consider emotions of the intended consumers, businesses will be in trouble due to a lack of purchases.

The main objective of this research is to design and develop an application that is capable of identifying human emotions of online users in the Facebook social network. To achieve the main objective, the system uses the sub-objectives, including the analysis of existing methodologies of collecting Facebook data, collecting Facebook comments according to the identified methodology, implementing a model that classifies the human emotions and identifying human emotions according to published Facebook comments accordingly.

This system mainly focuses on the comments which have been published in a particular business page. To understand the meaning of the text, it uses the techniques which come under Natural Language Processing (NLP). NLP can be used to process both speech and text. The proposed system uses text processing to capture the meaning of words that have been collected from Facebook comments. In order to identify the real meaning of the text, the system uses Semantic

analysis, which comes under Natural Language Processing (NLP). Semantic analysis describes the method of understanding natural language the way that humans interconnect based on meaning and context [6]. By this foundation, this research utilizes a strategy to categorize emotion based on text instead of the three essential semantic classifications including positive, negative and natural. This has been done with the help of collecting data through automatic semantic analysis. For the proposed research, Facebook social media platform has been used to gather user comments for the particular business page, which was already published by a user. The research specifically focused on Facebook comments written in the English language to narrow down the research problem.

II. BACKGROUND STUDY

Binali, et. al. [7] discuss computational approaches for emotion detection in text. The aim of this project is to present emotion theories that deliver a foundation for emotion models. They present how these models have been utilized by examining computational ways to deal with emotion detection. Apart from that, in order to detect the emotions, authors have proposed a hybrid-based architecture. They have used SVM algorithm for validating the suggested architecture and accomplishes an expectation precision of 96.43% of the web blog information. This hybrid-based architecture is proposed to fulfill two different requirements, including hidden phrase pattern detection and the emotion prediction. In hidden phrase pattern, syntactic and semantic details are used to identify the emotions in the subtle sentences, whereas emotion prediction is used to find the emotion being stated, based on information-rich linguistic resources and trained classifiers.

Yassine et. al. [8] present a framework for examining the relationship between friendship and emotion expression in online social networks. The main objective of this framework is to extract or highlight the emotional content of the text which is already available in online social media. In order to identify the writer's emotions using text, they have performed text mining techniques on comments which are retrieved from social media. The framework is developed using numerous significant steps, including raw data collection, lexicon development, feature generation, data preprocessing, creating data model, text classification and friendship classification. All those features help recognize whether the online post expresses any emotions or feelings. Since the social media comments have informal language, authors have faced difficulties in sentiment analysis. To overcome this issue, they have proposed a framework with special lexicons. Moreover, they have tested this framework with the help of the k-means clustering algorithm and they have got high accuracy outcomes for the model in defining subjectivity of texts and forecasting friendship.

Sewwandi et al., [9] propose a web application to recognize the personality of humans by using linguistic feature analysis. They have used Facebook social media data and information when conducting their research. There are three major dimensions that they have used when developing the application, including ontology-based text classification, questionnaire-based personality detection and linguistic feature vector matrix using Linguistic Inquiry Word Count (LIWC) features such as semantic analysis with supervised

machine learning algorithms. According to the authors, the main advantage of this application is its consideration of the personality when they are recruiting or promoting employees. They have conducted a feasibility study, covering the financial, technical and operational areas. From that, they have identified that there are no technical dependencies due to open source technologies. They have used a restful API to extract data from Facebook. As the first dimension of personality detection, the application uses an ontology-based approach. There, ontology is constructed using 'protege' software which has the OWL ontology developing language with the OWL-DL package. Moreover, ontology is designed along with a data set attainable by a recognized Physiologist in Sri Lanka.

Mashal et. al. [10] propose an approach to identify emotion, intensity using social media data. It consists of four major steps, including training data set, preprocessing, feature vector construction and emotion intensity prediction. As the first step they have trained the data set which covered four emotion categories including fear, anger, joy and sadness. Training data set included tweets and it has used real-valued score in the range of 0 to 1. That range is defined by the authors to represent a degree of emotion. After that they have done the preprocessing. According to the authors, preprocessing is very important because tweets are one hundred and forty words long in length and they state that there can be elements that mislead the classification process. They have done the preprocessing with the help of major steps, including conversion of words to lowercase, replacement of user mentioning with @user, replacement of letters and punctuations that are repeated and removing hashtags. As the third step they have constructed a feature vector. There, they have used WordNet files for each and every emotion type along with the physically interpreted words and a list of opposing words. To invert the meaning of the sentence negating words are used in the feature vector creation. Apart from those, they have trained the prediction model in the preprocessing. As the last step they have predicted the emotion by using the feature vector with the help of intensity which is given to the prediction algorithm as an input.

Rabeya, et. al. [11] have conducted a research to detect emotions in Bengali text at the sentence level. The research has used a backtracking approach which is based on the lexicon. In order to conduct the research, the authors have considered only two basic emotions; happiness and sadness. Apart from that, the research uses the most commonly used labeled databases; ISEAR and SemEval. The implementation of the approach is carried out in a few steps including data gathering and extraction, preprocessing, analyzing sentiment and emotion classification. To extract data for processing, authors have used sources such as journals, news, lines, textbooks, blogs, Facebook posts etc. Once the collected data has been extracted preprocessing is done for the tokenization of sentences. Tokenization process has been done for each and every input sentence generating the expression which can be used to calculate the sentiment as the next step of the methodology. To find out the emotions in the Bengali text, sentiment of each sentence has been identified by using the backtracking technique. There, the authors have used the expression from the back to find the sentiment. For instance, researches have considered that people have expressed their emotions at the end of the sentence and not in the position's

length -1 to -3. After that, the particular emotion is identified based on the sentiment associated with them. According to the authors, the proposed method resulted in an accuracy of 77.16%.

In a research conducted by Muljono, et. al. [12] identifies an approach to recognize the emotions using Indonesian text and they have evaluated the performance of four techniques including J48, K-Nearest Neighbor (KNN), Support Vector Machine-Sequential Minimal Optimization (SVM-SMO) and Naïve Bayes (NB). Authors have found that most Indonesians communicate with each other by using their own local languages, including Javanese, Sundanese and Balinese, but in formal situations Indonesian is used as the language. When conducting the research, they have collected data from different websites such as www.dongengceritakaty.com, www.dongeng.org, and www.pendongeng.com. Their datasets consisted of one to three sentences. WordNet Affect List has been used to gain the emotional tag in the sentences. Also, they have manually labeled sentences which are not included in the WordNet Affect List. To increase the accuracy of the research they have collected 1000 sentences and stored in text format file. As the next part of the research they have done the preprocessing. They have used several steps, including tokenization, case normalization, stop word removal, and stemming to achieve the preprocessing. There, Lucene library has been included to develop the code from scratch. After that they have done the term weighting by using the TFIDF (Term Frequency Inverse Document Frequency) which is a popular term weighting scheme. As the next step classification has been done by using the two techniques; machine learning and supervised learning, and unsupervised learning. This research compares four classification algorithms, including Naïve Bayes (NB), J48, K-Nearest Neighbor (KNN) and Support Vector Machine-Sequential Minimal Optimization (SVM-SMO) to identify the best classification algorithm for textual emotion recognition. The experiment of these classification techniques is carried out using WEKA (Waikato Environment for Knowledge Analysis). Finally, they have done the evaluation with the help of 10-fold and split validation. Moreover, data have been divided in to testing and training by cross validation. There, the authors have defined that 90% of the total data have been used for training and 10% have been used for testing. Based on the experiments, they have stated that SVM-SMO classifier gives the best performance when compared to other classification algorithms.

Sarakit, et. al. [13] has conducted a research to classify the emotions in Thai YouTube comments. The main purpose of this research is to add more value to YouTube. They have used comments written in Thai language for a particular video to identify the emotions. Authors have considered the six basic emotions, including anger, fear, happiness, disgust, sadness and surprise. Their research consists of several steps such as collecting YouTube comments, pre-processing, feature selection, term weight and classification. As the first step, the authors have collected data from the YouTube. To do that, they have considered 50 Thai video clips. Among them, 20 video clips are music videos and 30 are video clips from commercial advertisements. In order to collect the data, they have used the YouTube 2.0 API. According to the authors, they have gathered a total of 2771 comments from Music videos and 3077 comments from advertisements. After

that they have performed the pre-processing. Pre-processing includes the procedures segmenting comments into a set of word tokens, removing Thai and English stop words, Removing HTML tags and stemming words. Lexitron dictionary from NECTEC and LexTo: Thai Lexeme Tokenizer have been used for segmentation. As the third step, feature selection has been done by assigning a frequently threshold value for selected keywords. If the word has appeared more than the minimum threshold value, it is considered a keyword. If it has not, it is removed from the feature list. After that, they have assigned term frequency and term frequency inverse document frequency. Finally, emotion classification has been done by the standard classification methods including SVM, multinomial naive Bayes and decision tree namely J48. According to the authors, the highest accuracy is returned from the support vector machine (SVM). It generates an accuracy of 76.14% for commercial advertisements and 84.48% of music videos.

III. METHODOLOGY

In order to classify human emotions, the developed system includes a few major steps. As the first step, the system collects Facebook comments on a particular business page, which is already published. The next step is to preprocess the collected comments according to the common pre-processing procedures. As the third step, the system processes a critical step which is called classification. To do that, a classification model is created using the machine learning algorithm; Naïve Bayes. Apart from that, the model is trained using an existing data set. Finally, by comparing the keywords, the system classifies human emotions. The classification process is done for every hour and generates the highest occurring emotion as the output. Likewise, this process is continued for each and every hour to classify the emotions.

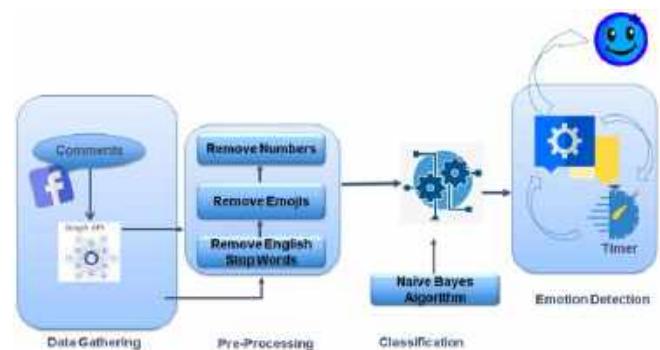


Fig. 1. High-level diagram

A. Data collection

The very first step of the research was to collect the data or comments of users from Facebook. To do that, Facebook Graph API has been used. Graph API, is an inbuilt API provided by Facebook for their developers. It is a primary way of getting data in and out from the Facebook platform. Moreover, Graph API [14] is based on HTTP and it can be used for the applications to programmatically query data, manage advertisements, upload photos etc. According to Facebook, they have allowed developers to extract the comments from the API only if he/she is an admin of a particular page or their

own comments on their own posts. Since Facebook wants to protect user data, they have followed this mechanism. To extract any information through the Graph API, developers need to log in to “Facebook for developers” website and get their access token.

B. Pre-processing

Once the data have been collected through Graph API, next step was processing them according to the common pre-processing procedures. Those procedures include, eliminating English stop words, removing numbers, removing emoji's, removing double quotations, slashes, underscores etc.

C. Classification

As the third step, the system has performed the critical process of classification. To do that, a classification model was created by using the machine learning algorithm Naïve Bayes. Naïve Bayes belongs to the probabilistic algorithm family and it uses Bayes's Theorem to predict the text category. Naïve Bayes classifiers are extremely scalable and require a number of parameters in a learning problem. The model was trained using the existing data set. When creating the model, tags or the labels should be created as the first step. The research has used the emotion tags, including Happiness, Sadness, Anger and Surprise.

D. Emotion Detection

Finally, preprocessed data has been sent to the model via API and by comparing the key words, the model the classified human emotions as the output.

IV. EXPERIMENTAL RESULTS

This chapter illustrates the results of the above steps. As the first step, the system has collected data via Facebook Graph API and comments have been returned as a Json object. After gathering the data, Json object should be deserialized in order to covert the Json string to a particular custom object. After that, the collected data has been assigned to a list which includes comments put up during the last hour. As the next step, preprocessing has been performed. The resulted values have been sent to the model as a byte array. Then, the model has classified the comments one by one and resulted in the output. Finally, among the resulted values, the system has returned the maximum emotion with an image, and all the results, including sadness, happiness, anger and surprise are displayed in a pie chart. The process is being repeated until it classifies the emotions in the next hour. Similarly, this has been continued until the user exits from the system.

The following figure shows the most recent hourly data gathered via Facebook Graph API. It has returned the unprocessed comments and stored them inside a Json object. After that, the Json object has been deserialized and converted to a particular custom format.

Once the data have been collected a list has been created by assigning all the comments which are published in the last hour. The comments have been preprocessed before sending to the model. Then, resulted values have been sent to the emotion classification model in order to classify the output. Once the model has returned the output, the system calculates the number of emotions expressed during the last hour.

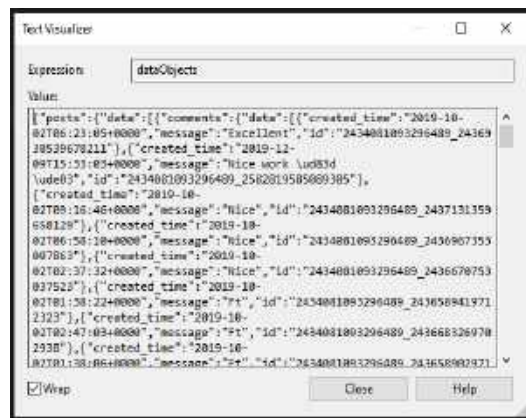


Fig. 2. Collected Facebook comments

Finally, among the values the system has returned, the emotion having the maximum count with an image and all the results, including sadness, happiness, anger and surprise have been displayed in a pie chart.



Fig. 3. Graphical representation of detecting emotion during the last hour

Once all the outputs have been generated, the process repeats the same steps and continues emotion detection until the user exits from the system.

According to the final outcome of the research, an accuracy level of 80% is achieved. To find the accuracy of the results, confusion matrix has been used. It has given the accuracy by comparing the actual and predicted classes. Figure 4 given below shows the classes of the confusion matrix.

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

Fig. 4. Graphical representation of confusion matrix

TP (True Positive) which describes the predicted values are correctly predicted as actual positive. FP (False Positive) defines the predicted values are incorrectly predicted as actual positive. For an example, negative values predicted as positive. FN (False Negative) defines that the Positive values are predicted as negative. TN (True Negative) describes the predicted values are correctly predicted as an actual negative.

For example, if the user has commented “Excellent job, poor work, amazing talent, prices are high and great work”, the system returned, the output emotions as happy, sad, surprise, sad and happy respectively. Therefore, those values have belonged to the TP class. Furthermore, there were some failed outputs as well. For example, if the user has published comments ‘Not happy, not good’, the system generates the output emotion as ‘happy’ instead of the actual result ‘sad.’ Therefore, this kind of outputs belong to the FP. The following equation (1) has been used to calculate the accuracy level.

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

V. CONCLUSION

Feelings or emotions play a critical role in human life and people express their feelings in everyday communications. Almost all the people in the world share their emotions on social media. The purpose of the research is to identify emotions of online users by using Facebook comments. Selecting Facebook from numerous social media platforms, the research specifically focuses on Facebook comments written in the English language in order to narrow down the problem. When it comes to marketing, emotions are what lead the onlookers to purchase. Therefore, identification of user emotions can be helpful for businesses in various scenarios. For example, when a user arrives on Facebook and if the technology can change the marketing advertisements dynamically according to user’s emotions, it will be a turning point for a business. Therefore, the main objective of the developed system was to identify emotions of the online users by extracting their comments. The system uses the techniques in semantic analysis and the major steps include data collection from Facebook, preprocessing, classification and emotion detection. To identify the emotions, a classification model has been created by using the Naïve Bayes algorithm. Finally, the output is graphically represented by using a pie chart.

REFERENCES

- [1] A. Varghese, J. P. Cherian and D. J. Kizhakkethottam, "Overview on Emotion Recognition System," in *International Conference on Soft-Computing and Network Security*, India, 2015.
- [2] "broadbandsearch," [Online]. Available: <https://www.broadbandsearch.net/blog/internet-statistics>. [Accessed: 14-01-2019].
- [3] B. Bot, "Escale Solutions," Top 20 Facebook Statistics in January 2019 - Escale Solutions, [Online]. Available: <https://www.escalesolutions.com/blog/top-20-facebook-statistics-in-january-2019/>. [Accessed: 28 -01-2019].
- [4] "Brandwatch," [Online]. Available: <https://www.brandwatch.com/blog/facebook-statistics/>. [Accessed: 28-11-2019].
- [5] "The Ultimate Guide to Emotional Marketing," hubspot, [Online]. Available: <https://blog.hubspot.com/marketing/emotion-marketing>. [Accessed: 11-01-2019].
- [6] "Expertsystem.com," Natural Language Process semantic analysis: definition, [Online]. Available: <https://www.expertsystem.com/natural-language-process-semantic-analysis-definition/>. [Accessed: 02-02-2019].
- [7] H. Binali, C. Wu and V. Potdar, "Computational Approaches for Emotion Detection," in *4th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST)*, Dubai, United Arab Emirates, 2010.
- [8] M. Yassine and H. Hajj, "A framework for emotion mining from text in online social network," in *IEEE International Conference Data Mining Workshops*, USA, 2010, p. 1136–1142.
- [9] D. Sewwandi, K. Perera, S. Sandaruwan, O. Lakchani, A. Nugaliyadde and S. Thelijjagoda, "Linguistic Feature Based Personality Recognition Using Social Media Data," in *6th National Conference on Technology and Management (NCTM)*, Malabe, 2017.
- [10] S. X. Mashal and K. Asnani, "Emotion Intensity Detection for Social Media Data," in *IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC)*, Tamil Nadu, 2017.
- [11] T. Rabeya, S. Ferdous, H. S. Ali and N. R. Chakraborty, "A Survey on Emotion Detection A lexicon based backtracking approach for detecting emotion from Bengali text," in *20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2017.
- [12] Muljono, N. A. S, Winarsih and C. Supriyanto, "Evaluation of Classification Methods for Indonesian Text Emotion Detection," in *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, 2016.
- [13] P. Sarakit, T. Theeramunkong, C. Haruechaiyasak and M. Okumura, "Classifying Emotion in Thai YouTube Comments," in *6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, Hua-Hin, Thailand, 2015.
- [14] "Facebook for developers," [Online]. Available: <https://developers.facebook.com/docs/graph-api/overview/>. [Accessed: 25-05-2019].

Keyword extraction from Tweets using NLP tools for collecting relevant news

Thiruni D. Jayasiriwardene*
Faculty of Information Technology
University of Moratuwa, Sri Lanka
thirudilushi@gmail.com

Gamage Upeksha Ganegoda
Department of Interdisciplinary Studies
Faculty of Information Technology
University of Moratuwa, Sri Lanka
upekshag@uom.lk

Abstract: Keywords play a major role in representing the gist of a document. Therefore, a lot of Natural Language processing tools have been implemented to identify keywords in both structured and unstructured texts. Text that appears in social media platforms such as twitter is mostly unstructured because of the character limitation. Consequently, a lot of short terms and symbols such as emoticons and URLs are included in tweets. Keyword extraction from grammatically ambiguous text is not easy compared to structured text since it is hard to rely on the linguistic features in unstructured texts. But when it comes to news on twitter, it may contain somewhat structured text than informal text does but it depends on the tweeter, the person who posts the tweet. In this paper, a methodology is proposed to extract keywords from a given tweet to retrieve relevant news that has been posted on twitter, for fake news detection. The intention of extracting keywords is to find more related news efficiently and effectively. For this approach, a corpus that contains tweet texts from different domains is built in order to make this approach more generic instead of making it a domain-specific approach. In fact, the Stanford Core NLP tool kit, Wordnet linguistic database and statistical method are used for extracting keywords from a tweet. For the system evaluation, the Turing test which has human intervention is used. The system was able to acquire an accuracy of 67.6% according to the evaluation conducted.

Keywords: Named Entity Recognition (NER), Natural Language Processing (NLP), Part of speech tagging, Stanford Core NLP, Wordnet corpus.

I. INTRODUCTION

A keyword is a word that succinctly and accurately describes the subject or the aspect that identifies the subject mentioned in a document [1], plays a major role as an indicator of important textual information which spread among the people as soft documents or hard documents. In scientific communication, words are used to communicate information unambiguously while highlighting the words as keywords that are focused on the communicated topic. On the other hand, keywords are essential for the reader to get a quick idea of the information contained as text, in the meantime, it is important when searching for the associated information for a particular topic. Therefore, Natural Language Processing (NLP) tools have been built in order to extract keywords from different types of documents or sources. Scientific documentation contains structured text which enables natural language processing tools to work with linguistic and syntactical features of the language. Unstructured texts are commonly used in social media such as Twitter because of its character limitation. A tweet contains a maximum of 280 characters. Therefore, people

tend to use words in an unstructured way. It doesn't contain linguistic features and sometimes provides ambiguity for the reader.

Several approaches such as the statistical approach, Rule-Based Approach, Machine Learning approach and Domain-specific approach [1, 2] exist to extract keywords from textual documents. Figure 1 depicts the classification of automatic keyword extraction.

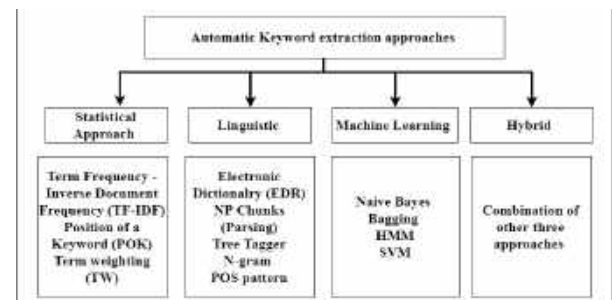


Fig. 1. Classification of automatic keyword extraction [2]

Method of extracting keyword differs based on the objective of the extraction; it could be carried out for text summarization, text categorization, information retrieval and question answering [2]. research had been carried out to extract keywords from unstructured text. It is not similar to extracting keywords from structured text such as abstracts and news articles due to lack of linguistic features, for instance, positioning of the word in a sentence and the number of occurrences of the word in the particular sentence. Unstructured texts are commonly used in social media because of the character limitation for a tweet post. Furthermore, it contains Abbreviations, links for websites, emoticons and images. Tweets are especially grammatically ambiguous so the accuracy of Part of Speech (POS) tagging may be less. Therefore, methods that are used to extract keywords from the structured text will not be accurate. Some of the statistical methods won't work on tweets as one word appears only once in many cases. But it can be used to filter 'most appeared words' in the corpus.

This paper proposes a methodology to extract keywords from a tweet to retrieve relevant news for a particular incident. This keyword extraction intends to collect associate news as much as possible. This paper is organized as follows. Related work for keyword extraction is overviewed in section II. The methodology is elaborated in section III and methods of evaluation are discussed in section IV.

II. RELATED WORK

Research has been conducted on extracting keywords from grammatically ambiguous texts such as tweets using NLP tools. This approach is a combination of machine learning and rule-based approaches [3]. A corpus that contains domain-specific words that needed to be rejected had been built to remove the noises from the tweet texts. Machine learning-based Stanford Core NLP [4] POS tagging is used since it has the highest token accuracy. Penn tree notation is used to tag the tokens and Tregex [5] notation is used to structure the tree. Matching of rejected words was done based on the Levenshtein Distance; [6] when the distance between the word and the rejected word equals zero then the particular word is rejected and eliminated from the list of words. This process enables us to filter the final set of keywords consisting of Noun Phrases (NP) and Verb Phrases (VP) from the given tweet. These keywords are further filtered inside the second parser to find the other keywords which are not tagged as either NP or VP. This system is developed using java as a standalone desktop application.

An approach for keyword extraction was proposed to address the problems such as frequent usage of lexical variants and the high variance of cardinality presented in each tweet. In this approach, they had provided a keyword annotated data set of 1827 records and a system using unsupervised fashion. Brown clustering and continuous word vector approach have been integrated for unsupervised feature extraction in tweets [6]. Brown clustering along with hidden Markov model assumptions have been used to cluster lexical variants while generating a hierarchical structure of clusters. Structured skip-gram goal function is proposed to use for extracting words as a vector [7]. In this approach, a method is included to learn and predict the number of keywords to be generated depending on the count of functional and non-functional words in the given tweet [3].

An attempt had been taken to extract keywords and generate headlines from the unstructured text using the hidden Markov model and other NLP tools such as POS tagging and clustering. In this approach, the first text preprocessing part had been done using the following steps. First, the input text is tagged using the Part of Speech tagging which implemented TreeTagger [8] a decision tree based on the probability. A binary decision tree is implemented recursively by using an improved version of ID-3 algorithm. A dataset of trigram is used and the probability was decided by traversing the tree. This tagging approach has achieved a higher level of accuracy than trigram tagger on PennTreeBank data. Then the tagged output was normalized and stemming was done, stop words were removed and merging similar content was done using WordNet. Then the keywords were extracted. The number of words to be extracted was reduced by merging the words with similar content [9].

A novel method has been invented to extract topical keywords from twitter. In this approach, topics are learnt from tweets to extract and organize key phrases using three standard steps known as Keyword ranking, candidate key phrase generation and key phrase ranking [10]. A modified topical page ranking method which introduces topic sensitive score propagation [11] was used to boost the performance. Candidate key phrase generation method combined with principle probabilistic phrase ranking method is proposed for

keyword ranking. For topic discovery using twitter, a modified author-topic model named Twitter-LDA was used under the assumption that a single topic exists for the entire tweet [10]. The original topical page ranking method was modified by including context-free co-occurrence edge weight to rank words that have co-occurrence. A different probabilistic scoring function based on two hypotheses "a good key phrase should be closely related to the given topic" and "a good key phrase should be interesting and can attract users" was proposed and implemented. When extracting key phrases, length preference was incorporated to avoid meaningless key phrases. According to the evaluation of results, proposed context-sensitive page ranking performed better than the standard method as well as the probabilistic ranking method boosting the performance of keyphrase extraction [10].

TwittDict is a novel approach to extract semantically associated words with a target word from a corpus of tweets [12]. This approach addresses the problem of extracting semantically related key phrases from microblogs while solving the matter of language creativity and noise such as nonstandard verbs or symbolic expression [12]. This model first recognizes the topics that have been mentioned in the corpus and checks for the association of those topics for the given words. Then starts mapping to find the semantically associated words. It assumes that a word in a tweet cannot represent its semantic but a set of tweets that contains the same word can give the meaning of the word. Latent Dirichlet Allocation (LDA) [13] is a method of text mining used to represent a document as a set of topics and to identify the topics.

Some of the commonly used approaches to extract keywords from text processing along with NLP tools have been mentioned below. These approaches have been used to extract keywords from both structured and unstructured texts.

A simple statistical approach is rough and has a tendency to work without a training set. It focuses on the statistics obtained from non-linguistic features of the document where these insights can be used to generate keywords. N-gram statistical data can be used to filter the keywords inside the text. It is known as TF-IDF – Term Frequency – Inverse Document Frequency [1, 9], The main criteria is the frequency of occurrence. This statistical information can then be used to find the support and confidence of the word. Later, the keywords are inferred using the Apriori technique which is generally used for frequent itemset mining and association rule learning [2, 14, 15].

Linguistic features of the words are focused on keywords extraction and detection in text documents using the Linguistic approach. It consists of lexical analysis, syntactical analysis and discourse analysis. Electronic dictionary, tree tagger, Wordnet, n-grams, POS patterns are the main resources that are incorporated for lexical analysis [16]. Noun phrases, noun chunks are used as resources for syntactical analysis. This approach is more accurate and computationally intensive. But the limitation of this approach is that it requires domain knowledge [1, 15].

Keyword extraction can be seen in the angle of a learning problem. Therefore, a machine learning approach can be used along with manually annotated training data and training models. Training models such as support vector machine

(SVM) naïve Bayes, bagging and Hidden Markov model (HMM) are commonly used [2,14]. Keyword Generation Algorithm (KEA) is one of the most popular and accurate algorithms which is built upon this approach. The KEA algorithm first converts words into nodes, whenever two words are laid on the same sentence then a vector graph is created by connecting the two nodes. A number of edges is converted into scores and are clustered accordingly. After that cluster heads are treated as keywords and they are categorized into two categories, keyword or not a keyword using Bayes classification [9]. There are two methods that can be used; supervised learning and unsupervised learning [14].

A hybrid approach is a combination of supervised and unsupervised learning or heuristics such as position, length, layout feature of the words, etc. It is designed to extract the best feature from the above-mentioned approaches [2, 14].

A considerable number of algorithms have been described and used for keyword extraction in documents. C4.5 algorithm which is an extension of ID3 algorithm is used in the keyword extraction process. It includes decision tree statistical classification which is more suitable for balanced class attributes [17]. KEA is a keyword generation algorithm which builds upon Naïve Bayes's learning approach and statistical approach called TF-IDF. This algorithm is based on two lexical features named Term Frequency and the position of the keyword. If the occurrence of the word is high in a document there is a high probability for that word to be a keyword. There should be a large data set in order to train this model well [14, 18]. KEA++ is an extension of KEA keyword generating algorithm and it uses three linguistic features to select the keyword. It uses a thesaurus to link the synonyms of the words and filter the words with a high node degree. KEA++ uses structures-controlled vocabulary. The advantage of KEA++ is its use of controlled vocabulary which eliminates the occurrence of meaningless and incorrect word extractions, its performance is dependent on the control's vocabulary [19].

A method is proposed to extract keywords from tweets by using brown clustering which is clustered by brown corpus and continuous word vectors. The clustering algorithm attempts to find clusters with the maximum likelihood. A tweet can have any length from 1 to 280 characters. Therefore, the number of keywords to be extracted must be proportional to the word count in the tweet. This fact is considered in the approach presented in this paper. In the evaluation, precision and recall hold a higher value for the algorithm MAUI which consists of both Browns clustering and Word Vectors than other existing methods [14]. Another method to extract keywords has been proposed by using Backpropagation Neural Network [20]. Corpora that had been used to train the model consist of journal articles. Each and every word is encoded with a set of features such as term frequency. Inverse document frequency (IDF) is not used in this approach since it requires the analysis of the whole system. In order to finetune the system, Backpropagation Neural Network is used through C4.5 algorithm. According to the evaluation results obtained from this approach, BNN is found to predict keywords with 90.11% precision, 59.50% recall and 0.717 F-Measure [21]. A model is proposed using Recurrent Neural Network (RNN) containing two hidden layers to extract keywords using tweets. The first layer captures the information of keyword while the second layer

extracts the key phrase by using information. Evaluation of results proves that this method performs better than the traditional state-of-the-art. [22].

III. METHODOLOGY

A. Dataset

A twitter data set that contains tweet text was obtained, in order to extract candidate keywords and for evaluation purposes. This data set is a combination of tweets from different domains such as health, sports, politics and education. More than 100,000 records from various news categories are available. The intention was to make the system work accurately in any domain. In other words, to make it a generic approach. Data were collected using the Twitter streaming API and directly downloaded from Kaggle.

B. Pre-processing

Fake news detection does not focus on a specific domain therefore, the corpus should contain tweets from different domains in order to find the candidate keywords from a tweet. There are existing corpora that contain twitter data but according to the experiments carried out, they do not contain enough tweets to find the candidate phrase. Therefore, it was decided to create a corpus by combining domain-specific corpora and some manually collected tweets using Twitter streaming API.

Data preprocessing plays a major role when creating a corpus because the proposed approach is implemented using unstructured text that is used on twitter. The process of data pre-processing carried out is depicted in figure 2.

A tweet consists of URLs, punctuations and emoticons which are needed to be removed in the phase of preprocessing. Furthermore, there are words such as pronouns, articles and prepositions that have been frequently used in natural language which must be filtered out at the preprocessing stage because the statistical approach used for keyword extraction takes the most frequent word in the text, so the impact of stop words must be removed. The list of stop words that is used for pre-processing is created by analyzing a previously used stop words list in NLP and some of the abbreviations that are commonly used in social media platforms such as LOL, GM, etc. Usually, those words are not used in tweets that contain news. But this step is included to avoid uncertainty, as the language of social media platforms is evolving and getting more informal according to the user base and the tweeter.

Then Lemmatization is carried out to convert a word to the base form. Otherwise, it will be difficult to identify the same word in a different format [9]. Stemming was not used because it leads to spelling errors and ambiguous words since it removes the last few letters of a word. Then strings are tokenized using NLTK library in order to continue with spelling corrections. There could be intentional or unintentional spelling errors in the tweet content. Python package pypellchecker provides features to find misspelled words and suggest possible corrections. The word with the highest probability will be selected as the corrected version of misspelled words. After following the preprocessing steps, the corpus will be created.

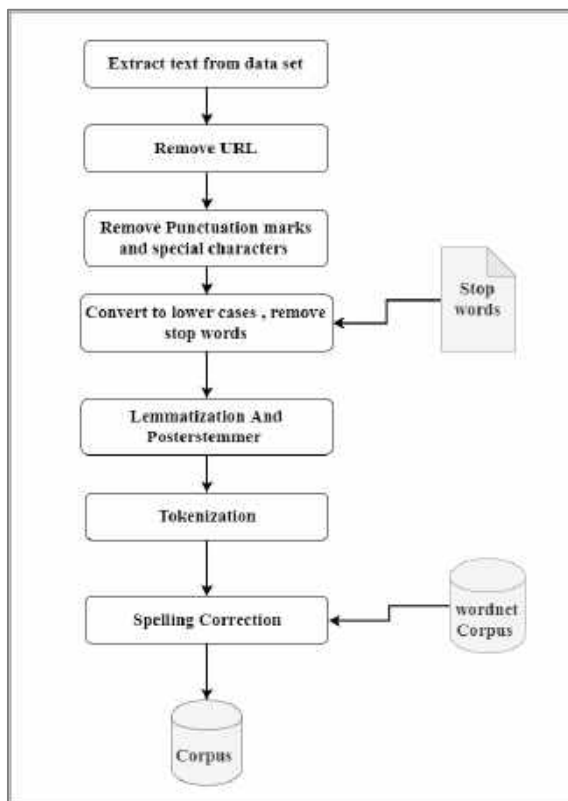


Fig. 2. Data pre-processing

C. Extracting keywords

Once the claimed tweet's text is given to the system it is preprocessed to extract the keywords. The preprocessing steps include removing URLs, special characters, punctuation and stop words. Then, lemmatization is done and the string is tokenized into words. This text might include words that are misspelled, thus spelling correction also has to be done. At the end of this preprocessing, the output will be a set of words containing the keywords.

After pre-processing the claimed tweet text, POS is needed to be done with the intention of using Named Entity Ranking (NER). It is required to apply POS tagging for the NER tagging. Even though there are libraries for POS tagging, every library will not give a high accuracy since it depends on the data set it was trained on. Since this research is on unstructured text, it is better to find a library which is trained on a twitter data [14] set. Stanford CoreNLP toolkit [23] is an extensible pipeline that facilitates natural language analysis and could be used for POS tagging as well as for NER tagging.

Figure 3 depicts the preprocessing steps for the text of the claimed tweet. After Applying the POS tagging, the NER must be applied. Stanford Core NLP toolkit facilitates labeling words into seven classes. Location, person, organization, money and percentage, date and time [24] are the classes of the model. When extracting keywords with the intention of collecting associate news to the given tweet text, location, person and organization play a major role in retrieving associate news that has been circulated in the social media platform (twitter). In fact, these seven classes are the most important keywords in a tweet.

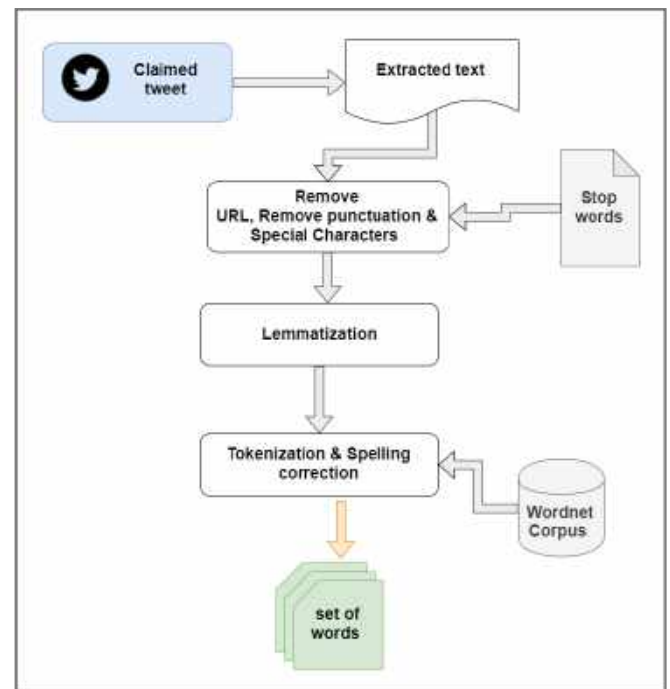


Fig. 3. Preprocessing for the text

Even though these essential keywords can be extracted from the given tweet using Stanford core NLP framework, some of the important keywords can be missed due to the unrecognition of the word. In addition, some of the new locations, persons and organizations will not be captured. As a solution for this issue, a methodology can be proposed to use TF-IDF statistical method along with NER to find the candidate keywords. TF-IDF is based on the frequency of occurrence of a word in the corpus. Since the corpus contains more than 100,000 tweets related to news from different domains, this approach will be a suitable way to reach the goal. The following figure depicts the flow of candidate keyword extraction from a given text.

After extracting the candidate keywords from the given tweet text, it is essential to find synonyms for the words that are not identified by NER tagging to expand the search for associate news posted on twitter. Different users will post the same news in different manners. Wordnet lexical database which has been used by major search engines can be used to retrieve similar words and the value for the similarity of two words is done by using word2vec using Gensim [25]. The words having more similarity can be taken as candidate keywords. To find words with more similarity a threshold value can be defined. to define a threshold value a statistical experiment can be conducted.

To generate key phrases using the above extracted keywords, n-gram can be used and it will give a large number of word combinations if a limitation for the number of keywords to be extracted is not given. Therefore, a threshold value must be decided by testing, to reach the expected output. this will make generating key phrases more effective and easier to handle. The keyword extraction from the given tweet is carried out to extract relevant news articles that have been posted on twitter for fake news detection; to make the relevant news collection more accurate and effective, in other words, to collect news which is truly related to the claimed news.

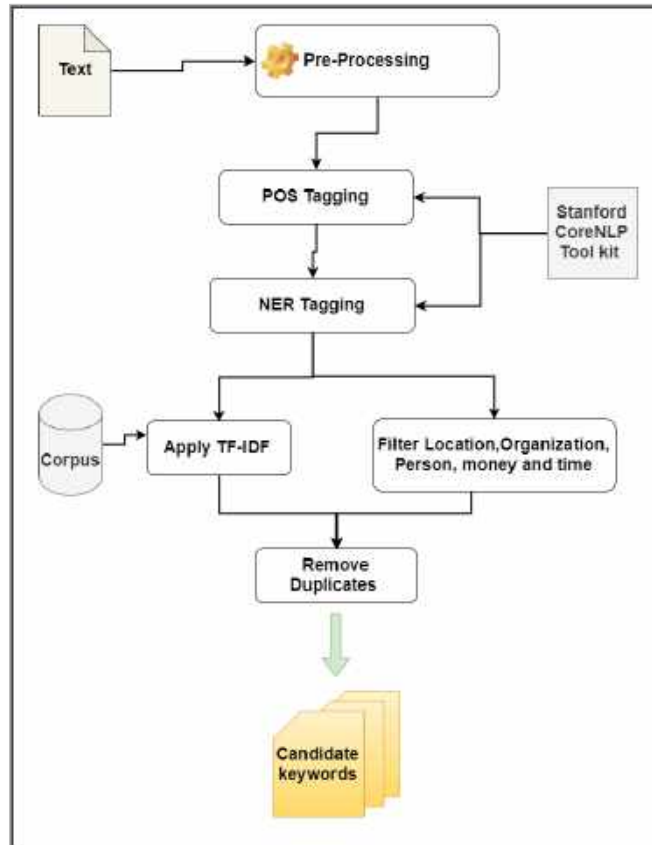


Fig. 4. The flow of candidate keyword extraction

After extracting key phrases, they can be used as parameters for the Twitter streaming API. Key phrases can be assigned to the track variable provided in the Twitter API streaming method. Extracted keywords are saved into an array and assigned to the track variable in the twitter API. Then it will retrieve all the tweets which contain those key phrases. In fact, it will receive tweets that contain at least one of the key phrases or there could be a combination of keywords in those tweets. At least ten thousand records will be collected from twitter API if it is possible. This will not guarantee that each and every tweet received is relevant to the claimed tweet. Therefore, further processing must be conducted to filter the news articles to find those that are more relevant to the claimed tweet.

This process focuses on the most important factors related to the claimed tweet such as person names, locations, organizations, countries and specifically mentioned dates or the received date of the tweet. Before starting the filtering process, it is essential to remove duplicate records. Therefore, tweets that contain the same important factors mentioned above, contain most of the extracted keywords in the text and has a tweet date close to the date on which the claim is made, are filtered from the extracted set of tweets.

IV. EVALUATION

After selecting the proper threshold value to determine and choose the words with more similarity, an evaluation was conducted by comparing the threshold value and the accuracy of the output. Table I depicts the selected values and accuracies.

TABLE I. COMPARING THRESHOLD VALUES WITH ACCURACY

Threshold value	Accuracy of the final output	Assumed Reason
0.5	51.4%	Selecting more synonyms which does not have much similarity
0.69	67.6%	Moderate and provide the most relevant synonyms
0.75	49.2%	Selecting less synonyms even though the similarity is high

According to the statistical experiment carried out, the similarity threshold value was set to 0.69.

Turing test [3] will be used to evaluate the extracted keywords from the given tweet text because there is no other evaluation method that is better than the evaluation carried out by human intervention.

As mentioned above, the Turing test is carried out with human intervention. The intervention of humans to the test is important and the output of the evaluation depends upon the humans who are going to be used for the testing. The humans who will participate in the Turing test will be experts in the language. Since the system is about news posted on twitter and the system is a generic model that can be applied to any domain, people who write English articles can be used. Social media platforms are commonly used by the younger generation of the country. Therefore, undergraduates involved in the area of language will be suitable for the Turing test.

Table II depicts the participants of the Turing test; human keyword generators used to build the test dataset.

TABLE II. PARTICIPANTS OF THE TURING TEST

Type of Participants	Justification	Minimum Qualification
English article writers	They have a vast knowledge on language manipulation for article writing on any area.	Active journalists
Graduates	Highly experienced professionals on using English in academics and professional fields	Completed a bachelor's degree (related to language)
Undergraduates	Knowledge and experience on using English language for academics	Reading for a bachelor's degree (related to language)
General public	Uses of Social media frequently	Daily twitter users

A test data set was created by manually annotating 500 tweets with keywords and the output for each tweet was taken through the proposed approach. Then a comparison was carried out and the accuracy of the proposed approach was determined.

When calculating the accuracy of the system there are two main factors to be considered.

1. Number of keywords
2. Extracted keywords including synonyms

Table III depicts the types of records used with the assumptions.

TABLE III. ASSUMPTION AND RECORD TYPES

Record type	Number of keywords	Number of matches	Status
1	Equal	All are matched	Identical
2	Equal	Less than 80%	Negative
3	Greater than the human generated keyword set	Greater than 80%	Positive
4	Greater than the human generated keyword set	Less than 80%	Negative
5	Less than the human generated keyword set	-	Negative

The accuracy was calculated using the following equation (1).

$$\text{Successful percentage} = \frac{X + Y}{N} * 100 \% \quad (1)$$

X = Identical records Y = Positive records

N = total number of records (tweet set)

According to the data set and the evaluation conducted there were 107 identical records and 231 positive records and the total percentage of accuracy was 67.6%. Since this keyword extraction consists of words that have similarities, the accuracy is dependent on the intelligence, knowledge and vocabulary of the people who were selected for the Turing test.

Moreover, this evaluation can be conducted separately for each type of participants, then the accuracy can be gained according to each category separately.

Evaluation by comparing existing methods to extract keywords from unstructured text such as using Stanford core NLP [3], automatic keyword extraction using brown clustering [14], Rapid Automatic Keyword extraction algorithm (RAKE) and existing methods to extract keywords from structured text such as Keyword Extraction Algorithm KEA cannot be conducted because those models do not extract keywords with the suitable synonyms. If those approaches are evaluated against the proposed approach a low accuracy will be given

The tweets that are collected using the extracted keywords and twitter API are filtered to extract the most relevant set of tweets for the claimed tweet.

V. CONCLUSION

It takes some time for a person to read the whole text and get an idea about its content, but, keywords have made this task easy. A keyword is a word that represents the whole idea of the text. Therefore, attention has been paid on keyword extraction methods from text using NLP tools. It is easy to extract keywords from structured documents by considering their linguistic features. But for unstructured text such as tweets, it has become difficult because linguistic features cannot be found due to character limitations. In fact, the language pattern is different from those informal documents.

This paper proposes a methodology to extract keywords from a given tweet text for the purpose of retrieving relevant news that has been posted on twitter to collect data for fake news detection. The proposed method uses Stanford core NLP, POS tagging, NER as well as TF-IDF statistical method for keyword extraction. In Addition, Wordnet lexical database has been used to find synonyms and Ginsim can be used along with word2vector to find synonyms for the words and the amount of similarity of words. Then the bi-gram technique is used to generate key phrases to increase the accuracy and efficiency for retrieving relevant news. Extracted keywords are used to gather the most relevant news tweets for the claimed tweet. In fact, the set of tweets retrieved were filtered and duplicates were removed to get a clean set of tweets to help detect fake news. The Evaluation can be done using the Turing test and more attention should be paid to the standards of the participants of the test.

A dedicated corpus has been implemented with more than 100000 tweet news from different domains such as sports, politics, etc. Stanford core NLP toolkit was selected because it's also built upon using a set of tweets. In the proposed method, essential keywords such as the name of a person, location, organization, date and time will not be missed in the array of candidate keywords even though they are not captured by the TF-IDF method. Most importantly, this method is a generic method which does not depend on a specific domain.

As future work, the method which is used to find synonyms and similarity of two words can be modified to give a more precise output which will lead to an increment of performance.

ACKNOWLEDGMENT

The author of this paper would like to express gratitude for the supervisor for her immense support and guidance to make this paper a success.

REFERENCES

- [1] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review", *International Journal of Computer Applications*, vol. 109, no. 2, pp. 18-23, 2015. Available: https://www.researchgate.net/publication/272372039_Keyword_and_Keyphrase_Extraction_Techniques_A_Literature_Review. [Accessed 3 June 2019].
- [2] S. K. Bharti and K. S. Babu "Automatic Keyword Extraction for Text Summarization: A Survey", *CoRR*, vol. 170403242, 2017. Available: <http://arxiv.org/abs/1704.03242>. [Accessed 5 June 2019]
- [3] T. Weerasooriya, N. Perera and S. Liyanage, "A method to extract essential keywords from a tweet using NLP tools", *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions*, pp. 29-34, 2016. [Accessed 30 May 2019].
- [4] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr.*, pp. 55-60, 2019. Available: <https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf>. [Accessed 7 October 2019].
- [5] R. Levy and G. Andrew, "Tregex and Tsurgeon: tools for querying and manipulating tree data structures", *5th Int. Conf. Lang. Resour. Eval. (LREC 2006)*, pp. 2231-2234, 2006. Available: <http://www.mit.edu/~rplevy/papers/levy-andrew-2006.pdf>. [Accessed 7 December 2019].
- [6] [Levenshtein distance, [Online] Available: https://rosettacode.org/wiki/Levenshtein_distance. [Accessed: 07-Oct- 2019].

- [7] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification", *arXiv:1607.01759v3 [cs.CL]*, 2016. Available: <https://arxiv.org/pdf/1607.01759.pdf>. [Accessed 7 December 2019].
- [8] [Online]. Available: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTag%ger.html>. [Accessed: 07- Oct-2019].
- [9] Amit Kumar Mondal and Dipak Kumar Maji "Improved Algorithms For Keyword Extraction and Headline Generation From Unstructured Text", p. 14. Available: <https://pdfs.semanticscholar.org/d177/faa6f8c92c19e9b54ab1eaa94b510482425a.pdf>. [Accessed 5 June 2019]
- [10] Zhao, Wayne Xin and Jiang, Jing and He, Jing and Song, Yang and Achananuparp, Palakorn and Lim, Ee-Peng and Li, Xiaoming, "Topical Keyphrase Extraction from Twitter", *Association for Computational Linguistics*, pp. 379-388, 2011. Available: <http://dl.acm.org/citation.cfm?id=2002472.2002521>. [Accessed 5 June 2019].
- [11] Y. Liu, K. Young, Y. Curtis, B. Aragona and Z. Wang, "Social Bonding Decreases the Rewarding Properties of Amphetamine through a Dopamine D1 Receptor-Mediated Mechanism", *J Neurosci*. 2011, pp. 7960-7966., 2011. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114880/pdf/zns7960.pdf>. [Accessed 7 October 2019].
- [12] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin and Craig G. Nevill-Manning "KEA: Practical Automatic Keyphrase Extraction", vol. 23, 2015. Available: https://www.cs.waikato.ac.nz/~ml/publications/2005/chap_Witten-et-al_Windows.pdf. [Accessed 28 August 2019].
- [13] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research* 3 (2003), pp. 993-1022, 2003. Available: <http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>. [Accessed 8 October 2019].
- [14] Marujo, Luís & Ling, Wang & Trancoso, Isabel & Dyer, Chris & Black, Alan & Gershman, Anatole & Martins de Matos, David & Neto, João & Carbonell, Jaime. (2015). Automatic Keyword Extraction on Twitter. 2. 637-643. 10.3115/v1/P15-2105. Available: https://www.researchgate.net/publication/283816235_Automatic_Keyword_Extraction_on_Twitter. [Accessed 5 June 2019]
- [15] Z. Zhu, M. Li, L. Chen, Z. Yang and S. Chen, "Combination of Unsupervised Keyphrase Extraction Algorithms", *2013 International Conference on Asian Language Processing, Urumqi, 2013*, pp. 33-36., 2019. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6645997&isnumber=6645979>. [Accessed 5 June 2019].
- [16] M. Abulaish and T. Anwar "A SUPERVISED LEARNING APPROACH FOR AUTOMATIC KEYPHRASE EXTRACTION", *International Journal of Innovative Computing, Information and Control*, vol. 8, 2012. Available: <https://pdfs.semanticscholar.org/e36e/2955cd30917d62091ac5ef15bf71eed84616.pdf>. [Accessed 26 August 2019].
- [17] Turney, Peter "Learning Algorithms for Keyphrase Extraction", *Inf. Retr.*, vol. 2, pp. 303-336, 2000. Available: https://www.researchgate.net/publication/220479857_Learning_Algorithms_for_Keyphrase_Extraction/citation/download. [Accessed 5 July 2019].
- [18]] I. Witten, G. Paynter, E. Frank, C. Gutwin and C. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction", *Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA*, pp. 254-255, 1999. Available: https://www.cs.waikato.ac.nz/~ml/publications/2005/chap_Witten-et-al_Windows.pdf. [Accessed 4 May 2019].
- [19] O. Medelyan, & I. Witten (2006). Thesaurus based automatic keyphrase indexing. 296 - 297. 10.1145/1141753.1141819. Available: https://www.researchgate.net/publication/224061561_Thesaurus_base_d_automatic_keyphrase_indexing [Accessed 4 May 2019].
- [20] Taemin Jo, Jee-Hyong Lee "Latent Keyphrase Extraction Using Deep Belief Networks", *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 15, pp. 153-158, 2015. Available: <https://www.e-sciencecentral.org/articles/SC000013961>. [Accessed 28 August 2019].
- [21] J. P. Tensuan, A. Azcarraga "NEURAL NETWORK BASED KEYWORD EXTRACTION USING WORD FREQUENCY, POSITION, USAGE AND FORMAT FEATURES", *Research Congress 2012 De La Salle University*, 2013. Available: <https://pdfs.semanticscholar.org/9302/03b85f789972107df4bc08da77632e477b84.pdf>. [Accessed 5 June 2019].
- [22] Wang, Yang & Gong, Yeyun & Huang, Xuanjing. (2016). Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter. 836-845. 10.18653/v1/D16-1080. Available: https://www.researchgate.net/publication/311990419_Keyphrase_Extraction_Using_Deep_Recurrent_Neural_Networks_on_Twitter. [Accessed 10 June 2019].
- [23] Manning, Christopher & Surdeanu, Mihai & Bauer, John & Finkel, Jenny & Bethard, Steven & McClosky, David. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52Nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 10.3115/v1/P14-5010. Available: https://www.researchgate.net/publication/272091377_The_Stanford_CoreNLP_Natural_Language_Processing_Toolkit [Accessed 10 June 2019]
- [24] T. Weerasooriya, N. Perera, S.R. Liyanage, (2017). KeyXtract Twitter Model - An Essential Keywords Extraction Model for Twitter Designed using NLP Tools. Available: https://www.researchgate.net/publication/319035921_KeyXtract_Twitter_Model_An_Essential_Keywords_Extraction_Model_for_Twitter_Designed_using_NLP_Tools. [Accessed 5 June 2019].
- [25] [Online]. Available: <https://www.guru99.com/word-embedding-word2vec.html>. [Accessed: 20- Nov- 2019]

Sentiment classification of Sinhala content in social media

Pradeep Jayasuriya*
Faculty of Computing
Sri Lanka Institute of Information
Technology, Sri Lanka
pradeep.jayasuriya@my.sliit.lk

Bihara Kumarasinghe
Faculty of Computing
Sri Lanka Institute of Information
Technology, Sri Lanka
bihara.kumarasinghe@my.sliit.lk

Sarith Ekanayake
Faculty of Computing
Sri Lanka Institute of Information
Technology, Sri Lanka
sarith.ekanayake@my.sliit.lk

Isuru Weerasinghe
Faculty of Computing
Sri Lanka Institute of Information
Technology, Sri Lanka
don.isuru@my.sliit.lk

Ranjiva Munasinghe
Faculty of Business
Sri Lanka Institute of Information
Technology, Sri Lanka
ranjiva.m@sliit.lk

Samantha Thelijagoda
Faculty of Business
Sri Lanka Institute of Information
Technology, Sri Lanka
samantha.t@sliit.lk

Abstract: In this study, we focus on the classification of Sinhala social media sentiments into positive and negative classes for a particular domain (sports). We have employed machine learning algorithms and lexicon-based sentiment classification methods. We also consider a hybrid approach by constructing an ensemble classifier in which we combine Machine Learning and Lexicon based methods. For individual methods, machine learning algorithms performed best in terms of accuracy. The ensemble classifier was able to improve performance further.

Keywords: Machine Learning, Natural Language Processing, Sentiment Analysis, Social Media, Supervised Learning

I. INTRODUCTION

Social Media has a major impact on the world today with global usage in 2018 estimated to be 2.65 billion. Social media has become the major platform where people share their opinions on various topics such as products, services, people, places, organizations, events, news, ideas etc. Many insights can be gained from understanding what is being said on social media – e.g. from a business perspective, social media is a great source for understanding where their products or services are positioned among the customers. Accordingly, social media sentiment analyzing researches have been conducted [1], [2], [3], [4] and tools have been developed for popular languages such as English (e.g. Social Studio, Hootsuite etc.) which can provide insights for businesses to improve their products and business processes. Social media monitoring is also important for monitoring social unrest [5].

In the local context, Sri Lanka has over 6 million social media users which represent close to 30% penetration. In particular, social media users expressing their opinions in the Sinhala language has also increased significantly. At present,

there are some research efforts at the state universities to use Natural Language Processing (NLP) on the Sinhala language, in particular focusing on morphological analysis on traditional texts and machine translations e.g.: Multilingual document generation. To the best of our knowledge, the work done on analyzing Sinhala content in social media is limited¹, particularly in polarity classification of sentiments i.e. classifying sentiments into positive negative and neutral classes. This study has classified Sinhala sentiments using a dictionary-based lexicon method [6].

Sentiment analysis is an area of study within NLP for extracting sentiments from text via automated techniques. Opinion mining and sentiment analysis are well-established in linguistic resource-rich languages (e.g. English). The success of an opinion mining approach depends on the availability of resources, such as special lexicons, coding libraries, and WordNet type tools for a particular language. Due to the lack of such resources, it is more difficult to analyze the sentiments of languages that are less commonly used like Sinhala [7]. Another challenge for NLP analysis for Sinhala (and other Indic languages) is due to their diglossic nature - whereby there are both formal and informal dialects of the same language which are very different. It is the informal language that is more frequently used in Sinhala content on social media. The domain is also important, as algorithms that are trained for one particular domain provide poor results in a different domain. Other challenges include the use of English loan words (written using the letters from the Modern Latin alphabet) and the use of 'Singlish' – where Sinhala words are spelled out phonetically in English. A more complete list of challenges in Indic languages can be found in [5].

There is two main sentiment analysis approaches Machine Learning (ML) based [8] and lexicon-based

¹ There are a considerable number of studies on Hate Speech and Racism detection for Sinhala on Social media

sentiment analysis. In this study, we have implemented both these methods. Lexicon-based sentiment analysis is generally not used for social media sentiment analysis due to the use of informal language in social media. We have proposed a new approach to lexicon library construction as a solution to this matter. We also employ an ensemble approach [9] to improve the predictive performance of the sentiment analysis, in which we combine both machine learning and lexicon-based methods as a hybrid sentiment analysis approach. Sentiment classification in all the methods we have employed in this paper are supervised approaches [10]. In particular, we work with a binary classification of sentiments into positive and negative (polarity) classes. YouTube is selected as the social media platform and ‘sports’ is the selected domain of this study. We have focused on comment-level sentiment classification where a comment contains one or several sentences that are considered a single entity by the sentiment analysis process.

The paper is structured in the following manner – we begin with the methodology section where we discuss the construction of the sentiment analysis model. In particular, we describe the dataset, data pre-processing, feature extraction and different analysis approaches taken. The next section discusses our results and findings. The paper ends with a summary and discussion of the current study and our planned future work.

II. METHODOLOGY

This section describes the construction of the sentiment analysis model for analyzing Sinhala social media content. It involves data tokenization, pre-processing, feature extraction and sentiment analysis. Python is used as the language for the development of this model.

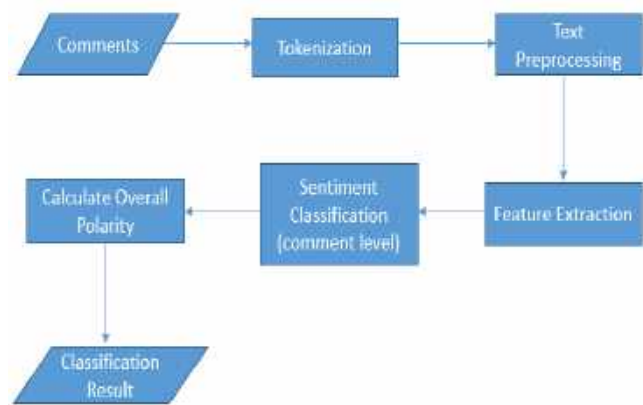


Fig. 1. Sentiment analysis flow chart.

A. Data-set description

Comments were obtained from sports-related videos (involving cricket, rugby, and athletics) from YouTube. The next step was to label these comments into classes (positive or negative) by sentiment, resulting in a dataset suitable for supervised learning. When creating the dataset, longer comments (comments with more than five sentences) were manually split in a way that a split contains a complete and independent sentiment. We also ensured the dataset allowed for stratified sampling. A total of 2210 comments were grouped as follows for training and testing purposes.

TABLE I. DATASET DESCRIPTION

	Train set	Test set	Total
Positive comments	830	275	1105
Negative comments	830	275	1105
Total	1660	550	2210

Source: [10]

The dataset consists of 2810 total sentences and 1346 of them are distributed in the 1105 positive comments and the remaining 1464 sentences are distributed in the 1105 negative comments. There is a total of 21,573 words in the dataset. They are distributed as 8389 words in positive comments and 13,184 words in negative comments.

B. Data pre-processing

The first step in this stage is tokenization. Text is first tokenized as comments and then into words for carrying out further analysis. Text cleaning follows tokenization, where only the main Sinhala characters were considered and all non-Sinhala characters, punctuation and numerical text were removed from comments.

After the initial cleaning, stop words were removed from the text. Stop word removal is an important task in sentiment analysis and was first introduced by Hans Luhn [11]. Stop words are common words with a high term frequency in a document that does not have any sentimental value. There are different methods available for stop-word removal [12], and the stop word removal greatly enhances the performance of the feature extraction algorithm [1, 13]. Removing stop words also reduces the dimensionality of the data sets. It will leave key opinion words which will make the sentiment analyzing process more accurate. Stop words are taken from a customized list of stop words for the particular domain. At the simplest level stop, words are iterated in a word list and removed from the text.

C. Feature extraction

Feature extraction is a very important task in sentiment analysis because the accuracy of the analysis depends on it. We have used the word N-gram method for feature extraction - in particular unigram, bigram and trigram features [14]. N-gram feature vectors are created as a bag of n-grams representations from tokenized and pre-processed text.

D. Machine learning-based sentiment analysis

In our approach we start by using different ML algorithms for our classification: Naïve Bayes classifier (NB), Logistic Regression Classifier (LR) and support vector machine classifier (SVM) [13]. Each algorithm is trained using unigram, bigram and trigram feature extraction methods to train 3 classifiers for each algorithm – resulting in 9 classification methods. In a subsequent section, we describe creating an ensemble classifier based on combinations of the above base learners and lexicon-based classifier.

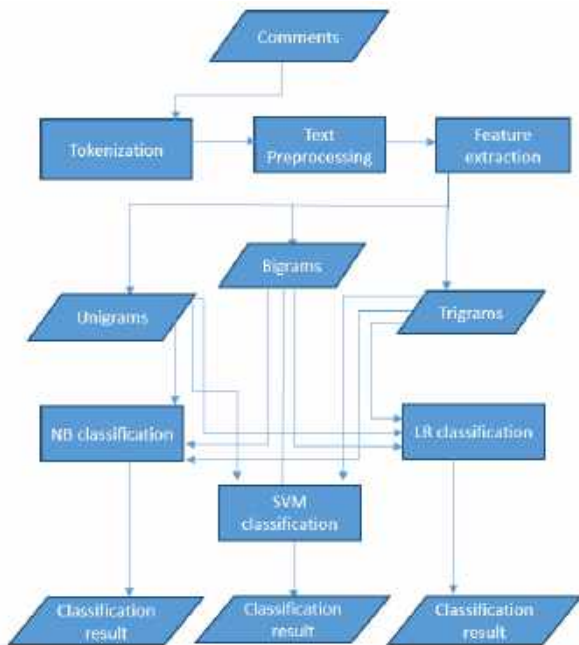


Fig.2..ML-based sentiment analysis: Base learner training

E. Lexicon-based sentiment analysis

We describe the lexicon-based method in a little more detail as we have combined existing approaches in a novel way. This section starts with 1) lexicon library construction which consists of two subsections: 1.1) subjective lexicon construction and 1.2) expanding subjective lexicon. Expanding subjective lexicon involves 1.2.1) string similarity analysis and 1.2.2) mismatches removal strategy. Finally the 2) lexicon-based sentiment classification is discussed.

For the lexicon-based analysis, we employ a Bayesian analysis method [15] and improve upon this by using the Levenshtein ratio analysis for string similarity analysis [16].

1) Lexicon library construction

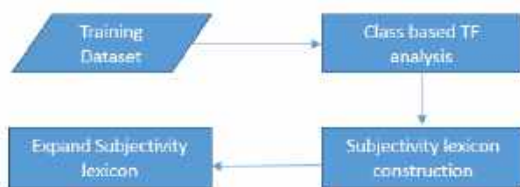


Fig.3. Lexicon library construction

In our approach of lexicon library construction, we have constructed a subjectivity lexicon [17] using the training data set and expanded it by using a Sinhala-English dictionary.

a) Subjective lexicon construction

In the literature, a list of words with the polarities assigned to each word in the list is called a subjective lexicon [18]. We calculate a positive and negative probability score for each word i.e. unigram feature in the training data set based on its term frequency (TF) in each class. The difference in each class for each word is also calculated to construct a difference score. These scores are then assigned as sentiment scores for a particular word. From the training dataset, we

obtained a subjective lexicon consist of 4389 words at the end of this analysis.

$$\text{Positive polarity score} = \frac{\text{Frequency of positive}}{\text{Total positive comments}} \quad (1)$$

$$\text{Negative polarity score} = \frac{\text{Frequency of negative}}{\text{Total negative comments}} \quad (2)$$

$$\text{Difference} = \text{Positive polarity score} - \text{Negative Polarity Score} \quad (3)$$

As the Difference score of a word increases, it has a higher sentiment value. The difference score is used for sorting the subjective lexicon so that the sentiment analyzing process can identify its significance. Bigram and trigram features will not be considered for lexicon-based analysis.

b) Expanding subjective lexicon using a Sinhala-English dictionary

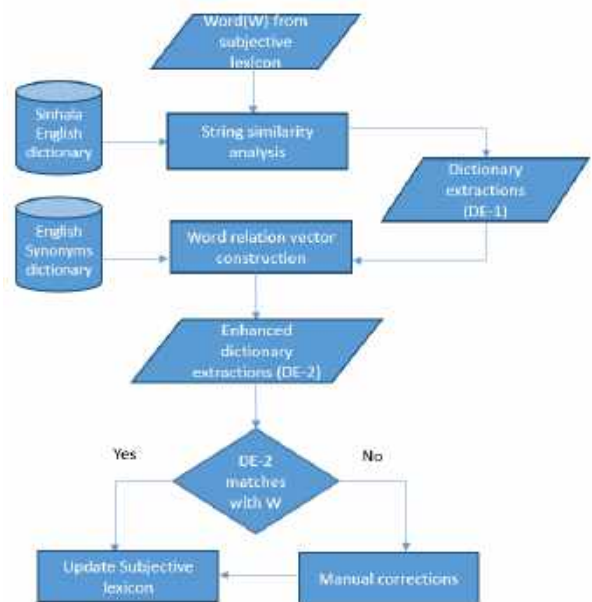


Fig. 4. Expanding subjective lexicon

Subjective lexicon is expanded using string similarity analysis upon a Sinhala English dictionary and by employing an additional mismatches removal strategy.

I. String similarity analysis

Each word in the constructed subjective lexicon is checked against a Sinhala-English dictionary for String similarity analysis. The Sinhala-English dictionary has a structure as each English word has one or more Sinhala meanings. The purpose of string similarity analysis is to extract synonyms and similar words (extracting different words of the same lemma) for words in the subjective lexicon from the dictionary. We have employed Levenshtein ratio analysis [16] for the purpose of string similarity analysis. Levenshtein ratio analysis provides matched entries from the

dictionary for a particular word in the subjective lexicon as a result.

e.g.: A word from the subjective lexicon: 'ආදරයි'

Dictionary extractions from the Levenshtein ratio analysis for 'ආදරයි':

- 'affection': 'ආවේදනාව', 'චිත්තය', 'අනුරාගය', 'ආදරය', 'ආබාධය', 'දයාව', 'රෝගය', 'ස්තේනය'
- 'love': 'ආදරය', 'ආදරය කරනවා', 'ආලය', 'ආලය කරන වස්තුව', 'පෙම්වතිය', 'මෙමන්රිය', 'ස්තේනය', 'සෙනෙහස'
- 'respect': 'ගෞරවය', 'ආකාරය', 'ආදරය', 'උපහාරය', 'ගරු කරනවා', 'විෂයය', 'සම්මානය', 'සැලකිලි දක්වනවා'

Underlined words are the detected words from Levenshtein ratio analysis. Sinhala synonyms from each dictionary extraction are collected to update the subjective lexicon. In this step, it is important to exclude duplicates that are already present in the subjective lexicon.

II. Mismatches removal strategy

Results provided from the Levenshtein ratio analysis sometimes can be inaccurate because of the following reasons.

- Collected synonyms from dictionary extractions may include irrelevant words due to language differences [6].

e.g.: Word from the subjective lexicon: ඉක්මනි (Means 'Quick' in Sinhala)

Dictionary extractions from String similarity analysis:

- කඩිසර, ශීඝ්‍ර (Means 'Quick' in Sinhala)
- උපවාසය, උපවාසයෙහියෙදෙනවා (Means a Hunger strike in Sinhala)

These errors will be corrected manually at the end of this analysis.

- String similarity analysis may detect irrelevant Sinhala words. Even if the two words are unrelated, Similarity in the structure (Letters and their order which constructs the word) of the two irrelevant words is the reason for this issue.

In order to overcome b, English words of the dictionary word extractions ('affection', 'love', 'respect' as for the above example) are leveraged. For a particular Levenshtein ratio analysis result of a word in the subjective lexicon, resulting dictionary extractions' English words are checked for English synonyms from an online English synonyms dictionary using python web automation. Each resulting English synonym set of each English word from the dictionary extractions is then analyzed by checking for common synonyms between them to construct word relation vectors. In this regard 1) If word A and B are having a common synonym and 2) B and C is having a common synonym A and C is also considered as having a common synonym for indicating their relation. As for this example, words A, B and C will be included in the same word relation vector. These word relation vectors

indicate the relations between the dictionary extractions. From the constructed word relation vectors, we considered the vector with the highest number of words as the mismatches removed vector. (Sometimes this hypothesis may cause inaccurate dictionary extractions. They will be manually corrected at the end of this analysis)

Finally, the Sinhala words are picked from the selected vector's dictionary extractions and they are assigned the same polarity scores of the Analyzed Sinhala word in the subjective lexicon.

e.g.: As for the above example since 'affection', 'love' and 'respect' are words which have English synonyms in common, all corresponding Sinhala words are selected and subjective lexicon is updated as follows.

ආදරයි 0.308 5.218 0.12 |ආවේදනාව|චිත්තය|අනුරාගය|ආදරය|

Fig. 5. Updating subjective lexicon

This analysis provided Sinhala synonym sets for 789 words in the subjective lexicon and 138 manual corrections were applied for them.

2) Sentiment classification using lexicon library

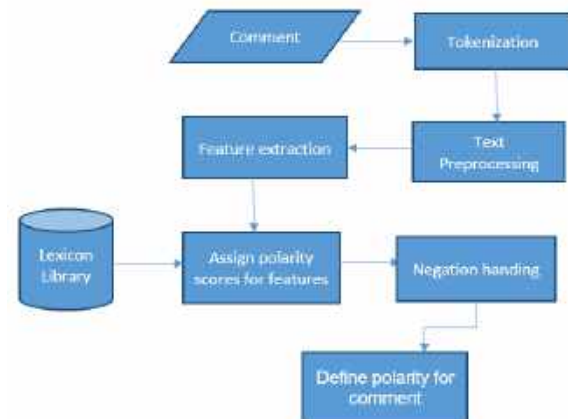


Fig. 6. Lexicon based sentiment analysis

Words in a comment are traversed through the lexicon library and if a particular word is present in the library it's positive and negative sentiment scores will be collected to construct a positive and a total negative score for the comment. These total positive and negative scores are then compared and the score with the higher absolute value is selected as the final sentiment score. If the higher value is positive, the comment is classified as positive and if the higher value is negative, the comment is classified as negative.

Negation handling is another important task which increases the accuracy of lexicon-based sentiment analysis. Negation shifters (Fig.7) change the polarity of the features around them. A negation shifter is associated with a feature word within its context distance and the polarity of the feature word is shifted by the negation shifter. We have considered context distance as the first 2 words before the negation shifter. A negation shifter will multiply the value of any feature within the context distance of the negation shifter by a factor of -2 in our analysis process as negation rules used for the Sinhala language in [18].

නැ, නැ, නැහැ, නැත, නැති (all means no or not)
බැ, බැ, බැහැ, බැරි, බැරිම (all means can't)
එහෙ (don't)

Fig. 7. Selected negation shifters used for lexicon-based sentiment analysis

F. Ensemble classifier

We combine ML and lexicon-based methods using a majority voting ensemble classifier. A comment is classified by each base learner. The list of base learners are as follows:

- 4 Unigram classifiers :
 - 1) Lexicon-based classifier, 2) NB unigram classifier, 3) LR unigram classifier, 4) SVM unigram classifier
- 3 Bigram classifiers :
 - 5) NB bigram classifier, 6) LR bigram classifier, 7) SVM bigram classifier
- 3 Trigram classifiers :
 - 8) NB trigram classifier, 9) LR trigram classifier, 10) SVM trigram classifier

The final classification is determined by the majority vote of the base learners incorporated in the ensemble. We compare different ensembles by considering different combinations of the base learners.

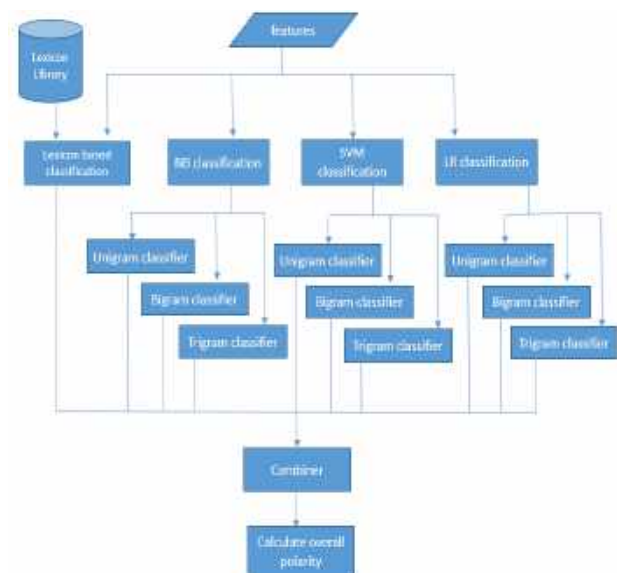


Fig. 8. Combining base learners using an Ensemble classifier.

III. RESULTS

We present our results which are the percentages of sentiments in the testing set that were predicted correctly by the model. F1-score and Accuracy are used as evaluation metrics. Both of these metrics range between 0 and 1, with higher values indicating better classification/prediction.

TABLE III. INDIVIDUAL ANALYSIS METHODS COMPARISON

Classifier		With stop word removal		Without stop word removal	
		F1 Score	Accuracy-Score	F1 Score	Accuracy-Score
Naïve Bayes	Unigram	0.73	0.78	0.71	0.77
	Bigram	0.66	0.61	0.64	0.54
	Trigram	0.61	0.53	0.60	0.50
	Unigram	0.74	0.73	0.74	0.73

SVM	Bigram	0.66	0.56	0.43	0.55
	Trigram	0.57	0.50	0.56	0.47
Logistic Regression	Unigram	0.77	0.78	0.76	0.76
	Bigram	0.67	0.60	0.43	0.59
	Trigram	0.62	0.54	0.61	0.51
Lexicon based analysis	Unigram	0.70	0.72	0.64	0.68

The results indicate that removing stop words improves prediction metrics. Bigram classifiers are the most positively affected classifiers from stop word removal.

For the majority voting ensemble classifier we found that even though the stop word removal improved the base learner prediction metrics, it reduced the corresponding metrics of the ensemble classifiers by 1-2%. The majority voting ensemble classifier performed the best (including improving upon the base learner performance) without stop word removal.

TABLE III. MAJORITY VOTING METHODS COMPARISON

Majority voting Ensemble classifier (without stop words removal)	NO. of base learners	F1 Score	Accuracy-Score
1. Unigram classifiers only	3	0.762	0.775
2. Unigram and bigram classifiers only. (LR bigram classifier excluded)	5	0.781	0.797
3. Unigram and bigram classifiers with Lexicon based classifier	7	0.786	0.800
4. Unigram, Bigram and Trigram classifiers only	9	0.768	0.756
5. one of three trigram classifiers excluded from 10 base learners	9	0.787	0.790

The results indicate the best f1 score and accuracy-score combinations were given by the following two combinations of base learners: 1) Unigram and bigram classifiers with lexicon-based classifier 2) Unigram, bigram and trigram classifiers with lexicon-based classifier.

IV. CONCLUSIONS & FUTURE DIRECTIONS

In this research, we focused on analyzing Sinhala sentiment by considering comments on sports-related videos on the YouTube social media platform. We employed both ML algorithms and lexicon-based methods for sentiment analysis. In addition, we considered a hybrid approach by combining the above methods using a majority vote ensemble classifier.

In terms of feature extraction, the unigram method was the most effective. N-grams (N>1) tend to improve language coverage [19] and performance when the corpus is larger [19]. Based on our results Logistic Regression unigram classifier with stop-word removal was the most accurate among individual ML algorithms according to the F1-score and accuracy metrics.

For the lexicon-based methods, the best performance was obtained by the use of Levenshtein ratio analysis with the

Sinhala – English dictionary to expand the subjective lexicon. This method also incorporated stop-word removal.

When considering individual methods, ML methods were more accurate than the lexicon-based methods.

For the majority vote ensemble classifier, we found that not removing stop words provides better performance. A possible explanation for this is that we are creating a strong classifier from an ensemble of weak base learners, whereas an ensemble of strong base learners may not show improvement over the strongest base learner. We also found that hybrid sentiment classification (the combination of ML and lexicon) was more accurate than ML or lexicon-based methods alone. The lexicon-based approach helps to improve the diversity of the base learners used for the ensemble classifier which helps for employing further ensemble techniques.

For future work, we hope to explore the following paths of research:

- Using other base learner algorithms (e.g. kNN Clustering, Random forest).
- Using further ensemble learning methods such as boosting (AdaBoost, Gradient boosting, extreme gradient boosting and light gradient boosting) and stacking methods.
- Employing character N-grams as features.
- Developing a combination of rule-based methods (e.g. POS tagging) and ML algorithms using linguistic features of the Sinhala language.
- Use of Deep Learning and other Neural Network-based methods.

APPENDIX

In each step of the lexicon construction, performance improvements were recorded as follows (with stop word removal).

TABLE IV. LEXICON BASED ANALYSIS

Steps of lexicon construction	F1 Score	Accuracy
1. Initial subjective lexicon (without expanding using Sinhala English dictionary)	0.60	0.62
2. Expanding subjective lexicon: without manual corrections	0.67	0.71
3. Expanding subjective lexicon : with manual corrections	0.70	0.72

ACKNOWLEDGMENT

We thank the Faculty of Graduate Research Studies - Sri Lankan Institute of Information Technology for their partial support for this research

REFERENCES

- [1] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," in *Proceedings of the 19th National Conference on Artificial Intelligence*, San Jose, California, 2004.
- [2] F. Neri, C. Aliprandi and M. Cuadros, "Sentiment Analysis on Social Media," in *International Conference on Advances in Social Networks Analysis and Mining*, 2012/08/28.
- [3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Language in Social Media (L4S) 2011*, Portland, Oregon, Association for Computational Linguistics, 2011-June, pp. 30-38.
- [4] S. Jayasanka, T. Madhushani, E. Marcus, . I. Aberathne and S. Premaratne, "Sentiment Analysis for Social Media," in *Information Technology Research Symposium*, 2013/11/22.
- [5] P. Bhattacharyya, H. Murthy, R. Munasinghe and S. Ranathunga, "Indic language computing," in *Communications of the ACM*, vol. 62 no. 11, November 2019, pp. 70-75.
- [6] N. Medagoda, S. Shanmuganathan and J. Whalley, "Sentiment Lexicon Construction Using SentiWordNet 3.0," in *2015 11th International Conference on Natural Computation (ICNC)*, Auckland, 2015.
- [7] N. de Silva, "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research," Cornell University, 05 06 2019. [Online]. Available: <https://arxiv.org/abs/1906.02358>. [Accessed 03 07 2019].
- [8] K. Al-Barzaji and A. Atanassov, "BIG DATA SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHMS," in *26 th International Symposium "Control of Energy, Industrial and Ecological Systems"*, Bankia , Bulgaria, 2018.
- [9] Y. Yang, "Chapter 4 - Ensemble Learning," in *Temporal Data Mining Via Unsupervised Ensemble Learning*, Elsevier, 2017, pp. 35-56.
- [10] J. E. T. Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, pp. 128 -138, 08 06 2017.
- [11] H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159-169, 1958.
- [12] J. Kaur and P. K. Buttar, "STOPWORDS REMOVAL AND ITS ALGORITHMS BASED ON DIFFERENT METHODS," *International Journal of Advanced Research in Computer Science*, Vols. Vol 9, No 5, pp. 81-88, 20 10 2018.
- [13] D. Ly, K. Sugiyama, Z. Lin and M.-Y. Kan, "Product Review Summarization from a Deeper Perspective," in *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL*, Ottawa, ON, Canada, 2011/01/01, pp. 311-314.
- [14] A. Esuli and . F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," in *5th Conference on Language Resources and Evaluation*, Genoa, Italy, 2006 - may.
- [15] T. Hastie, R. Tibshirani and J. Friedman, "8.3 Bayesian Methods," in *The Elements of Statistical learning*, 2 ed.
- [16] S. Zhang, Y. Hu and G. Bian, "Research on string similarity algorithm based on Levenshtein Distance," in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 25-26 March 2017.
- [17] S. Park and . K. Yanggon , "Building thesaurus lexicon using dictionary-based approach for sentiment classification," in *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2016.
- [18] N. Medagoda, "Framework for Sentiment Classification for Morphologically Rich Languages: A Case Study for Sinhala," Auckland University of Technology, Auckland, 2017.
- [19] A. Vlasblom, "Coursera Data Science Capstone Milestone Report," July 2015.

Ecosystem mapping technology to understand innovation challenges in MSME Sector: An analysis of the handicraft sector in Sri Lanka

Chathura Rajapakse
Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
chathura@kln.ac.lk

Abstract: This paper discusses an application of ecosystem mapping technology to visualize the innovation ecosystem in the Micro, Small and Medium Enterprises (MSME) sector. Ideally, actors in the MSME sector should function as a system, exchanging knowledge, information and other resources, to uplift innovation and create economic value to a country or a region. Visualizing this system enables to understand the deficiencies in the exchange of resources among actors; the deficiencies, which makes innovation and growth challenging to MSMEs. Based on this idea, the study presented in this paper focuses on the handicrafts sector in Sri Lanka and attempt to map its ecosystem based on the manually collected publicly available data on social media. The subsequent analysis of the support ecosystem map reveals a major deficit in the exchange of information technology resources, which hinders the digital innovation and transformation of the MSMEs in the handicrafts sector. The paper further discusses the implications and future research directions.

Keywords: Digital innovation, Ecosystem mapping, Handicrafts sector, Innovation ecosystem, Network visualization

I. INTRODUCTION

It is well accepted that visuals provide better and beautiful insights. In the big data world, visuals must go beyond merely being a conduit for information to offer some novelty: a fresh look at the data or a format that gives readers a spark of excitement and results in a new level of understanding [1]. Ecosystem maps are powerful and complex visual representations of actors in an ecosystem and their connectivity with other actors through the exchange of resources. They provide a novel and fresh view of the ecosystem, letting users excitedly see the unseen.

During the past couple of years, there seems to have an emerging interest among researchers to use ecosystem mapping technology to understand innovation ecosystems. Particularly in the domain of tech startups, some notable publications capture the key actors and their densities in the ecosystem, exploring opportunities for supporting the orchestration of innovation ecosystems [2]. Such studies are mostly supported by socially constructed data and provide system-level insights regarding the co-creation of value by the actors in the ecosystem. Moreover, such studies seem to have an immense potential to support policy-level decisions regarding national and regional innovation strategies.

This study is inspired by the increasing government sponsorship and the interference of various public and private sector institutions to foster innovation in Sri Lanka.

According to [3], the government has assigned high priority to the SME sector to strengthen the SME as the backbone of the economy and have taken steps to form an entrepreneurial development environment by introducing appropriate policy reforms and providing greater incentives for SMEs. Moreover, the country is orienting its economic growth model to be more open, private sector-led, knowledge-intensive, and focused on developing and exporting higher-value goods and services around the world [4].

However, in practice, it is observed a lack of proper coordination among institutions when it comes to innovation and entrepreneurship (I&E) related initiatives and activities. Furthermore, a notable priority is given to tech startups, especially by the software and business process management related institutions as well as academic entities. Moreover, it is not clear whether sufficient initiatives have been taken to empower MSMEs with technology. Especially, the adequacy of the transfer of information and communication technology-related capabilities is unclear despite its potential to digitally transform the MSME sector. According to [4], some of the key innovation and entrepreneurship challenges in Sri Lanka relevant for this study include;

- The SMEs exhibiting weak capabilities to adopt technology and to innovate
- An overall lack of coordination on I&E policies and programs and the low capacity of public agencies and officials to design, finance, implement and monitor such policies and programs
- Though pockets of research excellence exist, the public R&D system is fragmented and misaligned with the needs of enterprises and society and most institutions operate with little collaboration among themselves

Even though the priority given to tech startups is understandable amidst the need for enhancing technology-intensive exports, sustainable economic growth could be secured through the empowerment of MSMEs, which are mostly labor-intensive as of now. However, it is not clear at what level does that empowerment happen.

Ecosystem mapping could be a useful approach under these circumstances, in particular, to understand the lapses, misalignments, and imbalances in coordination, co-creation and resource exchange among ecosystem actors. Especially, the mapping of the innovation and entrepreneurship ecosystem in non-tech industries, in which the rural and

female communities are mostly engaged, is of paramount importance to devise sustainable policies and strategies to ensure competitiveness and economic growth. Therefore, this study mainly focuses on mapping the ecosystem of MSMEs of non-tech industries. This paper presents an initial structure of the ecosystem map of the handicrafts sector and some findings based on that. The subsequent chapters are organized as follows; Section II discusses the background and related work, Section III explains the research methodology, Section III presents the initial version of the innovation ecosystem map built, Section IV discusses the insights from the ecosystem map and Section V provides implications and concluding remarks.

II. BACKGROUND AND RELATED WORK

A. Network mapping and innovation ecosystems

Visualizing innovation networks is not a new idea. Over the years, researchers have used graphic depictions for the exploration and analysis of actors and their interactions with other actors in innovation networks. A notable introduction to the field is provided in [5], with the term ‘network map’, which is derived from the traditional sociogram and is designed to play a useful role in the portrayal and interpretation of innovation networks. According to [5], the network mapping approach exploits the visual capabilities of human cognition and the graphic proficiency of the personal computer, and hence, has the power to capture the diversity of actors, links and flows involved in the innovation process. [5] also introduces a set of conventions for network mapping and analysis.

Network mapping is a useful tool in the evaluation and timely monitoring of innovation activities. In other words, it provides a means of measuring innovation in an entrepreneurial ecosystem. The analog to the digital transformation taking place around the globe generate seas of socially constructed data, which can reveal previously unseen relationships between actors in innovation ecosystems. Based on this thinking, a novel approach to transform socially constructed data into useful visual indicators by the means of network maps is introduced in [6], which is named ‘the process of data-driven innovation ecosystem analytics’. The data-driven innovation ecosystem analytics process guides the conversion of the raw, socially constructed and open data into relational data tables, and thereby, the construction of visual structures and views. This approach enables the analysis of innovation ecosystems both qualitatively and quantitatively at different levels.

Extending on [6], [2] has discussed another important element in the study of innovation ecosystems, which is the *time*. The characteristics of innovation ecosystems such as the structural shape, number of actors and the density may change over time. Network mapping provides a powerful method to evaluate the changes over time visually. As quoted in [2], visualizations can reflect the structure of an innovation ecosystem at a single point in time, and they can also show the evolution of an ecosystem’s actors and their relationships over time. The data-driven approach can construct visualizations of a particular innovation ecosystem at several points in time, enabling us to see the respective changes visually and more conveniently. Commercial websites such as www.startupblink.com have also taken initiatives to map startup ecosystems across the world enabling the 1) ranking

of startup ecosystems around the world, 2) monitoring of their changes over time as well as 3) comparison of ecosystems against each other. As the collection of accurate data is challenging, they try to capture more accurate data through partnerships with local agents in each region of interest.

The adoption of ecosystem (network) mapping in the formulation of innovation and entrepreneurial policies and strategies is discussed in [7]. Most of the government and funding organizations currently have realized the limitations of formulating policies narrowly focusing on funding and training entrepreneurs; hence they focus more on the ‘entrepreneurial ecosystems’ at the national and regional level and develop broader policies and strategies. According to [7], the mapping ecosystem is an important strategic process in enabling entrepreneurial ecosystems. The other key strategic processes include; 1) favoring incumbents – i.e. currently existing dominant players – less, 2) listen to entrepreneurs, 3) think big, start small, move fast, 4) avoid artificially segmenting the community or strategies and 5) prepare to capitalize on crises.

B. Sri Lanka’s innovation and startup ecosystem

A few notable policies and strategic initiatives could be noticed in the recent past with the intention of creating an innovation-friendly entrepreneur ecosystem in Sri Lanka. [4] has identified the key government actors that play important roles in the innovation and entrepreneurial ecosystem of the country. [8] has introduced several strategic initiatives proposed by the Ministry of Science, Technology, and Research in Sri Lanka, to develop science and technology research to make Sri Lanka a world-class knowledge hub in science and technology. However, most of the goals mentioned there have not been achieved [4]. Another policy framework developed by the National Enterprise Development Authority (NEDA) is introduced in [9], which focuses on the growth and sustainability of SMEs through the creation of a conducive environment.

According to [4], A dynamic entrepreneurship ecosystem is emerging in parts of Sri Lanka, most notably in Colombo and Jaffna. This is well depicted by a recent ecosystem support map produced by the International Trade Center (ITC) pertaining to the entrepreneurship ecosystem in Sri Lanka [10]. As depicted by Fig. 1, their ecosystem map depicts the prominent actors in the startup ecosystem in Sri Lanka and their relationships. Even though this ecosystem map has several limitations, it could be considered as one of the pioneering studies reported so far regarding the application of ecosystem mapping in Sri Lanka. Few notable highlights from this ecosystem map includes 1) the dominance of ICT/BPO related actors, 2) bias towards Colombo-based tech startups ecosystem and 3) absence of key MSME related actors from other sectors such as agriculture, spice, and allied products, handicrafts, floriculture, and tourism, which contributes more to rural economies.



Fig. 1. Sri Lankan entrepreneurship support ecosystem map [10]

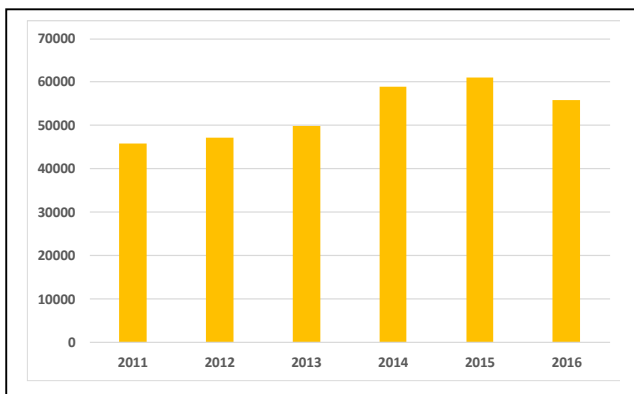


Fig. 2. Total Exports in the handicraft sector from 2011 – 2016. Figures in USD '000

C. Handicrafts sector in Sri Lanka

The handicraft sector was chosen for this study mainly due to its contribution to the rural economy. According to [11], except for a few key players, most of the enterprises in the handicrafts sector in Sri Lanka are cottage businesses, which are located mostly in rural areas all around the country. Though the investment in the last five years is small, as depicted in Fig. 2, the industry has shown some growth over the last five years with a slight decline in 2016. The production is mainly aimed at the local and niche export market and the cottage businesses are mainly supplying to the prominent players who do the end-sale and export. The key innovation and entrepreneurship support ecosystem actors include; 1) University of Moratuwa, 2) University of the Visual & Performing Arts 3) Academy of Design, 4) The Export Development Board 5) National Craft Council (NCC) and 6) National Design Center (NDC) [11].

III. RESEARCH OBJECTIVES AND METHODOLOGY

The key research question in this study is whether ecosystem mapping technology could be used as a tool to understand issues and challenges prevailing in the MSME sector. As the focus of this study is on mapping the innovation and entrepreneurship ecosystem in the handicrafts sector in Sri Lanka, the following research objectives were derived.

1. Identifying the key actors in the innovation and entrepreneurship ecosystem in the handicrafts sector in Sri Lanka

2. Constructing the innovation and entrepreneurship ecosystem map
3. Analyzing the ecosystem map to understand key challenges and issues in the crafts sector
4. Deriving conclusions on the use of ecosystem mapping technology to understand issues and challenges in the MSME sector
5. Determining future research directions for MSME studies with the help of ecosystem mapping technology

One of the biggest challenges in ecosystem mapping is the unavailability of data. The socially constructed data is available but the data is not readily available in the required format. Therefore, the data collection for this study was done manually from the data available on social media.

As the first step, the ecosystem actors were identified from desk research on the MSMEs in the handicrafts sector. There, the National Enterprise Development Authority (NEDA), the Small Enterprise Development (SED) division, the National Crafts Council (NCC), the Export Development Board (EDB), the National Design Centre, the Gem, and Jewellery Research and Training Institute and the Gem and Jewellery Authority were identified as the key actors in the innovation and enterprise ecosystem of the handicrafts sector.

As the second step, the social media posts on their official pages were collected manually for the period of 1-11-2018 to 30-10-2019. The posts were manually read to identify the other actors with whom they have interacted during the past one year. The interactions are classified as interactions for 'platform', 'knowledge', 'funding' or 'policy'. These classifications were purely the judgment of the researcher. A platform interaction is any interaction that creates a platform for some activity. For example, a competition or a workshop organized by the Ministry of Education for school children with the help of the National Crafts Council was considered as a platform interaction between the two entities. A knowledge interaction is a provision of a knowledge service from one entity to another. For example, Sri Lanka Institute of Marketing providing resource personal to a workshop organized by Small Enterprise Development (SED) division was considered as a knowledge interaction. Any interaction that supplied funding for a project was considered as a funding interaction whereas any interaction that supported the formation of a policy was considered as a policy relationship.

As the third step, the ecosystem map was created using the actors and links identified before. The data was fed into the statistical software package R and the ecosystem map was drawn using the *igraph* library. The generated ecosystem map was analyzed to derive conclusions.

IV. RESULTS

Fig. 3. illustrates the aforementioned ecosystem map generated based on the data.

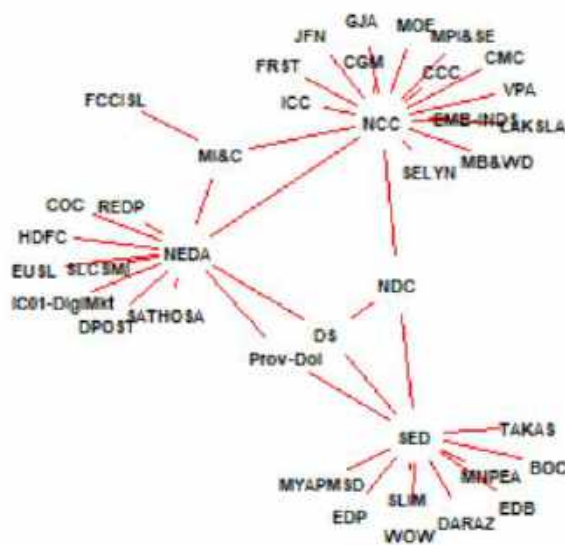


Fig. 3. The ecosystem map of the handicraft sector in Sri Lanka

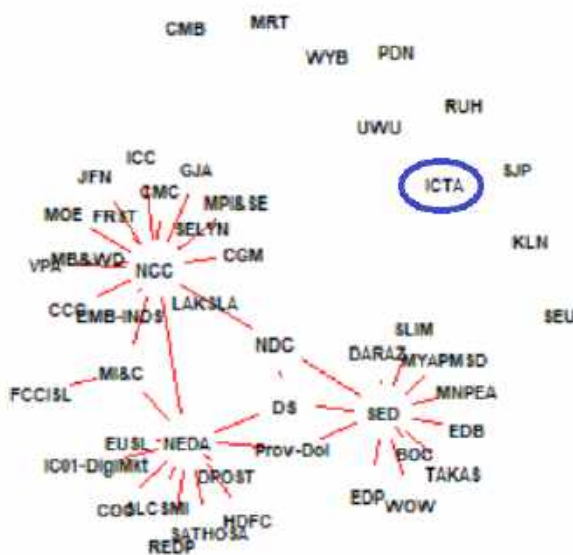


Fig. 4. An extended ecosystem map of the handicraft sector in Sri Lanka with some actors that are not connected, yet, with potential importance

V. DISCUSSION

The Fig. 3. It provides some basic, yet, important information about the handicraft sector. It could be compared with the ecosystem map depicted by Fig. 1. for a better understanding. Some of the important observations from the ecosystem map in Fig. 3 are as follows.

A. Less coordination among critical actors

According to Fig. 3., the National Enterprise Development Agency (NEDA), the National Crafts Council (NCC) and the

Small Enterprise Development (SED) division have the highest betweenness in the ecosystem map, which means they are the most prominent actors in the ecosystem. However, it further shows that there are no direct links between NEDA and SED as well as NCC and SED. The Divisional Secretariat (DS) offices and the Provincial Ministries of Industries (Prov-MoI) had been acting as the mediators between NEDA and SED. The two entities, which belong to two different ministries, hence appear to be operating independently on most occasions. The NCC seems to have connected with the SED through the National Design Center (NDC) only. This shows a clear lack of coordination among the important actors in the handicrafts sector.

B. The less active role of the Export Development Board (EDB)

Notably, the Export Development Board (EDB) is connected with SED directly, but not with NEDA. Moreover, the betweenness of the EDB in the ecosystem of the handicrafts sector is minimum compared to its significantly higher betweenness in the startup ecosystem depicted by Fig. 1. EDB has listed handicrafts as a key export category on its website, which contradicts this observation. Perhaps, the supply chain structure of the handicrafts sector is arranged in such a way that the EDB is more closely associating with established medium and large scale merchants but not with the micro and small scale enterprises that are served by NEDA and SED. However, this hypothesis needs further investigation.

C. Lack of ICT support and adoption

Compared to the startup ecosystem in Fig. 1, the Information and Communication Technology Agency (ICTA) of Sri Lanka are completely out of the picture in the ecosystem of the handicrafts sector. As depicted by Fig. 4, it has no single link with any of the key actors in the ecosystem. Moreover, according to Fig. 3, none of the key actors in the ecosystem of the handicraft sector have had any association with an entity known for ICT expertise during the last one-year period. However, according to [11], digital empowerment is critical in the handicraft sector to develop skills and capacity. The inactiveness of the ICTA depicted in the ecosystem map indicates a digital innovation challenge prevailing in the sector, which may affect the competitiveness of the products of Sri Lanka in the international market. Though understandably, the hand-made products would be more demanded in some instances, less adoption of information technology may hinder the growth of the sector.

D. Less involvement of the private sector

Except very few, most of the actors in the ecosystem map of the handicrafts sector are public sector organizations. In contrast, the startup ecosystem depicted by Fig. 1 has a nice mix of public and private sector organizations, which reflects a much higher growth potential. Since the public sector organizations are mostly funded by public money, they have limited financial capacity to support the growth of the sector. Thus, the lack of private sector involvement largely affects the resource accumulation and hence can hold back the innovation and growth of the sector

E. Less involvement of the universities and technical colleges

Universities and technical colleges have a significant potential to transfer critical knowledge for the growth of the MSMEs. As depicted by Fig. 1, the universities and technical colleges are playing an important role in the startup ecosystem in Sri Lanka. In contrast, as depicted by Fig. 4, the ecosystem of the handicrafts sector has shown no connection during the last year with the universities and technical colleges except the University of Jaffna (JFN) and the University of the Visual and Performing Arts (VPA). Partnering with the universities and technical colleges can expedite the growth of the sector through knowledge and technology transfer as well as research and development (R&D). However, the less involvement of universities and technical colleges indicates another innovation challenge in the handicraft sector.

VI. CONCLUSION

The above findings from the handicrafts sector in Sri Lanka strengthen the idea that the ecosystem mapping technology can support having a holistic view of a startup ecosystem to identify its issues and deficiencies in the exchange of resources. Such findings could be used to formulate effective policies and appropriate strategies to enhance innovation and growth in the MSME sector. However, the research presented in this paper has several limitations, which need to be overcome before the use of the findings to formulate or refine national or regional entrepreneurship policies.

A. Limitations of the study

Some of the major limitations of this study are as follows.

- The findings are purely based on a publicly available, manually collected, small sample of data. Hence, there is a chance that the findings are less accurate and incomplete. Lack of high-quality data and the difficulty to locate even the available data is a major limitation of this approach.
- The findings are not validated by experts in the field. Since this is ongoing research, validation of these findings is yet to be done. However, it is expected to interview some representative individuals from the handicraft sector to get these results validated.
- The network map used to model the ecosystem structure could be further improved for better insights. For example, the links between actors are not differentiated in terms of the type of resource being exchanged as well as the frequency of that exchange. Moreover, the nodes can be differentiated based on a criterion such as public or private sector organization.
- The ecosystem map presented in Fig. 3 and Fig. 4 is the ecosystem support map, which contains only the support actors that help the MSMEs in the handicraft sector to function. Though it is challenging, adding the actual MSMEs also into this ecosystem map would reveal much more interesting and useful insights.

B. Future Work

The potential of the findings from this research despite the above limitations indicates numerous avenues for future

research. Some of the notable future research directions are as follows.

- The study can be extended to other sectors such as agriculture, healthcare, and tourism, which could reveal unseen inter-sectoral patterns as well as deficiencies in the flow of resources to and among different sectors.
- The ecosystem mapping approach could be used to model large scale innovation and entrepreneurship ecosystems such as national and regional systems of innovation, which would indicate national or regional level innovation and entrepreneurship challenges.
- From a technical perspective, it is vital to develop a method to collect high-quality data to identify actors and their relationships. The manual data collection is cumbersome and error-prone, which needs a replacement with an automated method. However, it is highly challenging to develop such a method.
- Another important observation could be the time-based change of the structure and connectivity in the ecosystem map. As suggested by other researchers as well [2], it would be interesting to observe the change of the ecosystem map annually to understand growth (or decline)
- It is also possible to revisit the national innovation and entrepreneurship policies taking this approach with necessary improvements.

Based on the above limitations and the possible future work, further research will be carried out to improve this data-driven approach, for that to be a useful decision support tool for investors and policymakers.

REFERENCES

- [1] N. Iliinsky, "On Beauty" in *Beautiful Visualization: Looking at Data Through the Eyes of Experts*, J. Steele and N. Iliinsky, Eds. O'Reilly Media Inc., 2010, pp. 1–13.
- [2] K. Still, H. Jukka, M. G. Russell, and N. Rubens, "Insights for orchestrating innovation ecosystems: The case of EIT ICT Labs and data-driven network visualizations." *International Journal of Technology Management*, 2014.
- [3] J.P.D.R Jayasekara and A. Thilakarathna, "Government Policy and Strategy for SME Development", The Fourth IMF-Japan High-Level Tax Conference for Asian Countries, 2013.
- [4] "Innovation and Entrepreneurship Strategy of Sri Lanka 2018 – 2022", Ministry of Development Strategies and International Trade, Sri Lanka.
- [5] S. Conway and S. Fred, "Mapping innovation networks", *International Journal of Innovation Management*, Vol 2.02, 1998, pp. 223-254.
- [6] K. Still, H. Jukka, M. G. Russell, and N. Rubens, "Paradigm shift in innovation indicators-from analog to digital", *ISPIM Innovation Symposium*, The International Society for Professional Innovation Management (ISPIM), 2012.
- [7] P. E. Auerswald, "Enabling entrepreneurial ecosystems: Insights from ecology to inform effective entrepreneurship policy." *Kauffman Foundation Research Series on city, metro, and regional entrepreneurship*, 2015.
- [8] "Science, Technology and Innovation Strategy for Sri Lanka: 2011-2015", Ministry of Science, Technology, and Research, Sri Lanka.
- [9] "National Policy Framework for Small Medium Enterprise (SME) Development", Ministry of Industry and Commerce, Sri Lanka, 2016
- [10] "Mapping Sri Lanka's entrepreneurship ecosystem", *International Trade Center*, 2018
- [11] A. Chatterjee, A. Tiwari, P. Mohan Misra, and S. G. Dhande, "Designing Handicrafts using Information Communication Technology", *Indian Journal of Exports*, Vol. 12.1, 2008

Contribution of reflection in language emergence with an under-restricted situation

Kense Todo*
School of Computing
Tokyo Institute of Technology, Japan
k_todo@ali.c.titech.ac.jp

Masayuki Yamamura
School of Computing
Tokyo Institute of Technology, Japan
my@c.titech.ac.jp

Abstract: Owing to language emergence, human beings have been able to understand the intentions of others, generate common concepts, and extend new concepts. Artificial intelligence researchers have not only predicted words and sentences statistically in machine learning but also created a language system by communicating with the machine itself. However, strong constraints are exhibited in current studies. Models dependent on task settings, or supervisor signals and rewards exist, thus hindering the emergence of languages like the real-world. In this study, we improved Batali and Choi et al.'s research and attempted language emergence under conditions of low constraints such as human language generation. We developed a new language emergence agent that combines a language module and a visual module and included the bias that exists in humans as a "reflection function" into the new emergence algorithm. We used the MNIST dataset for language emergence. Irrespective of the function, messages corresponding to the label of MNIST could be generated. However, through qualitative and quantitative analysis, we confirmed that the reflection function caused pattern structuring in the message. This result suggested that the reflection function performed effectively in creating a grounding language from raw images with an under-restricted situation like the human language generation.

Keywords: Cognitive bias, Conceptual grounding, Language emergence, Multi agent description game, Reflection

I. INTRODUCTION

Humans, as well as animals, have emerged information and concepts through the exchange of languages [1],[2]. Languages enable the understanding of the intentions of others and a socially common system to be built [3]-[5]. In addition, languages can be used to recursively construct concepts and create new ones [6],[7]. The intelligence needs to formulate and systematize a language from contexts and situations. For artificial intelligence studies, it is important to build machines that can handle words and concepts[8],[9]. Deep learning, especially the attention mechanism, has improved the success of various natural language processing tasks [10]-[12]. Furthermore, studies on multi-modal information are progressing. Supervised learning based on large datasets regards languages as statistical pattern processing with discrete symbols.

Introducing the language system using the current datasets has been difficult in machine learning [13], [14]. Language intelligence involves more than statistically predicting words and sentences. Therefore, Researchers have built neural network agents to create the language system by communicating with each agent instead of language processing using statistical information [9],[15]. Complex tasks could be solved using these generated messages.

Constructed rules and messages that emerge owing to this communication are a system that differs from that of a human.

However, current language emergence studies have not achieved the autonomous grounding of mental images as those of humans and animals [16],[17]. Two issues remain: (1) The correspondence of generated messages to concepts is mediated by the architecture. Messages are not directly grounded to concepts. In some cases, a meaningless message was generated, and in other cases, a meaningful message was assigned to random noise images. (2) The "supervisor" role is essential because emergence agents do not gain a language system unless they are rewarded for communication. The supervisor method provided implicit indices such as "direction of correct communication for task resolutions" and "direction of a correct language system corresponding to a concept."

These issues occur because conventional language emergence models were considered based on "machine learning." "Machine learning" requires a "strong-restricted situation" such as a supervisor and uniquely defined concepts. On the other hand, the language emergence and learning of humans are different from the language emergence model of machine. Humans have prior knowledge of the concept [18,19]. Such the concept is not uniquely fixed. Humans can create language systems independently from the supervisor [20]. These mean that humans can make language emergence in an under-restricted situation. Little has been reported on language emergence under these conditions in which the agent possesses individual prior knowledge and is independent of the supervisor. Therefore, the need to introduce a human cognitive model is being discussed[21]-[23].

The purpose of this study is to confirm a compositional language emergence method, in which machines autonomously ground concepts to multiple agents in an under-restricted situation. We developed a method and architecture that can generate a language from discrete symbols. Using Description Game, agents must autonomously generate messages corresponding to cluster images that are not fixed at one point. To create compositional rules within these under-restricted situations, we introduce human cognitive bias into agents. Using the generated messages and the reconstructed images, we compared the characteristics of the agreed compositional rules by bias and discovered the emergent parameters necessary for the rules.

II. RELATED WORK

Artificial life studies of language emergence have discovered the dynamism of the generated language and the condition under which language systematization and formality

have emerged [24],[25]. Recently, a neural network model has been used to generate a complex language using reinforcement learning or supervised learning. It is classified into the task solution type and the concept correspondence type. In the task solution type, multiple agents are prepared, and the tasks, such as VQA problems, are solved by communicating each other's situations and behaviors [26]-[28]. The agents materialize words (instructions) through communication to generate appropriate behaviors and generate answers to specific questions. On the other hand, the concept correspondence type generate messages corresponding to objects recognized through a reinforcement learning(RL)-based approach [29]-[31]. This approach comprises different modules for generating and learning messages. It converges on a common language because the reward is transmitted to each module. However, the RL-based approach suffers from internal consistency [17],[32] [33].

Our study is inspired by the studies of Batali [34] and Choi et al [33]. Batali and Choi et al proposed the 'obverter technique' to generate compositional languages. The obverter technique is a commonly used message generation method [35]. A speaker cannot know the state of mind of a listener. Therefore, in the obverter technique, the speaker is assumed to have a similar mind as the listener. The speaker generates a message that maximizes his own understanding. Based on this assumption, the technique attempts to solve the problem of mutual communication. It can solve the problem of the RL-based approach "the problem of internal consistency in languages" and enables language emergence based on "the theory of mind [36]." However, in Batali's study, experiments were conducted under strong constraints that allowed the corresponding concepts to be uniquely determined[34]. In Choi et al.'s study, the supervised signal was required for generation[33].

The formation process and concept correspondence of human communication are useful as reference. Supervisors are not required for the process of human language acquisition. Furthermore, it is characterized by not learning concepts and words simultaneously [18],[19]. The emergence of sign languages in Africa [20] and the formation of languages such as the Pidgin and Creole languages demonstrate the unsupervised and asynchronous nature in language systems. Additionally, it is known that animals acquire different formal language depending on the species [37]. In this study, we refer to the above-mentioned studies to allow multiple agents to autonomously generate messages corresponding to concepts without a human supervisor.

We herein propose the multi-agent description game to establish a language system for multiple agents without a human supervisor. The description game consists of a teacher agent (speaker) that generates messages and a student agent (listener) that receives and learns the messages [30],[33]. The student agent receives the messages from the teacher agent and associates them with the presented image. Repeating this process creates a common language with messages corresponding to concepts between agents.

III. METHOD

A. Multi-Agent description game

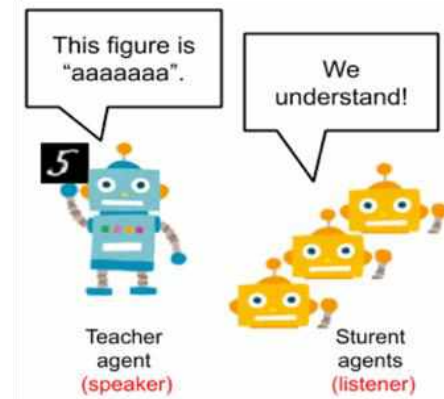


Fig 1. Multi-agent description game schema

The game proceeds as follows for each epoch. (1) One teacher agent is randomly selected from all agents. (2) The teacher agent randomly selects one image from a list and generates the message corresponding to the image. (3) The teacher agent sends the image and the message to the student agent, and the student agent associates the message with the image (Learning process). (4) (1) to (3) are performed a specified number of times. No image labels are presented between in-game agents. Therefore, the agent must associate the message with the representation read from the image. There is no guarantee that a message will completely correspond to a label. It is noted that "image" means raw pixel data that are provided to the agent, and "label" means "numbers (common concepts understood by humans)" written to the image.

In this study, one hundred agents were prepared and the image object used was the MNIST dataset. A training dataset was used during learning and generation, and a test dataset was used to evaluate the messages.

B. Model architecture and Obverter technique

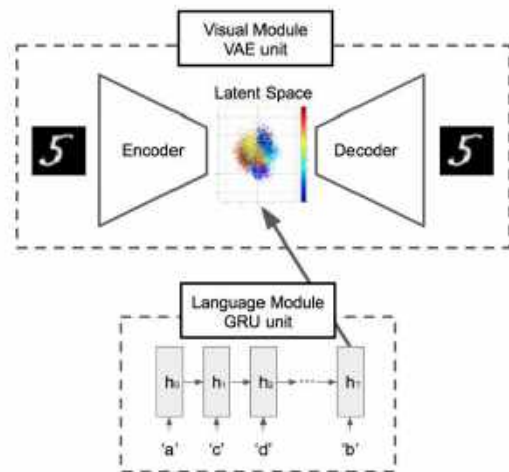


Fig 2. Agent model architecture (The visual module consists of the VAE unit. The language module consists of the GRU units. The VAE unit is composed of an encoder and a decoder. The encoder output is mapped to the latent space. Entering a message in the GRU units outputs a value corresponding to a point in the latent space.)

We herein propose an agent architecture for the multi-agent description game. An agent comprises a language module for reading and generating a message, and a visual module for processing an image and associating the image with the messages.

The visual module uses a variational autoencoder (VAE) [38] and the language module uses a gated recurrent unit (GRU) [39]. The visual module is given prior knowledge of MNIST. The language module outputs a numerical value corresponding to the latent space by inputting an arbitrary message that has been converted into a one-hot vector. This structure enables the generation of messages corresponding to the image and reconstruction of an image that corresponds to the messages. This module architecture differs from the pre-existing architecture in terms of some characteristics: (1) The proposed architecture enables language emergence by agent interaction without using human-prepared supervised signals and rewards. (2) The reconstructed image corresponding to the generated message can be directly confirmed.

An agent learns and generates a language using the obverter technique [34]. The below-mentioned procedure corresponds to self-supervised learning from the viewpoint of one agent. An agent set takes an emergent behavior that forms a conceptual pact without supervised signals.

1) *Generate messages*: In the teacher agent, the GRU model parameters are fixed. When a target image is input, the vision module encoder outputs a latent space value z corresponding to the image. After initializing the GRU hidden layer, the language module evaluates the output of all possible symbols. During this evaluation, the symbol with the minimum value between latent space z and the output of the language module is selected. The GRU hidden vector derived by the selected symbol is used by the next GRU unit. This procedure is repeated until the minimal value between z and the output value of the language module becomes less than a predefined threshold or until the max message length is attained. This procedure generates a message corresponding to the image.

2) *Learning message*: The student agent performs the learning so that the received message corresponds to the image. In the student agent, the GRU model parameters are updated. The GRU parameters are updated by backpropagating the mean square error loss between the value of the GRU unit output by inputting the message generated by the teacher agent and the latent space value z corresponding to the target image.

The details of the model architecture and parameter are described in Table 1. In this study, we used the same visual module agent group in each trial for comparison. Each agent in the agent group was learning a different latent space. The initial value of the language module was specified to be different for each agent.

TABLE I. PARAMETER LIST

Parameter	Value
Number of agents	100
Number of learnings in one Epoch	100
VAE structure	782(ReLU)-512-4-512(ReLU)-782(sigmoid)

Parameter	Value
VAE Optimization function	Adam(Lr=1e-3)
GRU structure	Hiddden128-Output4(Tanh)
GRU Optimization function	SGD(Lr=0.005)
Available symbols(Alphabet)	a,b,c,d
Maximum message length	10
Generation definition threshold	0.005

C. Reflection function

We introduced cognitive bias to the agents. In conceptual pacts, in some cases, humans may not agree with all the concepts presented [40]. Thus, human learning generally includes introspective elements. We improve and introduce this function so that the student agent compares messages generated by the teacher agent and rejects the message if it is significantly different from the self-concept. In this study, we defined this function as the “reflection function.”

The reflection function is realized by the student agent checking the message received from the teacher agent before updating the parameters. (1) The student agent receives a target image that the teacher agent used to generate a message. Each student agent self-generates its own message using an image by the method shown in § 3.2. (2) Each student agent calculates a degree of comparison similarity between the message received from the teacher agent and the self-generated message using gestalt pattern matching [41]. The degree of comparison similarity by gestalt pattern matching is output in the range of 0-1. The value of 1 means an exact match. (3-A) When the degree of comparison similarity is lower than the threshold value, the received message is regarded as an error and the learning process is not performed. (3-B) When the degree of comparison similarity is higher than the threshold value, the received message is regarded as correct and the learning process is performed.

In this experiment, three types were prepared for comparison. The first type was NORMAL: do not use reflection function. The second type was REFLECTION: the reflection function was incorporated from the beginning of the game. The threshold degree of comparison similarity was set to 0.4. The third type was SWITCH: to verify the strength of the constraints of the reflection function, we activated the reflection function during the game. SWITCH required parameters for the threshold degree and the activated point for the reflection function. The parameter used is the threshold value of 0.6/activated point at epoch = 600. We tested each type five times and evaluated the results.

D. Evaluation indices

We analyzed the language generation characteristics of each learning method. In particular, we investigated the correct answer rate of the image which the agent reconstructed from the messages, and the systematic rules of the messages. These points are necessary for the compositional properties of the emergence language.

1) *Basic information evaluation*: The first analysis was to confirm that learning has converged and to examine the similarity of messages between labels. We verified the mean square error loss between the output of the GRU unit and the

latent space z . We also confirmed the reconstructed image using the generated messages.

In addition, a quantitative evaluation of the reconstructed images was performed. The message for evaluation was obtained by the following method: Three arbitrary agents were selected from each agent set. 200 messages per agent were generated by the method proposed in §3.B. The breakdown of 200 messages is as follows: 100 messages were generated using a common image among the agents. Another 100 messages were generated using independent images for each agent. These messages allow duplicates. We entered the messages into the agents to generate the reconstructed image. The accuracy was finally calculated for each label by putting the image into the evaluators. The evaluators utilized five MNIST classifiers with over 98% accuracy. This evaluation method was based on Vedantam et al. [42].

2) *Compositional evaluation of lead symbols*: In the second analysis, we evaluated the characteristics of the relationship between the generated message and the label (concept) [43]. Specifically, the relationship between lead symbols and labels were quantitatively compared using lead symbol configuration for each label and usage rate of head symbol. Hereinafter, when the first symbol in the message is indicated or when a general concept is used, it is referred to as “Lead symbol”. The first two-digit symbol, which is frequently used in analysis, is referred to as “Head symbol”. We analyzed 600 messages generated in the above assessment. In lead symbol configuration for each label, the composition ratio of the lead symbol to one label was confirmed. If one label’s lead symbol consisted of only one or a few symbols, it is likely to be a clear ownership relationship between the symbol and the label. However, if it was composed of a plurality of lead symbols, it is likely that there is no clear symbol corresponding to the label. Specifically, the number of labels whose lead symbol constituted more than 70% of each label was counted.

In usage rate of head symbol, whether the expressed head symbol is biased to a specific symbol is evaluated as a variance. If the head symbol was generated without bias (i.e., using various symbols for the head), the variance became small and the symbol is considered to have a unique meaning. However, if a specific head symbol was mainly used (i.e., using only a single symbol), the variance became large and the symbol is considered to have multiple meanings. The variance formula is expressed as follows.

$$var = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (1)$$

Where x_i represents the number of uses of a certain head symbol and n represents the number of head symbol types.

3) *Attempts to formalize messages*: The third analysis confirmed the compositional rule for the message corresponding to reconstructed image. Language structure by recursion and multiple symbols is the basic element of “Compositionality” in language emergent research [44]. We entered a message to each agent set and confirmed the conditions of the message under which the image of each label was reconstructed. Therefore, we enumerated repeated patterns that appeared in the message by label as compositional rules.

Next, we simplified the compositional rules and analyzed whether a specific syntax appeared in the simple rules. Parentheses and repetitions have been omitted for simplicity. We analyzed the existence of overlap symbols in the structure of the simple rule’s sequence and first two-digit (head) symbols in simple rules. The comparison was based on the lead symbol. Using these indices, we can be confirmed of the basic symbol composition for reconstructing a certain label.

IV. RESULTS

A. Basic information evaluation

We confirmed the characteristics of the reflection types in the description game phases and the message generation. First, the distance between the message and target image was analyzed using the MSE. The loss tended to decrease during the description game in all reflection types (Fig. 3). SWITCH and NORMAL showed the same decreasing trend because they had no reflection function at the beginning of learning. In REFLECTION, a gradual decrease was shown compared with others. NORMAL showed an upward trend from epoch = 850. On the other hand, in REFLECTION and SWITCH, this tendency was not observed until epoch = 1100.

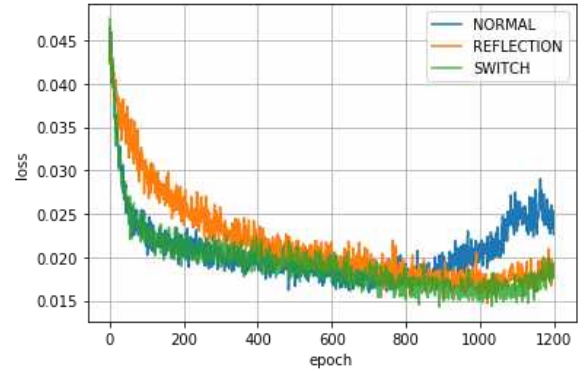


Fig 3. Transition of teacher’s loss function (The average of five trials is presented. Averages from epoch = 700 to epoch = 900 were 0.018, 0.017 and 0.016 for NORMAL, REFLECTION and SWITCH, respectively. Meanwhile, averages from epoch = 1100 to epoch = 1200 were 0.023 and 0.017 for NORMAL and INTERNAL, respectively. IN SWITCH, Averages from epoch = 1100 to epoch = 1200 were 0.017. A clear increase in loss function was observed only in NORMAL.)

Thus, these results exhibited different loss function transitions depending on the reflection types. Subsequently, we attempted to analyze epoch points that have a smaller loss function value for each reflection type. (NORMAL = 810, REFLECTION and SWITCH = 1200. For the analysis with NORMAL = 810, we created five new verification agent sets at epoch = 810. The decrease in loss function showed a similar trend.)

Fig. 4 shows the result of the reconstructed image of one agent group in each reflection type. As shown, all reflection types have the same ability to the reconstructed image. Some labels have low agent sharing for the messages. For example, labels “6,” “8” for NORMAL, label “6” for REFLECTION, and labels “2,” “6” for SWITCH showed this tendency. However, we confirmed that all the same messages reconstructed the images corresponding to the label.

Next, we confirmed the accuracy of the reconstructed image corresponding to the generated message (Fig. 5). The average accuracy was 0.49 (NORMAL), 0.47

(REFLECTION), and 0.47 (SWITCH). Statistically significant differences in the accuracy were not observed.

The clustering accuracy average of the VAE latent space for all agents was approximately 0.65 using the v-score. Therefore, the accuracy of the reconstructed image is lower than the clustering accuracy but did not present a significant decrease.

B. Compositional Evaluation of Lead Symbols

From the lead symbol configuration for each label and usage rate of head symbol analyses, the characteristics of the relationship between the generated message and the label are compared. The lead symbol configuration analysis for each label (Table II) showed almost the same value at the end of the description game in the reflection types. No statistically significant difference was observed. Meanwhile, in the usage rate of head symbol analysis, each reflection type showed different trends.

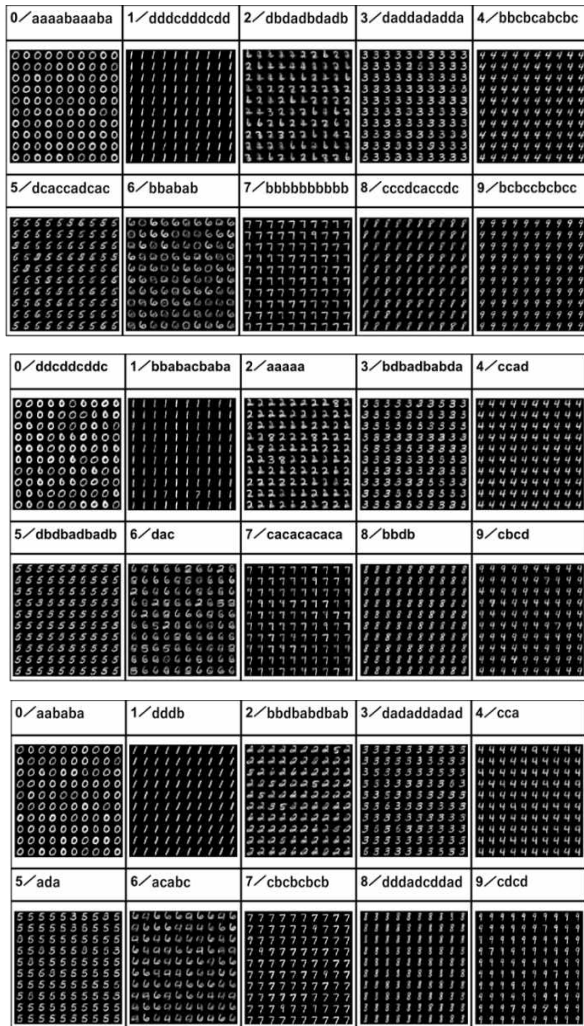


Fig 4. An example of the reconstructed image from messages (NORMAL, REFLECTION, and SWITCH, respectively, from top to bottom. The text above each figure is the message entered into the agents. The numbers in each figure represent the corresponding label numbers.)

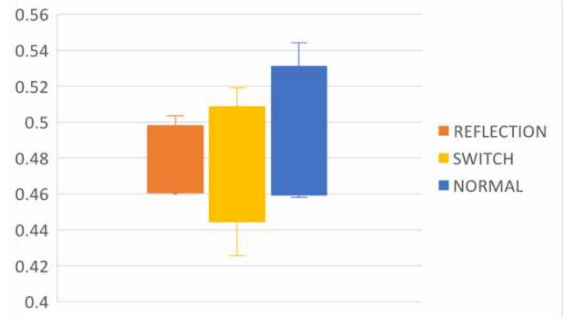


Fig 5. The accuracy of the reconstructed image in each reflection type (The vertical axis indicates accuracy.)

TABLE II. COMPOSITIONAL EVALUATION OF LEAD SYMBOLS.

	Reflection Type		
	NORMAL	REFLECTION	SWITCH
Lead symbol configuration	5.8	6	4.8
Usage rate of head symbol	1041.4	559.0	464.9

At epoch = 210, the variance was almost equal in all reflection types. As the game progressed, in NORMAL, the variance remained almost constant. REFLECTION showed a downward trend at the game progressed. SWITCH showed the same value as NORMAL before the reflection function was applied. At the end of the description game, a statistically significant difference ($p < 0.01$ in t-test) appeared in the value of the head symbol variance between NORMAL and the reflection function such as REFLECTION and SWITCH. The existence or nonexistence of the reflection function caused the lead symbol to exhibit a quantitatively different structure.

C. Attempts to formalize messages

1) *Decoding from human viewpoint of symbol compositional rule:* One agent group (Same group as Fig. 4) was extracted from each reflection type. We confirmed the compositional rules for messages according to Method 3.D.3. The compositional rules for each label are shown in Table III.

TABLE III. COMPARISON OF COMPOSITIONAL RULES.

WE ANALYZED THE MESSAGE IN WHICH THE RECONSTRUCTED IMAGE IS GUARANTEED FOR EACH LABEL. ONE AGENT GROUP WAS EXTRACTED FROM EACH REFLECTION TYPE. [xxx]* MEANS THAT THE xxx SYMBOL IS REPEATED. "y" SYMBOL OF [xxx]*y MEANS AN UNREPEATED SYMBOL. (z) MEANS THE SYMBOL THAT MAY BE REQUIRED DEPENDING ON THE REPETITION STATE OF THE MESSAGES. "WITHOUT S" MEANS AN EXCLUSIVE SYMBOL THAT CANNOT BE RECONSTRUCTED IF "s" SYMBOL EXISTS.

Label	Reflection Type		
	NORMAL	REFLECTION	SWITCH
0	[aaa]*[aab]* (a>=2)	[dda]*, [da]*	[a]*, [a(b)a]*
1	d>=5	(a)[bba]*b	[bd]*, [db]* (without "a")
2	[dcab]*, [dacb]*	(db)[aaa]*	bbd(a)(b)b (without "c")
3	[dad]*, [dda]*	(aa)[bdb]*a	d[ad]*, [dda]*
4	[bbc]*a	[cad]*, [ca]*d	[cca]*
5	(d)cccacca	d[bd]*	[aadd]*, [aad]*
6	[bba]*	d[aaa]*d, a[aaa]*d	[acb]*
7	b>=6	[cc(c)a]* (without "d")	[cb]*

Label	Reflection Type		
	NORMAL	REFLECTION	SWITCH
8	Ccd	[bbcd]*, [bbd]*	[dda]*(d)c
9	[bcbbc]*	[ccb]*, [cbc]*	[ccd]*, [cd]*

In all reflection types, each agent set formed its own compositional rule. A key symbol was created to generate each label's reconstructed image. A compositional rule having repeated patterns including the order and combination of symbols emerged in REFLECTION and SWITCH. Additionally, REFLECTION and SWITCH identified the presence of exclusive symbols that prevented the label from being reconstructed. Meanwhile, NORMAL had few labels with such a compositional rule. NORMAL did not show a clear repeated pattern on some labels and could reconstruct the reconstructed image by the minimum necessary key symbol.

2) *Parsing of simple rules*: After simplifying the compositional rule shown in Section 4.C.1, the syntax of the generated rule was analyzed (Table IV).

TABLE IV. SIMPLE RULES ALIGNED FOR EACH LEADING SYMBOL. THE ONLY NON-OVERLAPPING SYMBOL IN THE SIMPLE RULE OF THE SAME LEAD SYMBOL IS SHOWN IN ITALICS. IN ADDITION, THE PART WHERE THE COMBINATION OF THE FIRST TWO DIGITS (HEAD) OF THE SYMBOL DO NOT OVERLAP IS UNDERLINED.

NORMAL			
"a" series	"b" series	"c" series	"d" series
aaa/aab	bbbbb	ccca	dad/dda
	bba	ccd	dac <i>b</i> / <u>d</u> cab
	bbca		dddd
	<u>bcbbc</u>		

REFLECTION			
"a" series	"b" series	"c" series	"d" series
aaa	bbab	<u>cad</u>	daaad
aaaad	bbcd/bbd	cca	<u>dbd</u>
	<u>bdba</u>	ccb/ <u>cbc</u>	<u>dda</u> /da

SWITCH			
"a" series	"b" series	"c" series	"d" series
a/aa	<u>bd</u>	cca	<u>dad</u> /dda
aad /aadd	<u>bbdb</u>	<u>cb</u>	ddac
<u>acb</u>		ccd/ <u>cd</u>	<u>db</u>

- **NORMAL**: the use of the leading symbol was biased. The "b" series was especially used in the lead symbol. Three of the four were composed of the same head symbol. In addition, no unique symbol existed for each simple rule. The head symbol in the "c" series was the same. The unique symbol existed for each simple rules. In the "d" series, the unique symbol and the head symbol appeared on the same simple rule.

- **REFLECTION**: The lead symbol was allocated almost evenly. In the "b" series, a distinctive head symbol appeared, while the unique symbol appeared in one simple rule. In the "c" series, a distinctive head symbol appeared in two simple rules, while the unique symbol appeared in two simple rules. In the "d" series, a distinctive head symbol appeared in two simple rules, while the unique symbol appeared in one simple rule.
- **SWITCH**: SWITCH has a symbol syntax trend similar to that of REFLECTION. In "a" series, unique symbols appeared in two simple rules. A distinctive head symbol appeared in one simple rule. In the "b" series, although the symbols were similar, different head symbols appeared. In the "c" series, a unique symbol existed for every simple rule. In the "d" series, a distinctive head symbol appeared in two simple rules.

In REFLECTION and SWITCH, head symbol duplications were less than in NORMAL. The number of unique symbols appeared was slightly higher in types that had the reflection function.

In NORMAL, even in the "b" series, which requires a clear division, only one simple rule appeared clearly independent. Meanwhile, in REFLECTION and SWITCH, distinctive head symbols and unique symbols appeared in a balanced manner in each rule. Each rule had a proprietary tendency. It was confirmed that even simple rules not having a unique symbol exhibited a different pattern. Therefore, the tendency to form a unique symbol combination was confirmed in REFLECTION and SWITCH.

V. DISCUSSION

In this study, we proposed a multi-agent description game with 100 agents and a model architecture. The obverter technique was used for message generation. In addition, the reflection function, which imitates human language acquisition, was introduced.

As described in Sections 4.A, messages corresponding to the label were successfully generated, and we confirmed the reconstructed image from the message. It was found that by the proposed method, messages that correspond to concepts and allow an agreement between agents emerged without any human supervisor. However, as reported in Section 4.A, the decreasing tendency of the loss function was different between reflection types. As described in Sections 4.B and 4.C, the compositional rule between reflection types differed.

In NORMAL, the usage rate of head symbol showed a large value, it was considered that there was a bias in the use of lead symbols. On the other hand, in REFLECTION and SWITCH, the usage rate of head symbol showed a relatively small value, and it was considered that various lead symbols were used.

In NORMAL, a pattern with only one symbol or a label without clear repetition appeared. From the simple rules, the relationship between symbol patterns and labels was not clear. On the other hand, in REFLECTION and SWITCH, a label expression using a plurality of symbols mainly appeared, and a relationship between the specific pattern and the symbol was clear.

For SWITCH, the description game start point was NORMAL, and after some learning steps, the reflection function was activated. Notwithstanding, SWITCH showed the same results as REFLECTION. The reflection function seems to have broken the language structure at the time of NORMAL and generate a new structure. Thus, it was found that the compositional rules of messages that could be conceptually agreed by reflections were structured differently. It suggested that this characteristic difference was due to the reflection function.

We consider the features of the reflection function. NORMAL allowed students to learn all the messages generated by the teacher as a "positive example." In this situation, the "this and that" symbol was learned. Even a major difference was absorbed as a "correct" message. Therefore, the results of the loss function indicated that learning converged quickly. However, at the same time, it also indicated that this situation produced "over-learning [45]."

Meanwhile, in the reflection function, a learning constraint based on the agent's similarity evaluation existed. Because of this constraint, it is presumed that each agent creates a "negative example" in a pseudo manner by comparing itself with others. The positive and negative examples might have constructed a pseudo boundary among the labels. Therefore, it is assumed that the language generation has been constructed conservative. Consequently, a whole agent tends to a generate message having a common symbol for the same label. This situation, in which only common-term messages are generated, could be regarded as a bottleneck in information transmission between agent groups. Several studies have reported that bottlenecks produced syntactic structures [46], [47].

The bottleneck resulting from the reflection function is considered to have resulted in the build of the compositional rule having a pattern structure. This result suggests the existence of qualitatively different compositional trends is caused by the reflection function. Therefore, the reflection function can be expressed as an emergent parameter of a language system.

VI. CONCLUSION

In this study, we established a method for the emergence of a language system, in which consensus could be achieved in a group without a human supervisor. We proposed an emergence parameter, i.e., the "reflection function," which acquired structuring compositional rules; additionally, we confirmed its effectiveness. This suggests that it performs effectively under restricted situations similar to human language generation. Further studies are required to generalize the relationship between formalization and concept. An extension to a general data set in which a plurality of concepts such as a color, a shape, and a place are integrated can be considered. We will confirm the laws of systematization generated by machines using extended dataset and having "different concepts" in each agent.

REFERENCES

- [1] G. Lupyan and B. Bergen, "How Language Programs the Mind," *Top. Cogn. Sci.*, vol. 8, no. 2, pp. 408–424, Apr. 2016.
- [2] G. Lupyan and A. Clark, "Words and the World: Predictive Coding and the Language-Perception-Cognition Interface," *Curr. Dir. Psychol. Sci.*, vol. 24, no. 4, pp. 279–284, Aug. 2015.
- [3] P. Bloom, "How children learn the meanings of words.," *How children learn the meanings of words. The MIT Press*, Cambridge, MA, US, pp. xii, 300–xii, 300, 2000.
- [4] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning," *Psychol. Sci.*, vol. 20, no. 5, pp. 578–585, May 2009.
- [5] M. Tomasello, *Origins of human communication*. Cambridge, MA, US: MIT Press, 2008.
- [6] Robert C. Berwick and Noam Chomsky, *Why only us? Language and evolution*. 2015.
- [7] M. D. Hauser, "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? " *Science (80-.)*, vol. 298, no. 5598, pp. 1569–1579, Nov. 2002.
- [8] J. S. Bruner, "Intention in the structure of action and interaction.," *Adv. Infancy Res.*, vol. 1, pp. 41–56, 1981.
- [9] T. Mikolov, A. Joulin, and M. Baroni, "A roadmap towards machine intelligence," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9623 LNCS, pp. 29–61, 2018.
- [10] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.
- [11] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision," in *International Conference on Learning Representations (ICLR)*, 2019.
- [12] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "Visual Dialog," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, no. Cv, pp. 4466–4475.
- [13] T. Niven and H.-Y. Kao, "Probing Neural Network Comprehension of Natural Language Arguments," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, no. Table 1, pp. 4658–4664.
- [14] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 2368–2378.
- [15] L. Steels, "Evolving grounded communication for robots," *Trends Cogn. Sci.*, vol. 7, no. 7, pp. 308–312, Jul. 2003.
- [16] S. Kottur, J. Moura, S. Lee, and D. Batra, "Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2962–2967.
- [17] D. Bouchacourt and M. Baroni, "How agents see things: On visual representations in an emergent language game," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [18] M. Imai and T. Masuda, "The Role of Language and Culture in Universality and Diversity of Human Concepts," in *Advances in Culture and Psychology*, Oxford University Press, 2013, pp. 1–62.
- [19] K. Milligan, J. W. Astington, and L. A. Dack, "Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding," *Child Dev.*, vol. 78, no. 2, pp. 622–646, 2007.
- [20] A. Senghas, A. Özyürek, and S. Kita, "Children Creating Core Properties of Language: Evidence from an Emerging Sign Language in Nicaragua," *Science (80-.)*, vol. 309, no. 5743, p. 56c, 2005.
- [21] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building Machines That Learn and Think Like People," *Behav. Brain Sci.*, no. 046, pp. 1–101, 2016.
- [22] I. Rahwan et al., "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, 2019.
- [23] B. M. Lake, T. Linzen, and M. Baroni, "Human few-shot learning of compositional instructions," Jan. 2019.
- [24] R. Martin and P. R. R. White, *The Language of Evaluation*, vol. 2. London: Palgrave Macmillan UK, 2005.

- [25] S. Nolfi and M. Mirolli, *Evolution of Communication and Language in Embodied Agents*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
- [26] I. Mordatch and P. Abbeel, "Emergence of Grounded Compositional Language in Multi-Agent Populations," in *In Thirty-Second AAAI Conference on Artificial Intelligence*, 2017.
- [27] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra, "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2970–2979.
- [28] K. Evtimova, A. Drozdov, D. Kiela, and K. Cho, "Emergent Communication in a Multi-Modal, Multi-Step Referential Game," in *International Conference on Learning Representations (ICLR)*, 2017.
- [29] C. Liu et al., "Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols," in *Nips*, 2019.
- [30] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-Agent Cooperation and the Emergence of (Natural) Language," in *International Conference on Learning Representations (ICLR)*, 2016.
- [31] K. Cao, A. Lazaridou, M. Lanctot, J. Z. Leibo, K. Tuyls, and S. Clark, "Emergent Communication through Negotiation," in *International Conference on Learning Representations (ICLR)*, 2018.
- [32] R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, and Y. Dauphin, "On the Pitfalls of Measuring Emergent Communication," in *AAMAS '19: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent*, 2019, pp. 693–701.
- [33] E. Choi, A. Lazaridou, and N. de Freitas, "Compositional Obverter Communication Learning From Raw Visual Input," in *International Conference on Learning Representations (ICLR)*, 2018.
- [34] J. Batali, "Computational simulations of the emergence of grammar. Approaches to the evolution of language," *Social and cognitive bases*, 1998, pp. 405–426.
- [35] J. R. Hurford, "Biological evolution of the Saussurean sign as a component of the language acquisition device," *Lingua*, vol. 77, no. 2, pp. 187–222, Feb. 1989.
- [36] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behav. Brain Sci.*, vol. 4, no. 1978, pp. 515–526, 1978.
- [37] M. Spierings and C. Ten Cate, "Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment," *Proc. Natl. Acad. Sci. USA*, vol. 113, Jun. 2016.
- [38] M. Welling, "Auto-Encoding Variational Bayes arXiv: 1312 . 6114v10 [stat . ML] 1 May 2014," no. ML, pp. 1–14.
- [39] K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, vol. 28, no. 9, pp. 1724–1734.
- [40] S. E. Brennan and H. H. Clark, "Conceptual Pacts and Lexical Choice in Conversation," vol. 22, no. 6, pp. 1482–1493, 1996.
- [41] J. W. Ratcliff and D. E. Metzener, "Pattern-matching-the gestalt approach," *Dr Dobbs J.*, vol. 13, no. 7, p. 46, 1988.
- [42] R. Vedantam, I. Fischer, J. Huang, and K. Murphy, "Generative Models of Visually Grounded Imagination," in *International Conference on Learning Representations (ICLR)*, 2017.
- [43] H. Brighton, "Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings," vol. 242, no. 1, pp. 229–242, 2006.
- [44] S. Kirby, "The emergence of linguistic structure: An overview of the iterated learning model," in *Simulating the evolution of language*, 2002, pp. 121–147.
- [45] D. Rohrer, K. Taylor, H. Pashler, J. T. Wixted, and N. J. Cepeda, "The effect of overlearning on long-term retention," *Appl. Cogn. Psychol.*, vol. 19, no. 3, pp. 361–374, Apr. 2005.
- [46] P. Vogt, "The emergence of compositional structures in perceptually grounded language games," *Artif. Intell.*, vol. 167, no. 1–2, pp. 206–242, 2005.
- [47] H. Brighton, "Compositional syntax from cultural transmission.," *Artif. Life*, vol. 8, no. 1, pp. 25–54, 2002.

Profiling purchasing behavior of Generation Z

Nadeesha Kahawandala*

Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka.
lakmini1024@gmail.com

Suren Peter

Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka.
suren@kln.ac.lk

Hiruni Niwunhella

Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka.
hirunin@kln.ac.lk

Abstract: Generation Z has emerged as one of the most mystifying consumers as they are tech-savvy, digitally connected and educated users of technologies in the marketplace. These digital natives are predicted to account for around 40% of all consumer shopping by 2020. Their cognitive power and social media networking have made them the market mavens who possess a wide range of information and consumer knowledge about many dimensions of the markets. To be the leader in the marketplace where so many options are available due to the free trade economy, marketers have to escalate their knowhow about their customers if they want to capture the attention of this segment of the market. The chief objective of the study is to find the purchasing habits of Generation Z, specifically from the angle of social influence and technology. A sample of 42 respondents was approached for a survey using a structured questionnaire. The results of this initial study indicate that Generation Z shoppers are coming up with unprecedented shopping habits and preferences. Using cross-tabulation analysis, results showed that some of the most classical influential factors such as product features, price consciousness and family recommendations have a significant effect on the purchasing habits of this generation. However, the results indicate the influence of computer literacy, peer and social influence and social media identity on the purchasing decision of Generation Z. The results of this study can be utilized to assist in predicting potential consumer adoption behavior and in designing favourable shopping environments that are compatible with these specific consumer traits.

Keywords: Consumers, Generation Z, Purchasing behavior

I. INTRODUCTION

The lifting of restriction of trade across national boundaries and the evolution of free trade economy have delivered the consumer the opportunity to choose from a growing product range. Initially, marketers were successful enough to accomplish this dream. As the consumer is spoilt for choice, manufacturers face intense rivalry to grab the attention of the discerning customer. Since the products are made to cater to consumer's needs and demands, they should be carefully marketed for a fruitful achievement of the organizational goals.

Consumer purchasing behavior is defined by [1] as a series of activities conducted or perceived by people before buying products or services.[2] has analysed it via 3 basic steps. First, people perceive the situation, second, people use the power of reason to calculate whether taking one or another action will benefit their long-term interest and third people use the power of will to execute the decision. According to [3] consumer buying behavior is itself a complex, dynamic issue which cannot be defined easily and commonly.

But once consumer purchasing behaviors are analyzed by companies, they can better identify target markets and

formulate effective marketing strategies. So, the market researchers themselves are witnessing the success in marketing. Market research is also considered as an enabler of digital transformation. Organizations that leverage market research to drive key digital transformation programs such as DCX, and also leverage market research to assess and measure the readiness and success of digital transformation programs, will definitely have an edge over others. Being a key player in the digital transformation space also opens up new and unfolding opportunities for the practitioners of market research [4].

Companies will continuously develop new products or improve current products to meet the individual needs of consumers accurately. However, most often customers do not go through these logical processes every time they purchase a product [1]. The purchasing behavior of a consumer is influenced by factors like internal or psychological factors, social factors, cultural factors, economic factors and personal factors. [5] has discussed those factors with several sub factors. Under cultural factors culture, sub-culture and social class have been taken into account. Societies and tribes like the family, reference groups and roles who contribute to decision making were described under social factors. Price, income, distribution of income, utility and consumer preference were included under economic factors. Age, income and occupation were categorized under personal factors, while motivation and perception were combined with the psychological factors.

Age is extensively identified as a significant factor in defining consumer behavior. But comparatively, age acts as a boundary to separate the behavior of one set of consumers from another. That separation is evaluated as a generational cohort. That evaluation and summarization of the characteristics of that specific segment of generation are challenging and highly resource consuming. However, it is proposed that using generational cohorts is an even better way to capture the specific character and habits of certain consumer groups [6].

Generation is well-defined as people that are clustered within a certain range of ages, and have been shaped by alike conditions, technologies and life events which they have experienced at critical developmental stages [7]. Opinions, habits, and motivations of generation members, as well as their desire for certain products and buying behaviors are the determinants of the above-mentioned aspects. Gaining the attention of multi-generational consumers is an important aspect of marketing. But to do so, marketers require a deeper insight allowing them to understand diversified consumer behaviors of generational cohorts.

According to [8] there exist four generations of legal adults. Those four generations include the silent generation (those born between 1922 and 1945), the baby boomers (born from 1946 to 1964); Generation X (born from 1965 to 1977); and millennials (born from about 1977 to 1993) and at last, it is Generation Z (born from 1993 to 2005). Many researchers have defined the age gaps of these generations in different time frames which has minute gaps. For this study, we have chosen the summarization given below by [9].

TABLE I. GENERATION WISE COMPARISON OF CONSUMER BEHAVIOR BY [9]

Gen Z	Gen Y	Gen X
Born between 1995-2005	Born between 1980 and 1995	Born between 1965 and 1980
Social Natives	Book smart and Savvy	Street smart and skeptical
Connected	Confident	No drama
Global minder	Global minder and balance	Balance
Friends =Family	Close to parents	Prepared rule settings
the first generation to have Internet technology so readily available	expanded use of technology compared to previous generations	less use of technology

II. LITERATURE REVIEW

A. Profile of Generation Z

In this study, Generation Z was identified as the population who was born between 1995-2000. Gen Z has never known a world without smartphones and social media since technology has been readily available and accessible to them. They have developed fine motor skills to be able to manipulate a phone or tablet. Therefore, they are considered the most tech-savvy generation [10], [11]. For Gen Z diversity isn't an issue; they will only notice diversity in its absence. They have grown up in a period when the bi-racial and minority population has increased rapidly and are themselves more diverse. Gay marriage is a fact for them, not a debate [12].

Generation Z is not good listeners and they lack interpersonal skills. Communication with others generally consists of the use of the World Wide Web. Due to the curiosity in new technology, the members of Generation Z can usually be found at locations that offer the advantage of being hooked up to the Web [13],[14]. The financial hardships in the homes of many Generation Z youth might have led them to be highly conscious of the importance of money, and so saving might be of high importance to them [8]. As a group raised in constant war, contemporary youth may view the world with the belief that the world is "unsafe," yet at the same time, they may have greater global awareness as result. These consequences have resulted in a generation that potentially values fiscal responsibility, tolerance of others, education, employment flexibility, and networking abilities

In the Sri Lanka context, Generation Z has been identified as a generation that is driven by systems and great customer service. According to the same literature, Gen Z prefers systems and faster responses, would like to work in an environment that is systematic and helps them thrive. Most of them have highlighted that their networking skills are the most prominent. The Gen-Z's in Sri Lanka are

confident in communicating and making friends easily. At the same time availability of a peaceful post-war Sri Lankan environment, ability to access multiple devices: cheaper smartphones, Tabs, internet and their network with confident communication and networking skills have been identified as the key characteristics specific to Generation Z in Sri Lanka [13].

B. Generation Z and technology

Generation Z is significant because they were born into the digital age. So, they are also known as digital natives. With the web revolution that occurred throughout the 1990s, they have been exposed to an unprecedented amount of technology during childhood and adolescence. Further, because of their early exposure to large amounts of data and colourful graphics from the Web, they are very comfortable with maneuvering varying digital environments and engaging with new online applications [15]. No generation has demonstrated a level of proficiency or comfort with technology at such an early age as Generation Z [16]. So, their confidence in front of technological sources is predicted to be high.

They are the first mobile mavens in the world. Their preference for wireless, touch-operated screen iPads and smartphones rather than the ones limited with wired connectivity, showcase their familiarity with mobile technologies. Many in this group have hardly used comparative alternate research tools such as newspapers, journals or books because of their predisposition to the Internet. So, for this generation, it's only a matter of a click to emerge in the pool of knowledge when required. So, the cognitive decision-making ability based on the readily available knowledge sources is preferably high within this generation.

C. Generation Z and social interest

With advanced communication technology being an integral part of their lives, they are also more likely to communicate differently [17], [18]. This is the first generation to have Internet technology, so readily available. They take the Internet for granted and believe web sites such as Google, Facebook, Instagram and Snapchat as their community. While Facebook might have been the preferred social media outlet of previous generations, it seems Gen Z is more receptive to Snap Chat, Instagram, and Twitter, using short audio and video clips to complement their information intake. Within this community of cyberspace, a person can have many acquaintances without personally meeting anyone. Therefore, comparatively the social connectivity along with social pressure is predicted to be high within this generation.

So as per the previous studies similar to personal behavioral decisions, buying decisions are also significantly conditioned by one's generation; peers often guide product and brand choice, directly or indirectly. Thanks to tremendous technological advances and consumer connectedness, it is easy for companies to adapt their offering to the unique expectations of younger consumers.

Considering the rapid growth of this market segment as well as the significance of their demands and decisions, there arises a need to review and critique the mounting knowledge on consumer behavior towards Generation Z.

There have been some previous attempts to review some of the aspects of consumer behavior of Generation Z such as consumers' perceptions and expectations, as well as the potential impact of those expectations on the retailing industry in the years to come [19]. Payment options, perceived enjoyment, peer and social influence, product design and product brand as the key influential factors that control the buying decision of Gen Z [12]. Since this generation is yet to significantly enter into the good market there is a significant deficit of proper consumer behavioral analysis on this segment. The study aims to provide important insights into the habits and purchasing behavior of Generation Z.

D. Smart phones

The smartphone industry is a growing industry characterized by high competition and rapid technological advancement. Consequently, smartphones have undergone rapid innovations and even support advanced computing and communication features with: very high-speed data and network communication features, cutting-edge photo and video capturing and editing skills, exceptional personalization options, high-end designs, and superior processing competences. Not only this, but smartphones are becoming ubiquitous and expanding in functionalities to include: entertainment, payment, travel management, video and photo sharing, location services and much more. As the smartphone industry expands their product offerings have become more complex to include more than just the handset, but also the whole package from software systems, applications and network services.

The usage of the Internet, which has turned out to be groundbreaking in communication, has exhibited a rapid increase in the last decade. The effect of Internet programs, which have shown rapid growth during this time, on regular users' psychological and social lives, has become a subject of study [13]. Today, mobile phones have become an important part of the daily lives of individuals; they are accepted as indispensable tools to communicate with others; to call family members and friends, send text messages, connect to the internet, play games, listen to music and chill out [14]. The usage of smartphones is common among young people, more than other user groups, especially due to the design features. These features include emotional functions which are creating autonomy within the family and creating a source of entertainment [17]. Thus, designing relevant offerings and knowing what attracts customers becomes very important to increase sales and revenue.

III. METHODOLOGY

A structured questionnaire was used as a research instrument to collect primary data. The knowledge gained by the literature review and the information collected from the interviews was used to prepare a structured questionnaire. The initial section was used to gather general information (demographic factors and personality traits) about the respondent while the second section was used to measure the influence of different factors on the purchasing behavior of Generation Z consumers. The respondents are requested to rate the degree, to which the performance of each measure is actually achieved using a five-point Likert scale.

TABLE II. DEMOGRAPHIC PROFILE OF THE RESPONDENTS

Attribute	Grouping	Value	Value (%)
Gender	Male	67	45%
	Female	83	55%
Age	19-24	150	100%
Education level	O/L	6	4%
	A/L	17	11%
	Diploma	36	24%
	Undergraduate	88	59%
	Graduate	3	2%
Undergraduate Sector	Private	13	15%
	Government	75	85%
Employment Sector	Government	13	15%
	Self	12	14%
	Private	62	71%
Job Secotrs	IT	13	15%
	Apparel	14	16%
	Banking	12	14%
	Food & Beverage	13	15%
	Health	15	17%
	Transport	9	10%
	Other	11	13%

A sample of 150 respondents under stratified sampling methodology, within the age limit of (19-24) was chosen. The respondents are briefed about the objective and purpose of the survey beforehand and were requested to fill the questionnaire, which contained both open and close-ended questions. The study uses the context of purchasing mobile phones, in order to capture the purchasing behavior of the target market. Respondents included undergraduate and non-undergraduate students from both public and private universities, and also included those employed in the Western Province of Sri Lanka. Table II explicitly provides the demographic profile of the respondents

The data collected from these sources were statistically analyzed using SPSS. Validation of the results was undertaken with the supervision of academic and non- academic experts in the field.

IV. FINDINGS AND DISCUSSION

With the evaluation of results in the sample of Generation Z, several specificities of them were exposed which were

needed to be understood by marketing personnel before a product launch for this market segment.

Most importantly 66% of Generation Z have agreed that they research the product, prior to purchase. In this study, we have further examined the exact resources of information which they obtain for their product-based research. It was found that 42% of them receive information from family and friends while 18.7% of them stop at mobile phone retail stores for further information as per Table III below. Based on the ranks, family and friends, YouTube and retail stores are the three main strategic information sources.

TABLE III. CROSS TABULATION-1

When you purchased your current phone, where did you obtain the information? (1-Family or friends,2-YouTube,3-Magazines and papers,4-Mobile phone retail store,5-Facebook,6-Instagram,7-Other)				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	63	42.0	42.0	42.0
2	30	20.0	20.0	62.0
3	12	8.0	8.0	70.0
4	27	18.7	18.7	88.0
5	13	8.7	8.7	96.7
6	3	2.0	2.0	98.7
7	2	1.3	1.3	100.0
Total	150	100.0	100.0	

In this segment, 80% of respondents stated that on a daily basis they use the function of social networking on their mobile phones. From that social media networking, we have ranked the most frequently using social media through the survey. Based on that WhatsApp with 44% Facebook with 40.7% and Snapchat with 7.3% have ranked as the top 3 social connectivity for this segment. Therefore, according to the results, Facebook would be a significant platform for marketers to reach Gen Z. Significantly, the survey data on the level of influence from social media on product purchase revealed that 90% of Generation Z, believe that social media is a strong influence on the purchase decision. That indicates that both factors are very important and the most effective when trying to reach this segment. Marketing personnel needs to be mindful of this fact when they design marketing campaigns in social media platforms for Generation Z.

Seventy-eight percent of Generation Z respondents have stated that they work with the Internet daily while 85% of them have achieved a high degree of aptitude on computer literacy. This indicates a high level of technological know-how. This achievement has strengthened the confidence level of Generation Z where 60% of respondents are confident about their knowledge on mobile phones before product purchase. Surprisingly, there are no strong disagreements to this question. It could be concluded that Generation Z has a high level of computer literacy and has a high level of Internet literacy

TABLE IV. CROSS TABULATION-2

What is most frequently using social media? (1-Facebook, 2- Snapchat,3-WhatsApp, 4-Instagram, 5-Twitter,6-Other)				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	61	40.7	40.7	40.7
2	11	7.3	7.3	48.0
3	66	44.0	44.0	92.0
4	9	6.0	6.0	98.0
5	1	.7	.7	98.7
6	2	1.3	1.3	100
Total	150	100.0	100.0	

TABLE V. CROSS TABULATION 3

I have adequate knowledge about the features of the product (mobile phone) to compare and purchase the best mobile phone. (1-Strongly disagree, 2-Disagree, 3-Neutral, 4-Agree, 5-Strongly agree)				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
2	15	10.0	10.0	10.0
3	45	30.0	30.0	40.0
4	79	52.7	52.7	92.7
5	11	7.3	7.3	100.0
Total	150	100.0	100.0	

When it comes to decision-making power, 72% of Generation Z believes in their own ability before placing reliance on other ideas. Fifty-two percent have some influence from friends while 22.5% choose products while depending only on their opinions when purchasing the product and a similar number considers the recommendation from family.

Twenty-eight percent of respondents rely on media and advising before purchase and more than sixty-eight percent seek the recommendation of the friends in the product purchase than any other way. This indicates the level of influence from friends for this generation's decision-making process.

TABLE VI. CROSS TABULATION-4

Who influences most in the purchase of your mobile phone? (1-Family, 2-Friends, 3-Spouse, 4-Self, 5-Other)				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	29	19.3	19.3	19.3
2	78	52.0	52.0	71.3
3	8	5.3	5.3	76.7
4	34	22.7	22.7	99.3
5	1	.7	.7	100.0
Total	150	100.0	100.0	

Furthermore, 64.7% of Generation Z has agreed to the point where they can have better recognition from their friends if they listen to their friends in a product purchase. Because within that group 52% of respondents believe that they will gain this recognition by imitating the behavior of peer groups in purchasing while 50% of respondents decide to enhance their status within their peer group by using a purchase. The importance given for social media connectivity is relatively high within this generation according to the above-mentioned statistics. Most importantly it also indicates that this generation tries hard to maintain and develop their self-esteem and personalities in front of social media premises. Therefore, social media influence would be another important point to be considered for the successful launch of the marketing campaign for marketers.

TABLE VII. CROSS TABULATION -5

Please indicate the extent to which you agree or disagree with the following statements [ii. I can develop and enhance my status within my peer group/social media group if I purchase a mobile phone.] (1-Strongly disagree, 2-Disagree, 3-Neutral, 4-Agree, 5- Strongly Agree)				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	7	4.7	4.7	4.7
2	19	12.7	12.7	17.3
3	49	32.7	32.7	50.0
4	63	42.0	42.0	92.0
5	12	8.0	8.0	100.0
Total	150	100.0	100.0	

Better features compared to similar phones are the main reason for 56% of respondents to choose their current product. More than 84% of respondents have mentioned that they use the SMS function on a daily basis while 64% of respondents have mentioned that they use the function Gaming on a daily basis. Only 17.5% of respondents have mentioned that they use the function Video call on a daily basis while 41% of respondents and 46% of respondents have mentioned that they use the function Photo shooting and 52% for GPS respectively on daily basis. Surprisingly, 90% of the respondents have agreed that they use the function of web surfing while 89% of respondents use music listening function. Comparatively, surfing the web (90% of respondents), social networking (89% of respondents) and listening to music (89% of respondents) are ranked in the top three functions Generation Z interacts with on a daily basis.

Around thirteen percent of Generation Z has stated that price is a significant factor when considering the purchase. Surprisingly, the price factor seems less significant compared to other factors. Having said that, most of Generation Z (38.7%) has gone for the price band between Rs. 15,000 to Rs. 25,000, they do change their mobile phone approximately every 2 to 3 years and only 10% rely on the past experiences of the friends and family before a purchase decision.

TABLE VIII. CROSS TABULATION-6

What is the most influential reason behind choosing your current phone? (1- Priced competitively, 2-Catchy and attractive advertisements, 3-Better features compared to the similar phones, 4- Satisfied past experiences from family/friends, 5-Did not choose. It was a gift, 6-Other)				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	19	12.7	12.7	12.7
2	12	8.0	8.0	20.7
3	84	56.0	56.0	75.7
4	16	10.7	10.7	87.4
5	17	11.3	11.3	98.7
6	2	1.3	1.3	100
Total	150	100.0	100.0	

V. CONCLUSION

In conclusion, the Generation Z segment of the market differs in their characteristics, needs, attributes and work style from other generations. They are highly technically computer literate and primarily seek recommendations and concurrence of friends in product purchases. Family and friends, retail store and YouTube are the three main strategic information sources, while WhatsApp, Facebook and Snapchat are ranked as the top 3 social connectivity mediums for Generation Z. Therefore, according to the results, Facebook would be a beneficial platform for marketers to reach Gen Z. Comparatively, web surfing, social networking and listening to music are ranked as the top three mobile phone functions which they interact with.

Moreover, this generation tries hard to maintain and develop their self-esteem and personalities in front of social media premises. To cater to their demands with a proper strategy, it is important to investigate the determinant factors of this buyer's characteristics and actions. Implications are provided to assist in predicting potential consumer adoption behavior and in designing favourable shopping environments that are compatible with these specific consumer traits.

VI. LIMITATIONS

The main limitation of this study is the relatively small sample size. This was due to time and cost constraints. Research findings should be handled cautiously due to its level of generalizability. Future research is suggested to cover a larger size of the sample as the comparison with results based on a larger sample would help to validate the results. Further, the diversification of the sample across a geographic area, social strata, etc. could provide the opportunity to have a better understanding of the wider population.

VII. RECOMMENDATION FOR FUTURE WORK

The findings of the current study have significant implications for retailers and managers, suggesting that it is essential to deal with Generation Z consumers' demands regarding technology and social behavior. It is also evident that the level of technology usage and social media usage for the marketing of a product plays a significant role in shaping the

generation's consumer purchasing experience. This could have a possible influence on their decision-making processes. Overall results suggest that both manufacturers and sellers should focus more on enhancing product features (both soft and hard including design).

This research investigated gender differences very broadly which meant that it did not investigate the potential gender differences in depth or in detail. The author argues that a more detailed and in-depth investigation regarding gender differences would be interesting as it might suggest differences across such groups.

REFERENCES

- [1] M.Solomon, G. Bamossy, S.Askegaard, & M.Hogg, "Consumer as individuals," Consumer behavior: A European perspective, 5th edition. Harlow: , Pearson Higher Education, 2014.
- [2] D.Brooks, *The Social Animal*. New York: Random House Trade Paperbacks, 2012, p. 448.
- [3] R. Blackwell ,P. Miniard, and J.Engel, "Consumer behavior",Mason: Thompson, 2006.
- [4] A. Sankaran, Market Research, A Key Enabler of Digital Transformation, Retrieved from <https://greenbookblog.org/market-research-leaders/market-research-a-key-enabler-of-digital-transformation/search>. [Accessed October 25, 2018].
- [5] Sa, Mohamed ,Ali, and N. Ramya, "Factors affecting consumer buying behavior", 2016.
- [6] S. Lissitsa, O. Kol, "Generation X vs. Generation Y, A decade of online shopping," *Journal of Retailing and Consumer Services* 31, 304–312, 2016.
- [7] W.F.W Yusoff and T.S. Kian , "Generation differences in work motivation: From developing country perspective," *International Journal of Economy, Management and Social Sciences*, 2(4), 97–103, 2013.
- [8] P. Taylor and S.Keeter, "Millennials: A portrait of generation next." Washington, DC: Pew Research Center. Retrieved from <http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf>, 2010.
- [9] T. Spenny, "Millennial Shoppers Are Old News: Looking ahead to Gen Z, Financial services Consumer Trends Innovation and Digital Market Opportunities & Innovation Millennials", [Accessed September 25, 2014].
- [10] A. Turner, "Generation Z: Technology and social interest," *Journal of Individual Psychology*, 71(2), 103-113, 2015.
- [11] R. Zorn, "A new generation of graduate student, *The generation Z. College & University*", 92(1), 61-63, 2017.
- [12] American Academy of Child and Adolescent Psychiatry, "Facts for families: Multiracial children". Retrieved from http://www.aacap.org/App_Themes/AACAP/docs/facts_for_families/71_multiracial_children.pdf, 2011.
- [13] J. Chrishankar , "Sri Lankan 'Generation Z' - an Educators perspective Retrieved from: <https://www.linkedin.com/pulse/sri-lankan-generation-z-educators-perspective-janathanan/>, August 23, 2018.
- [14] R. H.Gouws and S.Tarp, "Information overload and data overload in lexicography," *International Journal of Lexicography*, 29 (3), 01-27, 2016.
- [15] S. Bradford, "Alternative Social Media as a Recruiting Tool for Generation Y and Generation Z", *International Journal for Innovation Education and Research* www.ijer.net Vol:-6 No-10, 2018 International Educative Research Foundation and Publisher © 2018
- [16] W.Palley, "Gen Z: Digital in their DNA. New York,"*Generation Z Goes to College*. Retrieved from http://www.jwtintelligence.com/wpcontent/uploads/2012/04/F_INTER_NAL_Gen_Z_0418122.pdf, 2012.
- [17] D.Bursch and K.Kelly, "Managing the multigenerational workplace", UNC Executive Development, 2004.
- [18] C.Turner, M. A. Hogg, P. J.Oakes, S. D.Reicher and M.S.Wetherell, "Rediscovering the social group: a self-categorization theory," *Handbook of Theories of Social Psychology* New York: Basil Blackwell, 1987.
- [19] S.Michelle, S.Violeta , T.E.Christopher , and D. Moira , "Exploring consumer purchase intentions towards traceable minced beef and beefsteak using the Theory of Planned Behavior", *An International Journal of HACCP and Food Safety*, 2018, No1.,33., 33-44, 2012.

Initiating customer relationship measurement at small and medium enterprises with low computational cost

Tharika Chalani Weerakoon*
Department of Physical Sciences and Technology
Sabaragamuwa University of Sri Lanka, Sri Lanka
thariweera@gmail.com

Kapila Tharanga Rathnayaka
Department of Physical Sciences and Technology
Sabaragamuwa University of Sri Lanka, Sri Lanka
kapila.tr@gmail.com

Abstract: Modern business entities re-engineer all core business and secondary business activities for customer satisfaction, thereby boosting profit margins. In re-engineering efforts, businesses require immense data processing and decision-making on customers and buying patterns. Still, SMEs face challenges when moving on to customer-centric marketing due to the less accumulated data and resistance towards investing in erudite decision-making tools. Hence, this current research study aims to provide a feasible data mining approach for SMEs in customer recognition with low computation complexity. Data accumulation has happened during the introduction of SME's mobile application to its customers. Hence, the dataset consists of demographic features age, gender, residency region and occupation of each customer. The proposed approach has two phases as follows; the first phase, the customer demographic data with the target variable of purchase value had subjected to data preprocessing. Null values and noise have treated with a binning method. In the second phase, feature engineering had carried out so categorical variables are in numerically manner. Therefore, the binary encoding was used for categorical. Finally, the dimensionality reduction of the processed data had done using Principal Component Analysis (PCA) to extract the most prominent and customer explanatory attributes within the SME. The PCA yielded 10% of a reduction in total explained variance percentage, meaning that the data had compressed. Using K-Means clustering 4 distinctive clusters were extracted. The usage of PCA had leveraged quick clustering and obtained 4 clusters represented the most impactful customers between the age of 0-17 and occupational level 10. With the implementation of PCA, the dataset narrowed down only with the most prominent features that an SME should care of and with this methodology SME can initiate the practising of more efficient customer data analysis using data mining and machine learning.

Keywords: Customer Relationship Management, Data mining, Principal Component Analysis, Small and Medium Enterprises

I. INTRODUCTION

Small and Medium Enterprise (SME) sector is an important contributor to many economies and has received special attention from government policy formulation bodies worldwide. More than 75% of enterprises are SMEs and according to World Bank reports, a quota of up to 60% of total employment and 40% of GDP is backed by SMEs in emerging economies like Brazil, India, China and South Africa. Furthermore, the World Bank emphasizes that the growth of SMEs substantially improves local economies of a particular country.

Technological advancements have influenced SMEs to engage in more competitive and dynamic business environments. Mainly, SMEs have focused on customer-oriented merchandising rather than traditional merchandising of goods and services. Hence, SMEs adopt mobile-based applications with online buying and selling approaches, accumulating data in databases. This accumulated data carry hidden information which can be used for effective and insightful decision making. To extract particularly hidden information, data mining concepts are being used.

Data mining yields productive outcomes for business acumen and SMEs and can be deployed to evaluate companies' own customer base. Usually, SMEs show resistance towards practising data mining at their enterprises due to lack of experience in data mining techniques and complexity in formulating the best data mining strategy at the enterprise level. Further, SMEs have the major concern of having small data sets compared to large scale organizations. Therefore, they are reluctant to invest since they perceive that data mining will not be effective in generating returns for their investment.

The study primarily aims to introduce an approach to initiating data mining at SMEs with respect to Customer Relationship Management (CRM).

II. LITERATURE REVIEW

A. Data Mining

Data mining is the excavating of useful and relevant information from considerably large data sets or databases. Data mining fundamentally objectify to provide more insight and understanding about the data structure and the important features of the dataset. The two goals of data mining are information description and prediction. SMEs can gain immense advantage within the industry by integrating data mining and knowledge discovery in accumulated data [1].

The process of data mining is a combination of an intensive set of steps from data gathering to knowledge extraction. Data will be gathered from different sources, primarily from data warehouses and secondary data from third-party data repositories. Then feature selection will be used to identify best data attributes which facilitate desired data-driven objectives set in the business organization. The target data will be formulated and pre-processing of data will commence since many inconsistencies may exist within the data. The pre-processed data will be transformed into tensors or datagrams, which are the forms that will leverage effective

data mining. At the data mining stage, the data will undergo pattern recognition, regression analysis and different statistical analysis along with machine learning. At the final stage of extracted information, the knowledge will be accumulated and will be used for the organization's decision-making process.

Data mining leverages enterprises to escalate business and financing competencies. The top management including the board of directors in business organizations requires information for better decision making in the marketing of products, modes and privileges for satisfying customers, competitor analysis and human resources management. The back-end of the business analytics function usually supports data mining. The business analytics front-end functions compose of executive reporting metrics and organized information. An effective and efficient data mining process execution becomes a core competency for the organization, thereby providing a competitive advantage over its rivals in a form of valuable business intelligence. The practice of data mining within the organization, assures better strategic and contingency plan formulation and execution at organizations [2].

Data mining induces core competencies within the business's internal operations and resources. This is the most suitable and probable cause for an SME to adopt data mining. SMEs also aspire to reach appropriate markets and to evolve the products' brand values within the industry they do business with. It is critical to understand how the core business functions of an SME adds value to its end products or services as per the perspective of customers. Hence, if SMEs can see the benefits that they can yield by data mining different sources of data, then they will be motivated to get started by articulating effective business intellectual inferences [3].

B. CRM

CRM is identified as a business model that integrates all business functions; primary and secondary as logistics, procurement, sales and marketing, and also emphasizes customer expressed satisfaction and unrevealed needs to assists business organizations to build loyal, profitable relationships with customers. It assists the monitoring process, along with, evolving the relationships for a better larger market share. The CRM concept first emerged among the vendor community in the mid-1990s and has since generated much interest among the academic community, afterward [4].

CRM has 4 different dimensions as Customer Identification, Customer Attraction, Customer Retention and Customer Development. These dimensions can be identified as the customer management life cycle. Data mining assists SMEs to analyze their datasets and model information and provides recommendations for each CRM dimensions. Research has reviewed and demonstrated the major data mining techniques that can be used in CRM. Accordingly, most deployed data mining techniques for a CRM are, Association rule, Decision tree, Genetic algorithm, Neural networks, and Clustering and Regression analysis. For an SME, which is initiating CRM, needs to identify its customers. Hence, the customer identification phase is vital to engage in. At the stage of initial customer identification, businesses tend to engage in TCA and customer segmentation. TCA involves seeking the profitable segments of customers through analysis of customers' underlying characteristics, whereas customer

segmentation involves the subdivision of an entire customer base into smaller customer groups or segments, consisting of customers who are relatively similar within each specific segment [5]. To commence TCA, businesses engage in Exploratory Data Analysis (EDA) with the aid of internally generated data, as well as third party data. Customer segmentation is identifying different clusters of customers according to enterprise defined objectives. For customer segmentation, there are several data mining techniques such as Principal Component Analysis (PCA), Association rule, Classification and Clustering.

C. Initial customer identification at SMEs

When initiating data mining at SMEs, there should be predetermined customer-oriented objectives, that are being clearly defined by the board of directors. Unless otherwise, a lack of the desired goal in data mining will retard the effectiveness of decision making.

Internally generated data at SMEs tend to have complex dimensionalities, which will create unwanted analytical burdens with regard to data attributes. More data attributes within a dataset will induce distractions. Hence, it is wistful to use dimensionality reduction techniques before the initial customer analysis.

SMEs face several challenges at the commencement of data mining, for initial customer identification. The most challenging factor that influences the retardation in initial customer identification is having small datasets. According to experiments on data mining, satisfactory results were only obtained with a large set of data. For small training sets, the performance and effectiveness in data mining approaches may not be good enough, or the learning task for a computer can even not be accomplished. As a result of this deficiency, the applications of neural networks get severely limited. When considering the root cause for why small datasets cannot provide enough information, the existence of gaps between samples can be noted [6].

The analysis will also be complicated due to a large number of attributes associated with the dataset of SMEs. When SMEs gather data, due to lack of experience, they gather unnecessary, misleading demographical attributes pertaining to customers. Analyzing the most relevant attributes is advantageous to SMEs in terms of computational costs. Hence, the PCA can be deployed to reduce dimensionality and thereby promote the analysis of the most important and relevant attributes of customers for SMEs.

D. PCA

PCA is a data mining approach which is used for dimensionality reduction in a dataset. Usually, customer data gathered at SMEs has lots of components and attributes as discussed. Hence, the dataset is bulky and tedious to be used for an effective data mining process. PCA is a statistical technique of data reduction and aims to produce a more significant, smaller quantity of derived variables representing a larger number of original variables. This analysis helps to simplify subsequent data analysis [7]. PCA is implemented to extract the most important information from the statistical data. It represents those important information as a set of newly formulated orthogonal variables which are known as principal components. PCA also focuses to display the similarity patterns that can be derived between the observations and variables as points in spot maps.

Mathematically, PCA has a greater dependence upon the positive semi-definite matrix Eigen-decomposition and also upon the rectangular matrix Singular Value Decomposition. More importantly, it is determined by eigenvectors and eigenvalues [8]. The main goals of PCA are extraction of the most important data from the dataset, compression of data, data description simplification, increase of the ability to engage in observation and variable structure analysis and data compression through dimension reduction. This would be achieved without a significant information loss and is used in image compression.

III. METHODOLOGY

The methodology was formulated considering the research objective of introducing feasible data mining for SMEs in customer identification, considering the need for keeping costs and complexity low. The methodology composed of two phases as TCA is done with the aid of the EDA data mining techniques. The second phase was carried out to identify clusters of customers within the SME using PCA.

The dataset used was accumulated as a result of the campaign hosted for the introduction of mobile applications. Mobile applications at its user registration gather demographical data for each customer. These demographic data were under the attributes of age, gender, the region of residence and occupation along with purchases made by each customer. All the demographic data are in categorical form, whereas the attribute, purchase, the target variable is in numerical form.

Initially, the extracted dataset was preprocessed to reduce the noise and to handle null values. Then the noise reduced dataset was used to engineer the features. Feature extraction conducts an attribute reduction process. It should be clarified that feature selection only ranks the existing attributes considering the predictive significance of each attribute, whereas feature extraction essentially transforms the attributes. The transformation of attributes or features generates linear combinations of the original attributes presented in the dataset [9]. The customer-related dataset mostly consisted of categorical data which need to be encoded for further data mining. Binary encoding was used to transform categorical data attributes to arithmetic-friendly values for further analysis. This feature was deployed with the aid of the Python data mining library, Scikit Learning.

Then the processed customer data underwent the first phase of the analysis, which is EDA. There, the overall customer data distribution for customer profitability was measured. This approach was used to identify the behaviour of customers for the organization's profit margins. Based on different demographic attributes, mainly focusing on age, gender, occupation and residency level, EDA was carried out. This enabled the linking of the profitable target market with customer attributes. Specifically, the univariate distribution of the target variable is considered, concerning purchases versus purchase count. Hence, at a glance, the indirect representation of the profitability distribution at the SME was illustrated. Then further analysis was carried out considering other customer demographic features, to understand the target customers in terms of the demography. The EDA was implemented with the use of Python data mining libraries; Numpy, Pandas, Matplotlib.

The second phase of the methodology was designed to identify the underlying customer segmentation within the SME. PCA was used since the dimensionality of the dataset gets reduced. This improves the analysis with the most prominent features that affect the SME's customer base, highlighted. The total dimensionality was reduced to 12 facets. As data being compressed, without a data loss, two meaningful and prominent features were claimed to be principal components for customer segmentation. Then K-Means clustering was deployed using R to identify the most relevant clusters with less computational efforts.

IV. RESULTS

A. Phase 1: EDA for TCA

The EDA conduct checks to ensure the quality of data, summary statistics calculation, and appropriate graph plotting. For SMEs, EDA is crucial, and considering the management viewpoint, objectives were defined. Accordingly, the following EDA results were obtained.

Fig. 1 illustrated below, shows Purchase versus the Purchase Counts, for the target variable, purchase distribution. The Pearson r is given as 0.27.

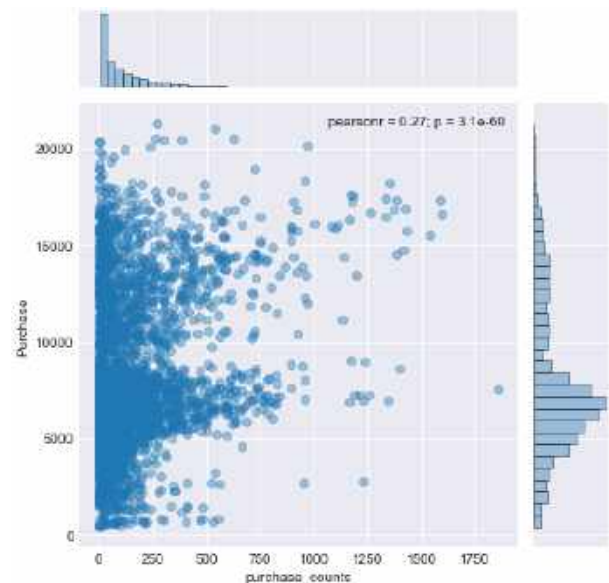


Fig. 1. Purchase vs Purchase count plot

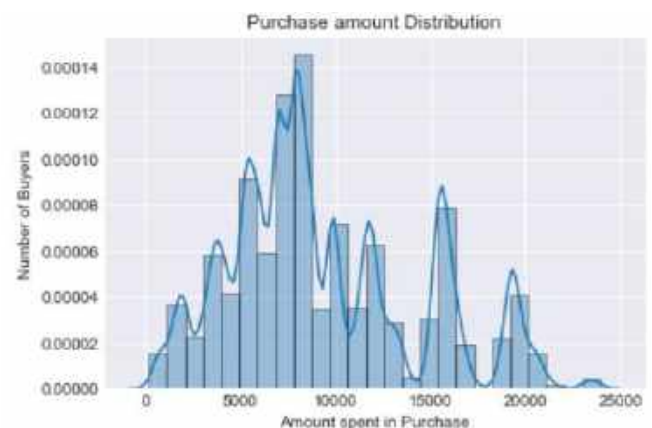


Fig. 2. Purchase amount distribution

Fig. 2, represents how to purchased values are distributed to buyers. It is revealed that purchase amounts between 5000 and 10,000 have attracted many numbers of buyers. This means that the products within the range of 5000-10,000 indicate a trend of fast-moving products.

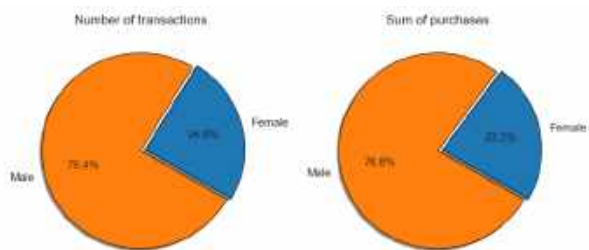


Fig. 3. Sales based on Gender

The above illustration in Fig. 3 shows how the target variable varies based on Gender. The male and female proportion in the number of transactions and sum of purchases show similar behaviours.



Fig. 4. Mean sales consideration based on Gender

The representation in Fig. 4 implies that the mean purchases of the two genders are almost equal.

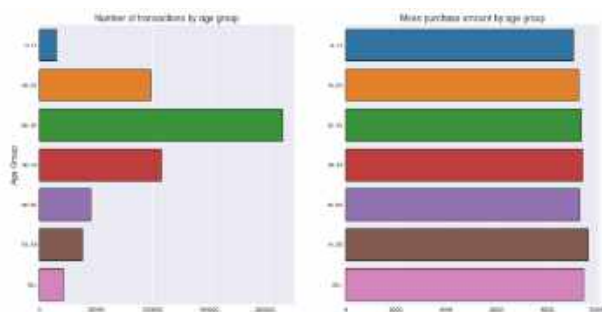


Fig. 5. Age-based sales distribution

As in Fig. 5, the number of transactions indicates wide fluctuations. But the mean purchase amount depicts that the age groups have contributed to purchases in similar amounts.

The illustration in Fig. 6 depicts how the existing customer base behaves within their demographic features for profit margins.

EDA has shown conflicting results which confuse and

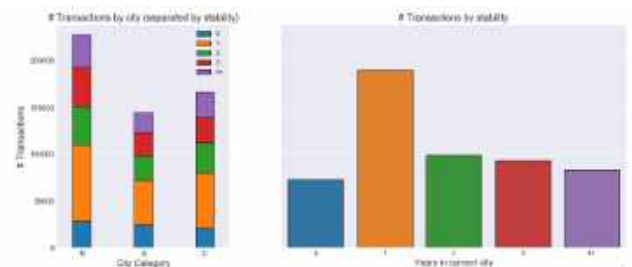


Fig. 6. Sales consideration on customer residency

retard the entire data mining approach within SMEs. Only with an EDA, SMEs get misleading business intelligence and may give up the opportunity to implement the most promising applications of machine learning and deep learning.

Hence, PCA can be used to infer the most effective demographical features that an SME should consider for CRM and target marketing. The following results were obtained via PCA conducted using R studio. According to the used dataset, the extended version of PCA, the MCA has used, so that the best mix of sub-demographical features can be extracted for further analysis.

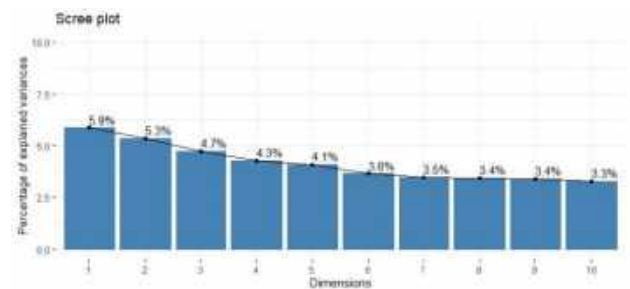


Fig. 7. Dimensionality scree plot

According to the scree plot in Fig. 7, the first data feature dimension holds 5.9% Multiple Component Analysis (MCA) and the second dimension shows 5.3% MCA. The aggregated MCA is around 10%. This means that the dataset does not hold distinguishably different data features. Hence, customers are not distinguishable or different from one another when it comes to purchases.

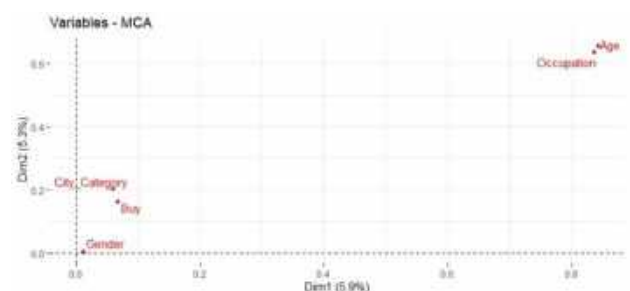


Fig.8. Dimensionality comparison with principal component

According to the representation in Fig. 8, Occupation and Age factors need to have greater focus as these factors will help SMEs to diversify its customer base. Here, the comparative association was considered between the most prominent, first two dimensions.

The following heat map (Fig. 9) shows how sub-attributes impact on the initial customer identification (the first phase of TCA) at SMEs.

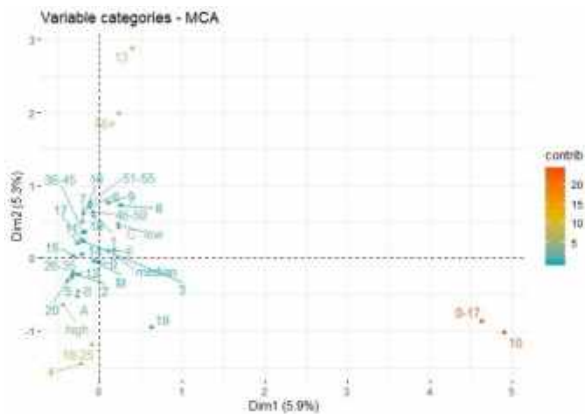


Fig. 9. Feature contribution

Features shown in red color are the directly, impactful features when comes to customer diversification along with the segmentation. Hence, Occupation 10 and Age range 0-17 need to be carefully considered when formulating strategic decisions.

Finally, customer segmentation and identification can be formulated and it is shown in Fig. 10 with 4 clusters.

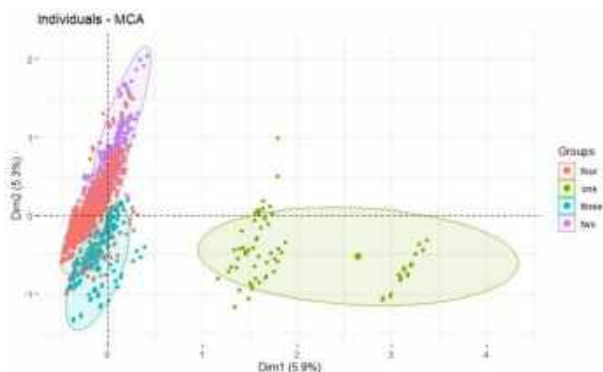


Fig. 10. Final customer segmentation

V. CONCLUSION

The study initially aimed to introduce a simple, yet robust data mining approach for SMEs to initiate fully pledged CRM. The most crucial step of CRM, which was identifying the existing customer base was achieved via two customer analytical approaches; TCA and customer segmentation.

For TCA, the data mining technique of EDA was used and effective decisions were drawn with simple computational efforts. The extracted, processed dataset yielded meaningful information that facilitates decision making at the SMEs.

The customer segmentation, identifying the most effective and promising customer base in terms of profits was done using PCA. The use of PCA technique compressed the dataset dimensionality and had facilitated to determine the principal components and sub-features which are impacting the customer behavioral and property-based segmentation.

Data mining techniques; EDA and PCA have improved the feasibility of data mining at SMEs in recognition of customers with low computation cost and low complexity.

The use of PCA has induced several theoretical implications. PCA analysis was based on the assumption of linearity. PCA aims to identify the orthogonal projections within the dataset, which has the highest variance, thereby determining the hidden linear correlations between variables. But when considering the data generated at SMEs, there can be non-linearly correlated variables within datasets. Hence, PCA is not adequate for the determination of the most prominent variable. PCA was also based on the assumption that the highest variance and lowest variance are the most important. This assumption holds implications when engaging in blind source separation. But in this study, PCA was used to determine the most prominent data attribute, which is a means of noise reduction. Therefore, this assumption is advantageous to study prediction precision. Theoretically, PCA is not a scale variant analysis, i.e. PCA makes rotation transformation within the dataset, which nullifies the effect of data scale or in other words, PCA does not normalize the dataset. This mentioned effect has little impact on the SMEs datasets, as the scale of data does not change predominantly over a while. But in a case of scale differentiation in variables, PCA will give quite a different set of components as principal components. This can only happen if the SMEs try to deviate from their core business activities or adopt a serious change within their organizations.

The main practical implication that can be encountered is the reluctance of management at SMEs to adopt data mining and machine learning approaches within the business intelligence formulation process. This is mainly due to the belief within SMEs, that predicting the future is not for the small or medium-sized businesses, but for large scale, multinational companies. SMEs need to adopt these approaches to sustain within the industry and also to gain a competitive edge than its rivalries.

REFERENCES

- [1] M. Baer, T. Ariyachandra, and M. Frolick, "Initiating and Implementing Data Mining Practices within a Small to Medium-Sized Business Organization," *J. Econ. Bus. Manag.*, vol. 1, no. 4, 2013.
- [2] P. M. Lee, "Use Of Data Mining In Business Analytics To Support Business Competitiveness," *Rev. Bus. Inf. Syst. - Second Quart. 2013*, vol. 17, no. 2, pp. 53–58, 2013.
- [3] S. Y. Coleman, "Data-Mining Opportunities for Small and Medium Enterprises with Official Statistics in the UK," *J. Off. Stat.*, vol. 32, no. 4, pp. 849–865, 2016.
- [4] N. Gordini, I. Sanpaolo, and V. Veglio, "Customer relationship management and data mining: A classification decision tree to predict customer purchasing behavior in global market," in *Handbook of Research on Novel Soft Computing Intelligent Algorithms: Theory and Practical Applications*, ch 1, Eds. Vasant Pandia, 2014.
- [5] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Expert Systems with Applications Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2592–2602, 2009.
- [6] R. Andonie, "Extreme Data Mining: Inference from Small Datasets," *Int. J. Comput. Commun. Control*, vol. V, no. 3, pp. 280–291, 2010.
- [7] S. S. Bhaduria and M. Pradesh, "Introduction to Principal Component Analysis in Applied Research," *New Man Int. J. Multidiscip. Stud.*, vol. 1, 2014.
- [8] U. Sarkar, F. Sciences, S. Datta, and F. Sciences, "Principal Component Analysis," *Int. J. Livest. Res.*, vol. 7, pp. 60–78, 2017.
- [9] C. Nithya and V. Saravanan, "A Survey of Feature Extraction and Feature engineering In Data Mining," *IOSR J. Eng.*, no. Iccids 2018, pp. 83–87, 2018.

Identification of factors and classifying the accident severity in Colombo - Katunayake expressway, Sri Lanka

M.A.K. Kushan*

Department of Statistics & Computer Science
Faculty of Science
University of Kelaniya, Sri Lanka
kalanakushan290@gmail.com

N.V. Chandrasekara

Department of Statistics & Computer Science
Faculty of Science
University of Kelaniya, Sri Lanka
nvchandrasekara@kln.ac.lk

Abstract: Sri Lanka's expressway system was launched in 2011 and now owns three major expressways. Many peoples choose expressways rather than normal ways due to the reasons of time, traffic, easy of driving, etc. According to police reports of highway main traffic police branch, in recent years the number of accidents occurring in expressways is increasing drastically. Nowadays, the rate of accident occurrence in Colombo-Katunayake Expressway is high compared to the other two expressways and there was no previous research has been done in Sri Lanka regarding accidents on Colombo-Katunayake expressway. Therefore, the objective of the study was to identify the factors contributing to accidents on the Colombo-Katunayake Expressway and to develop appropriate machine learning models to classify the severity of the accidents. In this study, 704 total accident cases were considered during the period 2013-2019. Chi-square test, logistic regression, and Kruskal-Wallis tests were used to identify the association between the accident severity and other influential variables found from the literature. Finally, seven variables: time category, driver's age category, vehicle type, the reason for the accident, number of vehicles involved, cause for accident and rainfall were identified as influencing variables to accident severity under 5% level of significance. Naïve Bayes classification algorithm and probabilistic neural network (PNN) were used in the study to forecast accident severity. A random under-sampling technique was used to overcome the class imbalanced problem persists in the data set considered in the study. The final models developed using the Naïve Bayes algorithm and PNN exhibit 72.14% and 74.29% overall classification accuracy respectively. Both aforementioned models can be considered as suitable models to forecast accident severity in the Colombo-Katunayake expressway where the PNN model exhibits slightly higher accuracy. The final models developed by this study can be used to implement safety improvements against traffic accidents in expressways of Sri Lanka.

Keywords: Classification, Class imbalance problem, Naïve Bayes algorithm, Probabilistic Neural Network (PNN), Road accident

I. INTRODUCTION

When we look back at the transport history of Sri Lanka, the British era comes to mind in a very specific way. They have made several changes by expanding the road infrastructure, increasing the vehicle population and by developing the railways to transport goods. After the independence and the new government, Sri Lankan roads were further improved. With the rapid growth of roads, the risk of accidents was increased. By considering the situation,

the first Traffic Act was enacted in Sri Lanka in 1934 and since 1938 the police have reported traffic accidents [1].

In Sri Lanka, there are 40,887 road accidents every year and an average of six deaths a day. Further, for a year nearly 9,000 motor vehicle accidents, over 5400 accidents due to speed occurred. The records reveal that over 740 pedestrians die every year and an average of two pedestrian deaths per day. Annually among the total road accidents, over 2470 people are faced with accidental deaths, 700 are left in serious injuries while some people with lifelong consequences [2].

With the introduction of the expressway system in Sri Lanka, people tend to use the expressway system apart from main roads for their long commutes due to the time, traffic and ease of driving. This leads to an increase the number of vehicles in expressways of Sri Lanka. As a consequence, the number of traffic accidents on highways also has been increasing dramatically in recent years. When considering all three expressways in Sri Lanka, there were 59 accidents in 2011, 449 accidents in 2012, 409 in 2013 and similarly, in 2018, 597 accidents have been reported. Overall, 3724 accidents were reported on the three major Expressways between 2011 and 2018. During this period 40 peoples died in fatal accidents [3]. Data reveals that the number of accidents has been increased over the last years s.

There are currently three major highways in Sri Lanka: The 'Southern Expressway (E01)', which was opened in 2011 and was the first expressway in Sri Lanka to cover 126.9 kilometers. The 'Colombo-Katunayake Expressway (E03)' was opened in 2013 and which connects the Colombo and Bandaranayake International Airport with a length of 25.8 kilometers and 'Outer-circular Expressway (E02)' was opened in 2014 from Kaduwela to Kottawa covering a distance of 18.9 kilometers [3].

In recent years, the incidence of traffic accidents on the Colombo-Katunayake Expressway is high compared to the other two expressways [3]. Apart from ordinary factors, climate factors may increase the number of accidents in the Colombo-Katunayake expressway. Therefore, it is imperative to identify the most likely causes of highway accidents in the Colombo-Katunayake expressway.

A motor traffic accident is any kind of vehicle accident that happens on a public highway. It includes a clash between vehicles and animals, vehicles and fixed obstacles, vehicles and pedestrians and also a single-vehicle accident without other road users [4].

Road accidents can be identified as a growing problem nationally or internationally. According to the World Health Organization (WHO) report, the number of accidents from non-communicable diseases increased from 28.1 to 49.7 in 1990 to 2020, with road accidents being the main cause of this increase [4]. Between 1938 and 1997, the number of road traffic fatalities in Sri Lanka increased tenfold and in 1997 there were 1835 deaths. Despite the need for prompt action to reduce this growing burden, successive governments at that time did not trust the country's road safety strategies and road safety research [5].

The main objective of most previous studies was to identify factors that significantly affect the severity of an accident [1,2,5]. Highway characteristics, atmospheric factors, and driving characteristics have been identified as factors contributing to the severity of a traffic accident, and usually one or more of the above-mentioned factors can contribute to the severity of the accident. [6].

According to the Institution of Highways and Transportations (IHT), there are numerous factors for road traffic accidents such as light conditions, weather, faulty design, vehicles with mechanical defects and Inadequate road infrastructure development [4].

While many researchers have not considered all the factors that contribute to an accident, but some believe there is a greater potential for the severity of accidents, such as time of the accident, age of the driver, gender of the driver, type of the vehicle, type of the accident, place, cause for accident. Other factors were not investigated due to significant limitations in the data obtained from accident reports.

In recent years, researchers have increasingly focused on identifying the factors that significantly affect the severity of road accidents [8, 9]. There are many techniques that researchers have used to study this issue including artificial neural network, cage logic processing, fuzzy ART maps, log-linear modeling, etc. [7, 9, 10]

The use of artificial neural network techniques to model traffic accident data reports helps to understand the drivers' behavior, road conditions, and weather conditions associated with the severity of various injuries. This will help decision-makers to develop better traffic safety control policies.

Many researchers used an artificial neural network to analyze the frequency of freeway accidents and pointed out that the artificial neural network system does not require a predefined underlying relationship between dependent and independent variables [8]. Studies have shown that artificial neural networks are a consistent alternative method for analyzing the frequency of freeway accidents [8,11].

Non-linear relationships between injury severity levels and crash-related factors were modeled by using a series of artificial neural. From those factors, it shows that artificial neural networks have better predictive power compared to other methods such as logistic regression, ARIMA, VAR, etc. [12]

Another research that focused on traffic accidents at signalized intersections, is divided the severity of the injury into three classes such as no injury, possible injury and disabling injury. And also founded that the multi-layered

perceptron (MLP) classification accuracy was higher than the Fuzzy ARTMAP [11].

Today road accidents become a vast problem not just for Sri Lanka but for all countries in the world. Therefore, it is imperative to take immediate steps to prevent accidents. The severity of the road accident on this study road is the greatest in property damages only accidents (86.22%) when compared with the Slight Injury (9.52%), Serious and Fatal injury (4.29%) [3]. Not a great deal of research has been conducted on expressways in Sri Lanka, and most of the research has been done using traditional methods such as logistic regression. Hence, this study is a novel approach to road accidents. The result of this research helps to place a policy of prevention for the police maker analyzing the road accident severity. This paper identifies the most likely causes of expressway accidents, builds the Naïve Bayes algorithm and Probabilistic Neural Network (PNN) to classify the severity of accidents on the Colombo-Katunayake expressway and compare the classification accuracy of these two models.

The organization of the paper is as follows: Section 2 consists of theories & techniques used for the study. Section 3 describes data pre-processing and preliminary analysis. Model building process and Classification of Accident severity are in Section 4 and Section 5 respectively. Performance measures explain in Section 6 whereas Section 7 contains results and discussion. Section 8 consists of conclusions followed by the acknowledgment and references conclude the article.

II. THEORIES & TECHNIQUES

Classification is an important data mining technology with a wide range of applications for classifying the various types of data used in almost every area of our lives. Classification analysis is the organization of data in a given class. This also known as supervised classification, uses the class labels provided to order data collection objects. Classification approaches usually use a training set that is associated with class labels that all objects are already known. The classification algorithm learns from the training set and builds a model. The trained model is used to classify new objects [13].

A. Naive Bayes classifier

The Naive Base algorithm is a simple probabilistic classifier that computes a set of probabilities by computing a combination of frequencies and values in a given dataset. The algorithm uses the base theorem and the assumption made in this classifier is that the predictors/features are independent. That is the presence of one particular feature does not affect the other. This conditional independent assumption is rarely true in real-world applications, and hence is characterized as naive, but the algorithm tends to perform well and fast in various classification problems [14].

B. Artificial Neural Networks (ANN)

When it comes to machine learning, it can train the computer to detect data patterns in different ways. This includes learning from experience, learning by examples, analogues learning, etc. With an increase in training time, the performance of any machine learning algorithm can be increased. The basis of the adaptive system is the machine learning mechanism, and Artificial Neural Networks (ANNs) are the most popular approach to machine learning [15].

The Artificial Neural Network (ANN) can be described as a technology that functions similar to that of the human brain. The human brain consists of a set of basic information processing units called neurons and there were approximately 10 billion neurons and 60 trillion connections between synapses [16]. By using multiple neurons simultaneously, the human brain can perform calculations, tasks, and decisions much faster than any high-performance computer present in today [15].

ANN models have the ability to mimic human information processing processes such as knowledge processing, prediction, and control. ANN system has gained much attention compared to other existing technologies, due to the reasons for learning spontaneously from examples, arguing about vague data, and responding to new information not previously stored in memory. ANN technology is now widely used in engineering, medicine, and many other fields. Therefore, this technology is gaining greater acceptance worldwide [15].

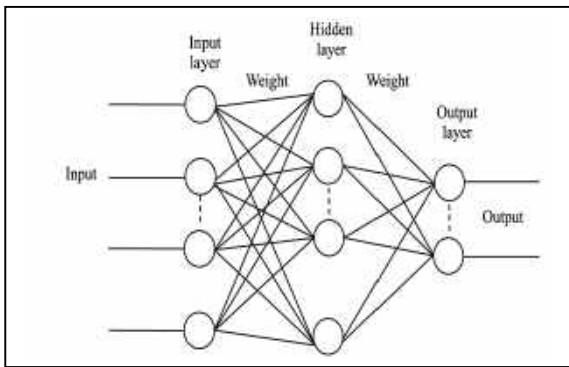


Fig. 1. Architecture of ANN

ANN is relatively new in the field of traffics engineering and accident analysis [15]. This new approach has rarely been shown in areas such as predicting traffic congestion [15, 16], determining truck attribute, and a few other applications. Further, the artificial network has been used to solve problems that have not been solved by statistical methods [8,12]. Researchers have done comparative studies of statistical methods with ANNs [15,17]. Their studies show that training in medium and large datasets can be very useful for ANN prediction [15, 16].

A neural network can learn both at the presence of teachers (Supervised learning) and without teachers (Unsupervised learning). Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. Supervised learning techniques mainly divide into two parts. They were regression and classification.

C. Probabilistic Neural Network (PNN)

The probabilistic neural network (PNN) is one of the most commonly used feedforward neural network techniques for classification and pattern recognition problems. Due to the ease of training and good statistical basis of Bayesian estimation theory, the PNN tool is a widely used technique to

solve many classification problems. When constructing a more accurate PNN model, care should be taken to determine the size of the network, the location of the pattern stratified neurons, and the value of the smooth parameter [18].

With the presence of a categorical response variable (accident severity), two main classification techniques (Naïve Bayes algorithm and PNN) were used in this research to classify the severity of the accident. Accident severity classification and performance analysis are described in the next sections.

D. Class Imbalance Problem

Data distributions are said to suffer from class imbalance when class distributions are very unbalanced. Most of the classification algorithms are more focusing on the classification of majority sample cases correctly while ignoring or misclassifying minority sample cases. Therefore, most classification learning algorithms provide low predictive accuracy for a class that has rarely occurred. Minority samples, except for the majority class samples, rarely occur but are very relevant to many real-world problems. There are different methods available to overcome the class imbalance in the dataset, which were the algorithmic approach, data-preprocessing approach, feature selection approach, etc. [19]. The multiple class imbalance problem arose because there were three classes (property damage, minor injury, serious injury) of the response variable in this study.

E. Random Under-Sampling

Under-sampling is one of the most used methods by many researchers to solve the class imbalance problem in data level approach. The most recently identified under-sampling method is random under-sampling, which is a non-heuristic method that attempts to balance class distributions by randomly deleting data from majority classes [20].

III. DATA PRE-PROCESSING & PRELIMINARY ANALYSIS

The accident data set used in this study was obtained from the expressway main traffic police branch in Kaduwela, Sri Lanka. Meteorological data (daily rainfall, daily average temperature) for the date of the accident have been obtained from the Department of Meteorology. All accidents on the Colombo-Katunayake Expressway from 2013 to 2019 were considered in this study which covers 704 cases. The police Accident Reporting System on the Colombo-Katunayake Expressway is poor and collecting data on the correct road accidents has been a great challenge in this study.

According to the variable definitions for the records of the general data estimation system, the dataset has drivers' records and vehicles' information. Some attributes which can be considered as input variables to models are time, driver's age, driver's gender, the ethnic group of driver, the total vehicle involved, cause for the accident, vehicle type, reason for the accident, injury severity, place category, rainfall, temperature. Additionally; the response variable of this research (accident severity) is also another attribute of a given accident record. The target variable, accident severity has three classes: property damages only (no injury), slight injury and serious and fatal injury. and Figure 2 illustrates the percentage of cases under each category.

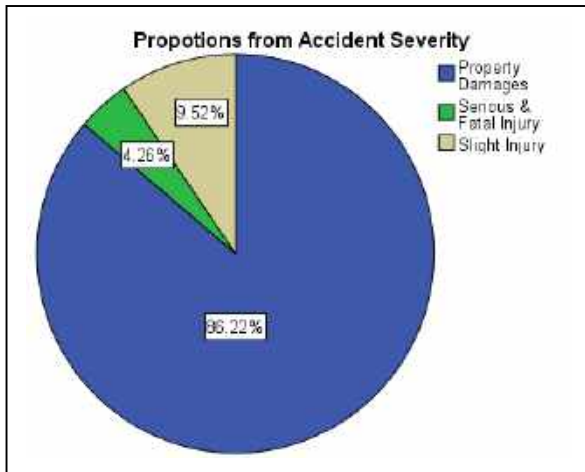


Fig. 2. Percentage of cases under each category of accident severity on the Colombo-Katunayake Expressway, Sri Lanka

In the original dataset, 86.22% of the cases belong to the class no injury (only property damages), 9.52% of the cases for slight injury and 4.29% of the cases for serious and fatal injury class.

The main objective of this research was to identify the major causes of accidents on the Colombo-Katunayake Expressway in Sri Lanka and to classify the severity of accidents. Developed machine learning models can be used to understand the relationship between driver characteristics, vehicle type, road conditions, and the environment and the severity of the accident. The results from this data analysis can provide critical information for the adoption of high technology in road accident prevention policy, especially in accident control mechanisms. The records in the dataset are arranged in pairs of inputs and outputs and each record has its own output. The supervised learning algorithms will map the input vector to the desired output for the given set of values.

The number of deaths due to road accidents on Colombo-Katunayake Expressway from 2013 to 2019 is very low. Hence, one response category was created as serious and fatal injuries which include all very serious accidents. Further, Expressway's main traffic police branch classifies the reason for an accident into eight main categories.

They are drink and drive, careless driving, rain, technical errors, hitting animals, sleeping, high speed and by other reasons.

Figure 3 illustrates the bar chart of the reason for accidents where the highest reason is other and the second highest reason is careless driving. Further, high speed and rain can be considered as potential reasons for an accident.

Eleven independent variables have been used to explain the variation in the dependent variable (accident severity). as: time category, driver's age category, driver's gender, the ethnic group of driver, number of vehicles involved, cause for the accident, vehicle type, reason for accident, place category, rainfall, the temperature in this study.

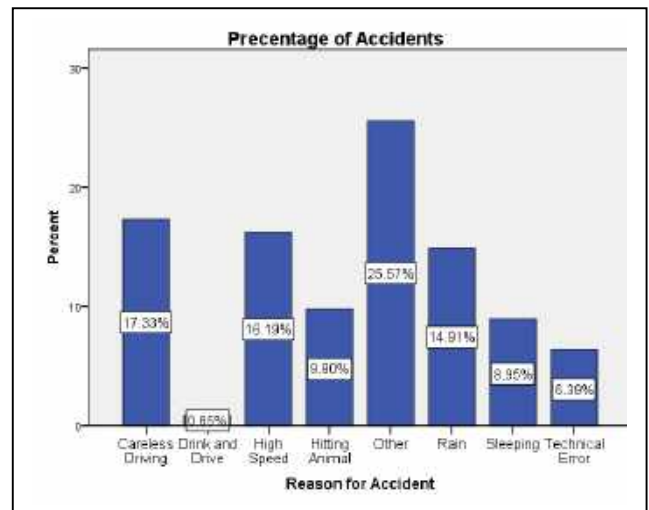


Fig. 3. Chart of the percentages of reasons for the accident

Among all independent variables, nine were categorical variables (non-climate factors): time category, driver's age category, driver's gender, ethnic group of driver, number of vehicles involved, cause for accident, vehicle type, reason for accident, place category and two were continuous variables (climate factors): rainfall and temperature.

Association between the response variable (accident severity) and categorical predictor variables were assessed using Pearson's chi-squared parametric test. Among nine categorical predictor variables driver's age category, time category, vehicle type, reason for the accident, number of vehicles involved and cause for the accident were identified as significant variables at a 5% level of significance. Furthermore, Logistic regression (parametric) & Kruskal-Wallis tests (non-parametric) were used to identify the association between the response variable (accident severity) and continuous predictor variables. Rainfall was found to be a significant variable in both tests.

Finally, nine variables: time category, driver's age category, vehicle type, reason for the accident, number of vehicles involved, cause for accident and rainfall were identified as an associated variable to accident severity under 5% level of significance.

IV. MODEL BUILDING PROCESS

The entire dataset was divided into two parts, 80% of data was used for model building and 20% of the data was used to validate the model. Since this is not a time series data, 80% of the data were selected in various ways for model building and the remaining 20% were used to test the built model.

Method 1: 80% (564) of the randomly selected data was used for model building and the rest 20% (140) data were used to test the model

Method 2: The first 80% (564) of the dataset was used to build the model and rest 20% (140) data were used to test the model.

Method 3: Considering the class ratio (property damage: slight Injury: the ratio of serious & fatal injuries is 86: 10: 4) of the original dataset,

(86/100) *564 = 485 'property damages only' accidents
(10/100) *564 = 56 'slight injury' accidents
(4/100) *564 = 23 'serious & fatal injury' accidents
(485+56+23) = 564 (accident records)

Altogether, 80% (564) of the data is randomly selected to build the model. The rest 20% (140) data were used to test the model.

Method 4: This was an improved version of the Method 3 with the under-sampling technique. By using the under-sampling technique considering the class ratio between property damages: slight injury: serious & fatal injuries which were started in *Method 3*, Out of 564 of data, 176 of the data was randomly selected as follows:

(86/100) *564*(20%) = 97 'property damages' accidents
(10/100) *564 = 56 'slight injury' accidents
(4/100) *564 = 23 'serious & fatal injury' accidents
(97+56+23) = 176 (accident records)

After the under-sampling technique, the original dataset ratio was changed to a new ratio (Property Damage: Slight Injury: the ratio of serious & fatal injuries is 55: 31: 14). The rest 20% (140) data which were used in *Method 3* were used to test the model.

Above data splitting methods were used in the model building process for both the Naïve Bayes algorithm and Probabilistic Neural Network (PNN).

V. CLASSIFICATION OF ACCIDENT SEVERITY

To clarify the severity of accidents, the above mentioned two classification techniques have been used and two techniques were compared using model accuracy.

A. Naive Bayes classifier

In Model 1, the model was constructed using the *Method1* data partitioning technique considering all predictor variables.

The key assumption of the Naïve Bayes Algorithm is all explanatory variables are needed to be independent with each other. Hence, it evaluated the multi-collinearity of all the above variables using the Variation Inflation Factor (VIF) value. Multi-collinearity was detected from two variables 'Vehicle type' and 'Number of vehicles involved'. Hence, Model 2 was created by dropping the two variables which exhibit multi-collinearity and using the *Method1* data splitting technique.

B. Probabilistic Neural Network (PNN)

Since the dataset is not a time series data, four processes described under Section 4 were carried out to select the 80% data to fit the model. When selecting the variables, the first evaluation was carried out using all eleven explanatory variables, but no significant classification accuracy in the built-in model which has been fit for the data. Then, the identified seven significant variables at a 5% significance level from Pearson's chi-squared test, logistic regression, and Kruskal –Wallis tests (as mentioned in Section 2) were used for further models.

In Model 3, *Method 1* data splitting technique was used to construct the model.

In Model 4, *Method 2* data splitting technique was used to construct the model.

In Model 5, *Method 3* data splitting technique was used to construct the model.

VI. PERFORMANCE MEASURE

A. Classification

Classification is one of the supervised learning techniques and means to group the output (response variable) inside a class. If the algorithm tries to label two different classes of input, it is called binary classification, and selecting more than two classes is called multiclass classification. The performance of the classification algorithm is usually checked by classification accuracy. Determining the best of the best classification depends on the user interpreting the problem. The classification accuracy depends on the number of data correctly classified into the relevant classes. This ignores the fact that the wrong class can also have costs associated with the wrong assignment and which should need to be determined [21,22].

B. Confusion matrix

The confusion matrix represents the accuracy of the solution to a classification problem. The n-classes classification problem generates a n x n confusion matrix, the number of values in Cij class indicates that the number of observations of the dataset that were predicted to class Cj but where the correct class is Ci. Obviously, the best solution is only when there are zero values outside the diagonal [13].

The confusion matrix contains information about actual and predicted classifications performed by a classification system. The following figure 4 shows the confusion matrix for a binary classification problem and the equation for calculating classification accuracy is displayed in Equation 1. This procedure is also applicable to the multi-class classification problem.

Figure 4 below shows the structure of the classification table and the equation for calculating classification accuracy.

	True Class	
	P	N
Hypothesized class Y	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Fig. 4. Structure of classification table / confusion matrix

Equation 1:

Classification accuracy of the table = (TP +TN) / (P+N)

VIII. RESULTS & DISCUSSION

The following tables illustrate the classification accuracy of accident severity from all 5 models discussed above.

TABLE I. SUMMARY OF CLASSIFICATION ACCURACY FOR MODEL FITTING FOR MODEL 1-MODEL 5

Model	Overall Accuracy	Accuracy of the class property damages	Accuracy of the class slight injuries	Accuracy of the class serious & fatal injuries
1	83.13%	93.76%	21.31%	19.05%
2	83.84%	94.39%	22.95%	19.05%
3	90.24%	98.76%	40.98%	38.10%
4	92.90%	98.82%	40.90%	38.46%
5	90.78%	98.35%	46.43%	39.13%

TABLE II. SUMMARY OF CLASSIFICATION ACCURACY FOR MODEL TESTING FOR MODEL 1-MODEL 5

Model	Overall Accuracy	Accuracy of the class property damages	Accuracy of the class slight injuries	Accuracy of the class serious & fatal injuries
1	80.85%	91.87%	11.11%	0.00%
2	82.26%	93.50%	11.11%	0.00%
3	70.0%	77.60%	16.67%	0.00%
4	60.0%	79.00%	17.39%	5.89%
5	65.71%	73.77%	18.18%	14.29%

By evaluating the results from the above tables, the discussion point was raised about the accuracy from slight injuries and serious and fatal injuries are been much lower than the accuracy from property damages. Hence, it can be identified that there is a problem with the dataset which has been used for this research. That problem is defined as the class imbalance problem.

According to the above dataset, the imbalanced ratio of classes of property damages: slight injuries: serious and fatal injuries is 86:10:4. The random under-sampling technique considered under the data processing approach was used to overcome the imbalance problem in this study. In Model 6, the Naïve Base classification algorithm was used to construct the model using the *Method 4* data splitting technique with significant independent variables.

In Model 7, the Probabilistic Neural Network (PNN) classification technique was used to construct the model using the *Method 4* data splitting technique with significant variables that were identified from the statistical tests. Following figure 5 show the optimum network architecture of the PNN network.

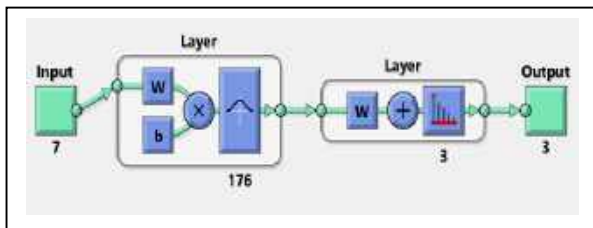


Fig 5: Network architecture of the Model 7 PNN

Seven input variables were used to build the PNN model with 0.1 spread. There were two hidden layers with 176 and 3 neurons respectively to classify the output category (Accident severity). The following tables are illustrating the classification accuracy of accident severity from Model 6 and Model 7.

TABLE III. SUMMARY OF CLASSIFICATION ACCURACY FOR MODEL FITTING FOR MODEL 6 AND MODEL 7

Model	Overall Accuracy	Accuracy of the class property damages	Accuracy of the class slight injuries	Accuracy of the class serious & fatal injuries
Naïve Bayes (Model 6)	60.23%	87.76%	27.78%	20.83%
PNN (Model 7)	99.43%	100%	100%	95.83%

TABLE IV. CONFUSION MATRIX FOR A TESTING DATASET FOR THE MODEL 6 (NAÏVE BAYES UNDER-SAMPLING)

Observed	Predicted			
	Property Damages Only	Slight Injury	Serious & Fatal Injury	Correct Percentage
Property Damages Only	99	18	4	81.82%
Slight Injury	11	2	0	15.38%
Serious & Fatal Injury	3	3	0	0.00%
Overall Percentage	80.71%	16.43%	2.86%	72.14%

TABLE V. CONFUSION MATRIX FOR TESTING DATASET FOR THE MODEL 7 (PNN UNDER-SAMPLING)

Observed	Predicted			
	Property Damages Only	Slight Injury	Serious & Fatal Injury	Percent Correct
Property Damages Only	99	18	4	81.82%
Slight Injury	9	4	0	30.77%
Serious & Fatal Injury	5	0	1	16.67%
Overall Percentage	80.71%	15.71%	3.57%	74.29%

In Naïve Base classification algorithm, 99 of the 121 (82%) 'Property Damages Only' are classified correctly. 2 of the 13 (15%) 'Slight Injury' are classified correctly and none of the 6 (0%) 'Serious & Fatal Injury' are classified correctly.

In Probabilistic Neural Network (PNN) 99 of the 121 (82%) 'Property Damages Only' are classified correctly. 4 of the 13 (31%) 'Slight Injury' are classified correctly and 1 of the 6 (17%) 'Serious & Fatal Injury' are classified correctly.

The Naïve Base classification algorithm predicts the severity of accidents with an accuracy of 72.14%, while the

Probabilistic Neural Network (PNN) classifies the severity of accidents with an accuracy of 74.29%.

Table 6 compares the performance of two techniques considered thought the study.

TABLE VI. CLASSIFICATION ACCURACY RELATED TO MODEL 6 & MODEL 7

Model	Technique used	Overall classification accuracy
6	Naïve Bayes	72.14%
7	PNN	74.29%

The classification accuracy of the final models (Naïve Bayes and PNN) which were created using the under-sampling data technique is greater than the accuracy derived from the original dataset.

By considering both models in Table 6, it can be said that the accuracy of both models is higher than 70%, which could be considered as approximately good models for classifying the accident severity of Colombo-Katunayake Expressway, Sri Lanka.

From the above two techniques, the overall accuracy of the Naïve Bayes classification model is similar to that of the probabilistic neural network classification model. By considering the accuracy of minority classes (Slight injuries, Serious and fatal injuries), the predictive accuracy of the Probabilistic Neural Network (PNN) classification model is better than the Naïve Bayes classification model.

Both models perform in good accuracy with the presence of significant variables. Hence, we can come up with a conclusion that the significant variables could be an effecting cause for being an accident of Colombo-Katunayake expressway. Further, it can be said that both the Naïve Bayes classification model and the PNN models are good for classifying accident severity of Colombo-Katunayake expressway while PNN exhibit slightly higher accuracy in classification when compared to Naïve Bayes classification model.

Some limitations of the study are as follows: Data collecting technique is more or less poor in Sri Lankan expressways, the study could have gone to a more realistic and accurate model if there are more variables to consider the accident severity; data collecting method of all the expressways of Sri Lanka are done in written format by human hand rather than computer systems. So it can take a lot of time converting data from written format to any computer format (E.g. Excel) for a person who is willing to do research on expressways of Sri Lanka and by using a multinomial regression model or any other statistical method, one can identify most influential factors for accident severity.

VII. CONCLUSION

This study aimed to identify the major causes of accidents on the Colombo-Katunayake Expressway in Sri Lanka and to classify the severity of accidents. After building the appropriate models to achieve that objective, the following conclusions were drawn.

- From the models considered in this study, both the Naïve Bayes classification algorithm and Probabilistic Neural Network (PNN) are good models for classifying the severity of accidents in Colombo-Katunayake Expressway with higher accuracy. Among them, PNN performs slightly better than the Naïve Bayes classifier.
- Considering the classification accuracies of the rarely occurred and more important 'serious & fatal injury' and 'slight injury' categories, the PNN model performed higher classification accuracy than the Naïve Base classification algorithm.
- The significant variables identified in this study: Time category, Driver's Age category, Vehicle type, Reason for the accident, Number of vehicles involved, Cause for accident and Rainfall can be considered as the most influential factors for met with an accident on the Colombo-Katunayake Expressway.
- Final Results obtained from this study can be used to implement safety improvements against the traffic accidents in expressways of Sri Lanka, Specifically in the Colombo-Katunayake Expressway.

REFERENCES

- [1] S. D. Dharmaratne, A. U. Jayatilleke, and A. C. Jayatilleke, "Road traffic crashes, injury and fatality trends in Sri Lanka: 1938-2013," *Bulletin of the World Health Organization*, 2015, 93, 640-647.
- [2] S. Renuraj, N. Varathan, and N. Satkunanathan, "Factors influencing traffic accidents in Jaffna," *Sri Lankan Journal of Applied Statistics*, 2015, 16(2).
- [3] Annual reports of expressways accidents from 2011-2018 [According to Police Report of Highway Main Traffic Police Branch in Kaduwela, Sri Lanka]
- [4] T. B. Tesema, A. Abraham, and C. Grosan, "Rule mining and classification of road traffic accidents using adaptive regression trees," *International Journal of Simulation*, 2005, 80-94.
- [5] S. D. Dharmaratne, and S. N. Ameratunga, "Road traffic injuries in Sri Lanka: a call to action," *Journal of the College of Physicians and Surgeons--Pakistan: JCPSP*, 2004, 729-730.
- [6] J. de Oña, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Analysis & Prevention*, 2011, 402-411.
- [7] J. Maiti and A. Bhattacharjee, "Predicting accident susceptibility: a logistic regression analysis of underground coal mine workers," *Journal of the Southern African Institute of Mining and Metallurgy*, 2001, 203-208.
- [8] M. M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using decision trees and neural networks", 2004, arXiv preprint cs/0405050.
- [9] Chong, M., A. Abraham, and M. Paprzycki, "Traffic accident data mining using machine learning paradigms," *Fourth International Conference on Intelligent Systems Design and Applications*, 2004, pp. 415-420.
- [10] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using machine learning paradigms," *Informatica*, 2005.
- [11] L. Y. Chang, "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network," *Safety science*, 2005, 541-557.
- [12] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident Analysis & Prevention*, 2006, 434-444.
- [13] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International*

- Journal of Intelligent Systems and Applications in Engineering*, 2019, 88-91.
- [14] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4.5 and a Naïve Bayes classifier for the prediction of lung cancer survivability," 2012, arXiv preprint arXiv:1206.1121.
 - [15] G.A. Ali, and C.S. Bakhiet, "Comparative analysis and Prediction of Traffic accidents in sudan using artificial Neural Networks and Statistical methods," Sudan University of Science and technology.
 - [16] H.S. Stem, "Neural networks, in applied statistic (with discussions)," *Technometric*, 1995, 205-220
 - [17] S. Peltzman, "The effects of automobile safety regulation," *Journal of political Economy*, 1975, 83(4), 677-725.
 - [18] R. D. Romero, D. S. Touretzky, and R. H. Thibadeau, "Optical Chinese character recognition using probabilistic neural networks," *Pattern recognition*, 1997, 30(8), 1279-1292.
 - [19] R. Longadge and S. Dongre, "Class imbalance problem in data mining review", 2013, arXiv preprint arXiv:1305.1707.
 - [20] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "Class imbalance problem," *Fourth international conference on natural computation*, 2008, pp. 192-201.
 - [21] M. H. Dunham, "Data mining: Introductory and advanced topics" Pearson Education India, 2006.
 - [22] J. COE, "Performance comparison of Naïve Bayes and J48 classification algorithms," *International Journal of Applied Engineering Research*, 2012.

Industry 4.0 readiness assessment for apparel industry: A study in the Sri Lankan context

Erandika Lakmali*

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
eralakmali95@gmail.com

Kasuni Vidanagamachchi

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
kasuniv@kln.ac.lk

Julian Nanayakkara

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
julian@kln.ac.lk

Abstract: Sri Lankan apparel industry is the most significant and driving contributor to the country's economy by constituting a large portion of GDP. In the highly competitive apparel world, manufacturers search solutions for problems such as worker inadequacy while minimizing the human impact. Therefore, there is a need for apparel manufacturers to enhance value chain processes with the latest technologies. Industry 4.0 is the fourth industrial revolution that transforms the physical production into a combined cyber-physical production environment with IoT and decentralized intelligence. It enhances the process functions from new product development to logistics by providing real-time visibility of the production flow. Existing literature mentions the applications of Industry 4.0 in the apparel industry, but these have not addressed the issue of assessing the readiness for its adaptation in the apparel value chain process. Hence this scrutiny proposes a model to assess the current level of readiness of the Sri Lankan apparel industry to adapt Industry 4.0 technologies and practices. The model was developed based on a systematic review of literature with the industry experts' guidance. The factors that determine the readiness for Industry 4.0 within an organizational context were classified under four categories; People, Process, Technology and Data which were defined as readiness dimensions. The proposed model consists of five readiness levels from 0 to 4 namely: Stranger, Beginner, Intermediate, Advanced and Elite. This model enables managers to measure the readiness for adapting of Industry 4.0 in selected apparel value chain processes by using the specified minimum requirements under each dimension and level. The outcome of this study indicates that Sri Lankan apparel industry is in "Intermediate" level in terms of overall readiness with a value of 1.91 in the predefined readiness scale from 0 to 4.

Keywords: Apparel 4.0, Apparel value chain Industry 4.0, Readiness assessment, Sri Lankan apparel industry

I. INTRODUCTION

Sri Lanka's apparel industry is one of the significant and driving contributors to the country's economy by contributing over \$5 billion to GDP [1]. This industry has achieved rapid growth rates over the past few decades despite the increasing competition. Today, the apparel industry as Sri Lanka's primary foreign exchange earner accounts for 40% of the total exports and 52% of industrial product exports [2]. Since apparel is a human-centric industry, it has become a challenge for a small country such as Sri Lanka when compared to other regional players. At present, the major challenges faced by the industry are labour shortage, a record number of labour turnover, increases in wages and also loss of preferential treatments such as GSP+. As a result of that, apparel manufacturing organizations have to implement

innovative solutions to overcome the challenges by enabling new technological developments. Therefore, these organizations have been left with the option of automating most of their processes [3]. Thus, the apparel manufacturers are in the initial step of transforming their factories into smart factories to find solutions to the aforesaid challenges by improving the productivity in both employees and operations.

The world has experienced three distinct industrial revolutions since 1800s. The first industrial revolution (Industry 1.0) was the usage of water and steam power for the invention of the steam machine. The second industrial revolution (Industry 2.0) was the period where the assembly line led to mass production and automation up to a certain extent. The third industrial revolution (Industry 3.0) was the rise of computer networks; robotics in manufacturing and the birth of the Internet, which is the game-changer in the ways that information is handled and shared [4]. The fourth industrial revolution (Industry 4.0) can be defined as an umbrella term for a new industrial paradigm that embraces emerging digital technologies: additive manufacturing, artificial intelligence (AI), autonomous robots, cyber-physical systems (CPS), Internet of Things (IoT), big data, augmented reality (AR) and creating the smart factory or in simple form the evolution of technology from embedded systems to cyber-physical systems [5]. Industry 4.0 empowers the businesses with improved control of the operation and allows leveraging real-time data to enhance productivity and processes improvements, which finally supports as driving forces for continuous growth within those industries.

According to industry experts' analysis, it is seen that when implementing Industry 4.0 in real-world enterprise environments, the problems faced are the lack of strategic guidance, poor acknowledgement on exceptionally complex Industry 4.0 ideas, the vulnerability of the results of Industry 4.0 applications with the matter of benefits and costs, failure to assess Industry 4.0 capability and lack of readiness of the company [6]. Concerning these issues, readiness assessment for Industry 4.0 becomes highly important, since a large number of organizations still struggle to initiate the transformation for Industry 4.0. An organizational readiness assessment uses a checklist, that is usually custom made based on the current situation at the organization and the parameters and requirements of the change or project that organization which desires to pursue [7]. Thus, an Industry 4.0 readiness assessment model helps organizations to decide their state of readiness in the adaption of Industry 4.0 technologies, distinguish the gaps and factors of

advancement for Industry 4.0 adaption, as well as paths for profitability improvement and advancement of achievable procedures and plans to perform outcome-based intervention projects.

The properties of existing Industry 4.0 readiness assessment models are different and there is no standard and well-accepted Industry 4.0 readiness assessment model [8], [6], [9], [45]. The existing models provide an analytical tool for evaluating an enterprise's current state of Industry 4.0 readiness, but most of the models do not contain a guide to upcoming steps within a certain roadmap to move up to higher readiness levels. None of the researches has fully elaborated sector-wide solutions for the apparel industry to assess the readiness for adaptation of Industry 4.0 in the apparel value chain process. Thus, the existing Industry 4.0 readiness assessment models demonstrate many weaknesses and drawbacks which trigger the need for the development of a new Industry 4.0 readiness assessment model. Thus, there is an urge for an Industry 4.0 readiness assessment model for the apparel industry, particularly in the Sri Lankan context.

The objective of this paper is to propose a readiness assessment model customized for the apparel industry with an assessment survey, in order to provide apparel manufacturing companies with a tool to measure the current state of readiness regarding Industry 4.0. Subsequently, this scrutiny measures the readiness for adapting Industry 4.0 within the Sri Lankan apparel industry using that proposed model and help to guide apparel manufacturers to move up the readiness ladder while maximizing operational performance. The remainder of the paper is structured as follows: the methodology applied for this study is described in the next section, section 3 elaborates the implementation of the readiness assessment model in detail, the analysis and the results are presented in section 4 while a comprehensive discussion of the findings is given in section 5. Finally, the closure of the paper is by presenting conclusions and an attempt to provide some perspectives on future research.

II. METHODOLOGY

The systematic review of the literature was based on the content analysis to gather the state of the knowledge on Sri Lankan apparel industry, Industry 4.0 and existing readiness assessment models in the context of Industry 4.0. The literature review was conducted according to the procedure proposed by Kitchen ham, [10] and the literature review protocol based on Popay et al., [11] to minimize the systematic error and bias in the screening of papers. Fifty-eight articles remained for qualitative synthesis and they were analysed. Through the systematic review of literature, the major challenges that the Sri Lankan apparel industry is currently facing when they compete with the global competitors were identified. Then a comprehensive review of innovative applications of Industry 4.0 in the apparel industry was presented. The existing Industry 4.0 readiness assessment models were evaluated in order to find the gaps that exist and to explore the strengths and weaknesses of each model using a set of evaluation criteria that were identified based on the literature and key points that need to be included in a standardized assessment model. Finally, the results of the systematic review of literature were used in the development of Industry 4.0 readiness assessment model for the Sri Lankan apparel industry. The readiness assessment model has been

validated based on the suggestions and criticism from the industry experts who are representing different functional areas within apparel manufacturing companies in Sri Lanka. Finally, the refined model was used to implement questionnaires and to continue the research process. Three leading large-scale companies and one medium scale company were selected. These large scale and medium scale companies were selected based on their capability to concentrate on Industry 4.0 implementations with enough financial strategies, and also a limited number of companies were selected since the assessment process included several value chains processes, dimensions and sub-dimensions which had a huge expansion. Since different plants within the same company are having different levels, several plants from each company were selected for the sample of the study. Data was collected from the management of each functional area of these apparel manufacturing plants. In quantitative data analysis of the answers collected from structured questionnaires for closed-ended questions, the relevant readiness level represented in each answer was identified based on a score value assigned for each readiness level.

III. INDUSTRY 4.0 READINESS ASSESSMENT MODEL

A. Industry 4.0 readiness assessment dimensions

A systematic review of literature was conducted in order to identify the readiness factors highlighted by different papers and readiness assessment models and to provide a comprehensive, overall view of readiness factors for Industry 4.0. Thereby the intention was to develop a standardized readiness assessment model for Industry 4.0 which can be used to determine the overall readiness of an organization /industry. The identified factors were further reviewed and classified into four major categories: People, Process, Technology and Data. The classified readiness factors with their references and categories are shown in Table I.

TABLE I. FACTORS DETERMINING THE READINESS FOR INDUSTRY 4.0

Readiness Categories	Readiness Factors	References
People	Organizational strategy on the internal and external environment	[5], [6], [8], [9], [12], [13], [14], [15]
	Organizational strategy on technology implementation	[6], [8], [9], [13], [14], [16]
	Customer integration	[13], [17]
	Supplier integration	[12], [17], [18], [19], [20]
	Employee adaptability to skills and technology	[5], [18], [21], [22], [23], [24]
	Top management commitment	[12], [18], [19]
Process	Elements of organizational strategy (Vision, Mission, Goals, Action Plans, KPIs)	[6], [25], [26]
	Self-optimizing processes	[5], [16], [21], [25]
	Product integration with CPS	[8], [13], [23], [27], [28]
	Product innovation	[3], [5], [8], [13], [17], [27], [28], [35]
Technology	Integration of CPS	[17], [23], [25], [29], [30], [31], [32], [33], [34]
	Machine-system integration	[12], [23], [31], [40]
	Compatibility of IT systems with processes	[16], [23], [31], [35], [36]

	Autonomously guided work pieces	[31], [35], [37]
	IT system security	[9], [16], [21], [38]
Data	Real-time data analytics and decision making	[9], [23], [32], [33], [36], [39], [40], [43]
	Advanced algorithms	[16], [21], [34], [38], [39], [41], [42], [44], [54]
	Cloud-based data management	[5], [9], [17], [23], [31], [33], [39], [43]

These factors have been used as readiness dimensions for Industry 4.0 readiness assessment model for the Sri Lankan apparel industry. The identified readiness dimensions are further described in Table II.

TABLE II. READINESS DIMENSIONS

Readiness Dimensions	Readiness Sub-dimensions	General Description
People	- Organizational attitude on the internal and external environment - Organizational strategy on technology implementation - Customer integration - Supplier integration - Employee adaptability to skills and technology - Top management commitment	A culture of continuous improvement in which smart technology is viewed as the solution. Specialized roles and responsibilities drive toward predictable outcomes. Strategies towards having a set of skilful workforces who support and use new technologies. Change champions who understand the benefits of technology and influence the others around them. Smart and strategic partnerships. Real-time customer and supplier engagement throughout the product life cycle.
Process	- Elements of organizational strategy - Self-optimizing processes - Product integration with CPS - Product innovation	Processes that integrated data visualization into decision making for continuous improvement. Self-optimizing processes. KPIs that are required to run processes are being tracked. Planned processes in a proactive setting for predicting, forecasting, and planning future requirements. Innovative product development. Smart management system to run business operations throughout product development and innovation.

Technology	- Integration of CPS - Machine-system integration - Compatibility of IT systems with processes - Autonomously guided work pieces - IT system security	Intelligent applications to monitor and visualize critical operational analytics. Digitalized system integration with external entities to enable operational predictability and enhance customer relationship. Cyber-Physical Systems to improve agility responsiveness and accelerate the product life cycle. Mobility to schedule tasks and display KPIs.
		Adequate protection for all information available as electronic data.
Data	- Real-time data analytics & decision making - ML algorithms - Cloud-based data management	Utilized data analytics and visualization for real-time decision making. Automated KPI reports. Data gathering to predict future demand, performances and requirements for standardized decision-making process. Centralized data management with an open source collaboration.

B. Industry 4.0 readiness assessment levels

Five readiness levels were identified based on the literature and guidance of industry experts. According to the accomplishment of readiness levels, a scoring method has been provided to measure the readiness level of each apparel value chain process and to analyse the overall readiness level within an individual organization among processes, compare among competitive organizations and assess the readiness level of the entire Sri Lankan apparel industry. A particular readiness level is determined based on the score range given to each level. Score value gained by a dimension is compared with the score range, and the readiness level on that dimension is assessed. The identified readiness levels are described in Table III.

TABLE III. READINESS LEVELS

Readiness Level	Score (x)	Definition
Level 0: Stranger	x=0	The level that Industry 4.0 is unknown/irrelevant to the functional area/factory and does not achieve any criteria for the development of Industry 4.0.
Level 1: Beginner	0<x<=1	The level that processes are primarily manual with semi-automated technologies and involved in Industry 4.0 by pilot projects. Investments are being allocated to implement Industry 4.0 strategies within a single part of the functional area/company. In-company information sharing is limited to a few parts of the functional area/company. Necessary skills required to expand Industry 4.0 are only found in a few groups within the functional area/company.

Level 2: Intermediate	$1 < x \leq 2$	The level that computer-based technologies are available, but human-machine interaction is also still available. A strategy with a scale to measure the progress of implementation of Industry 4.0 principles is being developed. Investments are being allocated to implement Industry 4.0 strategies within a few parts of the functional area/company. The first step is being initiated to enable the information exchange process with business partners. Necessary skills needed to expand Industry 4.0 are possessed in many groups in the functional area/company.
Level 3: Advanced	$2 < x \leq 3$	The level that automated processes without human intervention are available, but employees also interact as manual/semi-automated elements of the work with automated elements. A strategy with a scale to measure the progress of implementation of Industry 4.0 principles has been developed.
		Investments are being allocated to implement Industry 4.0 strategies within many parts of the functional area/company. The functional area/company is not yet integrated with its customers. Extensive efforts have already been made to expand the necessary skills needed for Industry 4.0 throughout the functional area/company.
Level 4: Elite	$3 < x \leq 4$	The level that functional area/company is powered by IoT, computer technologies, complex data analytics and virtual manufacturing with little or no human intervention through autonomous decision making and self-optimization. A strategy with a scale to measure the progress of implementation of Industry 4.0 principles has been developed and is already monitoring the status. Investments are being allocated to implement Industry 4.0 strategies throughout the company and support the above activities. A System-integrated data exchanging process is conducted both internally and externally with partners of business throughout the functional area/company. In-house expertise needs to move forward with Industry 4.0 are available in all critical parts of the functional area/company.

IV. RESULTS AND FINDINGS

First of all, it is important to highlight that not all results from the readiness assessment have been provided in this paper. The focus is to report the overall readiness of each manufacturing plant along with value chain processes. Finally, the overall readiness of the Sri Lankan apparel industry in adapting Industry 4.0 for each apparel value chain process was calculated and the overall readiness was assessed for Industry 4.0 in the Sri Lankan apparel industry. The mathematical equation used to calculate readiness values is mentioned below.

$$R_P = \frac{\sum_{i=1}^n R_{Di}}{n} \quad (1)$$

R – Readiness Value

D – Readiness Dimension

P – Value Chain Process

n - # of Readiness Dimensions

The readiness values calculated for apparel value chain processes of each manufacturing plant have been depicted in Table IV.

TABLE IV. READINESS ASSESSMENT IN APPAREL MANUFACTURING PLANTS

Plant #	Product Design	Supply chain	Planning	Manufacturing	Sales & Marketing	Readiness level of the Plant
Plant A1	2.07	1.38	1.92	1.78	1.31	1.69
Plant A2	2.13	2.08	2.21	2.19	2.24	2.17
Plant A3	2.54	1.47	1.77	1.84	1.59	1.84
Plant A4	2.56	1.75	2.69	2.11	1.67	2.16
Plant A5	2.36	2.21	2.46	2.53	1.99	2.31
Plant B1	1.83	1.96	2.29	2.11	1.64	1.97
Plant B2	1.75	1.96	1.58	1.62	1.97	1.78
Plant C1	1.66	1.65	2.00	1.84	1.73	1.78
Plant D1	1.59	1.40	1.78	1.20	1.55	1.50

The readiness assessment of the planning processes with readiness dimensions is presented using a radar chart in Fig. 1. The axis gridlines displayed in the radar chart present the readiness levels of the model as described in Table 2. People and Process dimensions have shown a huge variation in terms of readiness among manufacturing plants while Technology and Data dimensions have low variation compared to those two dimensions.

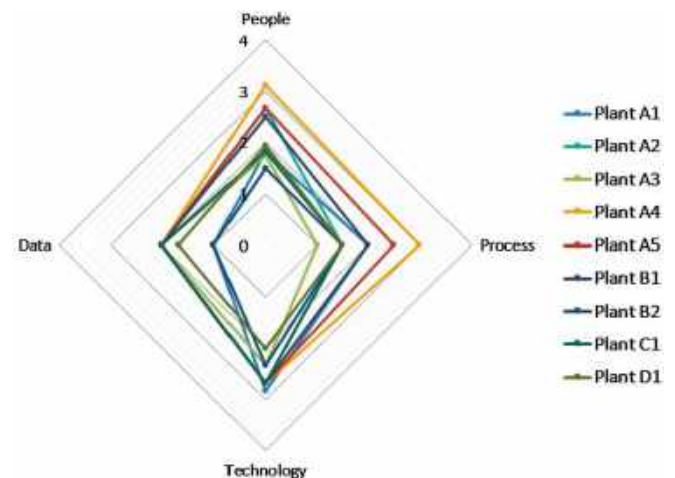


Fig.1. Readiness assessment of planning process

Hence, the variation of readiness among plants of the sample is displayed in Fig. 2. The figure describes the number of plants in the sample as a percentage against the variation of their readiness level in each dimension for the planning process.

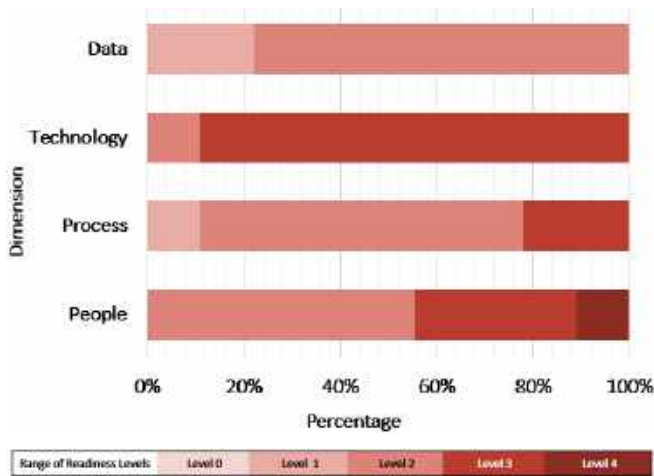


Fig. 2. Distribution of readiness for dimensions in planning process

V. DISCUSSION

A. Readiness assessment of value chain processes

The reason for selecting the planning process for discussion that it is the process which has achieved the highest readiness with a variation between four levels. According to the Fig. 2, all the plants have improved their readiness up to "Intermediate" level or above it from People readiness. 56% of the plants were in the "Intermediate" level, where employees have improved skills for making judgements and decision making, and also where employees communicate with stakeholders once in a while to follow through on requests. Then 33% of the plants were in the "Advanced" level by enabling communication with stakeholders once in a while to follow through on requests and update the plan in a critical situation with the skills of employees to solve complex problems. Consumer interaction is up to a medium level at the product design stage. Those plants have fully skilled employees, but research skills are at a low level. Out of the four dimensions, only People dimension had 11% portion of the sample who got the readiness level up to "Elite" where they have acquired the maximum readiness level with employees' skills for mathematical reasoning, analytical thinking and problem-solving. There is also communication with stakeholders early, often to follow through on requests and update the plan frequently.

Apparel manufacturing plants, in terms of Process readiness in the planning process, have shown up a variation between three levels. All the plants have improved their readiness up to "Beginner" level or above it. Most of the plants amounting to around 67% were in "Intermediate" level, where the production plan is optimized up to some extent with real-time information updates. Then 22% of the plants were in the "Advanced" level by enabling their planning departments' processes to optimize the production plan with real-time information updates. The least number of plants (11%) were at "Beginner" level by facilitating their plants to update the production plan only in critical situations with manual intervention.

The results of the survey indicated that Technology readiness within the planning process was up to "Intermediate" level or above it. 88% of the plants were in "Advanced" level, where integrative, automate and transparent planning and scheduling systems are used. Simulation is used to find the optimal arrangement of critical resources for selected product categories for profitable products. These plants practise several cyber security solutions with strategies to remove security gaps in technology. Only 11% of the plants were in the "Intermediate" level by preparing for Industry 4.0 with the use of systems which require little manual adjustment to generate plans. Simulation is used as a trial case to find the optimal arrangement of critical resources for the planning process for profitable products within these plants.

The lowest readiness among four levels has been shown by the Data dimension where it has variation in its readiness only between two levels: "Beginner" and "Intermediate". Most of the plants around 78% were at the "Intermediate" level, where demand forecasting is conducted using past data with simple applications, and then the past data on planning is used frequently for future planning and scheduling. Initial solutions already commenced in few areas for cloud-based computing, data warehousing and analytical purposes. Then 22% of the plants were at the "Beginner" level by enabling to forecast demand roughly using past data. Past data on planning are used by these plants for future planning and scheduling by computing manually. Initial solutions are being planned for cloud-based computing, data warehousing and analytical purposes.

The variation of readiness flows between four levels: "Beginner", "Intermediate", "Advanced" and "Elite" when considering the four dimensions in the planning process within the Sri Lankan apparel industry. Only the People dimension has improved their readiness up to "Elite" level and no other dimension has been able to move up to that level. The plants prepared up to the "Advanced" level were found only in the Technology dimension, followed by People dimension and finally in Process dimension. The highest number of plants were prepared up to "Intermediate" level and they were in Data dimension, following Process dimension, People dimension and finally in Technology dimension. Plants that were prepared up to "Beginner" were only found for Process and Data dimensions with the highest number of plants for Data dimension and then Process dimension.

B. Readiness assessment of Sri Lankan apparel industry

The overall readiness values of the value chain processes in the Sri Lankan apparel industry are presented using a radar chart in Fig. 3. The readiness of each value chain process is discussed below.

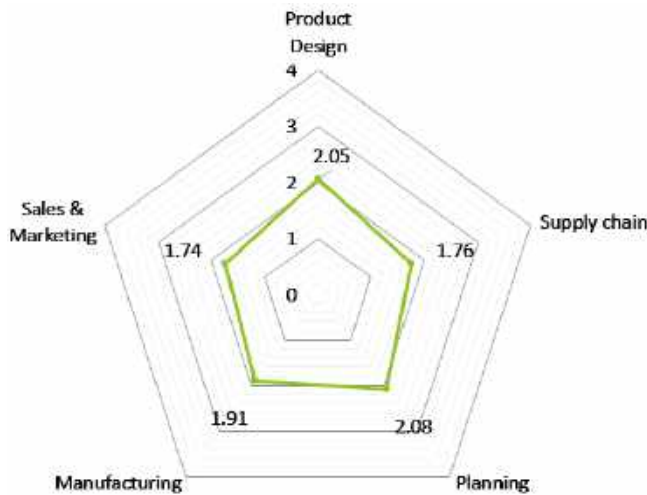


Fig. 3. Overall readiness in value chain processes

1) *Product Design*: The readiness level of the Product Design process is "Advanced" level (Level 3) with a readiness value of 2.05 which is the second-highest readiness compared with the other processes. It indicates that from People readiness, consumer interaction, product visualization are enabled at the product design stage. Employees have been fully skilled with creativity but research skills are at a moderate level. When considering Process readiness, most of the information, documents on new products are sent using cloud technology in a digital environment to other departments. Products show value from intellectual property (IP) licensing with smart clothing or wearable technologies in a considerable portion of the production. Sri Lankan apparel manufacturers develop prototypes using 3D printing up to a considerable level. Technology readiness is a critical point for each process and according to the current readiness, manufacturers use digital modelling for most of the garments. In terms of Data readiness, a high level of data usage can be seen for new product development processes. At the same time, pilot solutions have been implemented in a considerable proportion of the industry.

2) *Supply Chain*: The readiness level of the Supply Chain process is "Intermediate" level (Level 2) with a readiness value of 1.76. It indicates that from People readiness point of view, data transfer is among vital suppliers and customers. Employees are skilled in transactional tasks by the use of both manual and digital tools while top management has identified the possibilities for Industry 4.0 within supply chain functional area. From Process preparedness, these plants are maintaining supply chain processes that are partially incorporated among the organization, key vital suppliers and customers as far as data transfer. Improvements have been planned and commenced to reduce lead times for some materials. Sri Lankan apparel industry moderately responds to the changes in the market and shifts of customer needs. Technology readiness is a critical point for each process, and according to the current readiness enabling the visibility of site location, capacity, operations and inventory up to first level suppliers and customers is also critical. Supply chain applications are used by key suppliers and service providers. Data readiness in terms of Supply chain process indicates that initial solutions have been already commenced for cloud-

based software, data warehousing and analytical purposes in few areas.

3) *Planning*: The readiness level of the planning process is at the "Advanced" level (Level 3) with a readiness value of 2.08 which is the highest readiness value compared to the other value chain processes. Regarding the People readiness, it is highlighted on enabling communication with stakeholders once in a while to follow through on requests and update the plan in a critical situation with the skills of employees to solve complex problems. Consumer interaction is up to a medium level at the product design stage. Those plants have fully skilled employees, but their research skills are at a low level. From Process preparedness point of view manufacturing companies have enabled their planning departments' processes to optimize the production plan with real-time information updates. According to the current Technological readiness integrative, automated and transparent planning and scheduling systems are used by these plants. Simulation is used to find the optimal arrangement of critical resources for selected product categories for profitable products. In terms of Data readiness, demand forecast uses past data with advanced applications. They maintain a large volume of data that can be used as input for planning. Pilot projects are ongoing in a considerable proportion within this functional context.

4) *Manufacturing*: The readiness level of the Manufacturing process is "Intermediate" level (Level 2) with a readiness value of 1.91. It indicates that from People readiness point of view, employee training is conducted using many digital tools. Employees are skilled up with easy tasks, repetitive tasks, watch and call and contribute to standardization in these plants. The production line is balanced manually and most of the data are collected on employee performance at this level. Customers have the ability to track the status of the product only at limited stages of the manufacturing process since these plants are still at "Intermediate" level. The top management has identified the possibilities for Industry 4.0 within the manufacturing process where they are responsible for. From Process preparedness, these plants are not using self-optimising processes but pilots are available in some advanced areas of the manufacturing process. Large scale production strategy is still in action with limited differentiation and medium level customization of production. Products can be tracked as it moves between a few processes in manufacturing within these plants. These plants have prepared for Industry 4.0 in the quality assurance part with the use of real-time applications for in-line quality testing. Since the Sri Lankan apparel industry still represents "Intermediate" level, 3D printing technology is at the conceptual level in the manufacturing process. Systems and machines are interoperable to some extent and few systems and machines are being able to upgrade. Few autonomously guided machines are available but not in use. In terms of Data readiness, the required data is collected real-time in a few areas. Varieties of data are used to control processes. Initial solutions already have been commenced in few areas for cloud-based software, data warehousing and analytical purposes. At the same time, employee performance reporting is partially covered by real-time data collection on few parts within the manufacturing process in the Sri Lankan context.

5) *Sales & Marketing*: The readiness level of the Sales & Marketing process is “Intermediate” level (Level 2) with a readiness value of 1.74 which is the lowest readiness value compared to the other value chain processes. Regarding People readiness, the reengineering process of the organizational culture and structure has been initialized as suited for customer engagement. Employees are skilled up with few basic skills required for Industry 4.0 in sales and marketing, and knowledge sharing about Industry 4.0 is conducted within the sales & marketing functional area of these plants. From Process preparedness, these plants have enabled the communication channel integration and cooperation of them with accomplices up to a medium level. At the same time, these plants are maintaining a medium level of integration between campaign systems and sales channels. When considering the results of the study on Technology readiness few pilot projects have been initialized on triggering technologies for a few advanced operations in the sales and marketing process. IT security solutions are initialized for data exchange with business partners which is a crucial requirement that is needed to be fulfilled before adapting Industry 4.0. Data analytics tools are used at a medium level to measure sales performances and data-driven administrations are offered with least consumer integration. Initial solutions have been already commenced in a few areas for cloud-based software, data warehousing and analytical purposes within the sales and marketing process.

The overall readiness level of the Sri Lankan apparel industry is 1.91 calculated by taking the average value of the five apparel value chain processes which equally contributed to the readiness assessment. Thus, it revealed that the Sri Lankan apparel industry is in level 2: “Intermediate” level according to the developed readiness assessment model. Thus, the Sri Lankan apparel industry is in a situation where computer-based technologies are in use, but human-machine interaction is still available. A strategy with a scale to measure the progress of implementation of Industry 4.0 principles is being developed. Investments are being allocated to implement Industry 4.0 strategies within a few parts of the functional area/company. The first step is being initiated to enable the information exchange process with business partners. The necessary skills needed to expand Industry 4.0 are present in many groups within this apparel industry.

The Sri Lankan apparel industry further needs to move up in this readiness ladder in order to settle down with the highest readiness before adapting Industry 4.0 technologies and practices. Industry practitioners can improve this readiness up to the next level by establishing automated processes which are without human intervention and by training employees to develop manual/semi-automated elements of the work with automated elements. A strategy with a scale to measure the progress of implementation of Industry 4.0 principles has been developed. Investments are being allocated to implement Industry 4.0 strategies within many parts of the functional areas. Finally, they can expand the necessary skills needed for Industry 4.0 throughout the company with the extensive efforts of top management. It is essential to pay attention to all the readiness sub-dimensions, readiness dimensions and apparel value chain processes equally to comprehensively prepare for Industry 4.0.

VI. CONCLUSION AND FUTURE WORK

The study presented here aimed to develop a readiness assessment model customized to measure the readiness for adapting Industry 4.0 within Sri Lankan apparel industry with an assessment survey to provide apparel manufacturing companies with a tool to measure the current state of readiness regarding Industry 4.0. The model enables to assess the readiness up to different stages: the readiness of sub-dimensions, dimensions, apparel value chain processes, individual plant, a given apparel manufacturing company and finally the overall readiness of the Sri Lankan apparel industry as an average. The developed model is able to provide guidance to the apparel manufacturers, in order to identify their current state of preparedness and actions that should be taken to improve the readiness up to next level as a self-assessment tool, for five apparel value chain processes namely, Product Design, Supply Chain, Planning, Manufacturing and Sales & Marketing. This will be able to enhance the operational performance within this industry while solving the problems that the Sri Lankan apparel industry is currently facing when competing with the other regional players.

Since this study has considered a sample of nine apparel manufacturing plants which belong to four companies: including both large and medium scale companies, the sample can be further expanded to conduct more generalised readiness assessment. Future studies can aim at diversifying the proposed model to enhance the scope of readiness assessment with weighted readiness dimensions and to create comprehensive activity plans according to the companies’ current readiness level. Further, this model can be applied in other geographical contexts to assess the readiness for Industry 4.0 in the apparel industry.

REFERENCES

- [1] “Annual Report 2018: Central Bank of Sri Lanka,” Annual Report 2018 | Central Bank of Sri Lanka. [Online]. Available: <https://www.cbsl.gov.lk/en/publications/economic-and-financial-reports/annual-reports/annual-report-2018>. [Accessed: 29-Nov-2019].
- [2] “Performance Report of The Export Sector of Sri Lanka January - December 2013,” Sri Lanka Export Development Board. [Online]. Available: <http://www.srilankabusiness.com/exporters/export-performance-report.html>. [Accessed: 29-Nov-2019].
- [3] H. Jayatilake, and A. Withanaarachchi, “Industry 4.0 in the Apparel-Manufacturing Sector: Opportunities for Sri Lanka,” August 2016.
- [4] “(PDF) Industry 4.0 Revolution in Clothing and Apparel Factories ...”. [Online]. Available: https://www.researchgate.net/publication/332410328_Industry_40_R_evolution_in_Clothing_and_Apparel_Factories_Apparel_40. [Accessed: 29-Nov.-2019].
- [5] A. Rojko, “Industry 4.0 Concept: Background and Overview,” *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 11, no. 5, p. 77, 2017.
- [6] A. Schumacher, S. Erol, and W. Sihn, “A Maturity Model for Assessing Industry 4.0 Readiness and Maturity of Manufacturing Enterprises,” *Procedia CIRP*, vol. 52, pp. 161–166, 2016.
- [7] Z. Rajnai and I. Kocsis, “Assessing industry 4.0 readiness of enterprises,” *2018 IEEE 16th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2018.
- [8] K. Y. Akdil, A. Ustundag, and E. Cevikcan, “Maturity and Readiness Model for Industry 4.0 Strategy,” *Springer Series in Advanced Manufacturing Industry 4.0: Managing The Digital Transformation*, pp. 61–94, 2017.
- [9] E. Gökalp, U. Şener, and P. E. Eren, “Development of an Assessment Model for Industry 4.0: Industry 4.0-MM,” *Communications in*

- Computer and Information Science Software Process Improvement and Capability Determination, pp. 128–142, 2017.
- [10] B. Kitchenham, "Procedures for Performing Systematic Reviews," *Keele, UK, Keele Univ.* 33, 28, July 2004.
- [11] J. Popay, H. Roberts, A. Sowden, M. Petticrew, L. Arai, M. Rodgers, N. Britten, K. Roen, and S. Duffy, "Guidance on the Conduct of Narrative Synthesis in Systematic Reviews," *A Product from the ESRC Methods Programme*, April 2006.
- [12] "How virtualization, decentralization and network building ..." [Online]. Available: https://www.researchgate.net/publication/285495324_How_virtualization_decentralization_and_network_building_change_the_manufacturing_landscape_An_Industry_40_Perspective. [Accessed: 28-Nov-2019].
- [13] "Reinhard Geissbauer," *Industry 4.0 - Self Assessment*. [Online]. Available: <https://i40-self-assessment.pwc.de/i40/landing/>. [Accessed: 29-Nov-2019].
- [14] C. Santos, A. Mehraei, A. Barros, M. Araújo, and E. Ares, "Towards Industry 4.0: an overview of European strategic roadmaps," *Procedia Manufacturing*, vol. 13, pp. 972–979, 2017.
- [15] J. Basl and P. Doucek, "A Metamodel for Evaluating Enterprise Readiness in the Context of Industry 4.0," *Information*, vol.10, no. 3, p. 89, 2019.
- [16] C. Leyh, T. Schäffer, K. Bley, and S. Forstner, "SIMMI 4.0 – A Maturity Model for Classifying the Enterprise-wide IT and Software Landscape Focusing on Industry 4.0," *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, Feb. 2016.
- [17] "Paradigm Shift from Current Manufacturing to Social ..." [Online]. Available: https://www.researchgate.net/profile/Babak_Mohajeri/publication/298722427_Paradigm_Shift_from_Current_Manufacturing_to_Social_Manufacturing/links/56ea974608ae19f0283bdc76.pdf?origin=publication_detail. [Accessed: 29-Nov-2019].
- [18] L. Palazzeschi, O. Bucci, and A. D. Fabio, "Re-thinking Innovation in Organizations in the Industry 4.0 Scenario: New Challenges in a Primary Prevention Perspective," *Frontiers in Psychology*, vol.9, 2018.
- [19] R. Schmidt, M. Möhring, R.-C. Härting, C. Reichstein, P. Neumaier, and P. Jzinić, "Industry 4.0 - Potentials for Creating Smart Products: Empirical Research Results," *Business Information Systems Lecture Notes in Business Information Processing*, pp. 16–27, 2015.
- [20] W. Bauer, M. Hammerle, S. Schlund, and C. Vocke, "Transforming to a Hyper-connected Society and Economy – Towards an 'Industry 4.0,'" *Procedia Manufacturing*, 23-Oct-2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2351978915002012>. [Accessed: 29-Nov-2019].
- [21] "The smart factory," *Deloitte Insights*. [Online]. Available: <https://www2.deloitte.com/us/en/insights/focus/industry-4-0/smart-factory-connectedmanufacturing.html>. [Accessed: 29-Nov-2019].
- [22] A. Benešová and J. Tupa, "Requirements for Education and Qualification of People in Industry 4.0," *Procedia Manufacturing*, vol. 11, pp. 2195–2202, 2017.
- [23] K. Lichtblau, V. Stich, R. Bertenrath, M. Blum, M. Bleider, A. K. Millack, K. Schmitt, and E. Schmitz, "Industrie 4.0- Readiness," 2015.
- [24] M. Wolf, M. Kleindienst, C. Ramsauer, C. Zierler, and E. Winter, "CURRENT AND FUTURE INDUSTRIAL CHALLENGES: DEMOGRAPHIC CHANGE AND MEASURES FOR ELDERLY WORKERS IN INDUSTRY 4.0," *Graz University of Technology*. [Online]. Available: <https://graz.pure.elsevier.com/en/publications/current-and-future>. [Accessed: 29-Nov-2019].
- [25] K. Wang, Y. Wang, J. O. Strandhagen, and T. Yu, *Advanced manufacturing and automation V*. Southampton: WIT Press, 2016.
- [26] S. Erol, A. Schumacher, and W. Sihn, "Strategic guidance towards Industry 4.0—a three-stage process model," *International Conference on Competitive Manufacturing*, vol. 9, pp. 495–501, 2016.
- [27] C. Carfagna and P. Vincenzini, *Smart and Interactive Textiles*. *Trans Tech Publications*, 2013.
- [28] R. K. J. D. Silva, T. D. Rupasinghe, and P. Apeagyei, "A collaborative apparel new product development process model using virtual reality and augmented reality technologies as enablers," *International Journal of Fashion Design, Technology and Education*, vol. 12, no. 1, pp. 1–11, Nov. 2018.
- [29] "TOWARDS A MATURITY MODELING APPROACH FOR THE IMPLEMENTATION OF INDUSTRY 4.0" by Karan Menon, Hannu Kärkkäinen et al., Site. [Online]. Available: <https://aisel.aisnet.org/pacis2016/38/>. [Accessed: 29-Nov-2019].
- [30] A. Grieco, P. Caricato, D. Gianfreda, M. Pesce, V. Rigon, L. Tregnaghi, and A. Voglino, "An Industry 4.0 Case Study in Fashion Manufacturing," *Procedia Manufacturing*, vol. 11, pp. 871–877, 2017.
- [31] E. Oztemel and S. Gursev, "Literature review of Industry 4.0 and related technologies," *Journal of Intelligent Manufacturing*, 2018.
- [32] L. Monostori, "Cyber-physical Production Systems: Roots, Expectations and R&D Challenges," *Procedia CIRP*, 11-Jul-2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2212827114003497>. [Accessed: 29-Nov-2019].
- [33] M. Hermann, T. Pentek, and B. Otto, "Design Principles for Industrie 4.0 Scenarios," 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016.
- [34] R. F. Babiceanu and R. Seker, "Big Data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook," *Computers in Industry*, vol. 81, pp. 128–137, 2016.
- [35] E. Hofmann and M. Rüsch, "Industry 4.0 and the current status as well as future prospects on logistics," *Computers in Industry*, vol. 89, pp. 23–34, 2017.
- [36] J. Lee, H.-A. Kao, and S. Yang, "Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment," *Procedia CIRP*, vol. 16, pp. 3–8, 2014.
- [37] R. Molino, E. Carca, M. Zoppi, F. Bonsignorio, M. Callegari, A. Gabrielli, and M. Principi, "A Multi-agent 3D Simulation Environment for Clothing Industry," *Simulation, Modeling, and Programming for Autonomous Robots in Computer Science*, pp. 53–64, 2008.
- [38] "How ready is your company to connect ..." - Rockwell Automation. [Online]. Available: https://literature.rockwellautomation.com/idc/groups/literature/documents/wp/cie-wp002_en-p.pdf. [Accessed: 29-Nov-2019].
- [39] "(PDF) Big data in fashion industry - ResearchGate." [Online]. Available: https://www.researchgate.net/publication/320951232_Big_data_in_fashion_industry. [Accessed: 28-Nov-2019].
- [40] Y. Xu, S. Thomassey, and X. Zeng, "AI for Apparel Manufacturing in Big Data Era: A Focus on Cutting and Sewing," *Artificial Intelligence for Fashion Industry in the Big Data Era Springer Series in Fashion Business*, pp. 125–151, 2018.
- [41] C. Hsu, T. Lee, and H. Kuo, "Applying fuzzy theory based data mining to establish the female sizing systems," *Proceedings of the International Conference on Evolutionary Computation Theory & Applications*, 2009.
- [42] R. Nayak and R. Padhye, "Artificial intelligence and its application in the apparel industry," *Automation in Garment Manufacturing*, pp. 109–138, 2018.
- [43] T. Atobishi, S. Z. Gábor, and S. Podrúszik, "T. Atobishi, Z. G. Szalay, and S. Bayraktar, "Cloud Computing And Big Data In The Context Of Industry 4.0 : Opportunities And Challenges," *Proceedings of the IISSE Annual Conference, Sevilla, Spain*, 2018.
- [44] K. Kennedy, "Everybody Immersive Fashion Human-Computer Interaction in VR," *Artificial Intelligence on Fashion and Textiles Advances in Intelligent Systems and Computing*, pp. 37–43, 2018.
- [45] Jayatilake, Hasitha and Peter, Suren, A Road-map to Achieve Apparel 4.0; The Apparel Industry Perspective of Industry 4.0 (August 20, 2016). Available at SSRN: <https://ssrn.com/abstract=3440149> or <http://dx.doi.org/10.2139/ssrn.3440149>.

A solution to overcome speech disorder of patients using Brain Neuron EEG Signals

J. A. D. T Jayawickrama*
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology, Sri Lanka
dinuwan.adg@gmail.com

Samantha Thelijagoda
Business School
Sri Lanka Institute of Information Technology, Sri Lanka
samantha.t@slit.lk

Abstract: Speech disorders are neurodevelopmental disorders such as Stuttering, Dysarthria, Dysphonia and Aphasia associated with left inferior frontal structural anomalies that involve repeating or prolonging a word, syllable or phrase, or stopping during speech and making no sound for certain syllables. Most of the people who are suffering from speech disorders encounter difficulties in professional communication. Since people are busy with their day to day life, it is not practical to spend more time in consulting a doctor or do speech therapies for their medical issues. The speech therapist generally charges a significantly much higher rate for a single speech therapy practice, which the patient needs to practice at least twice or more for a week to get a better result. In an economy like Sri Lanka, people with average income cannot afford such an amount of money. Therefore, an innovative desktop application for speech disorder patients to overcome this problem has arisen. The main aim of this application is to reduce the speech imperative percentage of speech disorder patients via capturing the electroencephalogram feed of speech motor (Broca's area) using brain neuron O1, O2, C3, C4, F3, F4, F7, F8 electrodes and analyzing it to identify speech imperative issues. This system identifies the current impact on the left hemisphere of the brain (Broca's area) using EEG neurofeedback. Using speech voice analysis, the system provides the user to measure the articulation interference of the speech process. Self-Learning video tutorials are available for the clinical practices and treatments are available as prolong, relaxing, and humming exercises. Patients can track down the improvements daily or monthly by the rating system which makes the system unique among all other systems and the result can be directly sent to the desired consultant/neurophysiologist by the system itself. Patients can save time and the total cost of a therapy fee by using this system.

Keywords: Automated interface, Brain neuron waves, Self-learning video therapies, Speech disorders, Voice analysis algorithm

I. INTRODUCTION

The purpose of creating this concept is to improve people's speech disorder issues by using EEG brain wave signals. The algorithm basis audio recognition pattern contains self-learning video therapy tutorials. It is a challenging task, as it requires accurate EEG measuring electrode positions and algorithm patterns which can analyze the input voice and predict the outcome impact. Self-learning therapy-based video clips will help the patients to gain their skills. Many researchers have experienced speech/voice analyzing scenarios to measure the outcome of the voice, but none focused on analyzing the real effect of the speech motor function known as Broca's area [1] which controls and functions the speech activities in the brain [2]. Since there are various types of disorders such as aphasia, dysarthria and

dysphonia, there is a need for a system to identify the impact of these types of stages of the patients having speech disorders.

This system helps patients who are suffering from speech disorder problems to develop their vocal skills. The system should be operated by a patient or any third-party person. Using the system, it is possible to generate emails to doctors reminding the current activities of the patient. Between algorithm mechanisms in the system, one algorithm pattern analyses the changes of the brain's encephalography waves. The remaining speech-based algorithms help to identify the speech imperative, gaps between voices. The system provides self-learning videos for all the treatments, so patients can do treatments by themselves and, they can watch their improvements daily or monthly by the rating system which makes the system unique among all other systems. Besides, there is an algorithm-based voice pattern analyzing system to calculate the gap between the sentences and ensure the users' success using the system.

Due to the lack of such a system in the market place and based on the current studies, implementing a desktop application for people suffering from speech disorders can bring treatment to overcome their issues. This system is basically about a medical issue. Currently, many people are suffering from aphasia, dysarthria and dyslexia. This can be occurred due to many reasons such as normal speech development, inherited brain abnormalities, stroke or brain injury and mental health problems. Stammering is a speech disorder that associates repeating or prolonging a word, syllable or phrase or delaying during speech and making no sound for certain syllables. People who stammer know what they want to say/express, but have difficulty saying it. Children and adults who suffer from speech disorders get relevant medications or medications such as speech therapy and psychological consultancy to improve speech patterns.

Many researchers have gone through speech/voice analyzing scenarios to measure the outcome of the voice but none focused on analyzing the real effect of Broca's area which controls and functions the speech activities in the brain. In addition, using the video therapy tutorials, there is a chance of curing oral motor delay, tongue thrust and apraxia in children. Since there are various types of disorders such as aphasia, dysarthria, dysphonia there's a need for a system to identify such stages. [3] The main purpose of the research is to provide the most effective and efficient Information Technology based solution for people with speech disorders.

II. LITERATURE REVIEW

Jeffrey L. Fannin has conducted several studies on EEG behaviors and has discovered that the waking apprehension & reasoning wave is associated with normal waking apprehension beta (14-40Hz) [4] and an intensive state of alertness, logic and critical reasoning. Stress, anxiety and restlessness also can be translated by beta brain waves. The waves of deep relaxation are alpha (7.5-14Hz) and present in deep relaxations or during light meditation. It is an excellent time to program the mind for success and imagination, visualization, memory, learning and concentration also heightens by alpha waves. Sleep is detrimental to one's health in more ways than one. The insight wave is the Gamma wave and it is above the frequency of 40Hz. It is the fastest frequency above 40Hz. Most of the initial research shows that waves of Gamma are associated with bursts of insight and capable of high-level information processing. The slowest frequencies as known as the deep sleep wave are delta waves (0.5-4Hz) and are experienced in deep, dreamless sleep and in very deep, transcendental meditation where readiness is fully detached. The Light meditation & sleeping waves are theta waves (4-7.5Hz) all the while during deep meditation and including the all-important REM dream categories. Theta activity from 7Hz to 8Hz, where the optimal range for visualization, mind programming and using the beneficial power of the human mind begins. Children who stutter bear to demonstrate improvement in anomalies of Broca's area, such as less grey matter in the left inferior frontal gyrus, and disorganization of the white matter in the left Rolandic operculum. Researcher Dr. Christian A. Kell and her team [5] analyzed brain morphology like grey and white matter and activations during fluent speech management in PS (before and after a fluency-shaping therapy), RS, and in control, subjects using Magnetic Resonance Imaging (MRI). They have dissociated pathogenesis related anomalies from compensation effects by relating the magnitude of the neural anomaly (grey matter and white matter) to an individual degree of the symptom (off-line stuttering severity). Studies according to the past decade's neuroimaging, catered enough information on the neurophysiological changes paralleling processes like recovery from stuttering, such as treated recovery from stroke or motor learning, therefore, Roger J. Inghama, Janis C. Inghama [6] could not find a relative pattern to analyze brain activities with stuttering as well. However, they have used Deep Brain Stimulation (DBS/DES) a classical conclusion for Deep Brain Stimulation and both of stimulations are hypokinetic and hyperkinetic movement disorders with a primary subcortical origin, including Parkinson's disease, dystonia and essential tremor as well as neuropsychiatric disorders such as treatment resistant depression and obsessive-compulsive disorders, and Tourette Syndrome (TS), which is a neurobehavioral disorder. Common target structures are parts of the basal ganglia, namely the Globus Pallidus internus (GPi) and the SubThalamic Nucleus (STN). The procedure has some shortcomings that may likely decrease its potential application in stuttering.

Correspondingly within the area of deep brain simulation studies over the past twenty years, there has been a boundless boost in the number of published studies regarding the people who stutter, documenting structural and functional differences compared to their fluent peers. Andrew C. Etchella, Oren Civierb, Kirrie J. Ballardd and Paul F.

Sowman [7] have found out techniques like positron emission tomography (PET), functional magnetic resonance imaging (fMRI) and near infrared spectroscopy (NIRS) which rely on the flow or the magnetization of differences in blood and are indirect measures of neural activity. Functional Magnetic Resonance Imaging (fMRI) are comparatively better among than spatial resolutions (fMRI having the highest resolution of the three), being able to determine precisely which cortical (PET, fMRI, and NIRS) and subcortical (PET and fMRI) regions of the brain are active during a task. A set of researchers; Soo-Eun Changa, Michael Angstadt a, Ho Ming Chowa, Andrew C. Etchell a, Emily O. Garnett a, Ai Leen Choo b, Daniel Kessler a, Robert C. Welshc and Chandra Sripadaa [8], joined a large longitudinal neuroimaging dataset of the patients which includes the information of who do and do not stutter in children and analysis of whole-brain network in order to examine the changes in the intra- and inter-network connectivity associated with stuttering. The activity of the brain investigated by pediatric image technology during a developmental paradigm in which participants are presented with stimuli (e.g., picture naming, sentence reading, and word generation) and are instructed to provide a response.

A preparatory investigation on the neural oscillatory aspects of speech motor arrangement prior to dysfluent and fluent utterances in adults who stutter has been carried by Anna Mersova et al. in [9]. The capability of the speech in adults who stutter has begun to characterize speech-motor based preparation prior to study speech onset the neural mechanisms. Compelling evidence has suggested that stuttering is associated with atypical processing within cortical and sub-cortical motor networks, exceptionally in the range of beta frequency, that is effective before speech production even begins. Due to the low frequency of experimental in settings of stuttering, be that as it may, the literature has so far reported predominantly on fluent speech production in adults who stutters. Individually there are issues that arise when identifying damaged brain areas of stuttering patients. Therefore, Paul F. Sowman a, et Margaret., mentioned the differences of grey matter volume [10] of left caudate nucleus in stuttering patients by asking from the participants who suffers from stuttering to self-rate their current stuttering severity and the range of severities over which their stuttering could vary on a 10-point scale (1 = no stuttering, 10 = extremely severe). Age of stuttering onset, duration of stuttering, and information about any stuttering treatment were also recorded, along with any other relevant information offered by participants.

TABLE I. EXISTING SYSTEMS

Systems	Dysarthria	Apraxia	Oral Motor delay	Aphasia	EEG Based	Video Learning
Gemm Learning	Yes	Yes	No	No	No	No
Talk time with tucker	No	No	No	No	No	No
Articulation Pro	No	Yes	Yes	No	No	No
Webber hear it	No	No	Yes	No	No	No
Proposed Solution	Yes	Yes	Yes	Yes	Yes	Yes

III. METHODOLOGY

This solution is a practical world system rather than a system that solves theories. Using a prototyping methodology is the ideal solution to fulfill this requirement. Prototyping will test the system from time to time and will provide feedback on the accuracy of the system. In general, there are many types of speech disorders such as dysarthria, stuttering, aphasia and dysphonia. All of them are interconnected to the speech motor area called 'Broca'. Using the Encephalography signals, it will be much easier to capture the impact and the weight of the disorder which damaged the brain or its functions. There are brain waves called delta, theta, alpha, beta, and gamma which can be identified as Encephalography signals. By analyzing these signals which are captured by the device called "NeuroSky Mindwave" it will show the current state or the level of the patients which they are at. Using verbal sounds(speech) level will help to analyze the impact on the speech imperatives. The Encephalography sensor module will automatically start to read the outgoing signals from the brain. Using the voice inputs by the user, the system itself will analyze the piece of voice track and shows a result set the current progress of the user compared to a pre-recorded voice clip (it will always be kept as the default and will be only used for comparisons). Simultaneously, the electronic magnetic device (NeuroSky Mindwave Headset) will capture the outgoing brain waves receiving from the speech motor area from the brain. EEG signals will be categorized into parts and filtered using an algorithm. The system will forward the captured data of brain waves to a framework that analyzes the current impact on the brain. At the end of a session, a user can send a complete report to the desired consultant a copy of the analysis report generated by the system itself.

A. Planning

Based on many literature review findings, EEG brain neuron signals were found as one of the best methods to analyze the impact on brain among human voice frequency pattern analysis. Getting the report via interviewing the patient has been proven ineffective. The therapist should be board at giving the treatments. In this research, the current situation of the patient must be identified via the algorithm. This will allow the system to provide the best solutions and medical treatments through systematic ratings, tutorial videos and voice clips. This system will use EEG brain neuron wave analyzing patterns, speech recognition algorithms. The research will be carried out continuously to find out whether there is a better way to improve the accuracy of analyzing EEG brain waves. The main research tools of this research are EEG brain wave analyzing, vocal impact tests and self-learning medical treatment videos. As to established practicality of the research, it is mandatory to use delta, theta, alpha, beta, gamma electrodes to analyze the EEG signals. Analyzing vocal impact can be considered as a little bit impractical, but to make it most appropriate the researcher has found some methods which are Frenchay dysarthria assessment, Stuttering Hexogen and voice signal sampling that can be used as exercises as well.

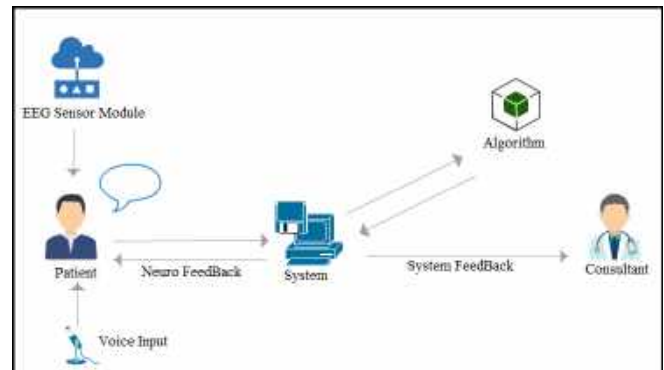


Fig. 1. System Design Diagram

B. Information gathering

Systems that assist speech disorder patients are increasingly used in web and mobile platforms. There are various types of desktop and mobile applications, but this innovation has already made way to a lot of comfort and convenience to several people having speech disorder issues. To gather primary data, meetings are arranged with the Head in Charge of Mental & Psychology department of General Hospital Colombo and Lady Ridgeway Hospital for Children, and interviews were conducted as a qualitative method of data gathering. Formal interviews have been carried out with doctors and consultants and informal interviews with patients. Interviews with neurophysiologists were conducted to gather information regarding the functioning of EEG signals and how O1, O2, C3, C4, F3, F4, F7, F8 electrodes are important to identify speech imperative issues with patients having voice disorders and what immediate reactions were made to overcome those issues by neurophysiologists.

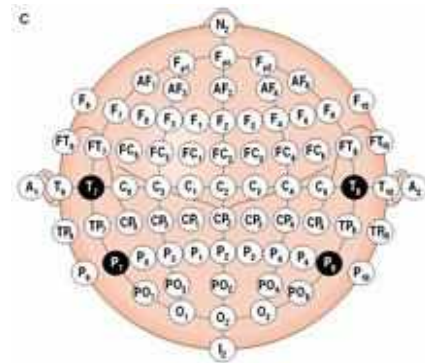


Fig. 2. EEG – Electroencephalography of Brain

Also, several meeting sessions were arranged with the physiatrist who does speech therapy such as fast track speech therapies and prolong therapies for stuttering patients. The findings from the interviews indicate that there are about 60 patients in Colombo General Hospital and collected data regarding all the symptoms and the history of their speech disorder issues. About 20 people have experienced speech disorders since birth. Some patients were stammering from time to time while some others were made by strokes. As primary data gathering, studies have been carried out for currently available voice recognition systems, identifying their methods. As for the secondary data gathering, a number of research papers were analyzed and that can be useful as a technological aspect of secondary data gathering.

C. Data Analyzing

In the present time, there is no system that analyzes speech disorders based on EEG brain signal electrodes, so developing this proposed system can be considered as a novel approach to compromise and align with the pre-designed process as its requirement analysis has been done by reviewing multiple research journals and interviews. The fundamental objective of the analysis phase is to figure out and document the needs of the business and the processing requirements of the new system. In the analysis phase of the project, requirement gatherings were carried out about the research and prioritized the requirements to achieve to desired goal step by step. Based on the studies, EEG signal analysis has been identified as the main requirement and voice recognition algorithms and self-learning video therapy tutorials as sub-requirements.

D. Designing

This will be the first phase in designing since the patient is interacting with the interface which has been already provided. The main interface contains registration, settings, view reports, video category, skills development, patient's info, encephalography analysis and voice algorithm analysis sub interfaces to allow the user to access the application in a more convenient way. The interface must be user-friendly and must fulfill all the needs of the patient and be easy to access.

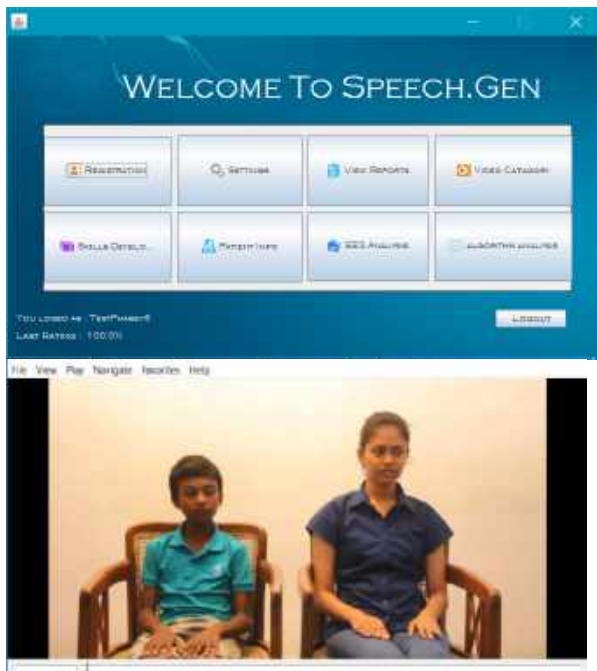


Fig. 3. Main Interface and Meditation Level signal

IV. IMPLEMENTATION

According to the information gathered, a prototype of the system was developed and shown to the responsible persons. In the EEG analysis phase, the system provides adjustable input fields for the delta, theta, alpha, beta, and gamma EEG waves. The main reason is not to set up with a static value to each signal leave because there will be various types of patients who use the system, so that frequency values depend on each patient's disorder conditions. The patient needs to be

wearing the NeuroSky Mindwave Headset when running the analysis. At first, it will read and save the data to an excel sheet when only the connect was made between the device and the human brain scalp. When the process is done, it will be shown as multi-type graphs displaying the frequency compared to the expected level and it will be sent to the neurologist as an email via the system.

TABLE II. EEG FEED ANALYSIS DIAGRAM

Signal Frequency							
Delta	Theta	Low alpha	High alpha	Low beta	High beta	Low gamma	Mid gamma
5180	248	44	47	72	81	11	11
880113	166214	4550	10565	9681	9366	3278	1447
301403	25462	4583	3550	689	951	457	167
233022	13483	4367	1251	2444	2610	883	213
454202	253160	24639	15882	27102	7756	4334	9711
2228099	123839	55359	53930	27389	8816	8734	7712

The speech analysis can be in divided into three sections. As for the first algorithm, it inputs the patient's voice via default microphone and it will be sent to a pattern analyzing method to identify the issues and the impact on the disorders of their speeches. As for the algorithm, it was constructed and filtered by the Department of Physiotherapy and Occupational Therapy under National Hospital Sri Lanka and will predict the output via calculating voice.

Voice Pattern calculation is one of the key calculations in the system. It calculates the reading time of a given paragraph by a healthy person and a stammering patient who are facing the test. According to the results of the patient, the current improvements of the patient are determined and ratings are provided using the following equation.

TPA 1 = Reading time of healthy person – Reading time of stammering Patient

- IF TPA 1 is between 1-5 – rating = 100%.
- IF TPA 2 is between 5-10 – rating = 90%
- IF TPA 3 is between 10-15 – rating = 80%
- IF TPA 4 is between 15-20 – rating = 70%
- IF TPA 5 is between 20-25 – rating = 60%
- IF TPA 6 is between 25-30 – rating = 50%
- IF TPA 7 is below 30 = “Display to improve”

The second speech analysis uses the Frenchay dysarthria assessment (FDA) which is an international treatment standard on speech disorders. Voice clips will be recorded according to the process. the patient must read out loud the shown sentences within the given time and then it will be sent to the desired speech therapist for further analysis of the verbally sound pronunciations.

For the third speech analysis, the patient's voice will be recorded and be instantly given an image output of a frequency [11] of the input voice file. Self-learning video therapy treatments will guide the patient to improve his/her speaking patterns.

Since it is the best clinical solution available, the researcher decided to record it as video clips and divide into

the seven categories such as breathing [12], fast prolong, prolong, humming, relaxation and shadowing exercises. All these videos were recorded under the direct supervision of a speech therapy specialist. Implementation of the system has been done using NetBeans IDE 8.0. Running the system requires MySQL Server, JDK 6 (minimum) platforms. The system database has been developed in a MySQL environment. Use of these languages and database environment were based on cost-effectiveness as all of them are open source resources. Integration with an SMTP server will be required to send email reports. Predicting the patient's voice-based rating system was done based on the voice pattern analysis algorithm.

V.RESULTS & DISCUSSION

The results of the analysis of EEG signals in the brain neuron signals are captured and displayed in charts. According to the system feedings, there are x & y-axis charts for Delta, Theta, Low alpha, High alpha, Low beta, High beta, Low gamma, Mid gamma. Assumptions levels of a sample neuro feedback raw data of a normal person were added as follows.

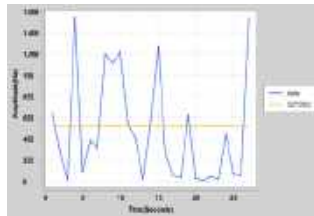


Fig. 4. Delta signal

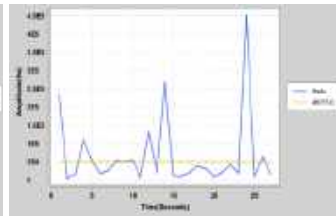


Fig. 5 Theta Signal

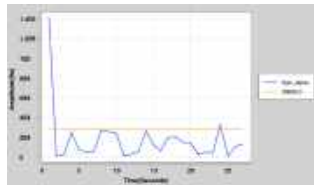


Fig. 6. High Alpha signal

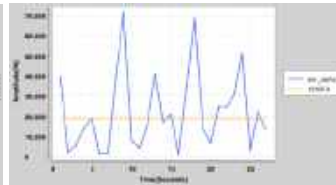


Fig. 7. Low Alpha signal

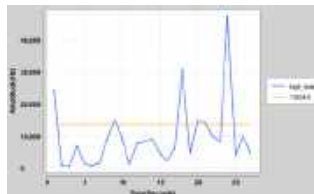


Fig. 8. High Beta signal

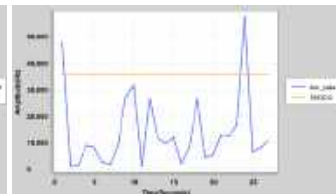


Fig. 9. Low Beta signal

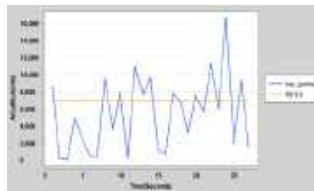


Fig. 10. Low Gamma signal

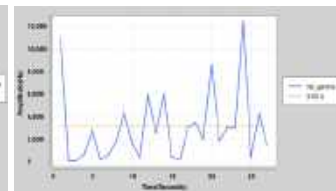


Fig. 11. Mid Gamma signal

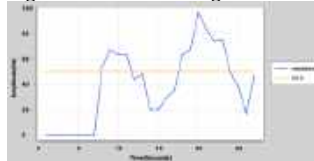


Fig. 12. Meditation Level signal

TABLE III. ELECTRODE FREQUENCY LISTING

Graph Figure	Electroencephalography	
	Signal	Frequency
Figure 4. Delta signal	Delta	527720
Figure 5. Theta signal	Theta	49175
Figure 6. High Alpha signal	High alpha	28666
Figure 7. Low Alpha signal	Low alpha	19105
Figure 8. High Beta signal	High beta	13604
Figure 9. Low Beta signal	Low beta	35430
Figure 10. Low Gamma signal	Low gamma	6910
Figure 11. Mid Gamma signal	Mid gamma	3100
Figure 12. Meditation Level signal	Meditation Level	50

Based on the shown data on the graphs, a neurophysiologist can measure the impact of the speech motor area of a speech disorder patient. Depending on the type of speech disorders, these input values may change according to the actions taken by a neurophysiologist.

High alpha signals are the sinusoidal signals of relatively high voltage/amplitude and develop from the human's brain at about 10 waves per second. High beta signals are formless waves of low volatility/amplitude and develop from the brain at a rate of 20 waves or more per second. Between these signals, the activity of the beta signal is seen as a disruption of the activity of alpha waves. An increase in the amplitude of alpha waves is an indication of reduced consciousness. A disruption of the alpha rhythms into formless beta activity is an indication of a focus of mental activity. The more alike any two brain areas left hemisphere and right hemisphere are in alpha patterns, the more indication of the lack of mental activity. This type of indication of the mental activity can be declared as reduced consciousness.

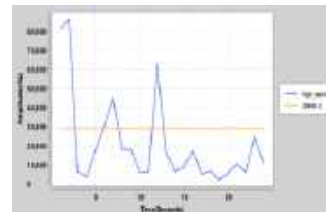


Fig. 13. High alpha of a Patient

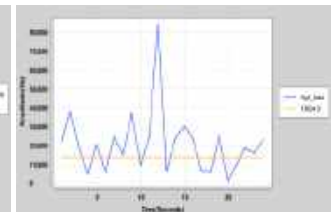


Fig. 14. High beta of a Patient

In regard to the size of the waves of speech disorder patient, it was found to have larger high alpha and high beta waves than normal person Fig.13, Fig.14. Following the criterion that larger waves, means a loss of mental specificity. Four of speech disorder patients [13] were used in order to test these scenarios. All the raw data was captured before and after the speech therapy, practical videos were watched by the patient. After attempting on several relaxations, breathing and humming exercise therapy videos with speech disorder patients, there was a slight incensement of the encephalography waves with speech disorder patients [14]. When the encephalography waves of speech disorder patient and the waves of normal person were compared, the normal person was discovered to have a stabilized amplitude and the speech disorder patient [15] was found with larger waves with high amplitudes and failed to maintain attention level as well. This was done only on the left hemisphere of the brain area.

As per the voice spectrum analysis, the finding indicates that there are differences in the normal person's voice

spectrum [16] against the patient's voice spectrum. It cannot be identified by the patient himself, but those spectrums will be sent as an e-mail to the neurologist/consultant via the system, in order to analyse further.

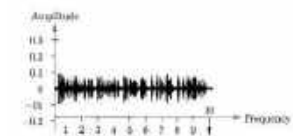


Fig. 15. Original Spectrum Signal

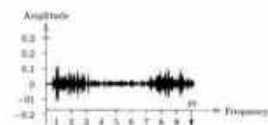


Fig. 16. Sample Spectrum Signal

The system was pre-integrated with a voice spectrum sample of a person [17] who does not have any speech disorders Fig 15, during the use of the system it compares with a voice spectrum sample of a speech disorder [18] patient Fig 16. Later, the test sample spectrum file will be sent to a doctor/neurologist for further analysis of the patient's sound wave condition in order to find defects of the speech.

VI. CONCLUSION

Current studies have found that using the EEG signal helped patients to monitor their level of impact, but not to identify the direct impact because the NeuroSky Mindwave device which I used to identify the EEG signals only covered the T4, A1 and Fp1 electrodes in the brain scalp and the device only had 1 channel to simulate the signals. Most electrode places were not covered by the electrode channels such as O1, F3, F4, F7, F8 electrodes which were only specialized for the speech motor area due to the lack of electrode channels of the device. However, there was a significant improvement in using the EEG signals for neurologists to maintain the patient's anxiety level using the mediation level provided by the system. Also, the Delta, Theta, Low alpha, High alpha, Low beta, High beta, Low gamma, Mid gamma signal levels were useful to monitor and track the brain activity while using the system multiple times. Fluency of speaking of the speech disorder patients has been improved using voice analysis patterns and with the help of self-learning video therapy tutorials as well. Therefore, the effectiveness of the suggested solution to overcome the issue of speech disorder patients using brain neuron EEG signal is clear as per the significant improvements of patients. Based on the studies, several factors and drawbacks were identified and those will be useful to inform the future studies in this type of speech disorder.

ACKNOWLEDGMENT

The work described in this paper was carried out for an M.Sc. Research Project. This paper is a result of the encouragement, support and guidance given by doctors, consultants and many others. Special thanks to Professor Raveen Hanwella (Senior Lecturer/Head-Department of Psychological Medicine Center, Colombo General Hospital), Doctor Chathurie Suraweera (Senior Consultant/Senior Lecturer in Psychiatry in Faculty of Medicine, University of Colombo), Doctor Anuruddha Padeniya (Pediatrics Physician-Lady Ridgeway Hospital) Doctor Piyara Ratnayake (Paediatric Neurologist- Lady Ridgeway Hospital), Ms. U. Tharanga (Speech Therapist-General

Hospital Kaluthara) and Mr. Lakshan Hirumal (Videographer- Lakshan Hirumal Photography).

REFERENCES

- [1] E. C. Gildersleeve-Neumann, K. J. Jakielski and R. Stoeckel, "Motor-Based Intervention Protocols in Treatment of Childhood Apraxia of Speech (CAS)". *Specific Language Impairment/Speech Sound Disorders*. 2014, pp.197-206.
- [2] "Speech and language difficulty commonly affects individuals with dementia and other neurological conditions". Weill institute for neurosciences. University of California. [Online]. Available From: <https://memory.ucsf.edu/symptoms/speech-language>[Accessed on: Oct. 10, 2019.]
- [3] L. Scott, The Florida State University. "Communication disorders and temperament". [Online] Available From: <https://www.stutteringhelp.org/communication-disorders-and-temperament>[Accessed on: Oct. 10, 2019]
- [4] L. J. Fannin, "Understanding your brainwaves". Center for Cognitive Enhancement.[Online] Available From: <http://www.wadelang.com/assets/understanding-brain-waves.pdf>[Accessed on: Oct. 10, 2019.]
- [5] A. C. Kell, "How the brain repairs stuttering". A Journal of Neurology. Brain Imaging Center, Department of Neurology. 2019, pp.2747–2760.
- [6] R. J. Ingham, J. C. Ingham, H. A. Euler and K. Neumann, "Stuttering treatment and brain research in adults: A still unfolding relationship". *Journal of Fluency Disorders*. 2018, p.109.
- [7] A. C. Etchellae, O. Civier, K. J. Ballard and P. F. Sowmane, "A systematic literature review of neuroimaging research on developmental stuttering between 1995 and 2016". *Journal of Fluency Disorders*. 2016, p.9.
- [8] Soo-Eun Chang, "Anomalous network architecture of the resting brain in children who stutter". *Journal of Fluency Disorders*, 2018, pp. 46-67.
- [9] A. Mersov, D. Cheyne, C. Jobst and L. D. E Nil, "A preliminary study on the neural oscillatory characteristics of motor preparation prior to dysfluent and fluent utterances in adults who stutter". *Journal of fluency disorders*, 2017, p.147.
- [10] F. P. Sowman, Grey matter volume differences in the left caudate nucleus of people who stutter. *Brain and Language*, 2017, p.10.
- [11] C. A. Kell, K. Neumann, M. Behrens, A. W. Von Gudenberg and A. L. Giraud, "Speaking-related changes in cortical functional connectivity associated with assisted and spontaneous recovery from developmental stuttering". *Journal of Fluency Disorders*, 2017, p.137.
- [12] R. Prestes, "Temporal processing and long-latency auditory evoked potential in stutterers". *Brazilian Journal of Otorhinolaryngology*. 2017, 83(2), p.144.
- [13] J. Chesters, K. E. Watkins and R. Möttönen, "Investigating the feasibility of using transcranial direct current stimulation to enhance fluency in people who stutter". *Brain and Language*. 2017, p.70.
- [14] J. Pálffy and J. Pospichal, "Pattern search in dysfluent speech. *IEEE International Workshop on Machine Learning for Signal Processing*". Santander, Spain, Sept. 2012.
- [15] M. Finsterwald, M. Vogel and K. Trondle, "Computer aided methods for diagnosis and therapy of speech breathing disorders". *Engineering in Medicine and Biology Society*, 2002, p.1.
- [16] S. S. Awad, M. W. Coreless and R. Merson, "Computer Assisted Treated for Motor Speech Disorders". *Instrumentation and Measurement Technology Conference*. 2002. p.599.
- [17] "Developmental Levels of Disfluency - Center for Stuttering Therapy", *Center for Stuttering Therapy*, 2020. [Online]. Available: <http://www.coloradostutteringtherapy.com/developmental-levels-of-disfluency/>. [Accessed: 10- Oct- 2019].
- [18] M. Danubianu, S. G. Pentiu, O. A. Schipor, M. Nestor and I. Ungureanu, "Distributed Intelligent System for Personalized Therapy of Speech Disorders". *Computing in the Global Information Technology*, 2008, p.166.

Extending use-case point-based software effort estimation for Open Source freelance software development

Dharshitha Srimal Senevirathne*
Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
senevira_im14045@stu.kln.ac.lk

Thareendhra Keerthi Wijayasiriwardhane
Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
thareen@kln.ac.lk

Abstract: Accurately predicting the software development effort is very crucial when delivering the software systems on time, within the budget and with the required functionality. Overestimation of the software development effort can lead to losing the projects whereas underestimation can cause budget and schedule overruns. The development effort of a software project depends on various factors and these effort factors associated with the freelance software development are different from those of traditional software development. Software development companies employ various proprietary tools in their projects for their planning, development, testing, etc. However, freelance software developers functioning under tight budgetary constraints are not in a position to afford them. As a result, they tend to use free and open-source tools for their software developments. There are various types of software effort estimation models proposed, published and practiced in the industry. However, there is no such software effort estimation model specifically proposed to estimate the effort of freelance software development. The main objective of this paper is to extend Use Case Point-based software effort estimation for the open-source freelance software development. Initially, details of open source software projects were collected from several freelance software developers. Based on the use case diagrams, Use Case Points counts are then calculated for each project. Taking other effort drivers associated with open source freelance software development also into account, we then estimate the effort of each software development. Our aim is to explore the viability of using Use Case Points as the main effort driver in estimating the effort of open source freelance software development.

Keywords: Freelance software developers, Open source software development, Software effort estimation, Use Case points

I. INTRODUCTION

Accurately predicting the software development effort of freelance software developers is very crucial when delivering the software systems on time, within the budget and with the required functionality. Overestimation of the software development effort can lead to losing the projects, whereas underestimation can cause for budget and schedule overruns. Since this research is based on the software effort estimation (SEE) of open source (OS) the freelance software development, it is really important for the freelance software developers to make accurate predictions of effort the software projects. There are no SEE models specifically proposed to estimate the effort of OS freelance software development.

A. Freelance Software Development

With the rapid growth of Information Technology (IT), the self-employment rate of the IT industry has started to increase. Most organizations and people tend to get away from the traditional methods of handling businesses. Since the businesses move towards technology-based solutions and systems, hiring a freelancer to come up with better solutions has become a more cost-effective option. It is really important for the freelance developers to properly estimate the effort of the software projects not only to land the projects but also to deliver the projects within the time and budget. Even though there is a rapid increment in the self-employment and freelancing rate the freelancers find it hard to estimate the effort of the projects accurately. Therefore, it is important for freelance developers to identify the factors that affect the effort of an Open Source Software (OSS) project. These factors are known as effort drivers. The effort drivers associated with freelance software development are different from those of traditional software development. In a software company, a software project is performed by a team, which consists of software developers, business analysts, a tech lead, a team lead, etc. However, freelancers tend to work alone or in very small groups.

B. Open Source Software Development

OSS developers are usually expected to use their own hardware and software tools. Similarly, OSS developers are typically not owned or controlled by any organization to monitor and manage their software development process [1]. Software development companies use various proprietary tools in their software projects in development, testing, planning, etc. phases. However, freelance software developers functioning under tight budgetary constraints are not in a position to afford them. As a result, they tend to use, free and OS tools for their software developments.

C. Use Case Points based model

There are various types of the SEE models proposed, published and practiced in the industry. Many of these models use software size as the main effort driver for their estimations. Among the various types of software size measures proposed, Lines of Code (LOC) is the most widely used measure in sizing software products. However, the LOC is only available in the latter stages of the development of a software product. Therefore, as an alternative, Function Points (FP), a measure based on the functionality of a software product has been introduced to quantify the software size in the early stages of software development. However,

the service of trained experts is required to get the FP count of software system accurately and thereby the freelance developers are not in a position to use FP for sizing their software products. Most of the modern software developments are carried out using Object-Oriented System Analysis and Design (OOSAD) methodology. In the early stages of the software development life cycle, OOSAD requires to draw use case diagrams to identify the requirements and the functionality of the system. As a result, an extension of FP, based on use cases called Use Case Points (UCP) has been introduced. Therefore, in this research, we discuss the viability of using UCP as the software sizing measure of each software project.

D. Software Effort Estimation

Among the various SEE methods, the analogy is considered as the most commonly used method. Analogy based estimation has shown better evaluation results compared to other machine learning and non-machine learning methods [2]. Since analogy based SEE models are able to learn from previous experiences analogy compares the effort drivers of the proposed software project to previous software project data to estimate the effort of most similar projects.

The goal of this research is to extend a SEE model to accurately predict the effort of OSS projects developed by freelance software developers. In this research, we look into OSS projects of freelance software developers. First, we identify the effort drivers of freelance OS SEE. Since UCP is selected as the software sizing measure of this research, we examine the use case diagrams of each software project. Based on the use case diagrams, the UCP count of each software project is calculated. The identified effort drivers with UCP count is applied to an analogy based SEE model to estimate the effort of each software development. Finally, the validation of the SEE model is done using n-fold cross-validation.

The structure of this paper is as follows: Section 2 presents an overview of related work. Section 3 describes the methodology that this research is conducted. Section 4 presents and discusses the results. Finally, in Section 5 we present the conclusions and future work.

II. RELATED WORK

There are different types of SEE in the industry by the practitioners. COCOMO, COCOMO II, WALSTON-FELIX MODELS, and SLIM are some of them. They use LOC as their main effort driver for their estimations. Since LOC is only available in the latter stages of software development, FP is used as an alternative. FP is a software sizing measure based on the functionality of software projects. It has been introduced to quantify the software size in the early stages of software development [3]. FP was originally proposed for the procedural systems, but it has now been extended to many other software development paradigms such as Object-Oriented systems [4], embedded systems and real-time systems [5], [6]. However, there is an international standard for counting FP and the service of trained experts is required to get the FP count of software systems and thereby the freelance developers functioning under tight budgetary constraints are not in a position to use FP for sizing their software products [7]. Nowadays, most of the modern software development projects are carried out based on

OOSAD methodology. In the early stages of the software development life cycle, OOSAD methodology uses use case diagrams to identify the requirements of the system. As a result, an extension of FP based on use cases called UCP has been introduced by Gustav Karner in 1993 and UCP is evolved from FP [3]. UCP requires the use case diagrams and use case descriptions of the software projects that we consider to calculate the effort [8].

A. Estimation by analogy

Effort estimation by analogy has become more popular within the software development research community because of its higher performance in prediction when different data types are used. The concept of this method has been simplified such that the effort of a new project can be estimated by reusing efforts about similar, already documented projects in a dataset, wherein a first step one has to identify similar projects which contain the useful predictions [9]. The predictive performance of analogy based SEE relies heavily on the selection of two interrelated parameters: number of nearest analogies and adjustment strategy [8]. There are many analogy-based SEE models that have been published by many researchers. However, ESTOR, ACE, and ANGEL are the most commonly used models in predicting effort of software development.

1) *Estor*: ESTOR is an early implementation of an analogy-based tool to estimate software project effort [10]. It was developed by Mukhopadhyay et al. in 1992 [11] in order to evaluate the feasibility of case-based reasoning in SEE. However, in accordance with the concept of case-based reasoning, ESTOR uses the case-based reasoning and the three best analogies to compute effort based on the Inverse Rank Weighted Mean (IRWM) [12].

2) *ACE*: ACE (Algorithmic Cost Estimator) was developed by Emilie Mendes et al. in the late 90s and it focuses on exploring the benefits of analogy-based estimation. It calculates the difference between the target project and each calculated project in the database. ACE principles involve the use of similarity functions which should be defined to be able to compute the similarity distance of each analogy with respect to the target project. Also, the similarity function helps in the ranking of analogies in terms of the most similar and least similar [10], [12].

3) *Angel*: Analogy based effort estimation methods generate new predictions based on the assumption that similar projects with respect to features description have similar efforts. The authors of [13] proposed a tool called ANGEL to predict the effort of a project by using the analogies and estimates provided by the completed projects. ANGEL uses historical effort data and the size of completed projects to predict the estimation of the new project [14]. There are many variables that affect the effort estimation. Not all the variables selected will be helpful when finding good analogies. Some variables may create noise. There, the authors [13] have automated the process and provided an environment in which data can be processed, analogies found and estimates produced. This supports the collecting, storing and identifying the most similar projects in order to estimate the effort for a new project. ANGEL is based upon the minimization of Euclidean distance in multi-dimensional space.

B. Model comparison

ACE is known to provide a lower degree of accuracy [11]. ESTOR and ANGEL are said to use the same principles in order to provide a list of most similar analogies though, the content of the list is not always accurate. Moreover, ANGEL is computationally expensive as compared to the other methods since it saves and computes similarity for all cases. ESTOR requires additional domain knowledge in order to succeed to accurately estimate projects from very different environments. Therefore, in this research, we extend the ANGEL model to estimate the effort of OSS projects developed by freelance software developers.

However, there are some limitations to the ANGEL tool [14].

- The values of the effort factors are standardized on the basis that each effort factor contributes equally to the value of the target effort factor.
- No specific method to handle out of range values.
- No specific method to handle "multiple-exact-match" conditions.
- No specific mechanism to handle "no-match" condition

C. Performance measures

There are many performance measures that researchers use in SEE such as MRE (Magnitude of Relative Error), MIBRE (Mean Inverted Balanced Relative Error), PRED(25) (percentage of predictions failing within 25%), MMRE (Mean of Magnitude Relative Error), etc. MRE measures the error ratio between the actual effort and the predicted effort and it can be explained from the equation (1) where E_i is the actual effort and E_i is the predicted effort.

$$MRE = |E_i - E_i| / E_i \quad (1)$$

However, MMRE is considered as the de facto standard of all the performance measures out of them [15]. The equation (2) presents the mean of all the MRE values.

$$MMRE = (\sum_{i=1}^n MRE) / n \quad (2)$$

III. METHODOLOGY

The methodology of this research is conducted via four phases. In the first phase of the study, a thorough review of the literature is done in order to identify the effort drivers of freelance OSS development. Next, we collect details about freelance OSS projects from several freelance software developers. The third phase of the study is focused on selecting an appropriate SEE model and extending it from traditional software development to OS, freelance software development. In the fourth phase, the outcome of the extended model is tested and validated.

A. Identifying Effort Drivers

In software companies, a software project is handled by a group of people, which consist of software developers, business analysts, a tech lead, a team lead, etc. [16]. However, freelancers tend to work alone or in very small groups. Expert judgment is one of the most popular effort estimation methods used in the software industry. This would be a costly option for a freelancer. Therefore, freelance software developers are incapable of hiring an industry expert. Software companies use various proprietary tools for software testing, coding, managing projects, etc. Freelancers

are not in a position to afford them. As a result, they tend to use, free and OS tools for their software developments.

1) *UCP Count*: In this research, UCP is selected as the software sizing factor of the freelance, OSS projects. First, the UCP count of each project is calculated. Most of the Technology Complexity Factors (TCF) such as special user training facilities required, access for third parties, response or throughput performance objectives, etc. and Environmental Complexity Factors (ECF) such as familiarity with Rational Unified Process, part-time workers, lead analyst capability, etc. are irrelevant and have no effect in freelance OSS development. Due to this reason, TCF and ECF can be omitted [17]. Therefore, Unadjusted Use Case Points (UUCP) is considered as the main effort factor in sizing the OSS products.

According to Karner [3], the actors in a use case model can be categorized as simple, average and complex.

- Simple: Weighting factor 1
- Average: Weighting factor 2
- Complex: Weighting factor 3

The total unadjusted actor weight (UAW) is measured by counting the number of actors in each category multiplying each total by the defined weighting factor and then adding the products

The use cases are also categorized into three categories and they are simple, average and complex [3]. This categorization depends on the number of transactions, including the transactions in alternative flows. However, included and extending use cases are not considered. A simple use case has 3 or fewer transactions; an average use case has 4 to 7 transactions, and a complex use case has more than 7 transactions. A weighting factor is assigned to each use case category:

- Simple: Weighting factor 5
- Average: Weighting factor 10
- Complex: Weighting factor 15.

The unadjusted use case weights (UUCW) are calculated by counting the number of use cases in each category, and then multiplying each category of use case with its weight and adding the products. The UAW is added to the UUCW to get the UUCP.

2) *Application Domain*: The software industry as such does not have a specific domain. Rather, this industry would provide services or enable other services by applying the right technologies. A software domain is nothing but the subject area in which a particular project belongs to. It can be scientific development, business development, health/medical industry, embedded systems, defense systems, etc. The demand for each domain can vary. Some domain may have high demand while the demands for other domains is low [18]. Some developers could be more comfortable in particular domains while some are new to those domains. This means some developers may have academic training or substantial work experience in a particular domain and they understand the design, architecture, domain rules, etc.

3) *Programming language*: The programming language of an OSS project is one of the important aspects to be considered. It should be selected based on the project requirements. Different programming languages are suitable

for different application domains. For example, Python is usually used for scientific systems, whereas PHP and Java are used for web developments. The choice of selecting a programming language for developers depends on the required functionalities of the software project [19].

The programming language selected for a given project should be sufficiently expressive to cover the requirements. On the other hand, an incorrect language selection may affect the solution from reaching the expected level. Some languages are expressive thanks to their verbose nature. It allows more expressivity and hence more choices to be able to make use of simple syntax. Effectively, the simple syntax would eventually result in the developer and reader's familiarity with the code quicker than having the reader being exposed to some new complex syntax [20].

4) *Developer experience*: Freelance software developers land projects through contacts of their previous projects. The experience in freelancing can't be measured with the time because some freelance software developers could have worked with fewer projects for a long duration of time while another set of developers would have worked with many projects in a short period of time. Therefore, the number of projects they have completed is used as the measurement of the freelance developer's experience. In online crowdsourcing platforms or online market places, they get projects based on their ratings and feedbacks. Rating and feedbacks are given by the project owners of previously completed projects. Therefore, having a good experience is a clear advantage to land more projects [18]. If a developer has previous experience it is easy to understand the requirements. When developing a particular feature or solving a particular problem it would take less time to come up with a better solution. Generally, specific skills and depth of experience matter a lot when dealing with complex tasks [18]. Experience helps the developers to select a better approach to provide a solution to a requirement.

B. Data Collection

We have considered OSS projects developed in multiple domains, by software development freelancers in Sri Lanka with different levels of experience. The domains include business applications, embedded systems, data science projects, etc. The complexity of these OSS projects is different and the time taken to complete each project is also different.

We have collected the details of 13 OSS projects from 9 freelance software developers. Based on the use case diagrams, the UUCP count is calculated for each project. There can be some inefficiencies in the use case diagrams. These inefficiencies will affect when calculating the UCP of those projects. Therefore, the total effort estimation of those software projects could vary from its real values. We have identified the erroneous diagrams by grouping the projects according to their domain and by comparing each project according to its UCP and the actual effort. Details of the collected OSS projects collected from freelance software developers are shown in Table I.

TABLE I. OPENSOURCE SOFTWARE PROJECTS OF FREELANCE SOFTWARE DEVELOPERS

No	Developer Experience	Language	Domain	UUCP	Effort (Hours)
1	17	Java	BA	247	201
2	19	Java	BA	304	265.5
3	21	Java	BA	258	264
4	25	PHP	BA	231	201
5	19	PHP	BA	216	209.5
6	23	PHP	BA	192	210
7	16	C	ES	177	224.5
8	19	C	ES	168	191
9	29	Python	ES	204	188
10	27	Python	ES	117	171
11	17	Python	SA	71	140
12	23	Python	SA	73	133
13	18	Pythons	SA	95	97.5

BA: Business Application, ES: Embedded systems, SA: Scientific Application

Since data values are on different scales, we standardize them using the equation given in (3) where X is the actual value, $X_{\min}$ is the minimum effort of the data set and $X_{\max}$ is the maximum effort of the data set.

$$X = (X - X_{\min}) / (X_{\max} - X_{\min}) \quad (3)$$

1) *Similarity Measure*: A similarity measure is used to calculate the similarity between projects based on how close the distance between projects according to the type of each attribute. There are many measurement techniques like Euclidean, Manhattan, and Minkowski that can be used to measure the distances between projects. Angel uses Euclidean distance to measure the distance (D) between two software project points in the plane with coordinates (x, y) and (x', y') is given by equation (4).

$$D = [(x - x')^2 + (y - y')^2]^{1/2} \quad (4)$$

2) *Number of Analogies*: The chosen analogy is defined by how many similar projects are used to compare to measure the effort of the OS software project. Basically, there are two ways to select analogies. They are fixed and dynamic analogy selection. This research adopts fixed analogy selection and suggests using one closest analogy ($K = 1$), two closest analogies $K = \{1, 2\}$, three closest analogies $K = \{1, 2, 3\}$. K can be defined through the following equation where N is the number of data elements. The equation (5) denotes the value of K .

$$K = (N)^{1/2} \quad (5)$$

3) *Analogy Adaptation*: The effort of the new project is calculated by using certain statistical techniques. There are four types of analogy adaptations that can be applied. They are the Closest Analogy (CA), the mean of closest analogies, the median of closest analogy, and IRWM of closest analogy.

4) *Analogy Adaptation Rules*: The effort estimation of a new project is done by dividing the actual effort of the old

project effort ($Effort_{old}$) with the size of old project ($Size_{old}$) then multiply with the size of new project ($Size_{new}$). Since the sizing factor of this model is UUCP, the equation (6) denoted the formulations of these adaptation rules [21].

$$Effort_{new} = (Effort_{old} / Size_{old}) \times Size_{new} \quad (6)$$

The n-fold Cross-Validation technique is one of the most used approaches by practitioners for model selection and error estimation of classifiers. The dataset is split into k number of subsets. Then, a portion of the split data set is used for testing while the rest of the data are used to learn the model. This is done iteratively [22]. Finally, by comparing the actual effort against the estimated effort, we validate our effort model using n-fold cross-validation. When a specific value for n is selected, it could be used in place of 'n' with the reference to the model, such as n = 3 becoming 3-fold cross-validation. The model is tested with a portion of data while holding the rest of the dataset as the training set. This process is iterated until the remaining groups as a training data set. Data sets are randomly divided into training data and testing data, with a percentage of 92% and 8% respectively. The accuracy is tested by using MMRE.

IV. RESULTS

We have tested the data with four analogy adaptation methods and the most accurate adaptation method is selected by comparing the MMRE of each method.

5) *Closest Analogy*: The closest analogy focuses on selecting from the closest project, which means the value of K=1. The estimated effort values for the closest analogy of each project with MRE values are shown in Table II.

6) *Mean of the Closest Analogy*: The mean of closest analogies is the adaptation analogy obtained by calculating the average effort driver from as many as K > 1 selected analogy. The estimated effort values for the mean of the closest analogies with MRE values are shown in Table III.

TABLE II. CLOSEST ANALOGY

Project No	Estimated Effort (hours)	Actual Effort (hours)	MRE
1	215.72	201	0.07323
2	311.07	265.5	0.17164
3	225.32	264	0.14649
4	252.66	201	0.25699
5	236.25	209.5	0.12768
6	167.06	210	0.20445
7	201.23	224.5	0.10364
8	213.08	191	0.11563
9	231.93	188	0.23366
10	213.16	171	0.24658
11	71.37	140	0.44671
12	73.38	133	0.44824
13	172.60	97.5	0.80739

TABLE III. MEAN OF THE CLOSEST ANALOGY

Project No	Estimated Effort (hours)	Actual Effort (hours)	MRE
1	252.74	201	0.25743
2	247.38	265.5	0.06823
3	209.95	264	0.20473
4	224.05	201	0.11467
5	175.77	209.5	0.16099
6	186.22	210	0.11323
7	177.93	224.5	0.20743
8	168.88	191	0.11579
9	371.67	188	0.97697
10	133.02	171	0.22212
11	129.36	140	0.00276
12	132.63	133	0.00275
13	173.08	97.5	0.81238

7) *Median of the Closest Analogy*: The median of the closest analogy is an adaptation analogy obtained by calculating the median effort driver from as many as K > 2 selected analogies. The estimated effort values for the median of the closest analogies with MRE values are shown in Table IV.

TABLE IV. MEDIAN OF THE CLOSEST ANALOGY

Project No	Estimated Effort (hours)	Actual Effort (hours)	MRE
1	223.01	201	0.10950
2	272.10	265.5	0.02486
3	218.04	264	0.17411
4	253.01	201	0.25877
5	214.30	209.5	0.02290
6	195.98	210	0.06674
7	206.51	224.5	0.080164
8	203.37	191	0.064739
9	410.32	188	1.18257
10	137.56	171	0.19553
11	128.83	140	0.00130
12	134.13	133	0.00851
13	196.78	97.5	1.06054

8) *Inverse Rank Weighted Mean of Closest Analogy*: IRWM is an adaptation analogy that gives the highest weight in the selected analogy most similar to other analogy. The estimated effort values for IRWM of the closest analogies with MRE values are shown in Table V. In this research the value of K is 3. Therefore, 3 of the closest analogies are selected, the first closest analogy (CA) is given a weight of three, the second closest analogy (SC) is given a weight of two, and the third closest analogy (TC) is given a weight of one [23]. The calculation of IRWM is formulated in equation (7).

$$IRWM = (3CA + 2SC + TC) / 6 \quad (7)$$

TABLE V. INVERSE RANK WEIGHTED MEAN

Project No	Estimated Effort (hours)	Actual Effort (hours)	MRE
1	225.54	201	0.12207
2	280.97	265.5	0.05827
3	219.12	264	0.17000
4	248.07	201	0.23416
5	215.19	209.5	0.02718
6	184.72	210	0.12039
7	199.98	224.5	0.10920
8	200.86	191	0.05161
9	344.42	188	0.83200
10	162.01	171	0.05259
11	109.77	140	0.14909
12	113.63	133	0.14561
13	184.77	97.5	0.93480

9) **MMRE Results:** The evaluation results of MMRE obtained by Euclidian distance had the lowest MMRE of 0.23131 with K = 3 using the IRWM. We have calculated the MMRE values for each type of analogy adaptations.

- Closest Analogy: 0.2602
- Mean of the Closest Analogy: 0.25073
- Median of the Closest Analogy: 0.25002
- Inverse Rank Weighted Mean: 0.23131

Closest analogy and the IRWM perform compared to other methods. IRWM adaptation provides the lowest MMRE value therefore it can be selected as the best adaptation method for this research.

V. CONCLUSION

There are no specific SEE models proposed to estimate the effort of OSS projects of freelance software developers. In this research, we have identified the effort drivers that affect the effort of an OSS project developed by freelance software developers. According to the literature, we have identified UUCP, developer experience, programming language, and application domain as the effort drivers of freelance OSS development. Accurately predicting the effort of software projects is very crucial for freelance developers. One of the most important issues of SEE is related to the estimation accuracy of the software projects. The accuracy of the estimated efforts is heavily reliant on the similarity of the historical data of the projects. Therefore, the similarity of the project to be estimated is very essential to improve the accuracy of the estimation. We selected the ANGEL SEE model, an analogy-based effort estimation model and we extended ANGEL to estimate the effort of OS freelance software development projects.

In this paper, we propose four analogy adaptation methods, namely the closest analogy, the mean of the closest analogies, the median of the closest analogies and IRWM.

The MMRE values of these four adaptation methods are respectively 0.2602, 0.25073, 0.25002 and 0.23131. Therefore, it is clear that IRWM performs better when adapting the analogies compared to the other 3 analogy adaptation methods. In this research, we have explored the viability of using UCP as the main effort driver and extension of the ANGEL model to estimate the effort of OS freelance software development.

REFERENCES

- [1] W.Scacchi., J. Feller, B. Fitzgerald, S. Hissam, and K. Lakhani, "Understanding Free/Open Source Software Development Processes", *Software Process--Improvement and Practice*, vol. 11, no. 2, pp95-105, March/April 2006.
- [2] A. Idri, F. A. Amazal, and A. Abran, "Analogy-based software development effort estimation: A systematic mapping and review," *Inf. Softw. Technol.*, vol. 58, pp. 206–230, 2015.
- [3] G. Karner. "Resource Estimation for Objectory Projects", *Objective Systems SF AB*, 1993.
- [4] G. Costagliola., F. Ferrucci, G. Tortora, and G. Vitiello, "Class point: an approach for the size estimation of object-oriented systems". *IEEE Transactions on Software Engineering*, no31(1), pp52 – 74, 2005.
- [5] C. Jones, *Applied Software Measurement: Assuring Productivity and Quality*: McGraw Hill, 1996 [
- [6] A. Abran, M. Maya, J. M. Desharnais and D. St-Pierre, "Adapting Function Points to Real-Time Software," *American Programmer*, vol. 10, 1997
- [7] M. Barlage, A. Born. and A. Witteloostuijn, "The needs of freelancers and the characteristics of 'gigs': Creating beneficial relations between freelancers and their hiring organizations", *Emerald Open Research*, August 2019.
- [8] M. Damodoran and A.N.E. Washington, "Estimation using Use Case Points", 2002.
- [9] M. Shepperd and C. Schofield, "Estimating software project effort using analogies", *Journal of IEEE Transaction on Software Engineering*, vol. 23, pp736–743, 1997.
- [10] J. Keung, "Software Development Cost Estimation Using Analogy: A Review", *Software Engineering Conference, ASWEC'09. Australian. IEEE*, 2009.
- [11] T. Mukhopadhyay, S. Vincinanza, M. J. Pietula., "Estimating the feasibility of a casebased reasoning model for software effort estimation", *MIS Quarterly*, 16:155, 1992.
- [12] E. Mendes, N. Mosley, S. Counsell, "Web effort estimation", *ICWE'07 Proceedings of the 7th international conference on Web engineering. Pages 90-104*, 2007.
- [13] K. Moløkken and M. Jorgensen, "A review of surveys on software effort estimation", *Empirical Software Engineering*, 2003. ISESE 2003.
- [14] S. Sridhar, "EXTENDED ANGEL: Knowledge-Based Approach For LOC And Effort Estimation For Multimedia Projects In Medical Domain", *International Journal of Software Engineering & Applications (IJSEA)*. vol. 2, no. 4, pp97-105, 2011.
- [15] P. Dan, and M. Korte, "Comparative Studies of the Model Evaluation Criteria MMRE and PRED in Software Cost Estimation Research", *Second International Symposium on Empirical Software Engineering and Measurement (ESEM 2008)*, 2008.
- [16] H.P. Patra, K. Rajnish, and U.S. Panda, "A New Software Cost Estimation model for Small Software Organizations: An Empirical Approach." *International Journal of Applied Engineering Research*. vol. 10, no. 15, pp. 36076-36082, 2015.
- [17] P. Mohagheghi, B. Anda, and R. Conradi., "Effort estimation of use cases for incremental large-scale software development," in *27th International Conference on Software Engineering. IEEE*, 2005.
- [18] A. Dubey, K. Abhinav, S. Taneja, G. Virdi, A. Dwarakanath, A. Kass and M.S. Kuriakose, "Dynamics of Software Development Crowdsourcing", *Proceedings of the 2016 IEEE 11th International Conference on Global Software Engineering (ICGSE)*, Irvine, CA, USA, 2–5 August 2016.

- [19] Z. Liao, B. Zhao, S. Liu, H. Jin., D. He, L. Yang,, Y. Zhang, J. Wu., "A Prediction Model of the Project Life-Span in Open Source Software Ecosystem", *Mobile Networks and Applications*, vol. 24, no. 4, 2019.
- [20] Dorn, J., A General Software Readability Model. University of Virginia, Charlottesville, Virginia. 2012.
- [21] A. Ardiansyah, M.M. Mardhia and S. Handayaningsih, "Analogy-based model for software project effort estimation", *International Journal of Advances in Intelligent Informatics*, vol 4,no. 3, November 2018.
- [22] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, "The 'K' in K-fold Cross Validation", *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2012.
- [23] Mendes, Cost Estimation Techniques for Web Projects. *IGI Global*, 2008.

A survey on applying machine learning to enhance trust in mobile adhoc networks

G. M. Jinarajadasa*

Department of Software Engineering
Faculty of Computing and Technology
University of Kelaniya, Sri Lanka
madhushikagihani@gmail.com

S. R. Liyange

Department of Software Engineering
Faculty of Computing and Technology
University of Kelaniya, Sri Lanka
sidath@kln.ac.lk

Abstract: Mobile ad-hoc networks (MANETs) play a vital role in the increasingly networked world where it has many applications in plenty of important fields including military sector, business applications, social networks such as Vehicular Networks (VANETs), and other intelligent systems. Because of the dynamic nature of mobile ad-hoc networks, they are more tent to be objected to the various malicious attacks. Over the recent past decades, a certain amount of researches has been done to increase reliable and trustworthy communications in a MANET environment. Over the proposed solutions, Machine learning applications have significant results. Hence based on those, a critical analysis of existing machine learning-based trust approaches for mobile ad-hoc networks are presented here. The focus of this survey is to classify and evaluate the existing trust mechanisms and to provide guidance for future research work in the area.

Keywords: Machine learning, Malicious attacks, Mobile Ad-hoc Network (MANET), Trust, Vehicular Ad-hoc Network (VANET)

I. INTRODUCTION

Because of the increased use of mobile devices, laptops, digital devices, and other IoT (Internet of Things) devices, wireless communication mechanisms have become popular throughout the recent past years. Mobile Ad-hoc networks are one of the most used class of wireless networks which have been taken extensively for the researches in the area over the recent years.

But the major problem in the MANETs is the security issue since it is having some challenges in data transmission because of its vulnerabilities. Because of having wireless links, limited physical protection, autonomous behavior, mobility and lack of centralized infrastructure, the security of MANETs is more challenging to acquire[1].

Also, this level of vulnerability varies upon the routing mechanism as well as with the mobility models. There are a set of mobility models defined and Random Waypoint Mobility Model is the most frequently used mobility model in MANETs[2].

Specific routing protocols are defined in the context of MANETs such as AODV (Ad hoc On-Demand Distance Vector), OLSR (Optimized Link State Routing), DSDV (Destination-Sequenced Distance-Vector), BATMAN(Better Approach To Mobile Ad hoc Networking), etc[3].

Although there are a lot of routing mechanisms, confirming the security of transferring data through a MANET, it is still a major problem due to the vulnerable nature of these networks where data can be misused or compromised. Since Trust is the major component that relies

upon when establishing reliable communications plenty of research works have been done considering the trust in the means of enhancing the security of MANETs.

This paper is a survey that gives a comparative explanation by a critical analysis upon the existing approaches for trust enhancement in MANETs and it especially focuses on the machine learning approaches and provides an idea for future research pathways.

Section II provides an idea about Trust in the context of MANETs and also gives some description of basic MANET challenges. The existing approaches as a whole are presented in section III and it describes guidelines for selecting the most suitable trust algorithm for MANETs based on trust properties. Existing Machine Learning approaches for MANET trust enhancement are described in section IV and provides differentiation between the machine learning techniques from different perspectives. Section V gives a summary of the presented work and the future directions for researches in the relevant area.



Fig. 1. Mobile Ad-hoc Network

II. TRUST IN MANETS

"Trust" concept is one of the most essential facts in networks when considering security services. In the context of MANETs, Trust can be declared as a node characteristic which can be a quantitative assessment of the reliability and accuracy of data received and transfer through a specific node in the perspective of the other network nodes [4][5]. Hence, a Trust Management mechanism is needed to establish security in a MANET environment by creating reliable communication links among the nodes. Establishing the trust within a MANET can be divided into several phases including calculation of trust, dissemination of trust, aggregation of trust, predicting trust and application of trust [4][5]. Figure 2 shows the primitive steps of the process of establishing trust in MANETs.

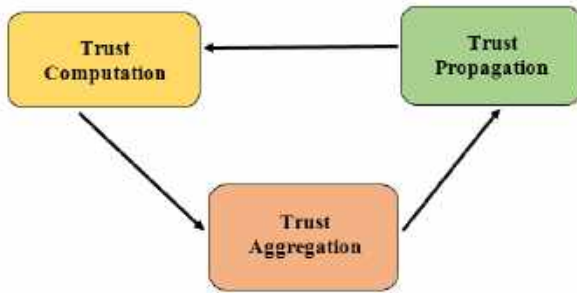


Fig. 2. Trust Establishment in a Network

A. Challenges in MANETs

Establishing trust or building security services in MANETs is more challenging since these ad hoc networks have a considerable amount of vulnerabilities than in the wired networks.

a) Wireless adhoc nature

MANETs do not have a central infrastructure or any other fixed infrastructure. Because of the lack of such infrastructure, nodes in a MANET play several roles including the actions of router, client and server. Also, the nodes have to communicate through multi-hops and data packets need to be travelling through different mobile nodes before arriving at their final destination. These may cause problems to arise like asymmetric links and the possibility of losing data by travelling through the malicious nodes [6][7].

b) Unreliability of wireless links

Since network connectivity is based on wireless links, MANETs are more susceptible to malicious attacks like eavesdropping, flooding, phishing, black-hole attack, wormhole attack, etc. Because of having lower bandwidth than wired networks and not wanting the physical access to make an attack, unlike in wired networks, attackers can easily exploit those features and make a severe threat to a mobile ad hoc environment [7].

c) Mobility and dynamic network topology

MANETs are consisting of a set of autonomous and highly mobile nodes where a new node/node can connect to the network constantly, and existing nodes can transit independently or leave the network. As a result of this, MANETs have a constantly changing network topology. Because of having such a dynamic environment, if some security issue occurs, it is difficult to identify the responsible nodes for such a malicious attack [6][7].

d) Energy and bandwidth restriction

MANETs have a severe resource constraints issue due to limitations in bandwidth, memory size, battery life and computational power. When considering a MANET, some or all of its nodes are relying on batteries or different exhaustible approaches for their energy. Those limitations make challenges in defining an assured routing protocol for a MANET [7][8].

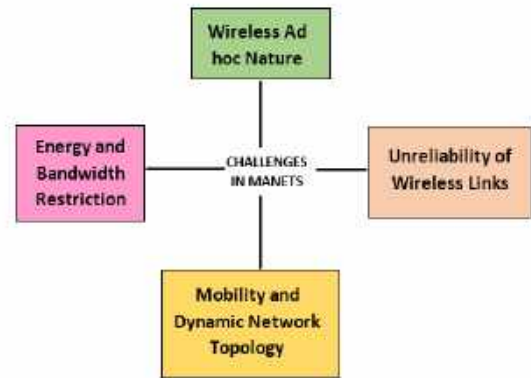


Fig. 3. Challenges in MANETs

B. Trust properties of MANETs

The following characteristics can be illustrated as the basic trust properties considering the unreliable and random nature of establishing communication links in MANETs.

a) Dynamicity

Due to constantly changing nature and the temptation of node failure MANET trust cannot be static, instead, it has to be dynamic which is changing accordingly [5][8].

b) Subjectivity

MANET node trust is subjective because of one node having different trust values concerning another neighbor node where it varies with the different experiences where the network topology is highly dynamic [5][8].

c) Not Inevitably Transitive

Trust in MANETs is not necessarily transitive which means having three nodes; node A, node B and node C where A trusts B and B trusts C; though it does not assure the expression A trusts C. To use the transitivity of trust, a node has to maintain 2 trust types which are direct trust and indirect trust or recommendation trust from the neighbor nodes [5][8].

d) Asymmetry

When considering the heterogeneous Mobile ad hoc network environments trust is unsymmetrical and not essentially reciprocal due to different capabilities such as a node with high computational power or battery capacity is not going to trust another node of low capabilities, but in a vice versa manner node which is having fewer capabilities is going to trust the same node with high capabilities [5][8].

e) Context dependency

MANET trust depends on context, which means different types of trust can be evaluated considering energy consumption, selfish behavior, computational power or efficiency, the occurrence frequency of malicious attacks, etc. [5][8].

III. SELECTING THE BEST ALGORITHM

When selecting the best algorithm, the above mentioned different challenges and trust properties have to be considered. Because of the physical distribution of the MANET information also have to be distributed among the nodes of the MANET. Thus the algorithms which can be distributed are

highly recommended for MANETs. The Distributed trust computing mechanisms for MANETs can be divided into 3 major types which are direct trust, indirect trust or recommendation based trust and hybrid trust[Fig.3].By considering that, plenty of researches is proposed in the area including the following approaches.

A. Swarm Intelligence

Swarm intelligence algorithms are a set of optimization algorithms that are created with the inspiration of nature; most of the time considering the behavior of social insects like ants, bees, etc. The "Antnet" and the "AntHocNet" are applications of swarm intelligence in MANETs where it utilizes the notion of Ant Colony Optimization(ACO) by finding near-best solutions to graph optimization problems[9]. Antnet and the AntHocNet are finding the near-optimal routes in a communication graph without global information. But the disadvantage of this approach is it creates additional communication overhead by the usual transferring of both 'forward ants' and the 'backward ants'[10][11]. Schoonderwoerd et al. have addressed the above-mentioned issue by proposing the solution called Ant-Based Control(ABC) which is very similar to Antnet where makes the communication overhead relatively smaller by using only the 'forward ants'[12].

Baras and Jiang have proposed a mechanism to manage the trust in MANETs in a twofold way which is, trust document distribution and distributed trust computation. They have proposed an approach called Ant Based Evidence Distribution(ABED) for trust document distribution schemes and random graph theory to evaluate new results in the context of distributed trust computation and establishment [13].

B. Probabilistic models

A fully distributed reputation system is proposed by Sonja et al. which used a modified 'Bayesian' approach and ratings of the trust are updated upon the compatibility of information taken by third party testimonies, with prior reputation ratings in context of neighborhood recommendations or watch of the third-party testimonies and also considering the CONFIDANT routing protocol [14] [15].

An information-theoretic framework for trust modelling and evaluation in the MANETs has been proposed by Yan Lindsay Sun et al. upon some axioms. Two types of trust models have been proposed by them, which are the entropy-based trust model and the probability-based trust model. Direct trust propagation is calculated directly in the entropy-based model and the integration and the multi-path trust dissemination are calculated in the probabilistic model using the probability values [16].

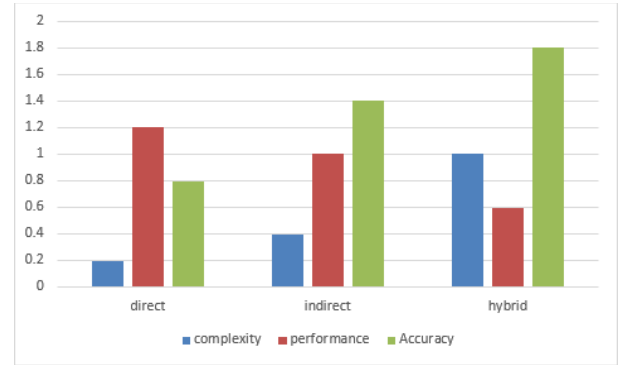


Fig. 4. Comparison of different distributed trust computing mechanisms

C. Mobile agents

In the context of networking, mobile agents can be referred to as small compositions of information or data packets that can migrate from one node to another node without any interaction with the environment. When considering the routing the applications of mobile agents such as Smart-Agents and the Ant-AODV approach updates the information about next hop or the path and find the optimal path by saving the state, transportation and resuming [17][18].

D. Machine learning approaches

Several Machine learning approaches have been proposed regarding enhancing the security of MANETs, including supervised learning, unsupervised learning, reinforcement learning and q learning. Q learning mechanism for routing enhancement has been proposed in the earliest research work in the area as well as a routing method called Q-map to secure multicast routing using the multi-agent reinforcement learning [19][20].

An approach that uses Team-partitioned, opaque-transition reinforcement learning (TPOT-RL) has been proposed for dynamic network routing which is dispersed and the implementation is not trivial. Though this achieved good results, it requires additional communication costs [21].

An implementation using Collaborative Reinforcement Learning(CRL) has been done to solve the issue of point-to-point routing in Mobile ad hoc networks [22]. A model called Trust and Q-learning based Security(TQS) is presented by Sivagurunathan et al. which can distinguish between trusted nodes and misbehaving nodes based on Q-learning [23].

A distributed reinforcement routing protocol called QLAODV (Q-Learning AODV) has been introduced to secure the information sharing in VANETs [24]. With the combination of Q-Learning and Fuzzy logic, a fuzzy constraint q learning algorithm is proposed over AODV(Ad-hoc On-Demand Distance Vector)routing protocol which is called PFQ-AODV [25].

Support Vector Machine(SVM) which is a supervised learning clustering mechanism has been applied in the area as intrusion detection agents, node clustering mechanisms and trust frameworks, which detects the misbehavior of the nodes [26][27][28].

E. Other theoretical models

Fuzzy logic-based trust computational model is presented by Ashish Kumar Jain and Vrinda Tokekar in order to enhance

the security of routing considering the AODV routing protocol. Also, they are presenting a mechanism to detect black-hole attacks in MANETs optimizing the fuzzy logic [29]. A dynamic model for predicting trust, which evaluates the trustworthiness of MANET nodes, has been proposed by Hui Xia et al. based on both historical behaviors and future behaviors where they used an extended fuzzy logic to predict future behavior. They have come up with a routing mechanism which is capable of finding the shortest path and by integrating the trust prediction model to that source routing mechanism they have introduced a novel routing protocol called Trust-based Source Routing protocol (TSR) which optimize the flexibility and feasibility of finding the shortest and optimal path [30].

Serdio et al. have used the watchdog method and the tool path-rater which runs on each node of the network where it can identify the misbehaving nodes and pick the most reliable path to transfer data [31]. CORE is a watchdog protocol-based reputation mechanism that identifies malicious nodes and enforces communication between nodes to prevent selfish behavior [32].

Further, a scalable maturity-based model for trust management which can provide recommendations about sharing data with the neighbor nodes has been proposed by Pedro et al. where they presented a protocol called Recommendation Exchange Protocol (REP) [33].

A route discovery protocol that prevents malicious attacks by providing accurate linking data has proposed by Panagiotis Papadimitratos and Zygumnt J. Haas [34].

A tool that is similar to Intrusion Detection Systems (IDS) which is separately situated on each node has been presented by the research work of A dynamic trust model for MANETs where it dynamically updates the trust value and improves the secure routing [35].

Further, ART and CAST are trusted management schemes for VANETs and context-aware trust framework for MANETs respectively [36] [37].

SEAD (Secure Efficient Ad hoc Distance Vector) is a novel routing protocol designed over the DSDV (Destination Sequenced Distance Vector) protocol that uses a one-way hash function to work against Denial-of-Service attacks and other misrouting issues [38].

Y. Shang et al have presented a heuristic search method for ad hoc networks. They have presented a distributed constraint-based routing approach which is consists of an efficient routing algorithm called CB-LRTA (Constraint-based Backpropagation LRTA) [39].

IV. MACHINE LEARNING APPLICATIONS IN MANETS

This section further describes the machine learning approaches applied to the trust enhancement of MANETs. Since mobile ad hoc networks are fully distributed systems when choosing the mechanisms also they should be distributed ones. From the above-presented paradigms Swarm Intelligence, Mobile Agents and Machine Learning Approaches are well suited to apply within MANETs. Moreover, Machine Learning algorithms are most suitable to apply within MANETs because of having the distributed manner, good prediction results, as well as the less computational power because of the collaborative manner of

its help to work as a multi-agent system or to only hold the relevant part of it within each node in the network.

When compared to other distributed techniques Machine learning approaches require a medium amount of memory and computational resources. Further, the accuracy of the results is high and the initial cost and the other additional costs can be vary depending on the type of reinforcement learning technique (Table I).

TABLE I. PROPERTIES OF DISTRIBUTED APPROACHES APPLIED TO MANETS

Property	Machine Learning	Swarm Intelligence	Mobile Agents	Heuristics
Memory requirement	Medium	Medium	Low	Medium
Computational requirement	Medium	Medium	Low	Low
Flexibility to topology	High	High	High	Medium
Accuracy of results	High	High	N/A	Medium
Initial cost	High/medium	High	Low	High
Additional cost	Low	Medium	Medium	Low

A. Supervised learning approaches

Support Vector Machines are one of the supervised learning algorithms which are mainly utilized in clustering and classification purposes as well as to other data mining tasks [40].

- An application of clustering which adapts Support vector machines to classify and cluster the mobile nodes in a mobile ad hoc network into two classes as a cluster head nodes and member nodes have been proposed earlier. The classification into two groups is done by making an SVM classifier to learn the results of the WCA (Weighted Clustering Algorithm). SVM reduces the time of clustering than WCA does, especially when the mobile nodes increase [27].
- SVM based two trust approaches called SMART and Sat have been proposed upon mobile ad hoc networks for trust automation and misbehavior detection respectively. These approaches have the advantage of outperforming than the previously proposed mechanisms in the area, handling a larger fraction of adversaries. Further, these approaches are more resilient when the nodes are highly mobile as well as they can detect malicious nodes that alter their behavior over time [28] [41].
- CEAP is a multi-decision intelligence detection model that complies with the highly mobile nature of VANETs (Vehicular Ad-hoc Networks). Vehicles are classified into smart vehicles and misbehavioral vehicles by analyzing the data exchange between two vehicles with the use of SVM and they have proposed a routing protocol called VANET-QoS OLSR over the OLSR (Optimized Link State Routing) protocol [42].

B. Q learning approaches

Q-learning is one of the best reinforcement learning algorithms in which a simple and powerful way for an agent to learn optimal actions in a controlled Markovian domain environment [43]. A set of Q-learning approaches have been proposed in the area of trust enhancement in the mobile ad hoc networks.

- Boyanand and Littman have proposed a Q-routing algorithm for packet routing by embedding the reinforcement module in each node of the network. They have tested the algorithm using a network that consists of 36 nodes and the particular algorithm can find the optimal routing policies without knowing the network topology or the traffic patterns previously [19].
- The Q-MAP algorithm is a mesh-based on-demand multicast routing scheme for mobile ad hoc networks with multi-agent Q-learning. It is well adapted to the dynamic topology and using distributed agent interactions it handles the scalability and tolerates the occurring of faults. Because of this, it gives accurate results by ensuring the reliability of the resource reservation in wireless ad hoc networks [20].
- An authentication mechanism based on Q learning has been introduced in early research work, called TQS (Trust and Q-learning based Security). The particular model detects the misbehaving nodes over the AODV (Ad hoc On-Demand Distance Vector) routing protocol by using their historical behavior of data transferring. TQS is especially focusing on detecting the black hole attack which helps to improve the trust calculation by giving immediate rewards [23].

C. Dual reinforcement learning

Dual reinforcement learning implies the concepts of DQNb (Deep Q Network) and double Q learning algorithms. This is as same as Q-learning and the reward function of dual reinforcement learning algorithms uses the optimal Q-values of the next state and the previous state as well. Although this is a bit complex than the Q-learning algorithms. This Double Q-learning reduces over-estimations by dividing the target's max operation into the selection of actions and evaluations [44]. Shailesh Kumar et al. has done some evaluation on DQR (Dual Reinforcement Q-Routing algorithm) and they identified that DRQ-Routing can sustain higher network loads than Q-Routing and non-adaptive-shortest-path routing [45].

D. TPOT-RL

Team-partitioned, opaque-transition reinforcement learning (TPOT-RL) is a distributed reinforcement learning technique that has been applied within multi-agent domains. This has been achieved good results in finding the best routing path, which is non-minimal but enhancing the security and the efficiency of data sharing in the network as a whole. The implementation of the TPOT-RL is not trivial but it requires additional communication cost [21].

E. Collaborative RL

'SAMPLE' is a Collaborative Reinforcement learning-based approach that has been applied in routing as a novel routing protocol. This is also originated upon Q-Learning and it presented a reinforcement learning algorithm which is designed to solve the point-to-point routing problem. SAMPLE optimize the routing behavior and adapts to dynamic behavior by learning the variations of the network properties from routing agents. Hence this is a Q-learning approach but the only difference is it is having a decay function which is very similar to the mechanism of the Ant Colony Optimization. This approach sustains low additional cost and tolerance to the network topology changes are high [22].

TABLE II. PROPERTIES OF MACHINE LEARNING APPROACHES APPLIED TO MANETS

ML Technique	Capturing Dynamicity	Additional Costs	Optimality of Results
SVM	Low	High	High
Q-Learning	High	High	High
TPOT-RL	High	Medium	High
Dual RL	High	Low	High
Collaborative RL	High	Low	High

According to the above information, it is identified that more than supervised learning and unsupervised learning, reinforcement learning approaches have been used commonly since its optimality of results is high as well as it is giving the accurate predictions because of its ability to capture the dynamic behavior instead of giving predictions based upon historical values. Moreover, it can be seen that Dual RL or DRQ routing is optimal in all aspects.

V. CONCLUSION

The presented work consisted of critical analysis about applying machine learning to enhance trust in MANETs. The paper itself demonstrating machine learning approaches and other existing approaches in the research area as well. When considering all the aspects it can be seen that machine learning mechanisms outperform other mechanisms. From the approaches apart from machine learning, Swarm intelligence seems good to be applied within the MANETs since it is flexible with the high mobility of the nodes. But when comes to energy consumption it is a bit difficult to apply those swarm intelligence methods due to the energy-restricted nature of MANETs. Therefore, Machine learning techniques are far better than swarm intelligence when considering all the possible issues. Among Machine learning approaches reinforcement learning gains more suitability for applied in mobile ad hoc networks since it gives more accurate results due to the ability to capturing the dynamic behavior easily as well as no need for historical data to give predictions where it can give predictions on newly joined network node also. Among those reinforcement learning techniques, Q-learning based Dual reinforcement learning gives more accurate results than the other techniques and it is the most optimal solution when considering the other facts instead of its complexity.

Hence, considering all this information the future research work should be launch in the area of machine learning; specifically, in the area of reinforcement learning according to the presented results of early work.

REFERENCES

- [1] D. Djenouri, L. Khelladi and A. Badache, "A survey of security issues in mobile ad hoc and sensor networks", IEEE Communications Surveys & Tutorials, vol. 7, no. 4, 2005, pp. 2-28. Available: 10.1109/comst.2005.1593277.
- [2] L. J. Madany, M. A. Madkour, and A. H. Al-Talhi, "Characteristics of mobility models for mobile ad hoc networks," 2009 IEEE International Conference on Signal and Image Processing Applications, 2009.
- [3] E. Royer and Chai-Keong Toh, "A review of current routing protocols for ad hoc mobile wireless networks", IEEE Personal Communications, vol. 6, no. 2, 1999, pp. 46-55. Available: 10.1109/98.760423.
- [4] K. Govindan and P. Mohapatra, "Trust Computations and Trust Dynamics in Mobile Adhoc Networks: A Survey", IEEE Communications Surveys & Tutorials, vol. 14, no. 2, 2012, pp. 279-298. Available: 10.1109/surv.2011.042711.00083.
- [5] R. Vijayan and N. Jeyanthi, "A survey of trust management in mobile ad hoc networks", Applied Engineering Research, vol. 11, no. 4, p. 2833-2838
- [6] A. Forster, "Machine Learning Techniques Applied to Wireless Ad-Hoc Networks: Guide and Survey," 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, 2007.
- [7] S. Gangwar, "Security threats in mobile ad hoc networks-A survey.", International Journal of Computer Science and Information Technologies, Vol 7, no.1, pp.74-77,2016.
- [8] Cho, A. Swami and I. Chen, "A Survey on Trust Management for Mobile Ad Hoc Networks", IEEE Communications Surveys & Tutorials, vol. 13, no. 4, 2011, pp. 562-583. Available: 10.1109/surv.2011.092110.00088.
- [9] M. Dorigo, M. Birattari and T. Stutzle, "Ant Colony Optimization", IEEE Computational Intelligence Magazine, vol. 1, no. 4, 2006, pp. 28-39. Available: 10.1109/ci-m.2006.248054.
- [10] G. Di Caro and M. Dorigo, "AntNet: Distributed Stigmergetic Control for Communications Networks", Journal of Artificial Intelligence Research, vol. 9, 1998, pp. 317-365. Available: 10.1613/jair.530.
- [11] G. Di Caro, F. Ducatelle and L. Gambardella, "AntHocNet: an adaptive nature-inspired algorithm for routing in mobile ad hoc networks", European Transactions on Telecommunications, vol. 16, no. 5, 2005, pp. 443-455. Available: 10.1002/ett.1062.
- [12] R. Schoonderwoerd, O. Holland, J. Bruten and L. Rothkrantz, "Ant-Based Load Balancing in Telecommunications Networks", Adaptive Behavior, vol. 5, no. 2, 1997, pp. 169-207. Available: 10.1177/105971239700500203.
- [13] J. Baras and T. Jiang, "Cooperative games, phase transitions on graphs and distributed trust in MANET," 2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601), 2004.
- [14] S. Buchegger and J. Le Boudec, "A robust reputation system for mobile ad-hoc networks", (No. REP_WORK), 2003.
- [15] S. Buchegger and J.-Y. L. Boudec, "Performance analysis of the CONFIDANT protocol," in Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing - MobiHoc 02, 2002.
- [16] Yan Lindsay Sun, Wei Yu, Zhu Han and K. Liu, "Information theoretic framework of trust modeling and evaluation for ad hoc networks", IEEE Journal on Selected Areas in Communications, vol. 24, no. 2, 2006, pp. 305-317. Available: 10.1109/jsac.2005.861389.
- [17] E. Bonabeau, S. Guérin, D. Snyers, P. Kuntz and G. Theraulaz, "Routing in telecommunications networks with ant-like agents", in International Workshop on Intelligent Agents for Telecommunication Applications, Springer, Berlin, Heidelberg, pp. 60-71.
- [18] S. Marwaha, C. K. Tham, and D. Srinivasan, "Mobile agents based routing protocol for mobile ad hoc networks," Global Telecommunications Conference, 2002. GLOBECOM 02. IEEE, 2002.
- [19] J.A. Boyan and M.L. Littman, "Packet routing in dynamically changing networks: A reinforcement learning approach," 1994 Advances in neural information processing systems, 1994, pp. 671-678.
- [20] R. Sun, S. Tatsumi, and G. Zhao, "Q-MAP: a novel multicast routing method in wireless ad hoc networks with multiagent reinforcement learning," 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOP 02. Proceedings., 2002.
- [21] P. Stone, "TPOT-RL applied to network routing", ICML ,pp. 935-942
- [22] Dowling, E. Curran, R. Cunningham and V. Cahill, "Using Feedback in Collaborative Reinforcement Learning to Adaptively Optimize MANET Routing", IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 35, no. 3, 2005, pp. 360-372. Available: 10.1109/tsmca.2005.846390.
- [23] S. S. P. K and T. A, "Authentication Using Trust to Detect Misbehaving Nodes in Mobile Ad hoc Networks Using Q-Learning", International Journal of Network Security & Its Applications, vol. 8, no. 3, 2016, pp. 47-64. Available: 10.5121/ijnsa.2016.8304.
- [24] C. Wu, K. Kumekawa and T. Kato, "Distributed Reinforcement Learning Approach for Vehicular Ad Hoc Networks", IEICE Transactions on Communications, vol. 93-, no. 6, 2010, pp. 1431-1442. Available: 10.1587/transcom.e93.b.1431.
- [25] C. Wu, S. Ohzahata and T. Kato, "Flexible, Portable, and Practicable Solution for Routing in VANETs: A Fuzzy Constraint Q-Learning Approach", IEEE Transactions on Vehicular Technology, vol. 62, no. 9, 2013pp. 4251-4263. Available: 10.1109/tvt.2013.2273945.
- [26] H. Deng, Q.-A. Zeng, and D. Agrawal, "SVM-based intrusion detection system for wireless ad hoc networks," 2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No.03CH37484), vol. Vol. 3, 2003, pp. 2147-2151.
- [27] H. Chen, R. Du, P. Li, and X. Li, "Clustering Application of SVM in Mobile Ad Hoc Network.", 2008 International Conference on Intelligent Computation Technology and Automation IEEE, 2008, pp. 924-926.
- [28] Li Wenjia, Anupam Joshi, Tim Finin, "SMART: An SVM-based Misbehavior Detection and Trust Management Framework for Mobile Ad hoc Networks", MILITARY COMMUNICATIONS CONFERENCE 2011-MILCOM, 2011, pp. 1102-1107.
- [29] A. Jain and V. Tokekar, "Security Enhancement of AODV Protocol using Fuzzy based Trust Computation in Mobile Ad Hoc Networks," Oriental journal of computer science and technology, vol. 10, no. 1, 2017, pp. 94-102
- [30] H. Xia, Z. Jia, X. Li, L. Ju and E.H.M. Sha, "Trust prediction and trust-based source routing in mobile ad hoc networks.", Ad Hoc Networks, vol 11, no. 7, 2013, pp.2096-2114.
- [31] S. Marti, T.J. Giuli, K. Lai and M. Baker, "Mitigating routing misbehavior in mobile ad hoc networks.", 6th annual international conference on Mobile computing and networking, ACM, 2000, pp. 255-265..
- [32] P. Michiardi and R. Molva, "Core: A Collaborative Reputation Mechanism to Enforce Node Cooperation in Mobile Ad Hoc Networks," Advanced Communications and Multimedia Security IFIP Advances in Information and Communication Technology, 2002, pp. 107-121.
- [33] P. Velloso, R. Laufer, D. O. Cunha, O. Duarte and G. Pujolle, "Trust management in mobile ad hoc networks using a scalable maturity-based model", IEEE Transactions on Network and Service Management, vol. 7, no. 3, 2010, pp. 172-185. Available: 10.1109/tnsm.2010.1009.i9p0339.
- [34] P. Papadimitratos and Z. Haas, "Secure message transmission in mobile ad hoc networks", Ad Hoc Networks, vol. 1, no. 1, 2003, pp. 193-209. Available: 10.1016/s1570-8705(03)00018-0.
- [35] Z. Liu, A. Joy, and R. Thompson, "A dynamic trust model for mobile ad hoc networks," in Proceedings of 10th IEEE International Workshop on Future Trends of Distributed Computing Systems, 2004. FTDCS 2004., 2004.
- [36] W. Li and H. Song, "ART: An Attack-Resistant Trust Management Scheme for Securing Vehicular Ad Hoc Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 4, 2016, pp. 960-969.
- [37] W. Li, A. Joshi, and T. Finin, "CAST: Context-Aware Security and Trust framework for Mobile Ad-hoc Networks using policies," Distributed and Parallel Databases, vol. 31, no. 2, 2012, pp. 353-376
- [38] Y.-C. Hu, D. B. Johnson, and A. Perrig, "SEAD: secure efficient distance vector routing for mobile wireless ad hoc networks," Ad Hoc Networks, vol. 1, no. 1, 2003, pp. 175-192.
- [39] Y. Shang, M.P. Fromherz, Y. Zhang and L.S. Crawford, "Constraint-based routing for ad-hoc networks.", Information Technology: Research and Education, in Proceedings of ITRE2003. International Conference, 2003, pp. 306-310.

- [40] L.Wang , Support vector machines: theory and applications ,Vol. 177, Springer Science and Business Media., 2005
- [41] S. Kumar and R. Miikkulainen, "Dual reinforcement Q-routing: An on-line adaptive routing algorithm.", Artificial neural networks in engineering Conference, 1997, pp. 231-238.
- [42] O. Wahab, A. Mourad, H. Otrouk and J. Bentahar, "CEAP: SVM-based intelligent detection model for clustered vehicular ad hoc networks", Expert Systems with Applications, vol. 50, 2016, pp. 40-54. Available: 10.1016/j.eswa.2015.12.006.
- [43] W. Li, A. Joshi, and T. Finin, "SAT: an SVM-based automated trust management system for Mobile Ad-hoc Networks," 2011 - MILCOM 2011 Military Communications Conference, 2011.
- [44] C. Watkins and P. Dayan, "Q-learning", Machine Learning, vol. 8, no. 3-4, 1992, pp. 279-292. Available: 10.1007/bf00992698.
- [45] H. Van Hasselt, A. Guez, and D.Silver, Deep Reinforcement Learning with Double Q-learning, AAAI, vol. 2, 2016, p. 5.

Predicting examination performance using machine learning approach: A case study of the Grade 5 scholarship examination in Sri Lanka

U. M. Ranawaka*

Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
umranawaka@gmail.com

Chathura Rajapakse

Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
chathura@kln.ac.lk

Abstract: Universal primary school education is a must requirement and one of the criteria that should be fulfilled by the developing countries according to the International development goals which are also recognized as “Eight Millennium Development Goals”. In the context of Sri Lanka, Government is mostly involved in primary education through government-controlled schools. The success of primary education is measured by conducting a scholarship examination. Those who are getting higher results are given opportunities to attend well-facilitated schools for secondary education. Due to that case, there is a massive competition for passing the examination. Limitless pressure for examination provides lots of issues to students. This paper uses data to investigate a model of academic performance as measured by past results of school tests of Grade 4 and Grade 5. 500 students from eight primary schools in the Gampaha district have been selected for collecting data. The Data on the above-mentioned students have been collected by conducting questionnaires to the teachers who in-charged the classes. Then the Logistic Regression model and Multiple Linear Regression model have been applied to predict students’ performances at the examination. The model depicts the likelihood of a student passing or failing the grade 5 scholarship examination and predicts the range of results that students will obtain in the examination. The accuracy of predictive models is measured using the results of students who have already faced the Grade 5 examination. Revealing the potential of students at the grade 5 examination is heavily benefited by teachers because they can provide personalized education for talented students and provide opportunities to other students to improve their talents. The initial architecture of the Grade 5 examination results’ predictive model is being discussed in this paper.

Keywords: Academic Performance Predicting, Logistic Regression, Machine Learning, Multiple Linear Regression

I. INTRODUCTION

This paper focuses on the capability of measuring academic performance to incorporate children’s primary education system with the possibility of reducing the pressure that applied to children’s minds and increasing the efficiency of primary education. The objective of this paper is to predict the results of the students before the examination and whether the student can pass or fail in the examination. Section (I) of this paper is an introduction about the current education in Sri Lanka, Grade 5 scholarship examination and Predictive analysis. Section (II) discusses deep insight into the current environment with related studies and identifies problems in current education environment. Section (III) considers about

how research happens and model is going to build. Data Collection, Data Preprocessing, Parameter Selection Method and proposed model are described here. The results and the accuracy of the model is discussed under the section (IV). Finally, in the section (V) it manifests the new implementations to the field along with the recommendations and improvements of the current model.

A. Education in Sri Lanka

Because of the free education approach (1947) and offering education through Sinhala, Tamil and English mediums to stand as the mechanism of training, Sri Lanka accomplished universal primary education in 1964 and eliminated the disparities in enrolment in education. This is a huge opportunity for talented students who are undergoing financial difficulties to stand out from the crowd as a well educated person and shape the future of Sri Lanka. Free education provides equity and equality of opportunity to every person. Social integration is influenced by education in terms of relevant resources in the field, opportunities, and consequences [1].

Schools in the Sri Lankan education system have been conducted and provisioned under five categories known as National, 1AB, 1C, Type 2 and Type 3 schools. Funds and facilities are provided differently for those school types. That reason brings some schools more popular than other schools. When schools become more popular, the demand for getting into school goes high.

Government spends nearly 4% of the Gross Domestic Product (GDP) on education. As a result, there is a significant development in literacy and school enrolment in both male and female students. Also, literacy rate is high in Sri Lanka (2012- Male 92.6% and Female 90.0%) [2].

B. Grade 5 scholarship examination

Primary education is one of the key stages in the Education Framework in Sri Lanka. Because, children between 5-10 years of age belong to primary education. Primary education is the first step of the whole education, first touch point and first impression of the society, teachers, and friends of a child. Primary education is the base of the children’s life and directly affects the growth and further education. If primary education becomes the worst nightmare of a child, that will affect the entire life of a student.

At the end of the primary education Department of Examinations conducts a scholarship examination to measure the education of students and provides a scholarship to go to

another school for further education. In order to obtain a scholarship, students should score better results than average results (more than 75%). There are a limited number of opportunities to enroll in well-facilitated schools, Therefore, Grade five scholarship examination is one of the highly competitive examinations in Sri Lanka.

There are arguments whether the Grade 5 examination is good or bad for a young student because it brings on undue pressure on them to obtain a good result. The President of Sri Lanka pointed out the pressure that students are facing at a young age because of exams [3]. The Ministry of Education suggests converting Grade 5 scholarship examination as non-compulsory [4]. Even though the Grade 5 examination brings difficulties to students, the government can't abolish it. Because it provides scholarships for talented students and there are no other alternatives to measure talent of students other than a written examination. Increasing the number of schools which have better facilities is the best solution. But countries like Sri Lanka will process many years to fulfil the target and cost a huge amount of money. Therefore, Education policy makers try to figure out alternate solutions for this problem.

C. Grade 5 examination at Gampaha District

Gampaha as one of the three districts in the western province, Sri Lanka consists of the second largest population in Sri Lanka. Colombo district is an adjacent district of Gampaha and both districts have many national schools and 1AB schools which have highest popularity among Sri Lankans like Royal College, Ananda College, Vishakha College, Devi Balika College and so on [5]. Students in Gampaha district have the opportunity to attend those schools by passing Grade 5 examination. Because of the popularity of those few best schools in Sri Lanka there are huge competition among students, parents of students as well as teachers and principals in primary schools. Parents want to attend their students to high rating schools because it improves their dignity in society. Teachers like to have more passed students in their class. It can measure the quality of primary teaching. So, there is a competition among teachers to be the best teacher who passed most students. Principals and other administrative staff consider school-wide passed students in exams. They want to be the best school who passed most students in Sri Lanka. It will increase the popularity of school and it will increase parents' demand to attend their child to that school for primary education. Therefore, all entities in the education of children have competition to pass the Grade 5 examination with good results.

Due to the competition, students must follow a tough schedule to get prepare for the examination. Other than schools there are private tuition classes, paper classes, and individual training to prepare students for examinations. When examination is near many parents go to supernatural beliefs to support their students in examination. Parents go to temples, Gods, Astrology services and follow so many worship methods to give blessing to their children in examination. Because the examination is difficult, a typical student prepares for Grade 5 examination since Grade 3, two years before examination.

Typical Students in Gampaha district try their best to get a good result, because there are lots of good schools near them. But the reality is every student can't pass the exam. Every

student doesn't have the same talent in academics. If teachers can identify students with other skills, teachers can prevent wasting time on unrealistic goals and guide them to improve their other skills. This will help to get rid of feelings that will hurt a student's mind like Toxic looser Syndrome, pressure, stress and depression. If not, students will lose their trust in their own talents. Students should be allowed to enjoy their childhood as much as possible.

D. Predictive analysis

Predictive models and analytics are the practices of extracting information from datasets to forecast future probabilities and outcomes. It is used to determine patterns and special behaviors in datasets in order to make decisions easily in future. Predictive analytics does not tell you what is going to happen within the future. Instead, it can give an idea of what may happen within the future with an appropriate level of reliability and includes what-if situations and risk assessment. The knowledge obtained through analysis can contribute to resolve many problems in the current system and improving the current system to a better level.

Predictive analysis in education and educational data is referred to as Educational Data Mining (EDM). EDM is defined as a process that transforms raw data generated by the education system into beneficial information that could be used by any relevant authority to take accurate actions [6].

II. LITERATURE REVIEW

A. Education in Sri Lanka

Education system of Sri Lanka is free for all students, but it is controlled by the Government. Therefore, there are both positives and negatives to may affect students' academic life. According to research paper [7], Sri Lanka had been able to register almost all the primary aged students into schools marking the net primary enrolment rate as 96%, and the primary completion rate as 95%. Gender equity also prevailed; male and female enrollment rates were equal at all educational levels. Unlike many countries in the world, Sri Lankan education hadn't gender discrimination, ethnic discrimination or other discriminations. Every child had equal rights to get education from schools. The teacher student ratio was 1:18.5 in 2010 in Sri Lankan education [8]. At the same time, the dropout rate had also been recorded as low. Among the students who started schooling in grade one, nearly 1.2% of pupils failed to complete grade 5 as the final year of primary level [9]. According to those facts, Sri Lankan education provides good opportunities to Sri Lankan students to get a quality education.

But child abuse and violence against in government schools are one of the key factors influenced student's performance negatively. Pressurizing students to pass examinations is a kind of abuse and violence of children's freedom. A circular had been issued by the Ministry of Education in restricting corporal punishments in schools, but still there were several incidents reported on the continuation of such violence [10]. According to a survey done by Business Times [11], it showed that "69.25% of Sri Lankans believed that country's education system was given the tough syllabus' and examinations and children were over-burdened with work leading up to the government examinations and private tuition had become essential rather than an option because of the present system (67.36%)".

Apart from the government public schools there are private tuition classes which give education by certified instructor, Undergraduate students or Advanced level students. In the research [12] pointed out that private education became an essential item to Sri Lankan economy. According to his survey, "More than 80% of students who study in grade 6 joined private tuition classes". In Sri Lankan education culture, private tuition is considered as an added advantage for examination. In that research paper said higher income and higher educational levels of the parents had been recorded as the factors which led to spend more on their children's' private tutoring. He also mentioned that the education level of parents had a positive relationship on household private tuition expenditure and employment status. A research paper [13] had observed the factors that led to the expenses of private tutoring and how it would affect the academic performance of the students in Vietnam. His results also suggested that there was an enhancement of the academic performance of the students because of the private classes.

Private tuition being a necessary item is not good for a country which has a free education system in government schools. Due to the competitiveness of examinations, Students were forced to participate in more additional classes to get more advantage in examinations. But private tuition classes might enhance the social differences and damage the free education policy [12]. The paper done by Daniel & Wang [14] demonstrated that school-provided after-school classes could reduce the need for private tuition and raise academic outcomes without harming student mental health in Korea's high school.

B. Grade 5 scholarship Examination

Grade 5 scholarship examination is introduced for Selecting highly potential and but underprivileged students for scholarships and to select the best students for the schools which are categorized as highly rated. Grade 5 examination is a great opportunity for children who have low-income families, because it gives opportunity to attend popular and well-facilitated schools and provide bursaries to the economically disadvantaged students.

Unfortunately, the Grade 5 scholarship examination has become a very competitive examination which has a negative influence on the students now. The Minister of Education pointed out that the Grade 5 scholarship examination caused several mental issues to students. A special parliamentary advisory committee for education had been appointed to investigate this matter. According to the educationalists and child psychologists, this exam had created an unwanted competition among the students without considering the variations in brain volume and growth in children. The time and energy of the students that could be used to explore much new knowledge had been limited by the examinations and cause to have schools which were not much productive, creative or innovative.

Grade 5 examination presently is a torment to kids, having supported unhealthy competition and over-ambitious parents. Mothers even more especially, got used to slave-drive their young children of Grades 4 and 5 to get more and more study. This was not so healthy for an impressionable kid. The students were forced to go to private tuition classes and to cram – even from the Grade one [15]. In order to pass the examination student should score at least 150 marks out of 200

(75%). In 2019, the Grade 5 Examination pass mark was 159 out of 200 (79%). If a student wanted to attend a good school with good facilities, he should score at least 170 marks (85%). In 2018 79.09% students obtained more than 70 marks and mean value was 117.98 out of 200 results in Grade 5 scholarship examination [16]. But the cutoff mark of Gampaha district for schools was 168 marks (84%). Therefore, many students who didn't get more than 168 marks couldn't go to a better school than their current school.

In Annual Performance Report [16] of the Ministry of Education said that ministry pays attention to find a way to minimize the pressure caused to the students due to making Grade 5 Scholarship examination a competitive examination. Also, the ministry paid attention to provide ample opportunities to the students of rural schools and the students of low-income families in examinations. Pressurizing students for grade 5 examination is a national level problem and many authorities in Sri Lanka has paid attention to mitigate that problem.

C. Educational data mining and Predictive models

Data mining is scanning for specific patterns within huge sets of data which makes a lot of possibilities for decision makers. By analyzing those patterns, better choices can be made in order to uplift the process of learning and assessment. To analyze patterns, various research papers use various data mining techniques.

In a research paper done by Mamcenko J. [17] summarized the level of student success in achieving the corresponding outcomes and objectives by incorporating clustering and association rules in data mining to analyze the data extracted from the electronic exam. This paper used unsupervised Machine Learning for analyzing. Researchers gave the easy exam questions and difficult exam questions to students and analyzed the relationship between time spent and correct answers during first pass. Using association techniques, he obtained students' knowledge level. In research paper done by Agasisti [18] used classification methods for their supervised machine learning for analyzing Schools and Higher educational data such as course scheduling, budgets, student transportation cost-benefit analysis and so on. In the journal Tanna [19] analyzed data using Decision Rules. Results were calculated from which a student could choose which stream and college he/she could select for further education based on Entrance Exam marks he/she had scored. This analysis used past inputs which enabled the user to get a more realistic result as compared to a system generating results based purely on the thresholds assigned.

Kumar V. [20] applied the CARM (Conversation Analytic Role-play Method) methods using a tool called Tanagra to improve university students' performance. Apriori algorithm was used to associate students' performances in general courses at the undergraduate and postgraduate levels, recognized associations, and then identified the factors that influence the students' opportunity to get success or improve, their failure, preferences, syllabus designing, and teaching and evaluation techniques. Baradwaj and Pal S. [21] performed a study using Bayesian classification on students from Awadh University in India. Generally, they applied the process through selecting three hundred students from several levels. The authors identified that the students' performance was strongly correlated with other elements, other than students'

effort such as students' daily habits, family income and different factors were mentioned in the paper. Kasih J. [22] suggested a model to predict students' final grade in the programming course. The prediction had been categorized in three options: "Extraordinary", "very satisfactory", and "satisfactory". A solution was granted based on the Apriori algorithm in order to understand the relationship between a programming course and other subjects' students took during the first four-semester of their study period. As a result, they discovered a high correlation between computer courses and math courses. Abdullah Z. [23] proposed a model to recognize appropriate programs for the students considering their interest rather than on availability. It was evident from the study that was conducted by using Computer Science students from Malaysia University, suggested that even though many students were learning computer science programs that were not under their field of interests.

G. Badr used a classification based on an association rules algorithm to develop a classifier to assess the student's performance in the programming course in research titled "Predicting Student's Performance in University Courses" [6]. These predictions would help the students to get to know how well they would perform in a course even before they register in it to avoid having to drop out. G. Badr tried to predict student performance in a programming course based on their performance in English and mathematics courses. He used two approaches. The first method used "students' grades in two English courses and two mathematics courses, which produced four rules with accuracy of 62.75%. The second method used only two English course results, which had accuracy of 67.33%". The results showed English courses had a significant effect on predicting performance in the programming course. In another research [24] said that integration of many noisy variables reduced accuracy of the prediction and would lead to overfitting of data.

There are many educational data mining applications in the world and many researchers' tries to improve the quality of education using data analysis.

D. Research Objectives

This paper introduces a model to predict performance at academics of students before the examination. Talented students in academics can be identified though this model. Teachers can give personalized education for talented students in academics. Also, teachers and parents can identify students who have other talents. Teachers can give essential education and give more time to develop their own skills like art, singing, sports skills and so on. Those students can become personalities that they deserved to be one day. Especially those students can spend their childhood very happily without unnecessary competition or pressure.

III. METHODOLOGY

Research approach of this development-oriented research could be defined under three major phases. In the first phase, there were interviews conducted with the respective authorities to understand the processes of the Sri Lankan education system. Also, a broad and extensive study of literature was done along with this. During the second phase, data from eight schools and 500 students were collected. This stratified sampling represented 1AB ("Maha vidyala"), Type 2 ("Kanishta vidyala") and Type 3 ("Primary") schools. Clustered sampling was used to select classes from a selected

school. In each school one or two classes selected as clustered sampling. In a class, every students' data was collected for research. If sampling method was applied in this step, there would be a discrimination of student talent. By collecting all students in class, best students, normal students and weak students could be selected without any bias. In the final phase, model was developed using those data, tested and validated with collected test sample.

A. Data Collection

Data was collected using a questionnaire that was given to teachers of students. This survey was done for teachers because teachers spent 30 hours per week with students and they had good knowledge in the academic environment of students. Teachers and parents nearly had equal knowledge of the academic talent of students. So, teachers were selected for the data collection. This questionnaire was based on research "The impact of school quality, socioeconomic factors, and child health on students' academic performance: evidence from Sri Lankan primary schools, 2011" done by Aturupane H., Glewwe P [7]. In that research, researchers used questionnaires to collect data of grade 4 students and their environment. With suitable modifications, the same questionnaire was used for this research.

Grade 5 students who sat for scholarship examination in 2019 were selected for the research. Their academic results in school tests, final grade and final marks were collected. Data consisted of 500 students who studied in eight different schools in Gampaha district. Both urban and rural schools were selected to reduce the biases of facilitated and non-facilitated school environments. Principals of those schools in Gampaha district were well-qualified, good experienced and they had sound knowledge in primary education. Principals provided maximum effort to give better education to students.

In selected schools, one or two classes randomly selected for data collection. There were well-experienced and well-talented teachers for Grade 5 and Grade 4 in primary sections in the context of Gampaha district. Teachers' experience played a major impact for students' education. Therefore, well-experienced teachers were selected for data collection in selected schools.

In the questionnaire, following data collected from selected entities.

- Gender
- Grade 4, 1st term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 4, 2nd term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 4, 3rd term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 5, 1st term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 5, 2nd term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)

These subjects are taught in every school at grade 5. The Grade 5 scholarship examination normally consisted of knowledge of Mathematics, Science (Environment), Sinhala and a little bit of English.

Other than the above data, teachers' experience and principals' experience were considered as control factors of research. In Gampaha district typically there were well-experienced teachers and principals for primary education who have more than 5 years of experience.

B. Data preprocessing

Data pre-processing has been performed to remove noise and to gain an accurate representation of data. This step is crucial in building the model since uncleaned data would affect the performance of the model. As the 1st step of the preprocessing, outliers of the dataset were removed. In the 2nd stage, uncompleted or unknown data entries were changed in the dataset. When teachers were filling out the questionnaire, they could mark some attributes that were unknown and left uncompleted. Mean value of selected subject had been used for treating uncompleted data entries.

C. Parameter selection method

There are 26 main attributes or columns in the data set. If all attributes were analyzed and used to create predictive models, it would reduce the performance of models and deliver overfit or underfit results. Therefore, it was vital to identify the most important features that would affect overall results of the model. Two methods were used to identify important parameters.

1. Principal Component Analysis (PCA)

PCA is a method of developing new variables (principal components), which are linear composites of the original variables. The values of principal components created by PCA were known as principal component scores (Table 1).

TABLE I. PRINCIPAL COMPONENT SCORES

Name of the Component	PCA values
PCA 1	0.71 (71%)
PCA 2	0.04 (04%)
PCA 3	0.03 (03%)
PCA 4	0.03 (03%)

Table 1 showed the first four PCA values out of many PCA values. PCA explained the variance of the dataset. PCA 1 held 71% of variance in the dataset. For this research first four PCA values were selected which hold 81% of variance.

1. Backward Elimination (Stepwise regression)

Backward Elimination is a method which includes beginning with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, removing the variable whose loss generates the most statistically insignificant deterioration of the model fit, and reiterating this process until no more variables can be removed without a statistically significant loss of fit. In this research P-values were calculated of each independent variable and compared with significance level. If the highest P-value was greater than the significance level, variable/feature was removed. This was an iterative process and the process happened until there was no highest P-values greater than significance level. After calculations, significant variables which affected results mostly were identified (Table II).

TABLE II. BACKWARD ELIMINATION RESULTS

Significance Level = 0.05	
Subject	P-Value
Grade 4, Term 2 – Sinhala	0.031
Grade 4, Term 2 – Mathematics	0.000
Grade 4, Term 3 – Science (Environment)	0.002
Grade 5, Term 1 – English	0.001
Grade 5, Term 1 – Mathematics	0.045
Grade 5, Term 2 – Sinhala	0.000
Grade 5, Term 2 – Mathematics	0.000
Grade 5, Term 2 – Science (Environment)	0.032

After filtering out above features/parameters were selected for further developments of model.

D. Model Selection and Development

All independent variables are quantitative data. Independent variables consist of term test results. Dependent variables have two categories.

- Grade 5 examination marks – Quantitative
- Grade 5 examination grade – Qualitative

Two separate methods were introduced in this research in order to predict both marks and grade.

1. Method 1 – Predicting Marks

In order to predict marks that students will get at the examination, Linear regression would be used. Linear regression is a linear approach to developing the relationship between a dependent variable and one or more independent variables. For more than one independent variable, the model is called multiple linear regression. Independent variables are features that are extracted from parameter selection methods.

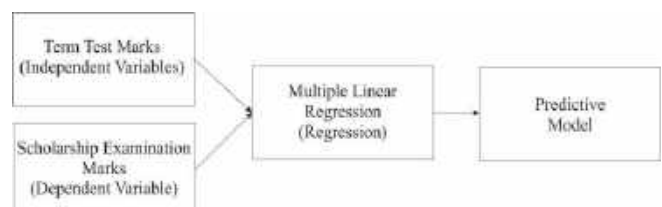


Fig 1 Predicting Marks – Method 1

The outcome of the model has predicted marks at given independent variable values.

2. Method 2 – Predicting Grade

In Grade 5 scholarship examination results consist of student grade at the examination. It is Pass the examination or Fail the examination. This is a dichotomous dependent variable. Logistic Regression is a supervised machine learning algorithm used in binary classification. Binomial Logistic regression was used to develop a model to predict whether the student would pass or fail at the upcoming examination.

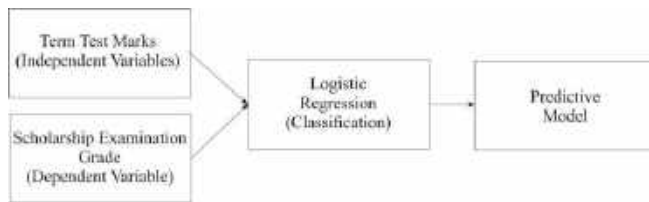


Fig 2 Predicting Grade – Method 2

The outcome of this model is whether students pass or fail at the examination. The threshold point was defined at 159 marks out of 200 marks. If students got below 159, they would be considered a failure. If students scored 159 or above, they would be considered as pass. Pass marks of the 2019 examination (159 marks) were used as binary classification.

IV. TESTING & VALIDATION

After developing the models, they should be tested and get an idea of how much applicability they derived and the performance of them in selected tasks. Accuracy of the model was a good measurement of the model's success in predicting student's performance at the exam.

In order to create a training data set and testing data set, the percentage split technique was used. Data set divided into 80% training data set and 20% testing data set. The model would train with training data and updated models' algorithms according to data. Then models' learning success is measured using testing data. All independent variables were scaled or standardized to minimize errors generated from high weighted values.

A. Method 1 – Predicting Marks

Multiple Linear regression was used to predict marks of Grade 5 scholarship examination. After successful training and testing data plots were developed to show the success of prediction (Figure 3).

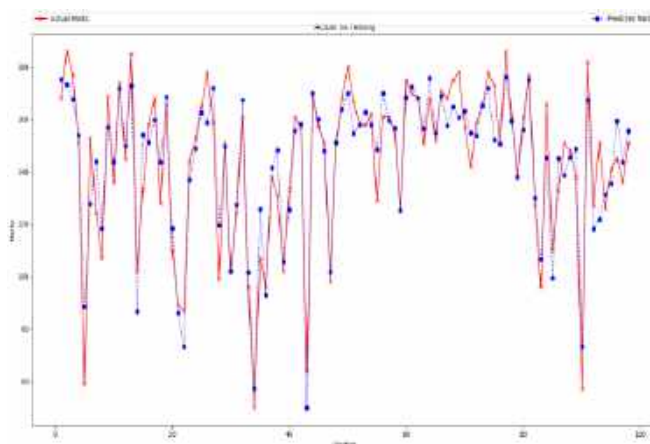


Fig 3 Multiple Linear Regression Results – Actual vs Predicted

Test data consists of 98 students and their actual marks and predicted marks which predicted using model shows in this graph. Solid line style and "x" marker are actual marks of test data set and dotted line style and dot marker are predicted marks predicted by the regression model.

TABLE III. STATISTICS

Intercept	22.90602
Root mean squared error (RMSE)	115.43913
Sum of Square of Residuals	11313.03512
R2 Score	0.87602

The mean absolute percentage error (MAPE) is used as a loss function for regression problems in machine learning. Using MAPE, Accuracy of the regression model can be extracted.

TABLE IV. ACCURACY OF REGRESSION MODEL

Total Error	6.33117
MAPE % (Mean Absolute Percentage Error)	6.46038
Accuracy %	93.53961

B. Method 2 – Predicting Grade

Predicting the grade that will be achieved by students was a classification problem. Therefore, logistic regression was used for developing this method. Confusion matrix can be used to calculate the accuracy of the model (Table V)

TABLE V. CONFUSION MATRIX

	Predicted Fail	Predicted Pass
Actual Fail	53	7
Actual Pass	6	32

Accuracy shows how the classifier model correctly predicted results. Error rate shows how often the prediction is wrong. Sensitivity shows when a student passes, how the classification model correctly identified it as passed. Specificity shows when a student fails, how the model identified it as fail. Precision shows the accuracy of the predicted pass. Prevalence shows How often the pass condition occurs in our dataset. (Table VI)

TABLE VI. ACCURACY OF CLASSIFICATION MODEL

Accuracy	0.86734
Error Rate	0.13265
Sensitivity / Recall	0.84210
Specificity	0.88333
Precision	0.82051
Prevalence	0.38775

A 10-fold cross-validation method used as a resampling method has been used to improve the accuracy of the model. The 10-fold cross-validation dataset was divided into 10 parts and tested each part with a model. The remaining nine parts would train the model. This process repeated for nine more iterations and more accurate results could be extracted. 10-fold cross-validation gave 10 accuracies for their 10 iterations. Table VII shows mean accuracy and Standard deviation of the resulting accuracy.

TABLE VII. 10-FOLD CROSS-VALIDATION RESULTS

Mean Accuracy	0.844079
Standard Deviation	0.047012

V. DISCUSSION

Techniques used in data mining allows the extraction of knowledge from student result data, presenting interesting

possibilities for the education domain. In this study, two models were created based on term test result variables gathered through a questionnaire approach. Some of the most influencing variables/factors were recognized and taken to predict the examination result and grade of a candidate. For predicting marks, Multiple Linear Regression was used (method 1). For predicting grade, the Logistic Regression method was used (method 2).

The Predicting model - method 1 achieved an accuracy of 94% for predicting student's marks. It shows that predicting model gives less error in prediction like the Mean Absolute Percentage Error is 6%. In simple terms, our prediction model can predict nearly correct marks

$$\text{Actual Marks} = \text{Predicted Marks} \pm 5 \text{ Marks}$$

The actual vs Predicted graph clearly shows that this equation is valid for most cases. Therefore, the obtained results from method 1 are good. Because predicted values are not much overfitting or underfit for actual values. In the educational environment, this model can suggest a range of results that will be taken by students. The range of results gave an idea about performance at the examination of students.

In the prediction model - method 2 which used to predict the grade of the examination had been achieved 84% accuracy using a 10-fold cross-validation method. For determining the grade, the threshold point/cut-off mark is defined as 159 marks. Using this model, a prediction can be done on whether a student can get more than 159 or not. Overall accuracy, Sensitivity, Specificity, and other measures are relatively good. But there is some noise in those models.

Even if those preliminary results derive satisfactory results, there is much room to improve accuracy, reduce the noise. Those two models were developed on 500 students' data. If there are more students' data accuracy of models will increase. Selecting sample schools all over the Gampaha district, increasing the number of sample schools and increasing the number of sample classrooms are good ideas to develop the model. As well as collecting data from many years is essential. In this research, all students are 2019 grade 5 students. It is essential to continue research for several years and collect students' data for several years. Those methods will increase the accuracy of prediction. Artificial Neural Network, Support vector machine, random forest algorithms can be used to develop separate models for these scenarios. Then the accuracy of all developed models can be compared and can figure out the most accurate model. Those models can be done fine-tuning to improve accuracy. As an example, changing classification algorithms, changing threshold values may increase the accuracy of the model.

It is better to consider social background, family background and mental background for developing predicting performance of the student. Parents' education level, parents' income, number of extra tuitions attended, extra-curricular activities and so many other factors have an impact on students' performance. As future work, those factors should be studied, and the prediction model should be recreated to make predictions more accurately.

Using educational data mining techniques and models developed in this research, teachers can get a sound knowledge of their students. Teachers can get an idea about

how teachers are going to improve students' skills effectively. This research doesn't mean to abandon students who don't have the academic talent and give focus only to talented students. But teachers should show the correct way to both talented and weak students in academics. Passing grade 5 examination is not the only way to create a bright future and get a better life. Teachers should encourage students to develop their talents rather than blindly studying for an examination and to be multi-talented students. With the help of these techniques, students can learn whatever they like and be talented happily.

REFERENCES

- [1] P. Colenso, *Education, and social cohesion: developing a framework for education sector reform in Sri Lanka*, vol. 35, 2005, pp. 411–428.
- [2] Research and Reports, "Sri Lankan Education" [Online], Available: www.unicef.org/infobycountry/sri_lanka_statistics.html.
- [3] Hiru news, "Government decides to abolish the grade 5 scholarship exam, says President" [Online], Available: <http://www.hirunews.lk/213156/government-decides-to-abolish-the-grade-5-scholarship-exam-says-president>.
- [4] I. Mudugamuwa, "Government issues circular making Grade 5 Scholarship Exam non-compulsory" [Online], Available: <http://www.dailynews.lk/2019/04/08/local/182549/government-issues-circular-making-grade-5-scholarship-exam-non-compulsory>.
- [5] Statical branch Ministry of education, *School Census Report_2017*, 2017.
- [6] G. Badra, A. Algobaila, H. Almutairia, M. Almuterya, *Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department*, 2016.
- [7] H. Aturupane, P. Glewwe, S. Wisniewski, The impact of school quality, socioeconomic factors, and child health on students' academic performance: evidence from Sri Lankan primary schools, 2011.
- [8] United Nations Development Programme, *Sri Lanka Human Development report*, 2012, pp.60.
- [9] World Bank, *Treasures of the education system in Sri Lanka: Restoring performance, expanding opportunities and enhancing prospects*, South Asia Region, Washington, DC, 2005.
- [10] S.Jayaweera, C.Gunawardena, *Social inclusion: gender and equity in education swaps in south asia sri lanka case study*, 2009.
- [11] Business Times, "Parents, administrators, civil society slam Lankan education system" [Online], Available: <http://www.sundaytimes.lk/110306/index.html>.
- [12] A. Pallegedara, Demand for private tuition classes under the free education policy. Evidence based on Sri Lanka, 2011.
- [13] H.A. Dang, *The determinants and impact of private tutoring classes in Vietnam. Economics of Education Review*, 2007, vol. 26, pp. 684–699.
- [14] C. Daniel. and L. C. Wang, The effect of after-school classes on private tuition, mental health, and academic outcomes: *evidence from Korea*, Monash University
- [15] I. M. K. Liyanage, *Education System of Sri Lanka: Strengths and Weaknesses*, 2014.
- [16] Ministry of Education Sri Lanka, *Annual Performance Report*, 2018.
- [17] J. Mamcenko, I. Sileikiene, J. Lieponiene, R. Kulvietiene, Vilnius, *Analysis Of E-Exam Data Using Data Mining Techniques*, Gediminas Technical University.
- [18] T. Agasisti, A.J. Bowers, *Data Analytics and Decision-Making in Education: Towards the Educational Data Scientist as a Key Actor in Schools and Higher Education Institutions*, 2017.
- [19] M. Tanna, Decision Support System for Admission in Engineering Colleges based on Entrance Exam Marks, *International Journal of Computer Applications*, August 2012, vol. 52, no. 11, pp. 0975 – 8887.
- [20] V. Kumar, A. Chadha, "Mining association rules in student's assessment data", in *2012 IJCSI International Journal of Computer Science*, vol 9, pp. 211-216.

- [21] B. Bhardwaj, and S. Pal, "Data Mining: A prediction for performance improvement using classification", in 2011 *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 9, no. 4, pp. 136-140.
- [22] J. Kasih, M. Ayub, and S. Susanto, "Predicting students' final passing results using the Apriori algorithm", in 2013 *World Transactions on Engineering and Technology Education*, vol. 11, no. 4, pp. 517-520.
- [23] Z. Abdullah, T. Herawan, N. Ahmad, and M. Deris, "Extracting highly positive association rules from students' enrollment data", in 2011, *Procedia - Social and Behavioral Sciences*, vol. 28, pp. 107-111.
- [24] Ransohoff, *Rules of evidence for cancer molecular marker discovery and validation. Nature Reviews/Cancer*, 2014, vol. 4, pp. 309-313

Intelligent changeover solution for a domestic hybrid power system

H. M. R. M. Herath*

Department of Electrical and
Telecommunications Engineering
South Eastern University of Sri Lanka
rroshanmadushanka@gmail.com

K. H. S. Tharindu

Department of Electrical and
Telecommunications Engineering
South Eastern University of Sri Lanka
supuntharindu93@gmail.com

K. D. M. Fernando

Department of Electrical and
Telecommunications Engineering
South Eastern University of Sri Lanka
dileepamaheshi@gmail.com

P. M. K. Alahakoon

Department of Electrical and Telecommunications Engineering
South Eastern University of Sri Lanka, Sri Lanka
mahindaalahakoon@yahoo.com

R. Hirshan

Department of Electrical and Telecommunications Engineering
South Eastern University of Sri Lanka, Sri Lanka
rajehirshan@gmail.com

Abstract: Electricity plays a major role in the modern world as almost all the equipment used is operated using electricity. Electricity demand of the world is increasing day by day and there should be a proper mechanism to meet the growing demand and to improve the efficiency of the power systems while continuously providing power with less environmental effects. This paper presents a changeover solution for a solar-grid hybrid power system that directly focuses on efficiently utilize the power sources by automatically selecting the power source according to the required power demand. Already available automatic power changeover switches in the market are only capable of selecting one source when the other source is not available. They cannot switch the power source considering the power demand. The novelty of this solution is, it can efficiently select the power source considering the apparent power demand of the house.

Keywords: Changeover switch, Hybrid system, Instant power supply, Microcontroller based changeover, Solar-powered house

I. INTRODUCTION

In the developing countries, it is a major issue to provide continuous power supply as the power generation is barely enough to satisfy the peak demand. It is very important to seek a sustainable solution as the electricity demand will only be increased. Nowadays, hybrid power systems can be considered as a better solution for the energy crisis. The hybrid power system is the most efficient system among other power systems which removes the weakness of one source by the other source.

Distributed power sources (on-site power sources) are becoming more popular due to technological innovations, fewer maintenance requirements, access to inexpensive energy resources, and other benefits. Among those renewables, solar power systems are now much more efficient and have become more affordable for residential and commercial customers. As solar energy can be harvested easily, and the system can be maintained easily it would be a better solution to use solar energy for household purposes. The solar panels can blend nicely with the roof, adding style and value to the home and it will not decrease the aesthetical value of the house. Solar energy is a better solution to minimize the electricity bill for most of the middle-class consumers in countries like Sri Lanka. Solar becomes more popular as it is

a truly renewable energy source and it is sustainable. Solar is an eco-friendly solution for electricity generation.

The main objective of the intelligent changeover solution for a hybrid power system is to increase the efficiency of the hybrid power system by automatically selecting the power sources according to the apparent power demand of the house. The specific objectives of the system can be classified as follows.

- Provide continuous power supply
- Reduce the number of units consumed by the grid
- Develop the introduced solution to be compatible with the existing power systems

II. LITERATURE REVIEW

As most of the industries require a continuous power supply for their operations, they use backup generators to ensure continuous power supply. If there is a failure in mains supply electricity, the changeover switch, switches the load to the auxiliary power supply, and return back to the mains supply when it is restored [1].

Although there are several types of changeover switches in the market, mainly it can be categorized as the manual or automatic. In a manual changeover switch, the user has to move a lever in the switch to change the power source. This is usually associated with time-wasting as well as it can happen hazards like an electric shock. In order to reduce the stress of manually switching and eliminates the delay, an automatic changeover switch was introduced. It ensures the smooth and immediate power changeover between the main supply and auxiliary supply by sensing the interruption in the main supply. There's no method to efficiently switch the power sources in the hybrid power system with considering power demand.

The hybrid power system is an integration of two or more conventional and non-conventional power sources. The main benefit of the hybrid system is it can mitigate the risk of power outages due to the diversity of power supply. Some examples of hybrid power systems are Solar and Wind, Solar and Grid, Solar and Generator, Solar, Wind and Generator, Grid and Wind, Grid and Generator, etc.

Solar and Wind hybrid system includes an aero generator and solar photovoltaic panels that offer several advantages over a single system. It can generate electricity during day time and also night time. When the sun is strong, the temperature is high and the wind is weak and vice versa. Solar and Wind hybrid system is suitable for city lights, communication base station, and monitoring systems and user's service in remote areas. Weaknesses of the Solar and Wind Hybrid are very complicated due to complicated control strategy, not efficient and not suitable for house premises [2]

[2] introduces an automatic power changeover switch that can automatically switch the alternative power source (generator or solar). If one supply is not available it will be switch to the other source and it will be switch back to the grid when the grid is restored. This is only a solution for power outages and it does not efficiently utilize power sources.

Basically, there are three types of solar power systems,

- Off-grid solar power system
- Grid-tied solar power system
- Hybrid solar power system

As the name implies, there's no connection between the off-grid solar system and the grid. The design of the off-grid solar system varies according to the load requirement of the place in which it is mounted. This system can be used as a solution to the power requirement of remote areas where grid transmission and other power sources are uneconomical. The main drawback of this system is batteries can only store a certain amount of energy. The people who are using off-grid solar power should change their power usage patterns to make off grid-solar more cost-effective solutions [3].

The grid-tied solar power system is a development model of a solar power system that is connected to the utility grid. When the system generates more power than the usage of the house, the surplus energy is sent back to the utility grid. Here power meter is used to measure the power flow in both directions (solar to grid and grid to the house). The electricity bill will be calculated only for the net electricity consumed or transferred during the total billing period. As grid-tied solar doesn't have storage facilities it is less expensive than the off-grid solar systems. The main weakness of this system is if there is a grid cut off, the system cannot be operated due to protection issues.

There were various methods propose to efficiently utilize grid-tied solar power systems. The controller-based grid-tied solar power system integrated with Instant Power Supply (IPS) ensures the continuous power supply through the Instant Power Supply module which provides power to the load when the main grid is not available. Here, solar energy has been considered as the priority and the battery will be charged using the power from solar. The surplus energy is coming from the solar provided to the grid, depending on the load variation. If solar power is not available, then it takes power from the grid. If grid supply is not available, it takes power from the instant power supply (IPS). People can get rid of load shedding because this system ensures continuous power flow[4].

Although this system ensures the continuity of the supply, some developing countries do not give provision for net metering[5].

Solar and grid hybrid system is a combination of off-grid solar and grid-tied solar systems. Hybrid solar systems generate power in the same way as a common grid tied solar system but use special hybrid inverters and batteries to store energy for later use. It has the capability to supply power during grid cutoffs. As there are two power sources, the weakness of one power source can be mitigated by the other power source. The solar-grid hybrid system is more efficient than the other power sources.

III. DESCRIPTION OF THE PROPOSED SYSTEM

The paper introduces a microcontroller-based changeover system to efficiently utilize the hybrid power system. The hybrid power system includes two power sources, grid supply, and solar power supply. According to the battery percentage, availability and apparent power demand of the house, changeover switch will switch to the best power supply at that moment. The introduced changeover system can be operated automatic mode or manual mode.

Automatic mode: Changeover switch will be automatically select the better source considering the power demand, battery level, availability. This solution will also operate as an instant power supply.

Manual mode: User can select the desired power source using a remote controller and he will be able to override the automatic mode. Instant power supply mode will also be operated in this mode.

The process of selecting the better power supply represent in the fig 1 flow diagram.

A. When the grid is available

If the battery level is more than 25%, the control unit will check the apparent power demand of the house. If it falls in between consideration range (higher than the lower limit and lower than the upper limit), the controller will turn on the inverter and supply solar power to the house. Otherwise, the controller will select the grid supply to the house.

The upper limit and lower limit differ from house to house. The lower limit value is chosen by considering the lower demand of the house when all the major electrical devices are turned off. By setting the lower value, the loss from keeping the inverter on for lower demand will be eliminated as the inverter will be turned off and the power will be taken from the grid. The upper limit value is chosen by considering the maximum output of the inverter. This value is set for the safety of the inverter.

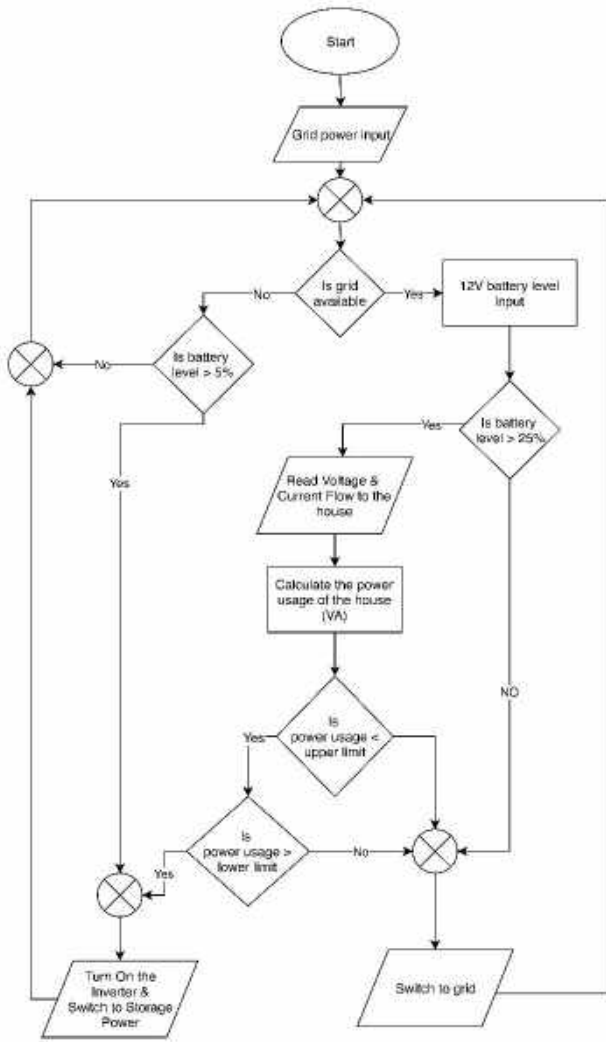


Fig. 1. Flow diagram of the unit

A separate port is given in the system to add external batteries to the system as per the user requirement.

B. When the grid is not available

When the grid supply is not available, the controller unit will select solar power (stored) to the house. If the user needs to connect a separate generator, it can be added to the generator connecting port in the system. If the power interruption from the grid lasts for a long time, an alarm will indicate the low battery level. Then the user can turn on the generator which is connected to the generator connecting port. The controller unit can switch the power source from solar to the generator without interrupting the supply to the house. Fig. 2 indicates the connection diagram of the entire unit.

- 1) **Solar panel:** The objective of the solar panel connecting to the system is to promote the green energy generation. The number of units generated by solar PV depends on the size of the solar panel, efficiency, weather, and seasonal variations. As per the diagram, the battery is charged by solar power. The solar regulator connected between the solar panel and battery in order to control the charging current of the battery. Battery capacity

should be adequate to store the daily harvested energy from the solar panels.

Daily energy harvested by solar panel

$$= W \times 3600 \times h \quad (1)$$

Total battery storage capacity

$$= E \times 12 \times 3600 \quad (2)$$

$$W \times 3600 \times h = E \times 12 \times 3600$$

$$W = 12 E / h \quad (3)$$

Where,

W = Wattage of panels (W)

h = Average solar energy harvesting hours per day (h)

E = Battery capacity (Ah)

This introduced unit offers three packages to the customer. Customers can select an appropriate package according to the power requirement of the house.

TABLE I. SPECIFICATION OF THE CUSTOMER SELECTION

Package	Battery capacity (Ah)	Wattage of solar panels (W)
Standard	90	200
Custom 1	180	400
Custom 2	210	460

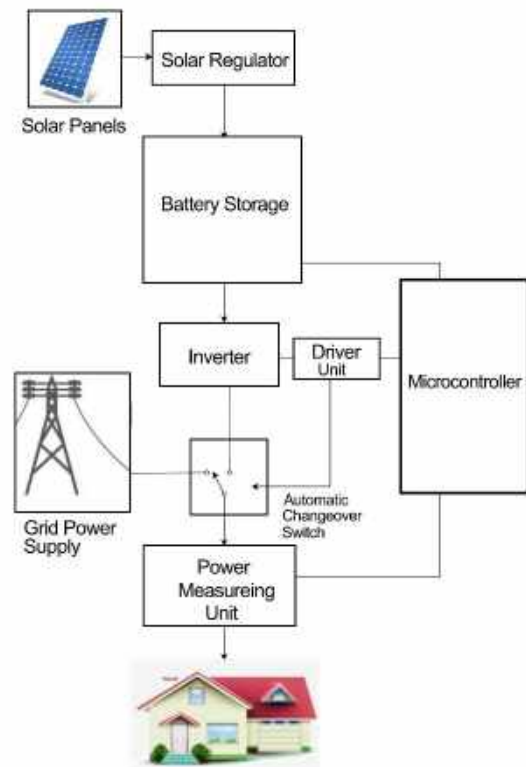


Fig. 2. Connection diagram

- 2) *Battery*: The system is tested by 12V, 90Ah lead-acid battery.
- 3) *Inverter*: The function of the inverter is to convert 12VDC into 230VAC. Because all the household appliances used in Sri Lanka operate at 230VAC.
- 4) *Power measuring unit*: The function of the power measuring unit continuously measures the current and voltage of the consumer end in order to calculate the apparent power demand of the house. The current and voltage measurements of the power measuring unit continuously feed to the microcontroller.
- 5) *Microcontroller*: The current and voltage of the house and battery voltage feed to the controller unit continuously and the controller unit calculates the battery percentage and apparent power demand of the house. Then the controller unit will analyze the apparent power demand of the house and the battery level to choose the best source for the home at that moment.
- 6) *Driver unit*: The function of the driver unit is to make the interface between microcontroller and changeover switch.
- 7) *Universal power adapter*: The introduced changeover switch can easily install to an already wired house without changing the wiring of the house through a universal adapter.

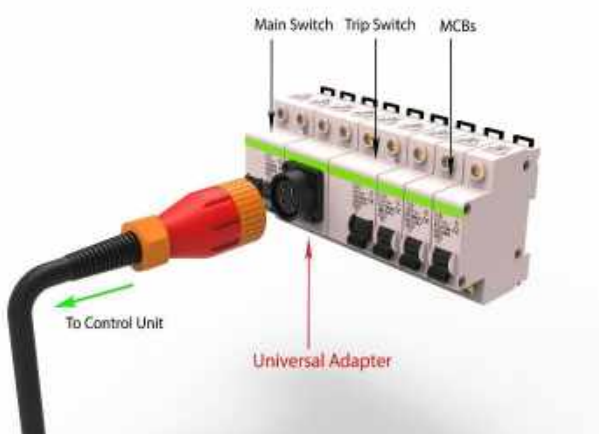


Fig. 3. Universal power adapter

C. Universal adapter installation to an already wired house

Fig. 4 shows the wiring diagram of the normal house. The universal adapter can be installed between the main switch and trip switch as shown in fig 5.

IV. IMPLEMENTATION

Considering the implemented separate units, they have to be installed as one unit to be attractive and increase the compatibility. Lead-acid battery should be installed within the system and that consumed considerable space in the final hardware unit. A separate transformer and an inverter also mounted in the unit. The casing of the unit should withstand for the weight of the component and for the easy installation process, separate adapter ports were mounted on top of the unit to connect the solar panel and the Grid supply. The separate transformer was mounted on the circuit board to

identify the grid supply is available. For the microcontroller and for the transistor switching circuits, separate 12V and 5V inputs were used. For the changeover purpose 12V, 5 pin relay was used that could withstand up to 50A current.

For the demand measurement, Hall-Effect current sensor was used and calibrated for 750W load. Using the current sensor and voltage sensor, the microcontroller will decide to change the power source to the best available power source to the unit. That will increase the battery lifetime and act as an instant power supply. Then the consumer will not face power outages due to the Grid Supply. As the solar panel is always connected to the unit, the 12V battery will be charged using solar power and all switching processes will be done as the system will automatically change the power supplying source, also the user could identify the mode of operation which the system has selected. For the separate modes of operation, three colored indicator light was mounted and programmed to display separate colors (RED, GREEN, BLUE) for each operating mode. The user could easily identify the working status by the indication color even he is far from the unit.

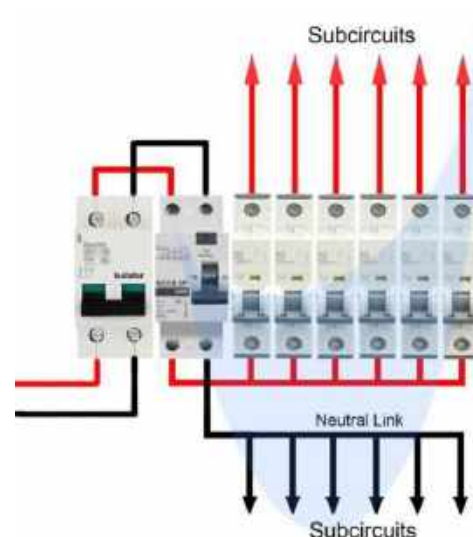


Fig. 4. DB Wiring layout before the installation

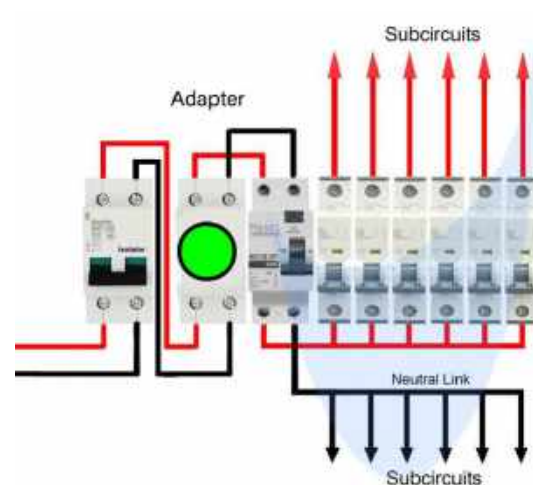


Fig. 5. DB Wiring layout after the installation

The switching mechanism is done using a relay between two power sources. The separate transformer has to be

installed to supply the 12V requirement for the controlling circuit. Live wire for the consumer is passed through a current sensor and the current value is checked using Arduino microcontroller. As the power is proportional to the magnitude of the current, the power sources can be changed considering the magnitude of the current (A) that flows to the consumer end. As the circuit has to be mounted in one unit, the pieces of equipment are assembled as follows.

After the assembling of the circuit, the unit was mounted on a plastic box as follows and the adapter ports, display, indicator light were mounted on the top surface of the box. 4-pin adapter was used to bring the power from the distribution board and feed them back to the Distribution Board (DB), after adding the proposed solution. A separate 2-wired port was added to connect a solar panel to the system and that can be selected by the user as the requirement and the cost. Apart from the above circuit components, the receiver for the remote control operation was connected with the micro-controller for the remote control operation.



Fig. 6. Main circuit board

TABLE II. IDENTIFICATION OF CURRENT STATE

State	Indicator color
When grid supply is available and working on the grid	Red
When grid supply is available and working on solar power	Blue
When grid supply is not available and working on solar power	Green

This solution could be operated using a remote controller. The supply will not be disturbed as the relay will be switched to the available power source when one source is not available. Users can only turn on or turn off the inverter only if the grid supply is available. The selected power source will be displayed on the LCD display which mounted on the unit. That LCD display will indicate the following operating details.

- Battery voltage
- Battery charging percentage
- Currently operating power source
- AC voltage
- AC current

The user can easily understand the operating status and the battery status from the LCD display. If the Grid Supply is not available for a longer period that could drain the battery, an alarm will be activated from a buzzer that was connected to the inverter. The user can identify the situation and connect a separate power source (Generator) if needed. The alarm will be automatically activated when the battery percentage is 20 % of the full capacity.

All the equipment was assembled into a plastic box (30cm*40cm*50cm) and installed the battery into the bottom half of the box. The introduced solution could be placed near the Distribution Board and connect to the house wiring system using the universal adapter that is placed in between the main switch and the RCCB. As the equipment is connected before RCCB, no safety issues will be there.



Fig. 7. The LCD display and indicator connection after installation



Fig. 8. Assembled complete unit (Included 90Ah Battery)

V. RESULTS

All the measurements were taken considering the standard package that consists of (200W solar panel, 90Ah lead-acid battery and 1kW inverter. The usable power supply as an instant power will depend on the required power

consumption. For the required consumption, usable hours can be calculated using the following equation.

$$W \times h = 90Ah \times 12V \quad (4)$$

Where,

W = Required power consumption

h = Usable hours

The changeover delay time was measured and observed that there were two delay times for two different conditions.

- Changeover delay when both supplies are available - 150ms
- Changeover delay when sudden disturbance to the grid - 200ms

VI. CONCLUSION

In the introduced system "Changeover Solution for a Domestic Hybrid Power System", two different power sources (grid supply and solar power supply) were considered and selected better power source accordingly considering the demand for a house. The best available option was selected considering the output current and the battery voltage from the 12V battery and the availability of the power sources. This can be used as an instant power supply. The user will not be disturbed by small power outages from the grid. As solar energy is used as a power source, the monthly electricity bill will also be reduced after the installation of the unit. This is a compatible upgrade for a house wiring system and the introduced unit should be installed as a separate unit using an adapter between main switch and RCCB in the distribution panel. The unit was tested using different loads up to 1kW. The load capacity depends on the capacity of the inverter. The introduced unit saves battery life as it does not allow large loads on the battery at normal conditions. Maintenance of the unit is also low and the system can be isolated from the domestic power system easier than the available power solutions. This unit can be remotely operated using a remote controller and override the manual operation.

REFERENCES

- [1] J. G. Kolo, "Design and construction of an automatic power changeover switch," *Assumption University Journal of Technology*, vol. 11, no. 2, pp. 1-6, 2007.
- [2] S. Durrani, H. Ali, A. Zahir, S. Khan, "Prototype Designing of Hybrid Power System for Industry with efficient design of Solar Power System," in *2018 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET)*, Islamabad, Pakistan, 2018.
- [3] R. K. Akikur, R. Saidur, H.W. Ping, K.R. Ullah, "Comparative study of stand-alone and hybrid solar energy systems suitable for off-grid rural electrification," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 738-752, November 2013.
- [4] M. Billah, S. K. Das, M. T. Islam, M. A. Haque, B. B. Pathik, "Design, simulation and implementation of a grid tied solar power controller integrated with instant power supply technology," in *2015 IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA)*, Bangkok, 2015.
- [5] K. Neupane, T. M. Undeland and A. Rouniyar, "Smart controller design for solar-grid hybrid system: Microcontroller based automatic

adjustable discretized solar-grid integration system," in *2014 International Conference on Intelligent Green Building and Smart Grid (IGBSG)*, Taipei, 2014.

Novel computational approaches for border irregularity prediction to detect melanoma in skin lesions

D. V. D. S. Abeysinghe
Department of Mathematics
Eastern University, Sri Lanka
dharaabeysinghe@gmail.com

S. Sotheeswaran*
Department of Mathematics
Eastern University, Sri Lanka
sotheeswarans@esn.ac.lk

Abstract: Medical image detection has been a rapidly growing field of study during the last few years. There are different challenges associated with it. Many works have been done in order to provide solutions for key challenges. This study of work is focused on melanoma detection by using Asymmetry, Border irregularity, Colour textures, and Diameter (ABCD) feature along with proposing two new approaches for border irregularity detection. The proposed two new approaches are distance difference method and gradient method, which follows the main concept as traversing along the continuous borderline of the lesion. Further, this study varies from the existing studies, since it has been taken counts of distances from the centroid to the borderline without considering the distance from the image border to the borderline of the lesion. It was able to achieve a classification rate of 79% and 78.5% using distance difference method and gradient method, respectively whereas the classification without the border irregularity feature achieved 78% of accuracy performing on PH² dataset. Further, this study can be stated as most appropriate to classify non-melanoma rather than melanoma. It is contributed by generating simple computer science-based approaches rather than complex mathematical methods to detect border irregularity and makes the medical image detection easy.

Keywords: ABCD features, Border irregularity, Distance difference method, Gradient method, Medical image detection, Melanoma and non-melanoma classification.

I. INTRODUCTION

Medical image detection or the task of identifying medical-based objects in a medical image is an important analysis to perform since medical images contain different characteristics rather than normal images. Many key challenges with regards to medical image detection exist such as preprocessing, segmentation, detection and extraction of features, classification into classes and finally to make crucial decisions about the prediction as shown in Fig. 1. These challenges can be highly impacting the dermoscopic images used in skin cancer detection when determining the region of interest and also in the extraction of features to predict skin cancers. Cancers are one of the most common diseases in the world that can develop almost everywhere in the body and over a hundred and more types of cancers skin cancers are a type that can be prevented and treated if found early. Among skin cancers, melanoma also known as malignant melanoma begins in cells called melanocytes that are in the upper layer of the skin which are responsible to produce a pigment called melanin to give the color to the skin. Some dermoscopic images of melanoma and non-melanoma are shown in Fig. 2.

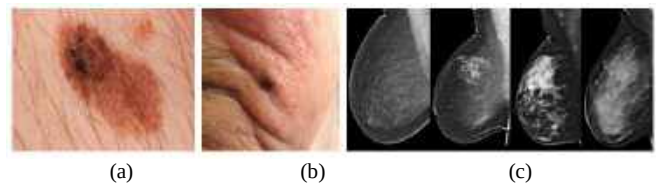


Fig 1. Challenges to medical image detection: (a) Skin cancer image with hair [1] (b) Dermoscopic image with wrinkles [2] and (c) Scans of Breasts with different densities [3].

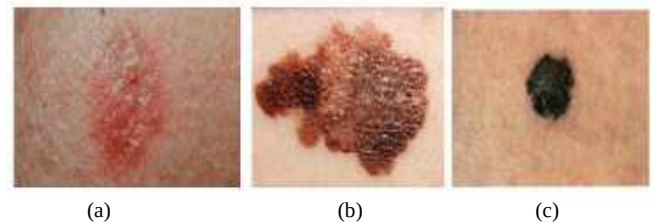


Fig 2. Melanoma and non-melanoma skin cancers : (a) Non-melanoma [4] (b) melanoma [5] and (c) melanoma [6] skin cancers.

In order to cure melanoma, it is an essential factor to enhance the prediction of skin cancers. Hence, to enhance the diagnosis of melanoma dermatologists tend to aid dermoscopy techniques where removal of unnecessary features of skin and get an improved visual effect of skin to produce more related details [7]. Melanoma looks tricky to identify at once due to the complexities in shapes, colors, size, etc. However, constant skin examinations may lead to earlier and correct detection. When focusing on the detection of melanoma, several criteria are available to identify a lesion as cancerous. Ugly duckling sign is one such property to locate a spot that is different from all other spots in the body. Further in dermoscopy, pigment networks, streaks, blue-whitish veil, regression structures, dots and globules, vascular structures, pink shades and starburst pattern etc. are used in the prediction of melanoma. ABCDE features or Asymmetry, Border irregularity, Colour textures, Diameter and Evolution of a mole are the most prominent set of features among those features. There can exist moles having one or two above properties or all of the above in it.

Hence this study has focused on improving the detection of melanoma, a quite common type of skin cancer, by providing two new approaches for border irregularity detection which is a feature among the set of geometric features, ABCD to predict melanoma.

The rest of this paper is organized as follows. In Section 2, we summarise different techniques that are closely related to

melanoma detection. Section 3 provides the background of skin cancer detection. In Section 4, the proposed techniques are described in detail. Section 5 describes the experimental setup. Section 6 discusses the results of the experiment. Finally, Section 7 concludes the paper with a discussion of the findings towards future extensions.

II. LITERATURE REVIEW

When considering skin cancer detection there are ample studies done related to it in the last decade. This Literature review is focused on the melanoma detection and the border irregularity skin cancer detection of medical images.

In [8], authors have proposed a computer-aided method for melanoma detection using Image Processing tools with ABCD parameters. Most importantly the proposed method pre-processes the images to enhance image quality. Then threshold values were generated by an automatic thresholding process and a binary mask has been created using the threshold values to locate the biggest blob to locate the lesion. Hence along with the location and with edge detection, the lesion has been segmented. Then standard geometry features such as area, perimeter, greatest diameter, circularity index and irregularity index were extracted and classification has been done to produce effective results. Overall this system has been concluded to be a user-friendly and robust system for images under any condition and diagnosis can be expected to be in a higher degree of accuracy.

In [9], authors have proposed a method to measure border irregularity of a skin lesion by locating indentations and protrusions along the lesion border and by creating a new area-based index called irregularity index for each indentation and protrusion. Initially, the scale-space has been extended from binary to a three-valued image and all the images have been preprocessed to remove hair using DullRazor and borders have been detected automatically. Then Gaussian smoothing has been performed and at every smoothing scale, the zero-crossings of the first derivative have been recorded. Finally, the sum of all the irregularity index values has been calculated for the overall prediction. Through this study, it has been able to pinpoint the problem areas of the lesions and also this method can be further used for other medical applications even.

In [10], authors have proposed an approach to detect and classify border irregularity of a skin lesion. This proposed method is consisting of four major steps as image enhancement, lesion segmentation, border irregularity detection and classification. Mainly, in the detection of border irregularity, the proposed study has implemented a four-step algorithm. In the algorithm, the first step computes the bounding box of the lesion. Then the boundary pixels upon the lines joining the center of mass with the vertices have been found. Thereafter, the distance between the border and the image edge has been calculated by soothing ragged edges, if there exists any using the Gaussian filter. Finally, to detect the border irregularity the derivative has been calculated to find the local maximums of the function by detecting the sign changes across the zero point. Then authors have performed a border irregularity classification by the overall score obtained from the scores of each segment of eight segments. This study has used 300 images taken from the Interactive Atlas of Dermoscopic image repository and

has achieved a detection accuracy of 79% and a classification accuracy of 90%.

In [11], the author has suggested a method based on lesion rotation and borderline division. Initially, black frame removal, hair detection and hair inpainting have been done to enhance the images. Secondly, the lesion has been segmented using a seeded region-growing algorithm. With the region growing techniques, the proposed method has been able to detect borders correctly even where edges are noisy. Then for the border irregularity assessment, the skin lesion has been rotated keeping the major axis of the lesion parallel to the horizontal line. Since happening a spatial transformation, one set of pixels are remapped to another. According to the proposed method, if the images are continuous then the remapping will not use interpolation, whereas if the images are discrete remapping requires interpolation. Also, for the process of interpolation, a bilinear method has been chosen after comparing it with the nearest neighbor method. After rotating the skin lesion borderline has been detected precisely and then the borderline function has been calculated and boundary pixels on the lines joining the center of mass with the vertices were found. With the use of those four points, the boundary has been divided into four unequal parts. Then the distance between the border and the image edge for each boundary part has been calculated. Further, the gap between the functions which got the results from the distance to the edges has been subtracted to get the border irregularity function. When detecting the border irregularities, a turning point or a point where the sign of the derivative changes has been taken into consideration. Finally, with those points, the border irregularity has been calculated. This proposed method has been tested on 350 dermoscopic images, 280 benign and 70 malignant cases that have been collected by two university hospitals. Further, this database contains 140 cases with border irregularity less than 3.0 while 210 cases have border irregularity greater than 4.0. Overall the system has achieved 91% accuracy and 89% precision.

Finally, can be stated that many studies have done preprocessing of images to remove unnecessary objects such as hair, black frame, ink marks etc. initially to achieve an accurate segmentation. And the use of geometric features such as ABCD features have been widely used and immensely supportive in melanoma detection. And border irregularity is a crucial factor to be determined and identification of borderline correctly will affect the results. Further, most of the border irregularity techniques have accounted for the distance between the borderline and the image edge to identify the irregularity and not have taken the distance between the centroid of the lesion and the borderline into the account.

Through this study, attempts to fill the gap between the requirements and the existing solutions. Typically, for melanoma detection, ABCD can be stated as a set of pioneer geometric features, and it always generates the requirement to have novel techniques to extract the features. Hence through this study, it focuses to enhance geometric features for melanoma detection (ABCD) by proposing novel methods for border irregularity. Further, this study varies from the existing border irregularity detection techniques since it has been taken counts of distances from the centroid to the borderline without considering the distance from the image border to the borderline of the lesion. Also, the study

uses continuous borderline without using the discrete parts of the borderline in irregularity prediction.

III. BACKGROUND

A. Skin cancer detection

In the detection of melanoma, some traditional medical methods like histopathological examinations and biopsies are used. Further, in the detection of non-invasive skin cancers, different other methods such as photography, ultrasound, confocal microscopy, Raman spectroscopy, fluorescence tomography, the multispectral imaging technique, thermography, electrical bio-impedance and tape stripping are used [12]. Moreover, with the advancement of technology detection using dermoscopic images with Computer-Aided Diagnosis (CAD) systems has been a promising method. In CAD, features are extracted and images are being analyzed to diagnose skin cancers with the aid of numerous computers science approaches such as image processing, machine learning, neural networks and data mining etc. Though CAD systems are highly sensitive for the detection, it has been found that the false-positive rate in CAD predictions is high and the traditional diagnostic methods cannot be still replaced [13]. However, CAD can be a perfect aid for a medical doctor to make his traditional predictions accurate since dermatologists' inspection of dermoscopic images is a complex task that requires ample time and expert skills. Some types of traditional biopsy methods are shown in Fig. 3.

B. Corner points detection

1) Harris corner detection

1. It is a combined algorithm to detect both border and corner points which have been developed based on Moravec's corner detector by revising it [14]. Harris corner detection algorithm relies on an autocorrelation matrix and then eigenvalues of the matrix values have been found to locate strong intensity variations among the neighborhood [15]. Initially, the images are smoothed using Gaussian filters and the gradient of the image has to calculate. Then the autocorrelation matrix must be generated. And using the autocorrelation matrix, eigenvalues for them have been calculated computing the corner strengths by Harris measure in (1). Then using non-maximum suppression the highest intensity points are being selected. Finally, among the selected points further, the output corner/edge points are chosen and subpixel accuracy has been calculated. (See Fig. 4).

$$R_H = \lambda_1 \lambda_2 - \kappa. (\lambda_1 + \lambda_2)^2 \quad (1)$$

Where,

λ_1, λ_2 : two real non-negative eigenvalues, κ : value between 0.04 or 0.06 and R_H : corner response [15].

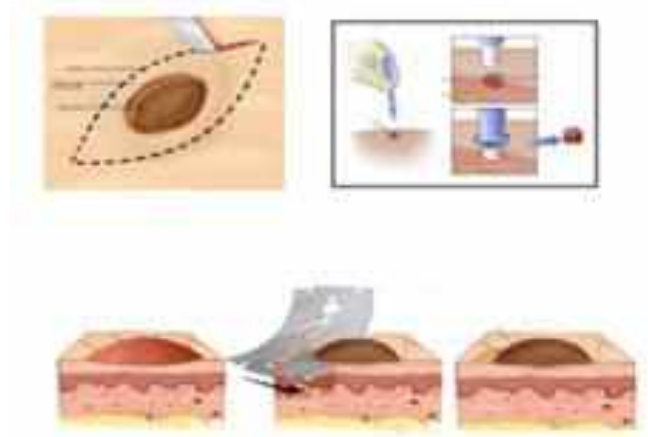


Fig 3. Types of biopsies: Excision biopsy, Punch biopsy and Shave biopsy [16].

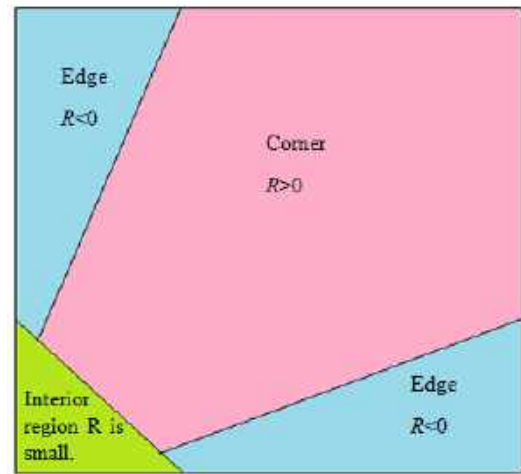


Fig. 4. Harris measure

C. Lesion classification

1) Naive Bayes classifier

1. Naive Bayes classifier is one of the most prominent classifiers that can be used for medical diagnosis which consists of a group of machine learning algorithms. It is a classifier that uses probability concept and classifying based on Bayes' theorem. The classifier is considered a naive approach as it considers all the data to be independent with each other. This classifier will get to know all the features of the relevant classes and will consider one by one feature for the classification. The Bayes' probability theorem for classification is given below in (2).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (2)$$

IV. METHODOLOGY

A. Binarisation

The obtained dataset has 3 classes of RGB color images of skin cancers along with binarised, well-segmented lesions of each of the images. Then the relevant images were separated as the RGB and binary images into different groups.

B. Border irregularity determination

Two different kinds of methods of the border irregularity were proposed and implemented.

1) Distance difference method

The centroid of the binary image and the corner points of the lesion were determined. Hence the distances between all corner points and the centroid have been calculated. Using the above-calculated distance, all height differences of the consecutive neighbor points have been calculated while traversing along with the detected corner points. Then considering a set of threshold values ranging from 15 to 30 (16 values) as T1, the extreme height differences above the thresholds were determined for each value. Then the percentage of the outward or inward points concerning the total number of corner points was determined. Another threshold value T2 of values 20, 25 and 30 were assigned to filter the highest percentages to determine the border irregularity. Images having the percentage greater than the second threshold value (T2), was considered to be irregular borders while others were considered as regular borders (See Fig. 5). Both threshold values were determined after observing the obtained results for the distance differences and then the percentages of distance differences.

2) Gradient method

1. In this approach, the gradients of the consecutive corner points were considered. Similar to the above method corner points were determined and the gradients of the neighbour corner points were determined. Using the gradient values rise and fall of the borderline was determined. Hence if the number of rises and falls in a particular image is above a set of threshold values (T), ranging from 45 to 55 (11 values), that was selected after observing the obtained results for gradients and then counting the number of sign changes to determine a local maximum and local minimum. Hence the threshold value T was considered as a benchmark predicting border irregularity. If an image has a ratio between no. of turning points to no. of corner points higher than T, that image was predicted as an image with an irregular border (See Fig. 6).

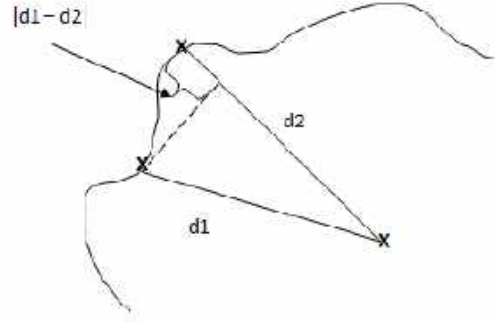


Fig 5: Distance difference method.

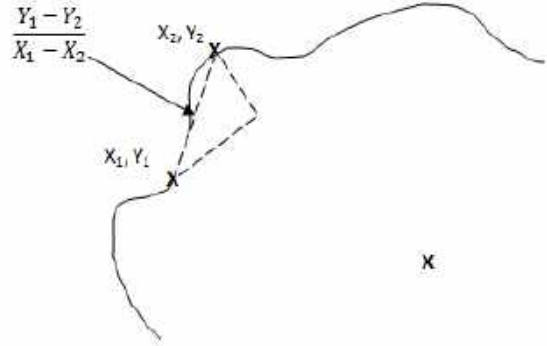


Fig 6: Gradient method.

C. ABCD features

The acquired dataset was having asymmetry, colour features along with it. The diameter of the lesions was also calculated in pixel values using 'region props' function in MATLAB and converted to millimeters. Then using the above-calculated border irregularity predictions, 48 predictions from the distance difference method by two threshold values ($T1 \times T2 = 16 \times 3$) and 11 predictions from gradient method by threshold value different set of ABCD feature values were generated. For each set of ABCD features the classifications were conducted to determine the most appropriate threshold values for each approach particularly for the used dataset. Further, a control set with only using Asymmetry, Colour features and Diameter (ACD), was also generated other than the ABCD sets to compare whether the obtained results were able to enhance the prediction or not.

D. Classifications

Naive Bayes classifier in WEKA tool was used to classify the dataset. The classification was done using ten-fold cross-validations making the whole dataset as training and testing to train and test the classifier. Initially with ACD features classification was performed and it was done as the control parameter. Then all features for the above generated 48+11 datasets from both methods were used to classify. Finally, the obtained classification results were compared with each other with respect to the control parameter. When comparing the results, the confusion matrices, number of

correctly classified and incorrectly classified instances, True Positive rate (TP), False Positive rate (FP), precision, Root Mean Square Error (RMSE) and also the contribution of border irregularity feature for the atypical, common and melanoma class classification were considered.

V. EXPERIMENTAL SETUP

A. Dataset

In this study, PH² dataset which includes 200 dermoscopic images in three classes as atypical nevus, common nevus and melanoma were used. The dataset contains 80 atypical nevus images, 80 common nevus and 40 melanoma images. All the images are in the same size and magnification (See Table I).

B. WEKA tool

Weka tool version 3.9 was used in the classification. It was operated in core i7 4500U CPU @ 1.80GHz 2.40GHz processor with 8GB RAM, 64-bit operating system.

C. MATLAB

In this study MATLAB 2015a, a 32-bit version was operated in a core i7 6700U CPU @ 3.40GHz 3.41GHz processor with 4GB RAM, 64-bit operating system.

TABLE I. CLASSES OF PH² DATASET.

No of images	Atypical nevus	Common nevus	Melanoma
	80	80	40

VI. RESULTS AND DISCUSSION

Through this study, it was able to obtain 78% of correctly classified instances using the ACD features. Further, using the distance difference method 79% of correctly classified instances and by the gradient method it was able to achieve the correct prediction of 78.5% (See Table II.) And when discussing the classifications, it was observed that the proposed methods work well with the non-melanoma classes, atypical nevus and common nevus identification. (See Table III-V). That can be stated because the classifications of atypical and common nevus are correct than the melanoma classification.

Further, according to the results, sample 44 (T1=29, T2=20) shows the best performance in the distance difference method. Moreover, the most accurate class prediction ratios of distance difference method 55/80, 69/80 and 34/40 for the three classes atypical, common and melanoma respectively can only be seen together in the confusion matrix of sample 44. Hence confusion matrix of sample 44 again proves that it has the best performance in the distance difference method. Similarly, sample 10 (T=54) shows the best performance in the gradient method. The most accurate class prediction ratios 52/80, 72/80 and 33/40 for the three classes atypical, common and melanoma respectively can only be seen together in the confusion matrix of sample 10, it again proves sample 10 has the best performance in the gradient method.

TABLE II. CLASSIFICATION RESULTS.

Method used	Accuracy	TP	Precision	F-measure	RMSE
Only with ACD features	78%	0.780	0.790	0.778	0.334
ABCD features with distance difference method	79%	0.790	0.796	0.789	0.336
ABCD features with gradient method	78.5%	0.785	0.795	0.783	0.335

TABLE III. BEST RESULTS OF CONFUSION MATRIX FOR DISTANCE DIFFERENCE METHOD

Class	Atypical	Common	Melanoma
Atypical	55	23	2
Common	11	69	0
Melanoma	5	1	34

TABLE IV. BEST RESULTS OF CONFUSION MATRIX FOR GRADIENT METHOD

Class	Atypical	Common	Melanoma
Atypical	52	26	2
Common	8	72	0
Melanoma	6	1	33

TABLE V. CONFUSION MATRIX WITH ONLY ACD FEATURES

Class	Atypical	Common	Melanoma
Atypical	51	27	2
Common	9	71	0
Melanoma	5	1	34

Overall, when illustrating the border irregularity predictions of skin lesions through the proposed methods, few sample results can be given as follows. The given sample images and respective edge distance difference patterns for consecutive borderline points are shown below in figures. (see Fig. 7-10). Considered threshold value T1=29 is the best T1 threshold value obtained by the distance difference method. Moreover, the following images also have been predicted as the same type as irregular and regular by both gradient and distance difference method.

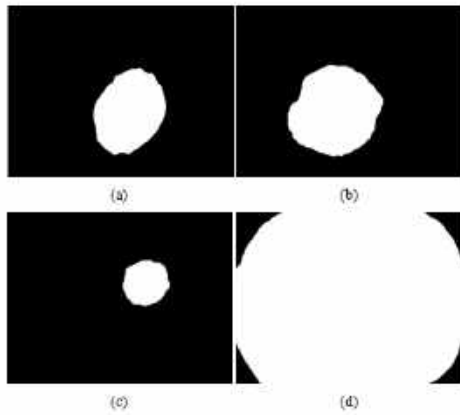


Fig 7: Sample images with regular borders.

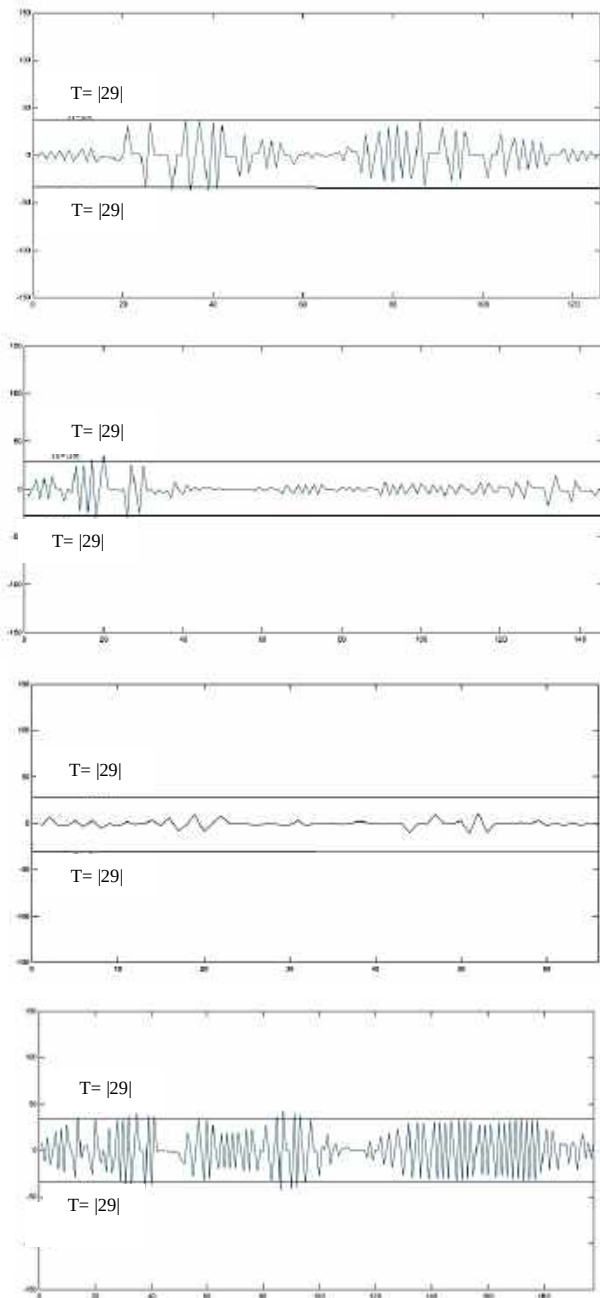


Fig 8: Distance difference of consecutive points for regular border lesions for irregular border skin lesions (a), (b), (c) and (d) in Fig. 7 respectively.

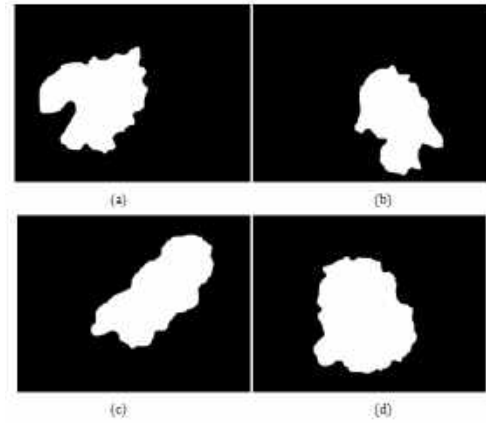


Fig 9: Sample images with irregular borders.

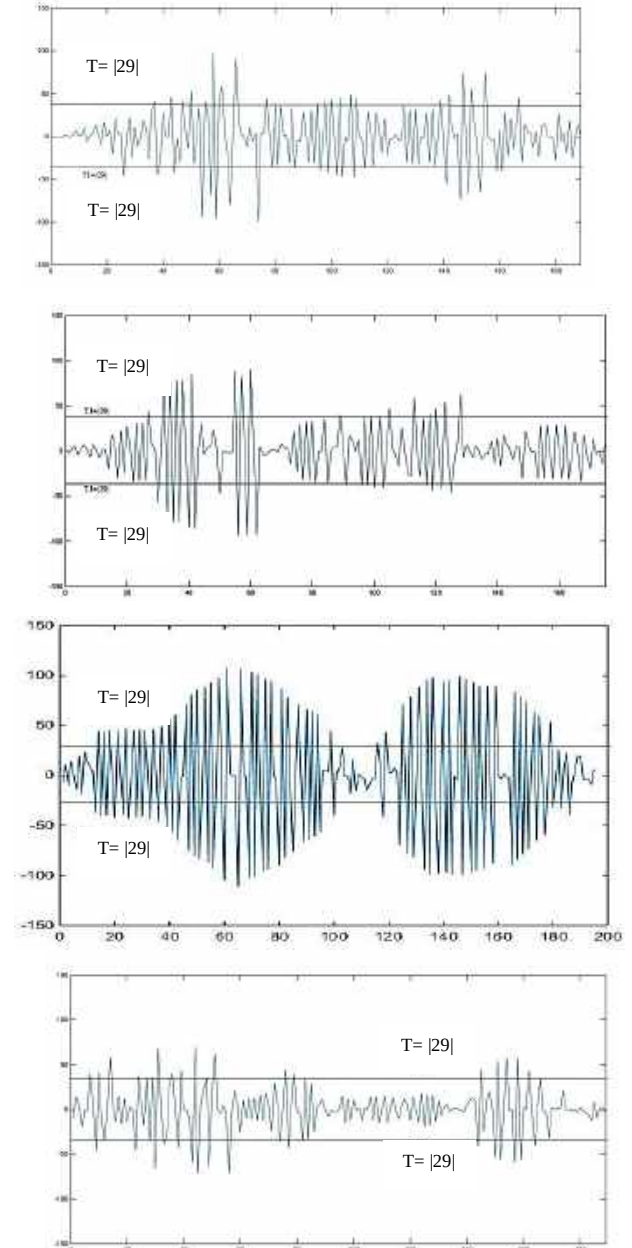


Fig 10: Distance difference of consecutive points for irregular border lesions for irregular border skin lesions (a), (b), (c) and (d) in Fig. 9 respectively.

The results can be varied with the dataset and also with the quality of data. Since this dataset has 200 images and most of the images are fallen to those two categories rather than melanoma it can be a reason for the above observation. Since the data have been segmented well and also the region of interest have been identified clearly, it has improved the precision of the results. However, the use of binary images can be in effect for the obtained results as the patch-based or color intensity features cannot be obtained rather than the number of colors was taken into account. Moreover, this study has obtained commendable results with only using the geometrical features.

VII. CONCLUSION

Through this study, it was able to propose two simple approaches to determine the border irregularity of a skin lesion. Further, the obtained results were used in melanoma classification to validate the obtained results. Overall from the study, it can be concluded that the classifications perform well in non-melanoma rather than melanoma classifications. Also, without the extracted border irregularity feature, the classification gives an accuracy of 78% whereas with the results it gives a slight improvement in classification as 79% and 78.5% for distance difference method and gradient method respectively. Further from the suggested methods, for the particular dataset distance difference methods shows remarkable performances in border irregularity detection and can be stated as the best method out of the two.

The future works of this study are commenced in two ways which are bag-of-features and deep learning approaches. The main intention of that work is to compare the success of using geometrical features and bag-of-features in melanoma detection. Further hoping to repeat the proposed approaches in detecting border irregularity for a large dataset and hence compare the performance with this study to validate the proposed methods.

REFERENCES

- [1] T. Allison, "6 Steps to Check for Melanoma." Healthcentral.Com, HealthCentral, 4 Oct. 2012, sOnline]. Available at: www.healthcentral.com/slideshow/6-steps-to-check-for-melanoma. [Accessed 6 October 2019].
- [2] "Another Pigmented BCC." Dermoscopy, [Online]. Available at: dermoscopy.wordpress.com/2014/09/15/another-pigmented-bcc/, 15 September 2014. [Accessed 6 October 2019].
- [3] "Breast Density: Are You Informed?," Imaging Technology News, [Online]. Available at: www.itnonline.com/article/breast-density-are-you-informed, 30 April 2014, [Accessed 6 October 2019].
- [4] Editorial Team, "Skin Cancer (Nonmelanoma).", Academic Association of Medicine, [Online]. Available at: aaom.org/health/skin-cancer-nonmelanoma/, 7 December 2017, [Accessed 6 October 2019].
- [5] Wikipedia Contributors. "Melanoma." *Wikipedia*, Wikimedia Foundation, [Online]. Available at: en.wikipedia.org/wiki/Melanoma, 12 October 2019, [Accessed 6 October 2019].
- [6] The Skin Cancer Foundation. Skin Cancer Information - The Skin Cancer Foundation. [Online]. Available at: <https://skincancer.org/skin-cancer-information> 2019. [Accessed 6 October 2019].
- [7] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks", IEEE Transactions on Medical Imaging, vol. 36, no: 4, pp. 994-1004, April 2017.
- [8] J. Shivangi, et al. "Computer-Aided Melanoma Skin Cancer Detection Using Image Processing.", International Conference on Intelligent Computing, Communication and Convergence (ICCC 2015), Procedia Computer Science, vol. 48, pp. 736-741, 2015, India.
- [9] T. K. Lee and M. S. Atkins, "A new approach to measure border irregularity for melanocytic lesions", Medical Imaging 2000: Image Processing, 2000.
- [10] J. Jaworek-Korjakowska and R. Tadeusiewicz, "Determination of border irregularity in dermoscopic color images of pigmented skin lesions," 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014.
- [11] J. Jaworek-Korjakowska, "Novel method for border irregularity assessment in dermoscopic color images," Computational and Mathematical Methods in Medicine, vol. 2015, pp. 1-11, 2015.
- [12] V. Narayanamurthy et.al. "Skin cancer detection using non-invasive techniques", Review Article, RSC Advances, vol. 8, no.49, pp. 28095-28130, 2018.
- [13] L. F. di Ruffano et.al. "Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults", Cochrane Database of Systematic Reviews, Apr. 2018.
- [14] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Proceedings of the Alvey Vision Conference 1988*, 1988.
- [15] J. Sánchez, N. Monzón, and A. Salgado, "An Analysis and Implementation of the Harris Corner Detector," Image Processing On Line, vol. 8, pp. 305-328, Mar. 2018.
- [16] B. U. W. Lei and T. W. Prow, "A Review of Microsampling Techniques and Their Social Impact.", Biomedical Microdevices, vol. 21, no. 4, 15 August 2019

Relationships between climatic factors to the paddy yield in the North-Western Province of Sri Lanka

Lasini Wickramasinghe*
Department of Electronics,
Faculty of Applied Sciences,
Wayamba University of Sri Lanka
Sri Lanka
lasini2013@gmail.com

Jeevani Jayasinghe
Department of Electronics,
Faculty of Applied Sciences,
Wayamba University of Sri Lanka
Sri Lanka
jeevani@wyb.ac.lk

Upaka Rathnayake
Department of Civil Engineering,
Faculty of Engineering,
Sri Lanka Institute of Information
Technology, Sri Lanka
upakasanjeewa@gmail.com

Abstract: Climate variation is one of the major impacting issues for paddy cultivation. It also highly impacts the harvest. Therefore, many researchers try to understand the relationships between climatic factors and harvest using numerous methods. Sri Lanka is still titled as a country with an agricultural-based economy and thus identifying the impact of climate variability on agriculture is very important. However, previous studies reveal a little information in the context of Sri Lanka on the impact of climate variabilities on agriculture. Therefore, this study showcases an artificial neural network (ANN) framework; that is an ordinary machine learning algorithm based on the model of the human neuron system, to evaluate the relationships among the climatic components and the paddy harvest in the North-Western province of Sri Lanka. This on-going study helps to analyze the relationships between the paddy harvest of the North-Western province and climate, including rainfall minimum atmospheric temperature and maximum atmospheric temperature. Correlation coefficient (R) and mean squared error (MSE) are used to test the performance of the ANN model. The results obtained from the analysis revealed that the predicted and real paddy yields have a significant correlation with rainfall, maximum temperature and minimum temperature.

Keywords: Artificial Neural Network (ANN), LM algorithm, North-Western province, Paddy yield, Rainfall, Temperature

I. INTRODUCTION

Climatic variations have a significant influence on the cultivation of crops such as rice [1]. Rice is a staple food in many countries and so does in Sri Lanka [2]. It is also one of the dominant agricultural products grown in the North-Western province in Sri Lanka. Many researchers work on identifying the relationships between climatic factors and paddy yield due to the importance of this staple food [3]. Among them, statistical methods [4] and artificial neural network (ANN) frameworks [5] are widely used in the literature. However, the climate is well understood as a highly non-linear phenomenon and thus, many researchers tend to use soft computing techniques including neural networks. Nevertheless, the research conducted in this field is less in the context of Sri Lanka even though it highly impacts paddy cultivation. Therefore, this paper explores the soft computing techniques (i.e. ANN work) to investigate the relationships among the climate variability and the paddy yield.

Paddy is the most widely planted crop in Asia [6] as it is the staple food of many countries. However, paddy uses more water than other cereal crops [7] thus making it potentially more vulnerable to droughts. Indeed [8], found that rice

production in India was more impacted by El Nino–Southern Oscillation (ENSO) than the production of the other crops (wheat, sorghum, and legumes). Several other studies were carried out to assess the impact of ENSO on rice production due to the importance of rice as a staple food and its potential susceptibility to drought [9-12]. However, they were carried out using the regression analysis.

Naylor et al. [10] have carried out research to identify the impact of climatic change on crop yielding in the mountainous region of Nepal. They considered the time series analysis of precipitation, temperature and yield of selected crops. However, the research too was based on the regression models.

Nevertheless, Sun et al. [5] have carried out research on rice yield prediction in Fujian Province, China under the typical Fujian climatic conditions by using ANN. They searched whether ANN models can be used to predict the Fujian rice yield, to identify the model's performance relative to variations of developmental parameters and to compare the results of multiple linear regression models with ANN models. Historical data of multiple locations were used with the climate data (such as rainfall, number of sunshine hours, daily solar radiation, temperature and wind speed).

Zubair [12] analysed the impact of ENSO on rice production in Sri Lanka. The analysis uses an extended time series and considers finer spatial units than presented in [6] and reports on the variability of Area cultivated with ENSO as well. It analyzed 60 years of PAY (Production, Area and Yield) data and report on correlations between seasonal Sea Surface Temperature Anomaly (SSTA) and seasonal production, area and yield. The findings of the analysis revealed that there is a significant influence on rainfall and rice production by ENSO.

There is a clear research gap in the research for the prediction of rice production based on the various climatic conditions in Sri Lanka. Therefore, an artificial neural network was trained for the purpose of predicting paddy yield to the climatic changes.

Section II of this paper explains the Lavenberg-Marquardt algorithm used in the development of the training process of ANN. Section III illustrates the mathematical formulations for the paddy yield with respect to rainfall, minimum temperature and maximum temperature. Section IV presents how ANN framework was applied to analyse data and section V presents the results obtained. Finally, the research findings are concluded.

II. LEVENBERG-MARQUARDT ALGORITHM

The literature suggests several algorithms to optimize the neural network. However, among the applicable algorithms, Levenberg-Marquardt (LM) algorithm in an artificial neural network is frequently used in the literature. The LM optimization algorithm is identified to be more powerful than the conventional gradient descent techniques and it is one of the most widely used optimization algorithms in the field.

The algorithm was developed many decades ago (in the 1960s) to solve non-linear least square problems [13]. Two minimization methods, named gradient descent method and Gauss-Newton method, are combined together in the algorithm. The summation of the squared errors is minimized in the gradient descent method. This is done by parameter variation in the steepest descent direction. However, the same errors are reduced in the Gauss-Newton method on finding the minimum of the assumed function which is locally quadratic.

The algorithm differs in the way the parameters are updated compared to the two methods discussed above and Levenberg developed the following relationship.

$$\left[J^T W J + \lambda I \right] h_{lm} = J^T W (y - \hat{y}) \quad (1)$$

The matrix notations are given below.

W	Weighting matrix
J	Jacobian matrix
λ	Damping parameter
h_{lm}	Parameter update
y	Measured data
$\hat{y}$	The curve-fit function $\hat{y}(t, p)$ of an independent variable t and a vector of n parameters p

Equation 1 depends on the Gauss-Newton update for small values of damping parameter (λ) whereas large values of λ deals with the gradient descent update [14]. Marquart's [13] updated the relationship as given in Equation 2.

$$\left[J^T W J + \lambda \text{drag}(J^T W J) \right] h_{lm} = J^T W (y - \hat{y}) \quad (2)$$

III. A MATHEMATICAL FORMULATION OF RELATIONSHIPS

Instead of paddy harvest, paddy yield was considered for the mathematical formulations. Harvest has a temporal variation not only with the climate but also with the harvested land area. Therefore, harvest per unit area (yield) was considered for mathematical modelling. Paddy yield was interpreted in terms of separate functions of rainfall, minimum temperature and maximum temperature. Equations 3, 4 and 5 illustrate the relationships.

$$\text{Yield}_{i,j} = \phi_1(\text{Monthly Rainfall}_{i,j}) \quad (3)$$

$$\text{Yield}_{i,j} = \phi_2(\text{Minimum Atmospheric Temperature}_{i,j}) \quad (4)$$

$$\text{Yield}_{i,j} = \phi_3(\text{Maximum Atmospheric Temperature}_{i,j}) \quad (5)$$

where ϕ_1 , ϕ_2 and ϕ_3 are the functions for corresponding relationships. Subscripts “ i ” and “ j ” stand for the location and season, respectively.



Fig.1. Locations of the study area (North-Western Province)

IV. ANALYZE THE IMPACT OF CLIMATE VARIABILITY

ANN framework was applied to the North-Western province of Sri Lanka, which consists of two districts; Puttalam and Kurunegala. Therefore, the climatic data for those two districts (refer Figure 1) were obtained from the Department of Meteorology, Sri Lanka. The resolution of the data scatter was monthly taken and seasonal data (Yala and Maha) were used for this analysis. The paddy yield data for the last 30 years were obtained from the Irrigation Department of Sri Lanka.

The data were then analyzed in the ANN framework developed in the MATLAB environment. As per the literature, several algorithms were applied as the optimization algorithm on the MATLAB based ANN. Levenberg-Marquardt (LM), Bayesian Regularization (BR) and Scaled Conjugate Gradient (SCG) algorithms are among them. In this research, LM was used in the ANN framework as it was used successfully by researchers in the field.

The ANN method was first used in artificial intelligence research that attempted to mimic the capacity to learn through biological neural systems. Many different types of neural nets are available and their structures are described in [14], [15] and [16]. The feed-forward network illustrated in Figure 2, shows a common ANN architecture used in the literature. This architecture requires comparably little memory and thus it is faster than the other models [17]. In addition, there is only one directional move in this architecture from the input layer to the hidden layer through the hidden layer, in contrast to the usual feedback architectures ANNs [5].

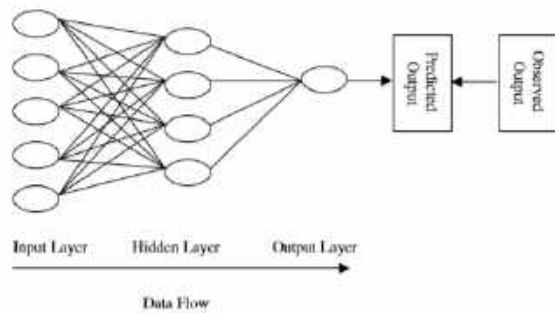


Fig.2. Layers and connections of a feed-forward back-propagating artificial neural network [5]

In this research, climate data were employed as the input layer and the yield data was used as the output layer. Two hidden layers were created automatically within these two layers while training the algorithm.

It is well noted that Sri Lanka has two major agricultural seasons (Yala and Maha) and the research was carried out based on those two seasons. 75% of the data were used for the training process of the neural network and 10% and 15% of data were used for the validation and testing processes, respectively for Puttalam Maha season rainfall and yield data. 80% of the data were used for the training process of the neural network and 5% and 15% of data were used for the validation and testing processes, respectively for Puttalam and Kurunegala Maha season maximum temperature, minimum temperature and yield data.

V. RESULTS AND DISCUSSION

Analysed results of yield data against three parameters namely rainfall, maximum temperature and minimum temperature in Kurunegala and Puttalam districts, respectively are given in this section. Figure 3 shows the correlation coefficient of predicted and measured yield from ANN against the yield with respect to the Yala rainfall on Kurunegala district.

There is an acceptable correlation to the predicted and measured yield (Figure 3). R equals 0.745 is an acceptable correlation coefficient to the climate studies as it is highly non-linear. However, there should be more data regarding the validation process. Nevertheless, the total correlation again shows a good match. Therefore, it can be presented here that (with further analysis), the developed ANN model can be used for the future prediction of paddy yield concerning the Yala rainfall.

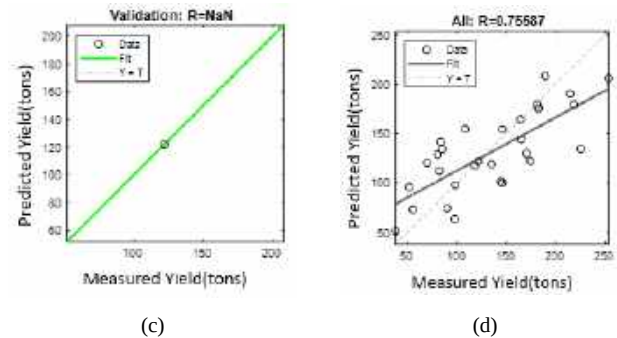
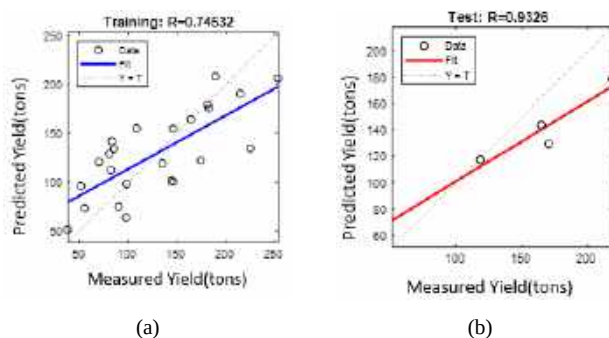


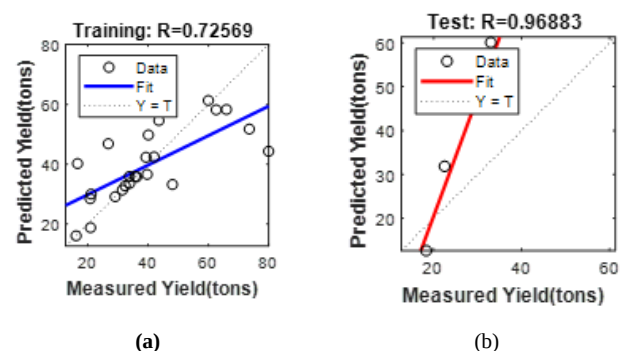
Fig.3. The correlation coefficient for Kurunegala district for Yala season Rainfall (a) For training (b) For Validation (c) For test (d) For All

Figure 4 presents the performance of the developed ANN architecture for the Yala season yield with respect to the corresponding rainfall. The best performance was reached at a cost of 13 epochs; thus, the system produces acceptable results at an acceptable computational cost.

Similar to Figure 3, Figure 5 shows the correlation coefficient of predicted and measured yield for the ANN developed against the yield with respect to the Maha season maximum temperature for Puttalam district. Similar results can be found compared to Figure 1. Overall, the analysis shows an acceptable correlation coefficient ($R=0.67$). However, there should be more data regarding the validation process. Therefore, the process is being modified further.



Fig.4. Mean squared error for Kurunegala district for Yala season rainfall analysis



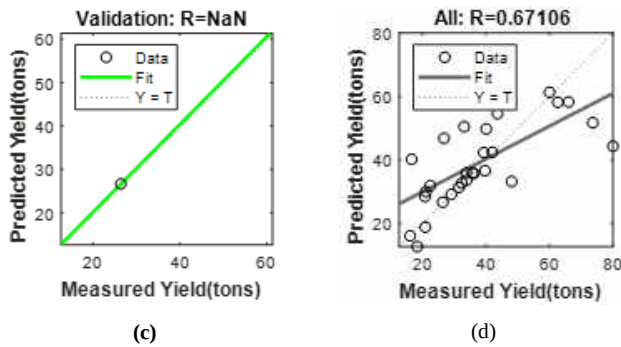


Fig.5. The correlation coefficient for Puttalam district for Maha season Maximum Temperature (a) For training (b) For test (c) For Validation (d) For All

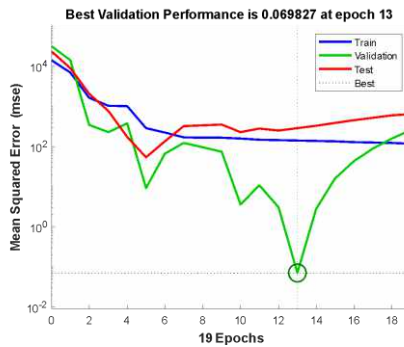


Fig.6. Mean squared error for Puttalam district for Maha season maximum temperature analysis

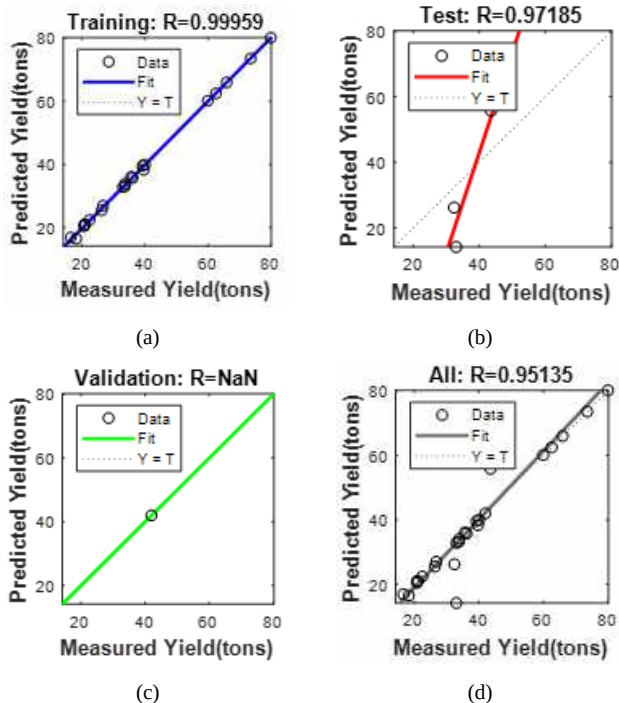


Fig.7. The correlation coefficient for Puttalam district for Maha season Minimum Temperature (a) For training (b) For Validation (c) For Test (d) For All

Figure 6 illustrates the mean squared values for Puttalam district Maha season yield concerning the maximum temperature. As it was already discussed under Figure 4, the

ANN model gives interesting results at a lower error (less than 0.1) at an acceptable computational cost (epoch of 13). This is important when the model is applied to real world cases.

Figure 7 showcases the correlation coefficient of predicted and measured yield for the ANN developed against the yield with respect to the Maha season minimum temperature for Puttalam district. Similar observations on the validation process can be seen in this case too. However, interestingly, the overall acceptability of the developed ANN model is significant, thus the coefficient of correlation of the predicted measured yield is 0.95. This is almost reaching 1.0; therefore, the ANN model satisfies the basic requirements for future predictions. Besides, similar computational costs were achieved for this analysis.

TABLE I. CORRELATION COEFFICIENTS FOR YALA SEASON

District	Climate variable	R
Kurunegala	Rainfall	0.76
	Minimum Temperature	0.79
	Maximum Temperature	0.37
Puttalam	Rainfall	0.92
	Minimum Temperature	0.84
	Maximum Temperature	0.47

TABLE II. CORRELATION COEFFICIENTS FOR MAHA SEASON

District	Climate variable	R
Kurunegala	Rainfall	0.63
	Minimum Temperature	0.85
	Maximum Temperature	0.61
Puttalam	Rainfall	0.81
	Minimum Temperature	0.95
	Maximum Temperature	0.67

Table I summarises the overall results obtained from the analysis of the Yala season to both districts with climate variables, while table 2 summarises that of the Maha season. They present the correlation coefficients for the ANN models developed with respect to climatic parameters. They show good agreement. Therefore, the ANN models are set for the analysis and the future climate data from various models can be sued to predict the future yield.

VI. CONCLUSION

An artificial neural network based on the climatic model is set to predict the rice yield in North-Western Province, Sri Lanka. The results obtained so far revealed that the developed models give acceptable model predictions. The models are based on rainfall, minimum temperature and maximum temperature. They show good agreements. More importantly, the results were produced at an acceptable computational cost to the minimum square errors. Therefore, the developed models can be easily used to predict the future yield with given predicted rainfall and temperature to Kurunegala and Puttalam districts. However, if the research is carried out for more climate parameters individually and as a set of climatic

parameters, it would be possible to build a holistic model to predict the paddy yield concerning the future climate changes.

ACKNOWLEDGMENT

The authors would like to thank the Department of Census and Department of Irrigation, Sri Lanka for providing the paddy yield and climate data for this study.

REFERENCES

- [1] R. B. Matthews, M. J. Kropff, T. Horie and D. Bachelet, "Simulating the impact of climate change on rice production in Asia and evaluating options for adaptation," *Agricultural Systems*, vol. 54, no. 3, pp. 399-425, 1997.
- [2] J. M. Bandara, A. K. Vithanage and G. A. Bean, "Occurrence of aflatoxins in parboiled in Sri Lanka," *Mycopathologia*, vol. 116, no. 2, pp. 65-70, 1991.
- [3] J. Felkner, K. Tazhibayeva and R. Townsend, "Impact of Climate Change on Rice Production in Thailand," *American Economic Review*, vol. 99, no. 2, pp. 205-10, 2009.
- [4] M. A. R. Sarker, J. Gow and K. Alam, "Exploring the relationship between climate change and rice yield in Bangladesh; An analysis of time series data," *Agricultural Systems*, vol. 112, pp. 11-16, 2012.
- [5] B. Ji, Y. Sun, S. Yang and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *Journal of Agricultural Science*, vol. 145, pp. 249-261, 2007.
- [6] D. Dawe, "Agricultural Research, Poverty Alleviation and Key Trends in Asia's Rice Economy," pp. 37-53, 2008.
- [7] B. Bouman et.al. "Rice: Feeding the billions. Water for Food, Water for Life: Comprehensive Assessment of Water Management in Agriculture," pp. 515-549, 2007.
- [8] R. Selvaraju, "Impact of El Niño–Southern Oscillation on Indian foodgrain production," *International Journal of Climatology*, vol. 23, no. 2, 2003.
- [9] W. P. Falcon, R. L. Naylor, W. L. Smith, M. B. Burke and E. B. McCullough, "Using climate models to improve Indonesian food security.," *Bulletin of Indonesian Economic Studies*, vol. 40, no. 3, pp. 355-377, 2010.
- [10] R. L. Naylor, W. P. Falcon, D. Rochberg and N. Wada, "Using El Niño–Southern Oscillation climate data to predict rice production in Indonesia.," *Climatic Change*, vol. 50, no. 3, pp. 255-265, 2001.
- [11] L. Zubair, "El Niño–Southern Oscillation influences on rice production in Sri Lanka," *Climatol*, vol. 22, pp. 249-260, 2002.
- [12] S. Z. Yahiya, S. Adhikari, M. Weerasekara and L. Zubair, "The Impacts of El Niño Southern Oscillation (ENSO) Events on Rice Production, Area and Yield in Sri Lanka".
- [13] D.W. Marquardt. "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11(2), pp. 431-441, 1963.
- [14] C. M. Bishop, *Neural Network for Pattern Recognition*, Oxford University Press, 1995.
- [15] B. D. Ripley, *Pattern Recognition and Neural Networks*, 1 ed., Cambridge University Press, 1996.
- [16] J. C. Principe, N. R. Euliano and W. C. Lefebvre, *Neural and Adaptive Systems: Fundamentals through Simulations*, John Wiley and Sons, 2000.
- [17] J. Lawrence, *Introduction to Neural Networks: Design, Theory, and Applications*, California Scientific Software, 1994

Aspect-based sentiment analysis on hair care product reviews

Malki Kothalawala*

Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology, Sri Lanka
malkinimesha@gmail.com

Samantha Thelijjagoda

SLIIT Business School
Sri Lanka Institute of Information Technology, Sri Lanka
samantha.t@slit.lk

Abstract: Nowadays, with almost everything being shared online, people are more verbal about their consumer experiences with products via reviews. Reviews can be vital for manufacturers to get insights into consumer opinions and consumers in their purchase decisions. Sentiment analysis, referring to the extraction of subjective opinions on a particular subject within a text, is a field within Natural Language Processing, that can convert this unstructured information hidden within reviews into structured information expressing public opinion. In regards to a specific product group like hair care products, certain brands are rising in the market due to their positive public opinion on particular aspects. While e-commerce websites facilitate users to view the reviews, they do not display which reviews contain which type of opinion on which aspect at a glance. This research aims to introduce an automated process that focuses on determining the polarity of online consumer reviews on different aspects of hair care products by using Aspect-based Sentiment Analysis. The system consists of processes like data gathering, pre-processing, aspect extraction and polarity detection and follows a sequential approach to achieve the intended goal. Consequently, by deciphering the aspect-wise polarity of the reviews, the implemented system demonstrates an accuracy of 85% from the test data for overall aspects, enabling consumers to get an at a glance idea about the public opinion and manufacturers to identify their strong and weak points.

Keywords: Aspect-based Sentiment Analysis, Natural Language Processing, Opinion Mining, Sentiment Analysis, Supervised Learning

I. INTRODUCTION

In a highly competitive business world, it is vital for brands and companies to get insights into consumer behaviors, their needs, likes and dislikes about a particular product, buying patterns and how consumers make decisions. Traditional businesses used to depend on practices like surveys in order to get insights into the consumer opinions. However, in the current modern world, businesses can utilize the power of the vast amount of both public and private information being generated on a daily basis due to the continuously growing number of users and devices connected to the internet. With an increasing amount of information being shared via the means of social channels, review sites, blogs and web forums, people tend to be more verbal about their consumer experiences online. Moreover, while traditional approaches like surveys often influence the responses of the consumers due to their use of target questions, reviews enable companies to get the natural opinions of the consumers. Furthermore, from the consumers' perspective, their purchase decisions are majorly influenced by the reviews and the experiences of the other consumers about the precise product which are available online. However,

consumers often have to read a long list of subjective reviews and opinions in order to decide whether the product fits their need since it is not visible at a glance and requirements are different from consumer to consumer.

Since general product reviews is a broad topic, it is specialized into a more specific type of reviews like the hair care product reviews. In regards to hair care product market, it has a wider scope with various types of hair care products being available. With different brands introducing various product lines intended for numerous hair goals for each category of hair care products, the market has become hypercompetitive. Due to factors like geographical limitations, ease of search offers and eliminating travel time and cost, consumers tend to opt for e-commerce websites like eBay, Amazon for buying their hair care products. When purchasing a hair care product via an e-commerce website, consumers come with different requirements in their mind and they often decide whether to purchase the product or not, by considering the reviews of the other consumers to validate against their hair goals. Moreover, once they use the product, they share their opinions so that those reviews can be helpful for the other consumers. For instance, a consumer might be interested in purchasing a certain shampoo from eBay and wishes to know the public opinion on how moisturizing the shampoo is, which is his or her hair goal. However, if most of the top reviews are talking about the scent of the shampoo, which therefore is not helpful for the consumer. As a result, the consumer often has to read a long list of reviews to find out whether the product matches his or her needs. Furthermore, in relation to the hair product manufacturers, they could be interested in knowing whether the consumers are satisfied with the ingredients of their new hair product line so that they can change their formulas according to consumer feedback. A mechanism of presenting the polarity of the reviews on the hair care product in terms of the aspects being talked about would be ideal in both of the aforementioned scenarios.

Sentiment analysis, a field within Natural Language Processing (NLP), referring to the extraction of subjective opinions on a particular subject within the text, can be utilized in comprehending most of the unstructured information available in review sites, social media and web forums, etc. Sentiment analysis aids in extracting information from a text about the polarity, which denotes whether the speaker.

expresses a positive or negative opinion, subject, which topic is being discussed and the opinion holder, the individual conveying the opinion. With the ever-increasing amounts of textual data in the world and most of them being unstructured, sentiment analysis is highly beneficial for commercial applications while enabling organizations to gain crucial

business insights by deciphering unseen meanings of the text in order to make better business decisions. While sentiment analysis can be assistive in deciding whether a person is expressing a negative, positive or neutral opinion, Aspect-based Sentiment Analysis (ABSA) is capable of uncovering which aspect or attribute of the product the person expresses the opinion on. It is significant to know precisely which feature of the hair care product is being talked about since consumers often look for reviews on different aspects.

Hence, the purpose of the system is to decipher the polarity of online consumer reviews on different aspects of hair care products using aspect-based sentiment analysis so that it can be beneficial for both the consumers and the product manufacturers.

II. RELATED WORK

A. Rule-based approaches

In a research conducted by Sonal Meenu Singh and Nidhi Mishra, it classified customer reviews on mobile phones as positive, negative or neutral based on manually extracted eight aspects including cost, size, battery, camera, OS, processor, storage and screen. The authors have conducted the research collecting 80 mobile phone reviews from commercial websites like www.amazon.com and www.cnet.com on Iphone 6, Moto G3 and blackberry Z10. The reviews collected have been pre-processed by removing stopwords, white spaces and special symbols and POS tagging has been performed on the pre-processed reviews with the assistance of Stanford parser. With the aid of Part-Of-Speech (POS) tagging, opinion words have been identified and stored from the reviews. SentiWordNet 3.0 has been used to identify scores for each aspect after the completion of pre-processing. In comparison with the online tools such as WordNet, SentiWordNet and MPQA, the system is depicted to have an accuracy of 75% and according to the results, the authors state that the system is capable of handling negation, intensifiers and synonyms more effectively than the compared tools. The drawback of the system is described as its inability to accurately identifying the polarity of the cost feature [1].

Paramita Ray and Amlan Chakrabarti have carried out research on sentiment analysis for product reviews using the Lexicon method. A framework for analyzing the sentiment of users on Twitter using the Twitter API has been proposed with the aid of R software. As the initial step, the authors have created a Twitter application in order to use Twitter API. After the completion of data collection, the tweets have been analyzed and pre-processed. As the following step, classification has been performed with the assistance of a lexical method or a dictionary-based approach. In order to test the results, tweets with the keyword 'iPhone' posted within a specific time period has been analyzed considering aspects like voice quality, battery quality, service, price, picture quality and size. Apart from being able to identify emoticons in tweets, the research was also done in both document level and aspect level analysis [2].

Moreover, I. K. C. U. Perera and H. A. Caldera have done a research on aspect-based opinion mining which specifically focuses on restaurant reviews. Reviews on 20 restaurants have been collected from the website www.zomato.com. For the purpose of extracting aspects, the pre-processed reviews have been tokenized with the aid of a POS tagger. Based on the

frequency of occurrence of words, aspects have been manually selected as place, food, service, time and staff. Using Stanford dependencies, which represent grammatical relations between words, opinion words have been identified and the filtered-out aspect words and opinion words have been sent to SentiWordNet. SentiWordNet assigns scores for the key value pairs and once all assigning scores for all opinion words against a certain aspect, the total of all scores has been calculated. Polarity has been decided based on the score value. The testing has been done both manually and systematically. In the systematic approach, the aspect polarity is checked by the system and in comparison to the manually checked results, the systems are depicted to show an accuracy of 70% [3].

B. Automatic approaches

A research conducted by Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain focused on sentiment analysis of consumer product reviews. The Statistical and Sentiment Analysis of Consumer Product Reviews (SACP) framework developed by the authors, is composed of two major components where one module handles Data Collection and Pre-processing and the other module works on Feature Selection and Analysis. The Data Collection module has collected more than 400,000 online reviews on around 4500 mobile phones from the e-commerce website Amazon.com. The data sets have been put to CSV file format and the data has then been pre-processed in order to remove white spaces, stop words, digits, punctuation and special symbols. The authors have used a package called tm for the purpose of text mining. As the initial process of the feature selection and analysis module, the extraction of related features from the data is performed. The module then carries out statistical analysis of the data in order to analyze the correlation between the features. As the final step of the module, the system performs sentiment analysis on the text in order to predict the sentiment as well as the polarity. The authors state that the classification proves to be efficient based on the 84.87% accuracy of Support Vector Machine (SVM) after cross validation [4].

Dhanalakshmi V., Dhivya Bino and Saravanan A. M. have conducted a research on opinion mining from student feedback data in 2016 using a supervised learning algorithm. The authors have used the tool Rapid Miner in mining and classifying the feedback provided by the students. The authors have used Support Vector Machine (SVM), Naïve Bayes (NB), K Nearest Neighbor (KNN) and Neural Networks (NN) as the supervised opinion mining algorithms. The authors have used the validation operator in Rapid Miner in order to concurrently train and test the classifiers. Using the Performance Operator in Rapid Miner, for each algorithm, accuracy, precision and recall values have been calculated for comparing the performance of the algorithms. From the results, the authors have identified that a 100% best result for precision is showed by K Nearest Neighbor (KNN) where a best of 97.07% and 99.11% for recall and accuracy has been shown by Naïve Bayes (NB) algorithm [5].

C. Hybrid approaches

Santhosh Kumar K L, Jayanti Desai, and Jharna Majumdar have conducted a research on opinion mining and sentiment analysis on online customer reviews. The system automatically extracts reviews from the Amazon website and classifies the reviews as positive or negative using algorithms such as Naïve Bayes classifier, Logistic Regression and

SentiWordNet. Based on the experimental results, the research concludes that Naïve Bayes classification happens to be the most efficient algorithm for opinion mining and that the system can be further extended with the aid of reviews from additional websites in order to decide the most efficient test classifier [6].

In research conducted by Francis F. Balahadia, Ma. Corazon G. Fernando and Irish C. Juanatas, they have proposed a performance evaluation tool for teachers with the aid of sentiment analysis. The system uses a dataset polarity created by the authors which contains a list of dictionaries of words. Using the Naïve Bayes algorithm, opinions have been extracted and analyzed in order to determine the polarity. As future work, the authors have intended to add a spelling and grammar checker to the system and also so enlarge the dataset dictionary they have created in order to improve the classification [7].

Through the literature review, it was identified that automatic or hybrid approaches prove to depict higher accuracy levels in comparison with the rule-based approach. Hence, a hybrid approach combining both machine learning algorithms and lexical resources was opted for the proposed system.

III. METHODOLOGY

The main goal of the system is to perform aspect-based opinion mining on the hair care product reviews. Since hair care products fall under various categories, the research specifically focused on hair products falling under the category of shampoo. Furthermore, due to shampoos being produced aiming different hair goals, there are numerous aspects to be considered. Therefore, four general aspects consisting of effectiveness, price, fragrance and formula have been opted manually after going through a range of reviews on hair shampoos. The system follows a sequential approach in order to achieve the objective. The entire system flow can be categorized into four major components including Data Gathering, Pre-Processing, Aspect Extraction and Polarity Detection. Fig. 1. illustrates a detailed process of the system architecture.

A. Data gathering

The preliminary step was to gather data or to collect consumer reviews, which act as the initial inputs to the system. In order to collect the reviews, the system facilitates users to enter a URL of an eBay product page. Using the provided URL, the system performs HTML parsing on the webpage with the aid of JSOUP library [8]. The system is capable of traversing through different pages of reviews and extracting all reviews on the product. The collected review data is then written to a CSV file. The system only collects the name of the person who wrote the review, the subject of the review and the review body on which aspect-based sentiment analysis is performed.

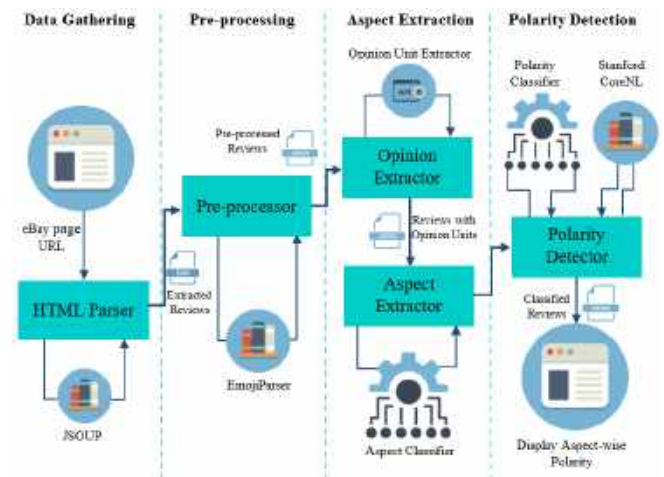


Fig. 1. High level diagram

B. Pre-processing

Once the data gathering stage is completed, the next stage is the pre-processing. The CSV file containing the reviews extracted from the data gathering stage is passed to the pre-processing module as the input. Since most of the consumer reviews tend to contain unnecessary noise such as smileys, emoticons, and capitalized words, which makes the reviews difficult to analyze, pre-processing is an essential step. The first step of the pre-processing module is to check whether the review is written in English. In case of it being written in any other language, that review is not taken for further processing. A regular expression is used to check whether the review is written in English language. Unnecessary special characters are eliminated and needlessly capitalized words are converted into simple letters. With the aid of EmojiParser library, emojis included within the review are removed [9]. Moreover, emoticons within the review are also eliminated by using a regular expression pattern.

C. Opinion extraction

Since a single sentence of a review can contain information on multiple aspects, opinion units within a sentence of a review are broken down with the aid of opinion unit extracting API provided by MonkeyLearn [10]. The API takes the text as an input and returns a JSON object containing the opinion units within the given text. The data is then written into a JSON file.

D. Aspect extraction

At the completion of pre-processing and opinion extraction, the extraction of aspects is started. In order to extract the aspects out of the opinion units, as the initial step, an aspect classification model has been created using Support Vector Machine (SVM). A csv file consisting of a list of 200 pre-processed reviews on shampoo has been used to train the model. The training data has been tagged with four manually defined aspects consisting of effectiveness, price, fragrance and formula. The aspect classification model has been tested using both Support Vector Machine and Multinomial Naïve Bayes algorithms since most of the related work has been done utilizing the two algorithms. SVM showcased a higher accuracy level, thereby SVM being finalized as the classification algorithm for the model.

E. Polarity detection

Following the extraction of aspects, the opinion units are passed to the Polarity Detection module. The modules utilize two mechanisms in identifying the polarity of an opinion unit. Initially, the system uses Stanford CoreNLP library in order to identify the polarity of an opinion unit [11]. By taking the opinion unit as an input, Stanford CoreNLP predicts the polarity according to fine-grained sentiment analysis. In other words, it predicts the polarity as very positive, positive, neutral, negative and very negative.

The system then uses the output given by a custom sentiment analysis model created using Support Vector Machine (SVM). The sentiment analysis polarity classification model has been created beforehand, and it has been trained using the same review dataset which was used to train the aspect classification model. In order to train the model, each review sentence has been manually tagged as positive, neutral or negative. Once again, the model has been created using both SVM and Support Vector Machine and Multinomial Naïve Bayes algorithms and SVM has been opted as the finalized algorithm due to it having a higher accuracy level for the test data used.

The two predictions from Stanford CoreNLP and polarity classification model are compared and a final polarity value is decided. The reason for using two predictions is due to Stanford CoreNLP being a general sentiment analysis library, at times providing incorrect outputs when certain hair care domain-specific words are used. Therefore, The JSON file containing the reviews along with the opinion units is then updated to include the aspect as well as the polarity for each opinion unit. With the assistance of the updated JSON file, the results are graphically represented through a JSP page by using web technologies like Ajax, jQuery and Bootstrap.

IV. RESULTS

The system was evaluated by providing multiple eBay product pages having single page reviews as well as multiple page reviews as inputs. The system proves to be able to clearly display aspect-wise polarity for the selected aspects for different opinion units of the reviews on shampoos.

The intended outcome of the initial data gathering process was a CSV file containing the author, subject and content of each review written for a shampoo listed on eBay. The results demonstrate that the system is capable of extracting the aforementioned information not only for single-paged reviews but also for multiple-paged reviews by using the product page URL. Fig. 2. demonstrates a sample CSV file output generated by the data gathering module.

sanford1997	Nice	It's a nice cowash. My curls/strands are left soft, clean and defined. I can finger de
rose_ridbet	Brilliant!	I have very dry, very thick curly hair and I am always having trouble finding the
January-austin	Yummy	I love using this cleansing conditioner. It smells yum & although it doesn't lather
robbydu-80	Hair felt clean	I was worried that co-washing for the first time would leave my hair feeling dry
karlaedison1	Perfect product to improve con	My hair has improved already! I am cutting down on conventional shampoo. I
super_12_vt	Perfect	Delivery was quick. I was surprised how much product you get, so great value!
clarklarbha	Washes well hair and doesn't d	It's perfect for co-wash. It really cleans the hair well, the fragrance is awesome
dorota6074	Just great!!	I love it for my daughter's African hair
kindr25	very good cowash	Perfect for my wavy hair! My hair is clean and soft and shiny after using it.
markele23488	Love this sooooo much a	
laurellewin	Conditions my hair beautifully!	
artist_pattier	Good product	Good cowash - protein free for curly girls
looney75	Feels like a much more natural	Leaves my hair feeling clean, healthier and thicker - I have fine dry hair.
lshidhr	Good for dry/dc	Excellent product for cleansing my Afro hair. Cleanses well without drying, sm
marseedak	Perfect cleanser	Made my hair silky and shiny.
tim_py	okay	okay

Fig. 2. CSV containing the reviews

Using the review data written to the CSV file, the pre-processing stage is capable of successfully identifying whether the review is written in English language, eliminating unnecessary special characters, converting needlessly capitalized words to simple and removing emojis and emoticons within the review. Fig. 3. shows how a sentence is pre-processed, completing the aforementioned tasks.

Before pre-processing : LOVE it! :) ☺
After pre-processing : Love it.

Fig. 3. A pre-processed review

The pre-processed data is then passed on for the extraction of opinions. As illustrated by Fig. 4, opinion units are separated within the sentence and data is written into a JSON file.

Reading the JSON file containing the opinion units for each review, aspects are identified using the aspect classification model. Then, the polarity for each opinion word is identified by combining the outputs of Stanford CoreNLP and polarity classification model created. Fig. 5. demonstrates the data written to the JSON file at the completion of aspect extraction and polarity detection.

```

"opinions": [
  {
    "opinion_unit": "I have very dry, very thick curly hair and i an always havin",
  },
  {
    "opinion_unit": "although it doesn't soap up in your hair like normal shampoo",
  },
  {
    "opinion_unit": "my curls are more bouncy and i don't need to put so much ser",
  },
  {
    "opinion_unit": "i will carry on using cleansing conditioners from now on.",
  },
  {
    "opinion_unit": "penc changer.",
  },
  {
    "opinion_unit": "want to now try their conditioners and other products."
  }
]

```

Fig. 4. Opinion units

```

{
  "opinion_unit": {
    "classifications": [
      {
        "category_id": 122849811,
        "probability": 0.744,
        "confidence": 0.967,
        "label": "price"
      }
    ]
  },
  "text": "i was surprised how much product you get, so great value for money.",
  "polarity": "POSITIVE"
}

```

Fig 5. Classified opinion units

Using the updated JSON file, the system is capable of graphically representing the results clearly as shown in Fig. 6. Tagged aspects are displayed for each review, denoted by the colours of green, yellow and red for positive, neutral and negative polarities respectively. The users have the ability to filter the reviews either by aspects or by aspect and the polarity.

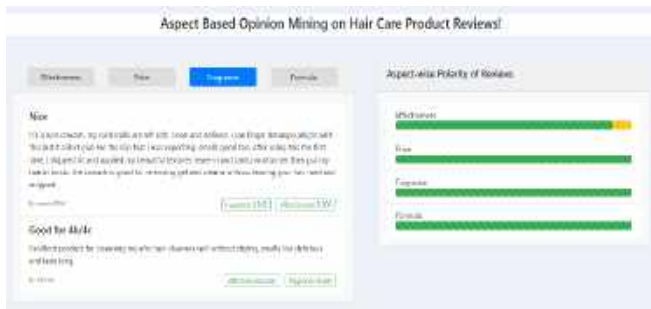


Fig. 6. Graphical representation of the output

V. DISCUSSION

The focus of the research was to decipher different public opinions on various aspects of hair shampoos expressed via reviews. Displaying the aspect-wise polarity was aimed since it is beneficial for both the manufacturers and consumers in gaining insights into consumer opinions and making purchase decisions respectively. Thus, the results of the system demonstrate that it is capable of dynamically displaying the aspect-wise polarity of reviews written on a given eBay product page.

The results received through test runs demonstrate that the system is capable of producing fairly accurate outputs for the given aspects. However, it takes a considerable amount of several seconds to produce the output due to the processes of web scraping, writing to and reading from files and the API calls involved as the number of reviews increases. Moreover, since it is a general sentiment analysis library, the Stanford CoreNLP library returns 'NEUTRAL' for certain opinion units like 'It makes my hair feel moisturized' whereas in an actual case, it contains a positive opinion. Therefore, the problem has been solved by integrating the polarity returned from the custom polarity classification model into the Stanford CoreNLP output. After the integration, the system proves to be able to accurately predict such a sentence as 'POSITIVE' as illustrated by Fig. 7.

Sentence : It makes my hair feel moisturized
Polarity from Stanford CoreNLP : NEUTRAL
Polarity from Classification Model : Positive

Fig. 7. Polarity prediction by CoreNLP and custom polarity classifier

However, the aspect classification model demonstrates to be incapable of correctly identifying sentences related to the aspect formula. As shown in Fig. 8., the phrase 'protein-free' refers to the aspect of the formula, but the classifier has only been able to identify the aspect of effectiveness denoted through the phrase 'Good cowash'.

```
"opinion_unit": {
  "classifications": [
    {
      "category_id": 122849862,
      "probability": 0.614,
      "confidence": 0.958,
      "label": "effectiveness"
    }
  ],
  "text": "Good cowash - protein free for curly girls",
  "polarity": "POSITIVE"
}
```

Fig. 8. Aspect prediction by custom aspect classifier

VI. CONCLUSION

The main objective of the research was to decipher the aspect-wise polarity of the hair care product reviews so that, it can be useful for both hair product manufacturers and consumers. Manufacturers can identify their strong and weak points in order to make future decisions and change products to fit the requirement of the majority of the consumers.

Consumers can get an at a glance idea about the public opinion on a certain hair goal they intend to achieve by purchasing the product. In order to achieve the aforementioned objective, the initial step was to study and analyze related work and different mechanisms used in sentiment analysis. Using the knowledge gained through the literature review, methods for data collection, pre-processing, aspect extraction and polarity detection were derived. Using the HTML parser library JSOUP, reviews were dynamically collected by taking an eBay product page URL as the input. The reviews were pre-processed by removing unnecessary special characters, emoji and emoticons. Opinion units were broken down from review sentences and two classification models were created to classify the aspect and polarity respectively using SVM algorithm. After training the models, the aspect-wise polarity of the hair care product reviews was determined and graphically represented using a web page as the final output. The system can be further improved to support more aspects of a hair care product. The research is based on manually selected aspects for products falling under the category of shampoo. Hence, a mechanism of dynamically detecting the aspects would be ideal a future improvement for the system. Moreover, the accuracy of the system in terms of both aspect and polarity can be enhanced with the aid of more training data.

REFERENCES

- [1] S. M. Singh and N. Mishra, "Aspect based Opinion Mining for Mobile," *2nd International Conference on Next Generation Computing Technologies (NGCT-2016)*, Dehradun, India, 2016.
- [2] P. Ray and A. Chakrabarti, "Twitter Sentiment Analysis for Product Review Using Lexicon Method," *International Conference on Data Management, Analytics and Innovation (ICDMAI)* Zeal Education Society, Pune, India, Feb 24-26, 2017.
- [3] I. K. C. U. Perera and H. Caldera, "Aspect Based Opinion Mining on Restaurant Reviews," *2nd IEEE International Conference on Computational Intelligence and Applications*, 2017.
- [4] Z. Singla, S. Randhawa and S. Jain, "Statistical and Sentiment Analysis of Consumer Product Reviews," *8th ICCNT 2017*, IEEE - 40222, July 3-5, 2017, IIT Delhi, Delhi, India, Delhi, 2017.

- [5] V. Dhanalakshmi, B. Dhivya and A. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms", *3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, Muscat, Oman, 2016, pp. 1 - 5.
- [6] K. L. S. Kumar, J. Desai and J. Majumdar, "Opinion Mining and Sentiment Analysis on Online." *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2016.
- [7] F. Balahadia, M. Fernando and I. Juanatas, "Teacher's Performance Evaluation Tool Using Opinion Mining with Sentiment Analysis", *IEEE Region 10 Symposium (TENSYP)*, Bali, Indonesia, 2016, pp. 95 - 98.
- [8] J. Hedley, "jsoup Java HTML Parser, with best of DOM, CSS, and jquery", *Jsoup.org*, 2019. [Online]. Available: <https://jsoup.org/>. [Accessed: 12- Dec- 2019].
- [9] "vdmont/emoji-java", *GitHub*, 2019. [Online]. Available: <https://github.com/vdmont/emoji-java>. [Accessed: 12- Dec- 2019].
- [10] "MonkeyLearn", *App.monkeylearn.com*, 2019. [Online]. Available: https://app.monkeylearn.com/main/extractors/ex_N4aFcea3/. [Accessed: 12- Dec- 2019].
- [11] Manning, D. Christopher, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit" in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55-60

Demystifying the concept of IoT enabled gamification in retail marketing: An exploratory study

Shalini Jayasooriya*

Department of Business Management
Sri Lanka Institute of Information
Technology, Sri Lanka
jayasooriya.shalini@gmail.com

Tharindhie Alles

Department of Information Management
Sri Lanka Institute of Information
Technology, Sri Lanka
tharindhie@gmail.com

Samantha Thelijagoda

SLIIT Business School
Sri Lanka Institute of Information
Technology, Sri Lanka
samantha.t@slit.lk

Abstract: The retail landscape is evolving rapidly as firms embrace innovative technologies in an attempt to stay ahead of the aggressive competition prevalent within the industry. Gamification is one such innovative technology that has been gaining popularity in recent times. This study aims to explore the application of Gamification in the context of Retail Marketing in Sri Lanka and ultimately propose a concept for a Gamified application that can be used by customers of modern-trade retailers. The study took an exploratory qualitative approach where intensive surveys of literature and in-depth interviews with a judgmental purposive sample of seven marketing professionals in the modern-trade retail industry were conducted to determine the current play of technology in retail marketing as well as the drivers & challenges of Gamification adoption. Further, in-depth interviews with the customers of such organizations were conducted in gathering user preferences and design recommendations for a Gamified app. Thematic analysis was carried out in deriving insights. Findings show that the retail firms currently employ several technologies in line with those discussed in existing literature such as loyalty card systems, digital signage, VR technologies, online Gamification amidst others in carrying out their marketing efforts. Gamification is predominantly applied in the online context as opposed to the offline (in-store) context. Furthermore, the key drivers that propel firms to implement novel technology like Gamification are to generate customer insights, enhance customer experience and achieve marketing related KPI targets. Conversely, inadequate technology infrastructure, justifying the focus on a niche crowd of tech-oriented customers and slow ROI pose as challenges in the process of Gamification adoption. Three main themes emerged upon exploring user preferences and design recommendations for a Gamified app and are identified as information at the touch of a fingerprint, automation & integration and use of game mechanics. Ultimately by incorporating these insights gathered, a concept for a Gamified app was proposed.

Keywords: Gamification, Marketing, Modern-trade retail, Technology

I. INTRODUCTION

The wireless sensor network technology has infiltrated many areas of our modern-day lives and the proliferation of each of these devices has successfully created the web of IoT [1]. Gamification deems to be one such component that is enabled due to the implications of IoT. This term in its most rudimentary essence elucidates how certain tasks are converted to be more game-like to motivate entities to complete it. It is also conferred that Gamification is “the use

of elements of game design in a non-gaming context” [2] and when properly articulated and utilized it can aid in achieving multiple advantages: improvement of processes, positive psychology and emotional affirmation, amplified relationships, increased engagement and producing feelings of accomplishment within individuals [3]. Gamification techniques comprise elements like; competitions, rankings, scoring schemes and incentives, and these act as a means to build brand loyalty, form connections and encourages consumers to return to the brand [4].

The current context of a retail environment is rather challenging, and marketers engage constantly in the pursuit of novel ways to offer customers a unique shopping journey and encourage traffic to physical outlets. Customers should be engaged through both emotional and behavioral approaches [5] and thus, the methods used to increase customer interest should be rich in emotional appeal as well as encourage interactions with the brands. In fulfilling this challenge, Gamification has joined the phenomena of artificial intelligence, big data, and crowdsourcing as a contemporary megatrend.

In the realm of marketing, with the use of this concept, marketers can compile customer-oriented games that propel firms closer to their aims such as boosting their brand image, an engagement that augments a loyal customer base. A few noteworthy and longstanding examples of Gamification are:

- Victoria Secret PINK Nation App: This is based on the millennial group and consumers can win offers just by playing concise games within the app.
- Bonobos’ Scavenger Hunt: This apparel eCommerce website initiated a digital scavenger hunt in their site—customers had to click on images and in turn, received thrilling rewards.
- NikeFuel Program: Nike was able to capitalize on their customers’ competitive nature and released a type of Gamification mechanism where users should post their best times, track progress and challenge peers to meet goals and enhance fitness and for which they all received early access to Nike products, free shipping and many more [6].

The concept of Gamification had been researched and tested by many scholars however there are very few researches concerning retail marketing and service marketing aspects, whereas the results derived through these stands to be

questionable in the context of a developing country like Sri Lanka. Hence, this research aims to provide theoretical contributions to the dearth of literature underlying the knowledge gap on the prospects of leveraging Gamification in retail marketing in Sri Lanka.

Before exploring the application of Gamification, in establishing an understanding of the existing play of technology within the retail industry, an investigation will be carried out on the current technologies employed by retail firms in the execution of their marketing efforts. Subsequently, this paper focuses on the drivers that motivate organizations to employ novel technologies like Gamification as well as those which emerge as challenges in the adoption process.

It is predicted that over 80% of Gamified applications are bound to fail due to poor designs incorporated [7]. In addressing this, the latter part of the paper focuses on determining the user preferences and design recommendations for a Gamified application in the retail context and drawing upon these findings, ultimately a concept for a Gamified application is put forth.

II. LITERATURE REVIEW

A. Current play of technology in the retail environment

Amidst the plethora of technology utilized by retailers Loyalty cards systems stands out as one of the most fundamental technologies and discussed in detail in the literature. Loyalty programs were established as a mechanism that offers incentives to regular customers intending to garner customer loyalty towards a particular retailer or brand [8]. Further, integrating Loyalty card systems to Electronic Point of Sale digital systems (EPOS) has enabled retailers to identify who bought what; when and where they bought it, which in turn capacitate retailers to merge their merchandising practices with customers buying behavior [9].

With the rising popularity of the World Wide Web, the world has infiltrated into an era where a substantial proportion of marketing communications are executed across the digital space, whereby the online presence of companies has become critical. Online channels of retailing such as websites enhance existing relationships with customers while complementing the offline, i.e. physical stores and thus firms being present online has become a strategic need nowadays [10]. Moreover, in the recent couple of decades, marketers' primarily used e-mail blasts, telemarketing, websites among other mechanisms to publicize a firm-specific or product-specific information, at present firms are shifting their presence in social media- such as Facebook, Instagram, Youtube- since customers are increasingly engaging over these platforms [11].

Moreover, the rise in the use of mobile phones has encouraged marketers to use mobile technologies in contacting customers about product information and promotions which had led to the coining of the term "Push based mobile marketing" referring to advertisements getting pushed to customer's mobile devices. Leveraging on the high penetration rates of mobile networks and the low costs associated with the medium, SMS marketing which preceded more recent Mobile Instant messaging (MIM) technologies like Whatsapp, Viber is predominantly in use [12].

The move to smartphones has revolutionized the shopping experience even further. While initially offering basic mobile apps, the efforts are now shifting towards integrating targeted offers via geo-fencing and scan-and-go-technologies whereby customers can scan products as they shop and proceed to make payments via the retailer's mobile app [13]. Gamification: the main focus of this paper, joins the recent developments in technology which has leveraged the popularity of mobile phones. Gameful experiences motivate particular behaviors such as making purchases, repeat purchases, and engagement and this has encouraged retailers into integrating Gamified approaches into their activities. It is predicted that by 2020, 87% of retailers anticipate incorporating Gamification features to their loyalty programs [14]. Probing on alternative means by which customer engagement can be achieved is of vital importance for the success of particularly online retailers' digital strategy. Gamification will facilitate in this, as it repositions the shopping experience as entertainment in an individual's cognition [15]. Thus, firms can leverage on Gamification to position the shopping experience as an enjoyable task within their customers, thereby attracting more customers to its stores.

Digital displays are used for advertising in-store by large retail chains to improve the in-store environment, promoting specific products and displaying other information. It is considered that Digital Displays/Signage augment approach behaviors of customers such as visiting frequency, the likelihood of revisits within a given time and likelihood of spending [16]. Moreover, Query-based Artificial Intelligence (AI) systems have been adopted whereby customer inquiries such as those on product functionalities and complementary product suggestions are being met by AI systems. The responses are tailored to complement a customer's existing knowledge through the analysis of historical purchase data followed by predictive analysis and this enables customers to be more informed and engaged [13].

Further, interactions between humans and the environment around them have altered with the development of interactive technologies such as Augmented Reality. AR technology integrates computer-generated graphics with the real world. Interacting and being occupied within holographic environments- a form of AR- offers customers a unique and entertaining experience [17]. Thus, in the retailing context, the technology is been utilized in presenting three-dimensional images of products to assist the customers with their purchase decisions.

B. IoT and Gamification

The paradigm of the Internet of Things (IoT) in layman terms is a system of physical objects embedded with sensor technology, electronics, connectivity and varied software to enable it to perform exceedingly well by the exchange of information with other connected devices [1]. The potential of IoT extends to aid in achieving goals and improvements in decision making through augmented intelligence, enhancing process optimization and efficiency via the vast amount of data collected through IoT devices and making improvements in the management & efficient flow of information [18].

The perception of Gamification has strong divided views – some deem it as a novel term for traditional marketing tools or rather as a modern method to exploit customers whilst others regard it as a sincere means employed by firms to

enhance value addition to customers [19]. Gamification in marketing refers to the application of game design rudiments to marketing activities, to produce a game-like experience for customers thereby enhancing customer service value and achieving optimum value creation [20]. Due to the continual improvements and advancements in technology, customers too are heavily selective of how they desire to spend their time and money and this, in turn, coerces companies to invest in various methodologies to keep them engaged [21]. As such, Gamification is one such strategy used by firms as they envision Gamified marketing activities to be an effective form of promoting products and branding, while also providing the customers a fun experience [4], as previously highlighted as well.

Literature also converses on the Real-world applications of Gamified efforts. An example of such a Gamified effort stands out to be a program designed by Alipay- whereby users accrue energy through means of walking, consuming and online payment and trades the collected energy to trees in the desert. This Gamified process presents users with a sense of accomplishment which in turn makes them preferentially use Alipay to collect energy and thus form a sense of devotedness towards the brand [22]. In terms of In-store Gamification, it is emphasized how Gamified in-store applications for customers enable retailers to combat against revenue denigration (Showrooming amid other retail challenges) [23]. In the same study, the scholars initiated and proposed a system where shoppers obtain a VR headset and a mobile phone running a Smart Phone-based Mixed Reality Application (SPMRA) through which they can walk around the store and view the environment in an augmented manner, thus observing more than what is physically present. This solution received positive feedback from both shoppers and the retail stores as it increased the engagement and the fun factor. Besides, Gamification has effects on the inclination to act on offers gained at the point of purchase through eye-tracking, smartphone activity logging and choice technical apparatuses [14].

C. Gamification elements

In referring to the literature under this segment, one can find varied means to classify the elements of Gamification: DMC framework, MDA framework, FDD framework, Elemental tetrad among multiple others. The more commonly known Gamification elements are points, badges, and leaderboards and stand out as a useful initiation point for Gamified efforts [24]. As per the same scholars, the DMC framework is the most widely used and it incorporates three main echelons; Dynamics (promote Gamification behaviors of users- emotions, narratives, constraints, relationships, and progression), Mechanics (used to drive Gamification process and the participation of users through competition, chances, challenges, corporation, rewards, acquiring resources and feedback) and Components (achievements, points, badges, leaderboards) [25].

Gamification elements provide an emotional experience for customers and lead to the completion of tasks and subsequent behaviors [26]. An illustration of this would be how customer purchasing behavior is manipulated by the element of the challenge during Gamified shopping; for example, unlocking an offer by completing a set challenge, successfully increases the preferences to act on that particular offer [27].

Further, in implementing a successful Gamified effort three main aspects are to be considered: (1) understand target audience profoundly, (2) determine the objectives of the system in a systematic manner and (3) utilize proper and appropriate game elements to motivate players as that is the overall purpose of using game elements in a non-gaming context. [28].

III. METHODOLOGY

Qualitative studies are appropriate for research on little-known phenomena or innovative systems [29], and thus a qualitative exploratory approach was taken considering the lack of literature and the novelty of the concept of Gamification within the Sri Lankan Context.

Considering the numerous retail stores available across several industries, the scope of this research was on modern-trade retailers, specifically, FMCG supermarket retailers in Sri Lanka as the type of Gamified app ultimately proposed in the study is most suitable to be introduced in such stores (easiest to set up the technological infrastructure needed) and as it's more likely that tech-savvy customers for whom this type of app is better suited for would visit these stores. The scope was limited to the modern-trade retail organizations in Sri Lanka.

Data collection was carried out through semi-structured interviews since this method support in gathering not only first-hand valid and reliable data, but it will also assist in refining ideas [30], which was vital considering the nature of the study.

Judgmental purposive sampling method was employed and in gathering insights about the current play of technology and Gamification, considering selection criteria of a minimum of five-year experience in marketing within the modern-trade retail industry and the potential interviewee's job role, a sample of seven marketing professionals was taken till theoretical saturation was reached. Open-ended questions about the current technologies used by the firms, those which they hope to use in the future as well as the purpose of using each of these technologies were probed on.

Similarly, in gathering insights on user preferences and design recommendations for a Gamified application in the retail context a purposive sample of 24 respondents was taken until theoretical saturation was reached. The selection criteria for the respondents were determined as being a customer of any of the above modern trade retail stores and those who are using mobile applications, specifically those incorporated with Gamification mechanisms. Open-ended questions regarding overall experience with the mobile apps (likes, dislikes, challenges, etc.), user expectations of mobile apps, potential features that generate interest to use Gamified apps specifically designed to be used in retail environments and usage scenario preferences (in-store/ offline app).

TABLE I. CURRENT TECHNOLOGIES EMPLOYED AND THEIR USE OBJECTIVES

Technology	Use Objective
Loyalty card systems	Encouraging customer loyalty to the store. Generating customer data (specifically buying patterns, contact information) which in turn is used to send targeted communications.
E-Commerce	As a channel of service delivery. For communications (offers and promotions, service offering, etc.)
Social media	For communications. Driving leads to the website. Driving customer engagement.
E-mails	For communications.
Digital tabloids	For communications (related to product offerings and offers) to drive leads to website/ encourage store visits.
SMS	For targeted communications to drive sales (customers segmented as high spenders/low spenders/lapse customers).
Digital Signage	The medium of product/brand advertising.
TV systems	The medium of product/brand advertising.
Checkout couponing	Presenting customized offers to encourage repeat visits to the store.
Online Gamification	Driving engagement.
VR technologies	Driving customer engagement.

Data gathered were analyzed through content thematic analysis technique using the NVivo 12 Plus Software as a facilitating tool. An inductive approach (absence of a predetermined structure or framework in data analysis), was taken as it is widely used when little is known regarding the research phenomenon [31]. A seven-tier process was followed in conducting thematic analysis laid out as Data Organization, Immersion in data, Generation of case summaries and potential themes/ categories, Data coding, presenting interpretation via analytic memos, exploring of alternative interpretations and lastly the report writing or other forms of presentation of results [29]. Themes and categories derived were either theory-generated codes (codes which are derived from previous literature) or in-vivo codes (the codes which arise from the data collected).

The process concluded with the findings being compared and contextualized to that of existing literature. In presenting the final concept for the Gamified app, Flowcharts and System Architecture diagrams were compiled.

IV. RESULTS

A. Current technology usage in retail marketing

Technology corresponding to those laid out in the prevailing literature is employed in achieving multiple marketing objectives by the retail organizations as detailed in Table I. Loyalty programs have evolved over the years to the

most contemporary format being coined with the term Loyalty 3.0. Loyalty 1.0, the most basic format of the program comprised of frequent-flyer programs, cash-back credit cards, and punch cards while Loyalty 2.0 focuses on customer segmentation through data available from Loyalty 1.0 programs and directing targeted direct mail and emails in generating customizable customer experience. Loyalty 3.0 builds upon major building blocks- human motivation, big data and Gamification [32]. At present, Sri Lanka is implementing Loyalty 2.0 programs with targeted offers sent via SMS, Facebook or Checkout couponing, leveraging on the information such as purchase history captured through Loyalty programs. Implementing the concept for a Gamified app put forth ultimately in this paper will advance firms towards Loyalty 3.0 programs. Furthermore, some modern-trade retail organizations belonging to diversified conglomerates were in the process of evolving into “group

loyalty” which allowed them to gather data on a specific customer across multiple domains such as healthcare, financial services and leisure which will enable them to further refine their targeted marketing activities. Referring to Gamification, all the professionals were acquainted with the term and its applications, with certain firms already incorporating Gamification into their marketing efforts.

A noteworthy illustration of such Gamified efforts is the Keells Supermarket chain’s (a leading player in the modern trade retail sector in Sri Lanka) launch of a free-to-play game named “Keells Power shopper” in 2018 which is accessible on iOS and Android devices and reportedly have been played by 10,000 players [33]. This is an online form of Gamification where the players engage in a simple arcade-style adventure game with the player’s Avatar running within an environment that is similar to the Keells supermarket collecting power-ups which in turn allows them to earn points. As such it employs game mechanisms of points, leaderboards, and rankings, avatars with rewards of PlayStation 4 gaming consoles being awarded to players who have scored the highest points at the end of a certain week. However, the game was launched within a separate gaming platform which had 50 other mini-games and 25,000 active users, and a player would have to download the gaming portal before playing the game. This was seen to be a massive negative point consumers had faced which had resulted in the game being not exceedingly successful.

As such it should be noted that Gamification is mainly applied in the online context and in-store Gamification is minimal: typically executed in collaboration with FMCG brands like that of Unilever Sri Lanka. Further, Mobile Instant Messaging and Autonomous shopping carts emerged as two technologies that are expected to be employed shortly. On probing about the use of Hologram technologies and In-store Gamification, it was revealed that firms are receptive to the idea, yet faces several challenges in the implementation as expelled out ahead in this paper. The use objectives of potentially employing each of these are laid out in Table II.

B. Drivers for Gamification adoption

Further, analysis of data exposed three recurrent themes which drive retail organizations to use technology as a strategic choice as described below:

1) *To generate customer insights:*

Unorthodox way of observing buying behavior, tracking the movement of customers and these insights will act as a valuable input to the decision-making processes, facilitate in designing more appealing offers/ sending targeted messages and make internal adjustments within stores.

TABLE II. TECHNOLOGIES TO BE EMPLOYED IN THE FUTURE AND THEIR USE OBJECTIVES

Technology	Use Objective
Mobile Instant Messaging	As an alternative/complement to the use of SMS.
Autonomous shopping carts	Gathering customer insights For targeted communications (specifically for offers intended at upselling, brand switching)
Holograms	Driving Customer Engagement
In-store Gamification	Driving Customer Engagement For Communications (specifically for offers)

2) *To enhance customer experience:*

Innovative mechanisms are essential in satisfying the contemporary customers' segments, especially the younger generations: articulates upon how generation Y consumers think differently and are always seeking an environment which is relaxed, flexible, interactive and engaging as opposed to those of other generations like baby boomers [34].

3) *To achieve marketing KPI targets:*

As with any other department within a firm, the marketing department has its own set of goals and objectives to be achieved such as building brand awareness, engagement, generating qualified leads and their respective Key Performance Indicator (KPIs) targets. Thus, if an objective relates to creating engagement or capturing the top of mind within a customer, disruptive technologies like Gamification is resorted to.

C. *Challenges for Gamification adoption*

Three recurrent themes emanated as the challenges faced by firms in the process of adopting digital initiatives like Gamification.

1) *Slow return on investments:*

Although retail revenues are high, profit margins were reported to be low thus making firms focus more on ROIs. As ROIs are not brought immediately in digital platforms, top managements are reluctant to spend on digital media/technology. However, interestingly some other firms were of an alternative view: that blurred ROIs are not of concern when focussing on creating a value proposition and cutting down on the Above –The-Line budgets (Television/radio Advertisements, etc.) to invest on more innovative platforms is worthwhile.

2) *Niche target crowd:*

Gamification is still at its onset, bringing about the challenge, especially within well-established retail trade organizations whose focus is on mainstream markets, to convince the top management about catering to a niche market segment which Gamified apps will appeal to.

3) *Technology infrastructure readiness:*

Certain firms believed that they were not ready in terms of the infrastructure necessary to implement and deliver customers a seamless experience, specifically in terms of in-store Gamification. Conversely, other firms declared that the

required infrastructure does not necessarily pose a challenge. Interestingly these were retailers who were late movers to the industry and focusing on leveraging technology in gaining a competitive advantage as opposed to storing location and number factor which more bigger players focussed on.

D. *User preferences and design recommendations for a Gamified application in the retail context*

When exploring on preferences and design recommendations for a Gamified app, analysis of data materialized three main themes that users focused on. The codes for the first two themes were derived from a previous similar study [35] as the data gathered reflected the same, while the third theme was in line with the mechanic's aspect of the MDA framework [35]. These are detailed out below:

1) *Information at the touch of a fingerprint:*

A central, accessible resource for information regarding the promotions available, loyalty schemes offered, newly launched products and outlet information as and when needed.

2) *Automation and integration:*

Smart functions which can passively capture data, require less input and be highly relevant, responsive and engaging. For example functions like Smart shopping lists which recommend offers available for products relevant to the items in the compiled list, which can be also be shared with family members across devices. Regarding automation, apps that can automate the process of shopping; for example, queue-less shopping solutions with the use of QR code scanning were preferred.

3) *Game mechanics:*

Inclusion of Avatars, badges, challenges / quizzes to collect points which in turn offer rewards either as loyalty points or discount offers for products relevant to a customer.

E. *The concept for a Gamified app*

Drawing upon other insights and specifically, those classified under the theme of Game mechanics discussed above, the most rudimentary level of the concept design for a Gamified app was designed as illustrated in Fig. 1.

- Users preferred both an online version and a version to be played in-store.
- Online Mode: Similar to the Gaming apps of Temple run and Dialog Mega Run, this option allows a user to play a game from anyplace, whereby users can earn points that can be converted into loyalty card points.
- In-Store Mode: Users actively spoke of how engaging and paradigm-shifting the app Pokémon Go was. Thus similarly, the in-store mode-normal version allows consumers to hover over products or the aisle to catch hidden daily deals and other offers (facilitated by beacons). Further as suggested by users, a mode specially tailored to kids is also incorporated.

The notion to link the app with the loyalty card was to easily facilitate keeping in track of progress and hassle-free when checking out as opposed to a unique code generator that generates codes each time the game is played.

The above concept is depicted in a flowchart as shown below in Fig. 2 to aid visual clarity and further understand the process and logic of the system involved. Further, to get a final overview of the proposed App, the System Architecture diagram is presented in Fig. 3.

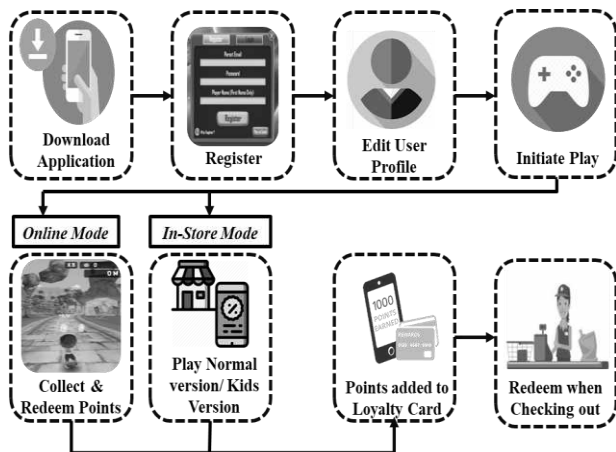


Fig.1.The basic concept for the proposed App

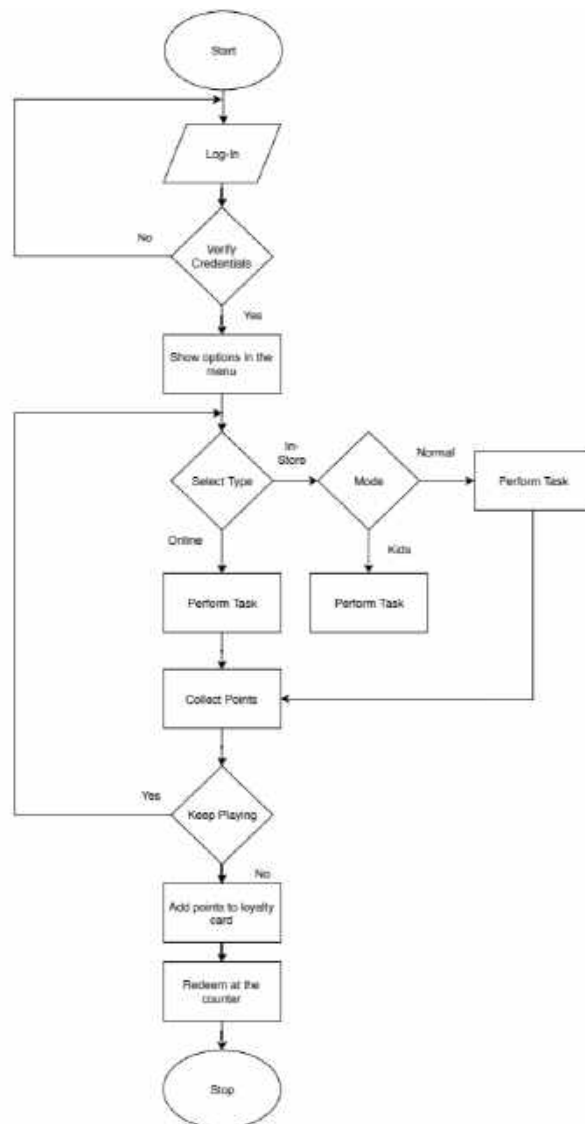


Fig.2. Basic concept for proposed App

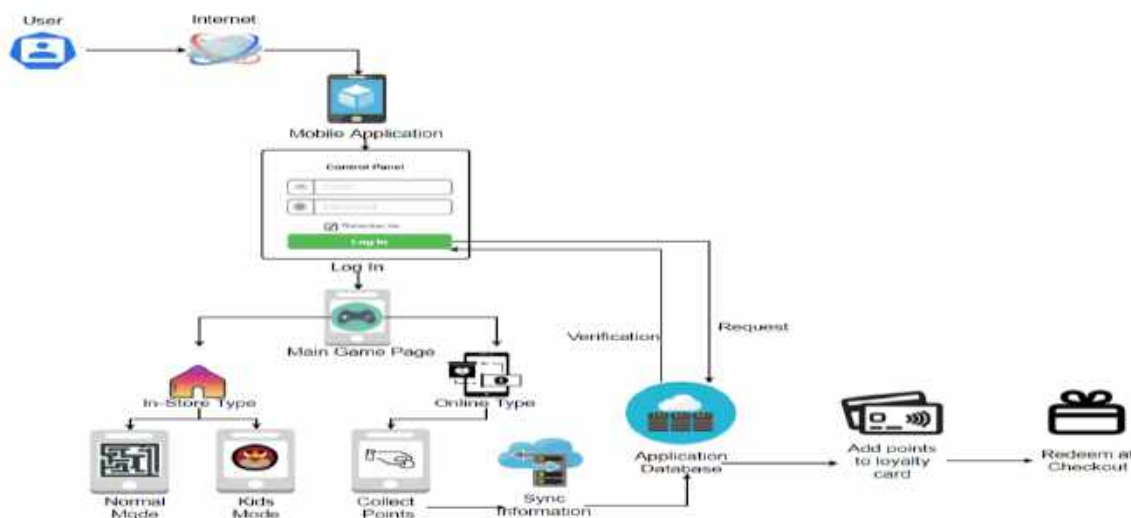


Fig. 3. System Architecture Diagram of proposed App

V. CONCLUSION

Upon capturing a holistic overview regarding the current technology used by retail marketers, the paper focused on the adoption of Gamification. Although the concept is at its infancy, it acts as an innovative way in offering customers a unique shopping experience within the highly competitive modern retail environment and despite that this mechanism is being resorted to by many retail giants across the globe, it is yet to be fully exploited within the Sri Lankan retail sector. Research has discovered that people make purchases such as food/alcohol at the locations they visit while playing games like Pokémon Go and therefore emphasize the possibility of monetization of location-based Gamified apps [36]. This accentuates the potential success of the Gamified solution put forth in this paper.

This research centered around the modern-trade retail context specifically because a mechanism which can stimulate enjoyment in an experience which is perceived to be a tedious task such as grocery shopping, seemed worthy of exploration. However, since the core of a Gamified experience is deriving entertainment, the concept might be more suitable for shopping experiences which likely to be more relaxing like that experienced when shopping for clothes. Thus, future research work can explore on the use of Gamification mechanisms in other formats of retail, specifically, Softline retail stores - those that sell goods generally having a consumable lifespan less than three years - such as apparel stores.

Further, because the exploratory nature of the study driving the use of a qualitative approach, as well as a relatively smaller sample size poses challenges in generalizing the results. While this stands as a limitation, it should be noted that this technique was deemed best fit in meeting the objectives of the study which required gathering data from information-rich cases. However, future research can focus on quantitative approaches where the findings of this research can be further verified and generalized.

In conclusion, the main points of discussion in the study: current play of technology within the retail marketing context, the overview of the drivers and the challenges behind adopting Gamification, the understanding of user preferences for a Gamified app together with the proposed concept for a Gamified app; equips the reader with the knowledge to initiate the process of Gamification in this previously untouched domain of retail marketing.

REFERENCES

- [1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions", *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645-1660, 2013. Available: 10.1016/j.future.2013.01.010.
- [2] Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness", *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, 2011. Available: 10.1145/2181037.2181040.
- [3] Park and S. Kim, "Patterns Among 754 Gamification Cases: Content Analysis for Gamification Development", *JMIR Serious Games*, vol. 6, no. 4, p. e11336, 2018.
- [4] Lucassen and S. Jansen, "Gamification in Consumer Marketing - Future or Fallacy?", *Procedia - Social and Behavioral Sciences*, vol. 148, pp. 194-202, 2014.
- [5] Sachdeva and S. Goel, "Retail store environment and customer experience: a paradigm", *Journal of Fashion Marketing and Management: An International Journal*, vol. 19, no. 3, pp. 290-298, 2015.
- [6] L. Ufford, "6 Examples of Retail Gamification to Boost Engagement and Sales", *Shopify*, 2017. [Online]. Available: <https://www.shopify.com/retail/6-examples-of-retail-gamification-to-boost-engagement-and-sales>. [Accessed: 21- Jul- 2019].
- [7] Burke, *Gamify: How Gamification Motivates People to Do Extraordinary Things*, 1st ed. New York: Routledge, 2014.
- [8] J. Bojei, C. Julian, C. Wel and Z. Ahmed, "The empirical link between relationship marketing tools and consumer retention in retail marketing", *Journal of Consumer Behaviour*, vol. 12, no. 3, pp. 171-181, 2013. Available: 10.1002/cb.1408.
- [9] [9]C. Hart, N. Doherty and F. Ellis-Chadwick, "Retailer adoption of the Internet - Implications for retail marketing", *European Journal of Marketing*, vol. 34, no. 8, pp. 954-974, 2000. Available: 10.1108/03090560010331441.
- [10] Toufaily, N. Souiden and R. Ladhari, "Consumer trust toward retail websites: Comparison between pure click and click-and-brick retailers", *Journal of Retailing and Consumer Services*, vol. 20, no. 6, pp. 538-548, 2013. Available: 10.1016/j.jretconser.2013.05.001.
- [11] M. Tiago and J. Verissimo, "Digital marketing and social media: Why bother?", *Business Horizons*, vol. 57, no. 6, pp. 703-708, 2014. Available: 10.1016/j.bushor.2014.07.002.
- [12] Khare and S. Rakesh, "Mobile marketing in Indian retail: a preliminary investigation of relationship and promotional endeavours through Short Message Service", *International Journal of Business Competition and Growth*, vol. 2, no. 2, 2012. Available: 10.1504/ijbcg.2012.046285.
- [13] Grewal, A. Roggeveen and J. Nordfält, "The Future of Retailing", *Journal of Retailing*, vol. 93, no. 1, pp. 1-6, 2017. Available: 10.1016/j.jretai.2016.12.008.
- [14] J. Högberg, P. Shams and E. Wästlund, "Gamified in-store mobile marketing: The mixed effect of gamified point-of-purchase advertising", *Journal of Retailing and Consumer Services*, vol. 50, pp. 298-304, 2019. Available: 10.1016/j.jretconser.2018.07.004.
- [15] Insley and D. Nunan, "Gamification and the online retail experience", *International Journal of Retail & Distribution Management*, vol. 42, no. 5, pp. 340-351, 2014. Available: 10.1108/ijrdm-01-2013-0030.
- [16] Roggeveen, J. Nordfält and D. Grewal, "Do Digital Displays Enhance Sales? Role of Retail Format and Message Content", *Journal of Retailing*, vol. 92, no. 1, pp. 122-131, 2016. Available: 10.1016/j.jretai.2015.08.001.
- [17] Poushneh, "Augmented reality in retail: A trade-off between user's control of access to personal information and augmentation quality", *Journal of Retailing and Consumer Services*, vol. 41, pp. 169-176, 2018. Available: 10.1016/j.jretconser.2017.12.010.
- [18] S. Fattah, N. Sung, I. Ahn, M. Ryu and J. Yun, "Building IoT Services for Aging in Place Using Standard-Based IoT Platforms and Heterogeneous IoT Products", *Sensors*, vol. 17, no. 10, p. 2311, 2017. Available: 10.3390/s17102311.
- [19] J. Hamari, "Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service", *Electronic Commerce Research and Applications*, vol. 12, no. 4, pp. 236-245, 2013. Available: 10.1016/j.elerap.2013.01.004.
- [20] R. Conaway and M. Garay, "Gamification and service marketing", *SpringerPlus*, vol. 3, no. 1, 2014. Available: 10.1186/2193-1801-3-653.
- [21] Y. Yang, Y. Asaad and Y. Dwivedi, "Examining the impact of gamification on intention of engagement and brand attitude in the marketing context", *Computers in Human Behavior*, vol. 73, pp. 459-469, 2017. Available: 10.1016/j.chb.2017.03.066.
- [22] P. Yang, T. Xu, Y. Feng, Y. Zhao and X. Wang, "The impact of gamification elements on the evaluation of marketing activities", in *International Conference on Electronic Business, China*, 2018.
- [23] L. Meegahapola and I. Perera, "Enhanced in-store shopping experience through smart phone based mixed reality application", *2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2017. Available: 10.1109/ictcr.2017.8257810.

- [24] K. Werbach and D. Hunter, *For the win: How Game Thinking Can Revolutionize Your Business*. Philadelphia: Wharton Digital Pr., 2012.
- [25] K. Werbach and D. Hunter, *The Gamification Toolkit: Dynamics, Mechanics, and Components for the Win*. Wharton School Press, 2015.
- [26] T. Harwood and T. Garry, "An investigation into gamification as a customer engagement experience environment", *Journal of Services Marketing*, vol. 29, no. 67, pp. 533-546, 2015. Available: 10.1108/jsm-01-2015-0045.
- [27] Jae Choi, "The Effect of Shopping Motives on Consumer's Emotion and Store Loyalty: Comparison with E-Shopping and In-Store Shopping", *Journal of Distribution and Management Research*, vol. 15, no. 1, pp. 23-33, 2012. Available: 10.17961/jdmr.15.1.201203.23.
- [28] Cheong, F. Cheong and J. Filippou, "Towards the Gamification of Learning: Investigating Student Perceptions of Game Elements", *Journal of Information Systems Education*, vol. 25, no. 3, 2014.
- [29] Marshall and G. Rossman, *Designing qualitative research*, 6th ed. SAGE Publications Inc., 2015.
- [30] M. Saunders, A. Thornhill and P. Lewis, *Research methods for business students*, 5th ed. Harlow: Pearson, 2009.
- [31] P. Burnard, P. Gill, K. Stewart, E. Treasure and B. Chadwick, "Analysing and presenting qualitative data", *British Dental Journal*, vol. 204, no. 8, pp. 429-432, 2008. Available: 10.1038/sj.bdj.2008.292.
- [32] Y. Andonova, "Loyalty 3.0: How big data and gamification are revolutionizing customer and employee engagement", *Journal of Marketing Analytics*, vol. 1, no. 4, pp. 234-236, 2013. Available: 10.1057/jma.2013.19.
- [33] "Arimac Digital launches 'Keells Power Shopper' - The Morning - Sri Lanka News", *The Morning - Sri Lanka News*, 2019. [Online]. Available: <http://www.themorning.lk/arimac-digital-launches-keells-power-shopper/>.
- [34] S. Poole, E. Kemp, K. Williams and L. Patterson, "Get Your Head in the Game: Using Gamification in Business Education to Connect with Generation Y", *Journal for Excellence in Business Education*, vol. 3, 2014.
- [35] M. Hilliard, A. Hahn, A. Ridge, M. Eakin and K. Riekert, "User Preferences and Design Recommendations for an mHealth App to Promote Cystic Fibrosis Self-Management", *JMIR mHealth and uHealth*, vol. 2, no. 4, p. e44, 2014. Available: 10.2196/mhealth.3599.
- [36] Colley et al., "The Geography of Pokémon GO", *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 2017. Available: 10.1145/3025453.3025495.

Author Index

A

Abeyasinghe, D.V.D.S.	216
Alahakoon, P.M.K.	210
Alles, T.	234
Anuradha, K.S.	7
Athree, M.	25

C

Chandrasekara, N.V.	166
Chathumali, E.J.A.P.C.	124

D

Dasanayaka, D.D.S.S.	45
De Silva, W.D.M.	31

E

Ekanayake, S.	136
---------------	-----

F

Fernando, A.	103
Fernando, A.I.R.	75
Fernando, K.D.M.	210

G

Ganegoda, G.U.	96, 129
Ganepola, G.A.D.	111
Gunarathna, P.S.T.K.	57

H

Harishchandra, N.N.	117
Herath, H.M.R.M.	210
Hirshan, R.	210

J

Jayabodhi, L.W.D.C.	68
Jayakodi, L.A.	117
Jayalal, S.	17, 31
Jayalath, S.A.	82
Jayalath, T.	37
Jayasinghe, J.	223
Jayasiri, A.	25
Jayasiriwardena, T.D.	129
Jayasooriya, S.	234
Jayasuriya, P.	136
Jayawardena, B.	51
Jayawickrama, J.A.D.T.	182
Jinarajadasa, G.M.	195

K

Kahawandala, N.	155
Kalansooriya, P.	111
Kalingamudali, S.R.D.	57
Kanishka, P.P.D.G.	57
Kothalawala, M.	228
Kumarasinghe, B.	136
Kushan, M.A.K.	166

L	
Lakmali, E.	174
Liyanage, S.R.	195
M	
Madusanka, T.	117
Munashinghe, L.	62
Munasinghe, R.	136
N	
Nanayakkara, J.	174
Nandasena, P.K.D.M.	57
Niwunhella, H.	155
P	
Pabasara, H.M.U.	17
Peiris, P.P.C.	51
Perera, M.D.R.	75
Perera, P.S.H.	57
Peter, S.	155
R	
Rajapakse, C.	51, 68, 82, 142, 202
Rajapaksha, R.M.I.U.	57
Ranawaka, U.M.	202
Ranaweera, A.L.A.K.	57
Rathnayaka, K.T.	161
Rathnayake, U.	223
S	
Samankula, W.G.D.M.	90
Senanayake, J.M.D.	68, 82
Senevirathne, D.S.	188
Shibata, M.	1
Sotheswaran, S.	216
Sugahara, N.	1
T	
Takahashi, M.	1
Telijjagoda, S.	7, 37, 124, 136, 182, 228, 234
Thalagala, T.S.	111
Tharindu, K.H.S.	210
Todo, K.	147
V	
Vidanagamachchi, K.	174
Vithaksha, L.G.C.	90
W	
Warnajith, N.	45
Weerakoon, T.C.	161
Weerasinghe, I.	136
Wickramasinghe, L.	223
Wijayasiriwardhane, T.K.	103, 188
Wijesekara, M.	96
Y	
Yamamura, M.	147
Yasarathna, T.L.	62

Combining the fields of Management and Information Technology, the Department of Industrial Management which has always been an exciting, stimulating and fun place to learn and develop to your full potential in order to launch a successful professional career, continues to meet the demands of global corporates by providing access to the frontiers of knowledge by way of this International Research Conference on Smart Computing and Systems Engineering.

Technical Co-Sponsors:



The Chartered
Institute of Logistics
and Transport



SLASSCOM
THE KNOWLEDGE AND INNOVATION CHAMBER

Published by:
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
dim.kln.ac.lk | im@kln.ac.lk | +94 112 914 482

ISSN 2613-8662



9 772613 866007