



SMART COMPUTING & SYSTEMS ENGINEERING

International Research Conference

Colombo, Sri Lanka | 16th September 2021

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 2021)

16th September 2021

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 2021

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	
Dr Mats Isaksson	xi
Professor Nirmalie Wiratunga	xiii
List of Papers	xv
Smart Computing	1
Systems Engineering	137

Editorial Board

International Research Conference on Smart Computing and Systems Engineering 2021

Chief Editor : **Dr. Suren Peter**

Committee : **Prof. (Mrs.) Annista Wijayanayake**

Dr. Keerthi Wijayasiriwardhane

Programme Committee

International Research Conference on Smart Computing and Systems Engineering 2021

Prof. Takao Terano	Chiba University of Commerce, Japan
Prof. Athula Ginige	Western Sydney University, Australia
Prof. Darshana Sedera	Southern Cross University, Australia
Prof. Setsuya Kurahashi	University of Tsukuba, Japan
Prof. Pradeep Abeygunawardana	Sri Lanka Institute of Information Technology, Sri Lanka
Prof. Koliya Pulasinghe	Sri Lanka Institute of Information Technology, Sri Lanka
Prof. Vojtěch Merunka	Czech University of Life Sciences, Czech Republic
Prof. S. Vasanthapriyan	Sabaragamuwa University, Sri Lanka
Prof. Janaka Wijayanayake	University of Kelaniya, Sri Lanka
Dr. Julian Nanayakkara	(Retired) University of Kelaniya, Sri Lanka
Prof. Prasad Jayaweera	University of Sri Jayewardenepura, Sri Lanka
Prof. Annista Wijayanayake	University of Kelaniya, Sri Lanka
Assoc. Prof. Masakazu Takahashi	Yamaguchi University Management of Technology, Japan
Asst. Prof. Ganga Hewage	Bryant University, USA
Asst. Prof. Shihan Wang	Utrecht University, Netherlands
Dr. Antonio Hyder	Hackers and Founders Research, USA
Dr. Raj Prasanna	Massey University, New Zealand
Dr. Prem Samaranayake	Western Sydney University, Australia
Dr. Niroshinie Fernando	Deakin University, Australia
Dr. Irvan Mhd	Tokyo Institute of Technology, Japan

Dr. Mohammad Ali Tareq	University Teknologi Malaysia
Dr. Malathi Sajeewani	Baker Heart and Diabetes Institute, Australia
Dr. Muhammed Badruddin Khan	Al-Imam Mohammad Ibn Saud Islamic University, KSA
Dr. Harsha Kalutarage	Robert Gordon University, Scotland, UK
Dr. Lalith Goonatilake	(Former Director) Trade Capacity Building, UNIDO
Dr. Suren Peter	University of Kelaniya, Sri Lanka
Dr. Indika Perera	University of Moratuwa, Sri Lanka
Dr. Ruwan Wickramarachchi	University of Kelaniya, Sri Lanka
Dr. Thabotharan Kathiravelu	University of Jaffna, Sri Lanka
Dr. Shantha Jayalal	University of Kelaniya, Sri Lanka
Dr. Suneth Pathirana	Uva Wellassa University, Sri Lanka
Dr. Keerthi Wijayasiriwardhane	University of Kelaniya, Sri Lanka
Dr. Ajantha Athukorala	University of Colombo School of Computing, Sri Lanka
Dr. Dilani Wickramarachchi	University of Kelaniya, Sri Lanka
Dr. Nithyanandan Pratheesh	Eastern University, Sri Lanka
Dr. Chathura Rajapakse	University of Kelaniya, Sri Lanka
Dr. Jeewanie Jayasinghe	University of Ruhuna, Sri Lanka
Dr. Amila Withanaarachchi	University of Kelaniya, Sri Lanka
Dr. Dhammika Elkaduwa	University of Peradeniya, Sri Lanka
Dr. Chathumi Kavirathne	University of Kelaniya, Sri Lanka

Organizing Committee

International Research Conference on Smart Computing and Systems Engineering – 2021

Conference Chair	Prof. Janaka Wijayanayake
Program Committee Chair	Dr. Ruwan Wickramarachchi
Track Chairs	Dr. Shantha Jayalal Dr. Amila Withanaarachchi
Conference Co-Secretaries	Ms. Hiruni Niwunhella Dr. Chathumi Ayanthi
Conference Asst. Secretary	Ms. Anushika Fernando
Members	Dr. Suren Peter Prof. (Ms.) Annista Wijayanayake Dr. Keerthi Wijesiriwardana Dr. Dilani Wickramaarachchi Dr. Chathura Rajapakse Mr. Buddhika Jayawardana Ms. Mahikala Niranga Mr. Janaka Senanayake

Keynote Speech

An Industry 4.0 Approach to Plastic Repair

Dr Mats Isaksson

Swinburne University of Technology, Australia
mats.isaksson@gmail.com

Industry 4.0 refers to the fourth industrial revolution. While the third industrial revolution involved the introduction of robotics and IT, the fourth industrial revolution focuses heavily on interconnectivity, reconfigurable autonomous automation, machine learning, and real-time data. Industry 4.0 applies to the entire life cycle of a product, including repair and recycling. It addresses the challenges of designing automation solutions that can adapt to changing conditions and achieve highly customized production. These challenges are particularly common in repair applications, where automatic repair of different product variants with various defects requires an extremely flexible automation solution.

A car headlight housing is priced between \$300 and \$6000. After a collision, the headlight housing often suffers only minor damages, such as one or a few broken plastic lugs; however, the entire headlight housing is typically replaced while the discarded unit ends up in landfill. Aside from the environmental impact, replacing a headlight housing has several other issues, including long lead times or cost and space issues if all headlight models would be stored. Manual repair is sometimes an option; however, it has multiple issues, including a lack of skilled workers and difficulties in achieving consistent quality and visually pleasing result.

In this presentation Dr Isaksson will describe a recently concluded collaboration between Swinburne University, PlastFix, and Innovative Manufacturing CRC (IMCRC) targeting automation of plastic repair. The presentation will showcase how advanced robotics, 3D printing, 3D scanning, and the development of a novel polypropylene composite filament were integrated to create a demonstrator for automatic repair of car headlight housings.

Keynote Speech

Learning to Personalise Human Activity Recognition

Professor Nirmalie Wiratunga

Robert Gordon University, United Kingdom
n.wiratunga@rgu.ac.uk

Innovative, person-centred strategies are required to monitor and predict physical activity and exercise behaviours, to scan and anticipate environmental barriers to activity, and to provide social and motivation support. Integrated Human Activity Recognition (HAR) and assistive technologies promise play a key role in this regard by enabling people to live their life well regardless of their chronic conditions. HAR is the classification of human movement, captured using one or more sensors either as wearables or embedded in the environment (e.g., depth cameras, pressure mats).

State-of-the-art methods of HAR rely on having access to a considerable amount of labelled data to train deep architectures with many train-able parameters. This becomes prohibitive when tasked with creating models that can personalise to nuances in human movement, such as when performing physical activities and exercises. In addition, collecting training data that can cover all possible subjects in the target population can be prohibitive. Instead, what we need are methods that can learn personalised models with few data for HAR research.

Recent advances in meta-learning provides interesting opportunities for similarity learning and personalised recommendations. Rather than learning a single model for a specific task, meta-learners adopt a generalist view of learning-to-learn, such that models are rapidly transferable to related but different new tasks. Unlike task-specific model training; a meta-learner's training instance, referred to as a meta-instance is a composite of two sets: a support set and a query set of instances.

In our work, we introduce learning-to-learn personalised models from few data. We extend the meta-instance creation process where random sampling of support and query sets is carried out on a reduced sample conditioned by a domain-specific attribute; namely the person or user, in order to create meta-instances for personalised HAR.

I will present our recent work on learning personalised HAR models with few data and motivate our contribution through an application where personalisation plays an important role, mainly that of human activity recognition for self-management of chronic diseases.

LIST OF PAPERS

Smart Computing	1 - 136
Systems Engineering	137 - 275

Theme 1: Smart Computing

SC-01	Autism spectrum disorder diagnosis support model using InceptionV3 Lakmini Herath, Dulani Meedeniya, M.A.J.C. Marasingha, Vajira Weerasinghe	1
SC-02	Smart technologies in tourism: A study using systematic review and grounded theory Abdul Cader Mohamed Nafrees, F.H.A. Shibly	8
SC-03	Architectural framework for an interactive learning toolkit Shakyani Jayasiriwardene, Dulani Meedeniya	14
SC-04	Temporal preferential attachment: Predicting new links in temporal social networks Panchani Wickramarachchi, Lankeshwara Munasinghe	22
SC-05	Technology-enabled online aggregated market for smallholder farmers to obtain enhanced farm-gate prices Malni Kumarathunga, Rodrigo Calheiros, Athula Ginige	28
SC-06	Automatic road traffic signs detection and recognition using ‘You Only Look Once’ version 4 (YOLOv4) W.H.D. Fernando, S. Sotheeswaran	38
SC-07	Forecasting foreign exchange rate: Use of FbProphet Fanoon Raheem, Nihla Iqbal	44
SC-08	Novel deep learning approaches for crop leaf disease classification: A review E.M.T.Y.K. Ekanayake, R.D. Nawarathna	49
SC-09	Thought identification through visual stimuli presentation from a commercially available EEG device M.P.A.V. Gunawardhana, C.A.N.W.K. Jayatissa, J. A. Seneviratne	53
SC-10	LYZGen: A mechanism to generate leads from Generation Y and Z by analysing web and social media data Janaka Senanayake, Nadeeka Pathirana	59
SC-11	A tree structure-based classification of diabetic retinopathy stages using convolutional neural network M.S.H. Peiris, S. Sotheeswaran	65

SC-12	Exploiting optimum acoustic features in COVID-19 individual's breathing sounds M.G. Manisha Milani, Murugaiya Ramashini, Krishani Murugiah, Lanka Geeganage Shamaan Chamal	71
SC-13	A community-based hybrid blockchain architecture for the organic food supply chain Thanushya Thanujan, Chathura Rajapakse, Dilani Wickramaarachchi	77
SC-14	Implementation of a personalized and healthy meal recommender system in aid to achieve user fitness goals Chamodi Lokuge, Gamage Upeksha Ganegoda	84
SC-15	Deep Learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka Siventhirarajah Sangeevan	94
SC-16	What makes job satisfaction in information technology industry? Nimasha Arambepola, Lankeshwara Munasinghe	99
SC-17	Feature selection in automobile price prediction: An integrated approach Sobana Selvaratnam, B. Yogarajah, T. Jeyamugan, Nagulan Ratnarajah	106
SC-18	Estimation of the incubation period of COVID-19 using boosted random forest algorithm P.P.P.M.T.D. Rathnayake, Janaka Senanayake, Dilani Wickramaarachchi	113
SC-19	Student concentration level monitoring system based on deep convolutional neural network U.B.P. Shamika, P.K.P.G. Panduwawala, W.A.C. Weerakoon, K.A.P. Dilanka	119
SC-20	TrackWarn: An AI-driven warning system for railway track workers M.I.M. Amjath, S. Kuhanesan	124
SC-21	Application of AlexNet convolutional neural network architecture-based transfer learning for automated recognition of casting surface defects Shiron Thalagala, Chamila Walgampaya	129

Theme 2: Systems Engineering

SE-01	An exploratory evaluation of replacing ESB with microservices in service oriented architecture L.D.S.B. Weerasinghe, Indika Perera	137
SE-02	Comparison of supervised learning-based indoor localization techniques for smart building applications M.W.P. Maduraga, Ruvan Abeysekara	145
SE-03	Solution approach to incompatibility of products in a multi-product and heterogeneous vehicle routing problem: An application in the 3PL industry H.D.W. Weerakkody, D.H.H. Niwunhella, A.N. Wijayanayake	149
SE-04	Model to optimize the quantities of delivery products prioritizing the sustainability performance A.P.K.J. Prabodhika, D.H.H. Niwunhella, A.N. Wijayanayake	154
SE-05	A MILP model to optimize the proportion of production quantities considering the ANP composite performance index N.T.H. Thalagahage, D.H.H. Niwunhella, A.N. Wijayanayake	161
SE-06	Reduce food crops wastage with hyperledger fabric-based food supply chain Dewmini Premarathna	168
SE-07	Application of Game Theory on financial benefits and employee satisfaction: Case study of a state bank of Sri Lanka D.D.G. Trevince Jayasekara, A.N. Wijayanayake, A.R. Dissanayake	177
SE-08	A novel approach for weather prediction for precision agriculture in Sri Lanka Using Machine Learning techniques J.S.A.N.W. Premachandra, P.P.N.V. Kumara	182
SE-09	Design and development of pump based chocolate 3D printer R.R.A.K.N. Rajapaksha, Dr. B.L.S. Thilakarathne, Yashodha G. Kondarage, Rajitha De Silva	190
SE-10	Theoretical framework to address the challenges in microservice architecture Dewmini Premarathna, Asanka Pathirana	195
SE-11	Challenges for adopting DevOps in information technology projects J.A.V.M.K. Jayakody, W.M.J.I. Wijayanayake	203

SE-12	Modelling and validation of arc-fault currents under resistive and inductive loads Yashodha Karunarathna, Janaka Wijayakulasooriya, Janaka Ekanayake, Pasindu Perera	211
SE-13	Decision-making models for a resilient supply chain in FMCG companies during a pandemic: A systematic literature review B.R.H. Madhavi, Ruwan Wickramarachchi	216
SE-14	Simulation analysis of an expressway toll plaza Shehara Grabau, Isuru Hewapathirana	223
SE-15	Docker incorporation is different from other computer system infrastructures: A review W.M.C.J.T. Kithulwatta, K.P.N. Jayasena, B.T.G.S. Kumara, R.M.K.T. Rathnayaka	230
SE-16	Vibration analysis to detect and locate engine misfires Prathap V. Jayasooriya, Geethal C. Siriwardana, Tharaka R. Bandara	237
SE-17	Identify the interrelationships of key success factors of third-party logistics service providers Theruwanda Perera, Ruwan Wickramarachchi, A.N. Wijayanayake	244
SE-18	A decentralized social network architecture Tharuka Sarathchandra, Damith Jayawikrama	251
SE-19	Framework to mitigate supply chain disruptions in the apparel industry during an epidemic outbreak M.A.S.M. Perera A.N. Wijayanayake, Suren Peter	258
SE-20	Solution approaches for combining first-mile pickup and last-mile delivery in an e-commerce logistic network: A systematic literature review M.I.D. Ranathunga, A.N. Wijayanayake, D.H.H. Niwunhella	267

SMART COMPUTING

Autism spectrum disorder diagnosis support model using InceptionV3

Lakmini Herath*

Postgraduate Institute of Science
University of Peradeniya, Sri Lanka
lakminiherath0@gmail.com

M. A. J. C. Marasingha

Department of Radiography/Radiotherapy
Faculty of Alide Helth Science
University of Peradeniya, Sri Lanka
janaka@ahs.pdn.ac.lk

Dulani Meedeniya

Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
dulanim@cse.mrt.ac.lk

Vajira Weerasinghe

Department of Physiology
Faculty of Medicine
University of Peradeniya, Sri Lanka
vajira54@yahoo.com

Abstract - Autism spectrum disorder (ASD) is one of the most common neurodevelopment disorders that severely affect patients in performing their day-to-day activities and social interactions. Early and accurate diagnosis can help decide the correct therapeutic adaptations for the patients to lead an almost normal life. The present practices of diagnosis of ASD are highly subjective and time-consuming. Today, as a popular solution, understanding abnormalities in brain functions using brain imagery such as functional magnetic resonance imaging (fMRI), is being performed using machine learning. This study presents a transfer learning-based approach using Inception v3 for ASD classification with fMRI data. The approach transforms the raw 4D fMRI dataset to 2D epi, stat map, and glass brain images. The classification results show higher accuracy values with pre-trained weights. Thus, the pre-trained ImageNet models with transfer learning provides a viable solution for diagnosing ASD from fMRI images.

Keywords - epi images, fMRI, Inceptionv3, stat map images, transfer learning

I. INTRODUCTION

The current motivation of psychiatric neuroimaging research is to identify objective biomarkers to diagnose neurological disorders like Autism Spectrum Disorder (ASD) and Attention deficit hyperactivity disorder (ADHD). Recent advances in the field of biomedical imaging and deep learning provide efficient diagnostic and treatment processes to identify different brain-based disorders [1][2][3]. This paper proposes a novel technique for automatic identification of ASD by applying Transfer Learning (TL) on Functional Magnetic Resonance Imaging (fMRI) data.

ASD is identified as a common multifactorial neurological disorder that affects the development of the brain, causing numerous disabilities. ICD-10 WHO (World Health Organization) [3] and DSM IV APA (American Psychiatric Association) [5], have specified main features in human interactions and behaviour of the patients, which can be used to diagnose ASD. Individuals diagnosed with ASD typically suffer from speech and communication difficulties, issues in social interaction, and lack of eye

contact [2].

In 2018, the Centers for Disease Control and Prevention (CDC) in USA have shown that the ratio of Autism patients to the general population is 1 to 59. This is twice as grater as the ratio reported in 2004, which was 1 to 125 [6]. Generally, there is a higher tendency of males being diagnosed with ASD than females, where the reported ratio is 4 to 1. According to the WHO report, it is estimated that 1 in 160 children has an ASD, worldwide. The study conducted in 2009 found that the prevalence of ASD among 18–24-month children is 1.07% in Sri Lanka [7].

Diagnosing ASD is a subjective and difficult task since there is no specific medical test. Usually, symptoms of ASD begin to appear before the age of 3 and it can prevail throughout the entire lifetime of a person, even if the severity may decline over time. Physicians and clinicians diagnose ASD by observing the patient's behaviour and development, considering the child's family history, genetic details, the progress of the development, and the skills they have in their lifestyle [2]. Current diagnostic processes may be carried out by involving several professionals from different disciplines with special skills t to identify the ASD-specific characteristics **Error! Reference source not found.** Lack of experience and training may lead to misdiagnosis of the children suffering from ASD. Research shows that early detection of ASD can lead to better results, enabling various ways to minimizing the symptoms and maximizing abilities [1].

The exact cause behind ASD is still unknown. A recent hypothesis in neurology has identified unusual neural activities in the brain of ASD patients. The cause has been discovered as the irregularities in neural patterns, disassociation, and anti-correlation of cognitive function between different regions, that affect the global brain network [9]. Thus, the fMRI data can be used to identify the abnormal neural pattern between brain regions to identify ASD.

The novelty of the paper was in carrying out a study to investigate the adaptation of pre-trained ImageNet weights on the Inception v3 model to classify ASD form controls using raw fMRI data. The Inception v3 model forms layers in parallel, whereas other models arrange layers in stacks. Thus, the Inception v3 model consists of a lesser number of parameters and generally provides higher accuracy compared to other models like VGG16, ResNet50. Therefore, the proposed approach uses Inception v3

architecture as the backbone deep learning technique. Most of the past research has used already preprocessed fMRI images like C-PAC (Configurable Pipeline for the Analysis of Connectomes) or various techniques [1] to preprocess fMRI images. In these studies, they have used features like functional connectivity (FC), specific brain regions related to ASD [10][11] to recognize ASD. The proposed method converted the raw 4D fMRI image to 2D epi images, stat map images, and glass brain images, while considering the sagittal, coronal, and axial views of brain volumes. The Inception v3 model was trained with and without ImageNet weights to investigate the weight transferring of different target domains. Thereafter, Section II gives a brief introduction to the background. Section III explains the workflow and the methodology followed by the research. Section IV discusses the model evaluation and, finally, Section V concludes the paper.

II. BACKGROUND

Functional MRI is a non-invasive technique that measures brain activities by detecting variations associated with blood flow [12]. It identifies high neural activity based on the fact that cerebral blood flow and neural activity are correlated, so that blood flow is high in the brain where neurons are highly active. The functional relationship that occurs in different brain regions at resting or task-negative state is measured by resting-state fMRI. It allows the observer to identify the abnormalities of the brain function easily due to the absence of added task-related brain functions. [13][14]. Thus, it is one of the popular techniques used in the identification of neuro-developmental disorders by observing the associations between brain function and phenotypic features [3][14].

Machine Learning (ML) is used to perform recurring and tedious tasks using feature extraction methods on raw data or with the features learned by other machine learning models. However, some issues associated with the medical image database have caused some limitations of using ML. For instance, the incompleteness by missing parameters and the lack of publicly sufficient labelled databases [16]. Furthermore, the performance of ML in medical image classification is far from the practical standard while the feature extraction and selection are time-consuming. The trending branch of ML, Deep learning (DL), can autonomously extract the prominent features from the raw input data, through a hierarchical sequence of non-linear transforms. DL is being used to identify patients with normal groups and it is further enhanced as a model to foresee the risk of developing disorders and predicting responses to different treatment procedures [11][14]. The fundamental goal of applying DL to neuroimage analysis is to remove the cumbersome and ultimately limiting feature selection process.

Moreover, Convolutional Neural Networks (CNNs), which are prevalent Deep Neural Networks (DNN), have shown significant performance in image classification [10][17][18]. DL models that use CNN are highly accurate because CNN extracts and learns features directly from images during the training process of the network.

As a result of the small sample size and high dimensionality of the fMRI dataset and the lack of interpretability of DL models, the application of whole-brain fMRI data is still limited [19]. Generally, a small number of ASD patients undergo fMRI scans as most seek

different other types of diagnosis for the moment. This leads to the unavailability of large sample datasets. However, many fMRI samples (volumes) are recorded for a single subject. These volumes can contain hundreds of dimensions, known as voxels. That results in lesser samples, but many dimensions in fMRI datasets. Thus, DL methods, as well as traditional machine learning methods, struggle in the learning curve, resulting in overfitting. In that case, TL is the key solution to this challenge.

Transfer learning is a way of gaining and storing knowledge from solving a problem of one task and applying it to a different but related task. It has become a popular concept in recent years and has been applied in a diverse set of domains. Pretraining is the first phase of TL in which, the network is trained using a large dataset consisting of highly varied labels/categories, representing many different areas. Then, the pre-trained network is 'fine-tuned' using a specific dataset from a field of interest. With this two-stage method, the high resource and time-consuming pre-training operation can be conducted only once, and then the results can be used in many different areas by fine-tuning.

In the field of medical image analysis, the current trend is to fine-tune an existing model with its architecture, with its pre-trained weights. ResNet [20], and Inception [21][22] are the few popular pre-trained DL models, which are trained on ImageNet datasets that are extensively applied in medical TL applications [23]. However, there is a considerable difference between ImageNet classification and medical image analysis in various ways.

In medical imaging problems, large images are represented a bodily region of interest which are used to identify the nature of the disease by recognizing the variations. On the other hand, in natural image datasets such as in ImageNet, the entire subject can be found within an image [23]. Further, ImageNet is a large dataset consisting of more than a million images that are smaller in size, while those of medical imagery is larger in size, but the number of images in the dataset is small. In addition, ImageNet is being trained for thousands of classes, while medical images are classified into few classes, less than 20, for instance. Moreover, the higher layers of the ImageNet architecture consist of many parameters, hence, is not the finest model for medical image classification.

Many CNN-based methods have been proposed to solve the problem of diagnosis of ASD using fMRI data, which remain unsolved and challenging. Related studies have addressed different approaches of pre-trained CNN networks like VGG 16, ReNet50, and Inception v3 with ImageNet weights with different input images. Husna et al. have applied a DL method from CNN variants of VGG-16 and ResNet-50 to identify ASD patients and extract the robust characteristics from fMRI. An accuracy of 63.4% and 87.0% has been achieved respectively [24].

In order to detect ASD, Dominic et al. have used a pre-trained InceptionResNetV2 model with TL on the augmented dataset. This has been generated by converting 4D resting-state fMRI into 2D data, where a validation accuracy of 57.75% has been achieved [25]. Ahmed et al. have developed an image generator, which developed single volume brain images from preprocessed fMRI images that are available in ABIDE dataset. The images were classified using ensemble classifiers which are combined with four different types of pre-train networks DenseNet, ResNet, Inception v3, Xception, and a CNN. The study has used

VGG16 as a feature extractor and gets an overall accuracy of around 82.7% [26].

Chen et al. have developed a VGG19 based CNN model with TL which provides 74.5% accuracy. The model predicts the cognitive assessment of infants using a brain structural connectome constructed by Diffusion tensor imaging (DTI) [27]. A deep multimodal proposed by Tang et al. have used two types of connectome data offered by fMRI scans. In Phase, I, the feature extractors, multilayer perceptron (MLP) and ResNet-18 were separately trained as independent networks. In phase II, an end-to-end model was obtained by combining MLP and ResNet-18 model with four fully connected layers as their output layer. The resulting multimodal network has been trained from scratch and classification accuracy of 74% has been achieved [28].

In another point of view, a few studies have used EEG signals and thermal images to diagnose ASD using machine learning-based classification techniques [29]. **Error! Reference source not found.** Haputhanthri et al. have utilized a correlation-based feature selection method to select relevant features and the necessary number of EEG channels. They have achieved an accuracy level of 93% by using Random Forest and Correlation-based Feature Selection **Error! Reference source not found.** The Accuracy of both logistic regression and multi-layer perceptron classifiers was able to be increased to 94% by integrating EEG and thermographic features [29].

III. DESIGN AND METHODOLOGY

A. Dataset

The Autism Imaging Data Exchange (ABIDE I/ II) dataset was used to carry out the proposed study [30] [31]. ABIDE is an online sharing consortium that provides Resting state fMRI (rsfMRI) data of ASD and controls participants' data with their phenotypic information. The ABIDE datasets consist of 17 different imaging sites. Out of the total dataset, a sample group aged between 0-12 is selected.

The sample dataset consists of 69 ASD individuals and 69 matched controls belonging to Kennedy Krieger Institute (KKI) data. The proposed ASD identification workflow is involves data preparation by converting 4D data to 2D images, feature extraction, followed by TL using pre-trained DNN model (InceptionV3), and evaluation as shown in Fig. 1.

B. 4D to 2D image transformation

The 4D fMRI image was transformed to a 2D image set, by slicing it along the sagittal, coronal, and axial directions. NIFTI is a file format for neuroimaging. As illustrated in Fig. 2, the 4D NIFTI image consists of a series of 3D volumes along the 4th axis; the time. The shape of the 4D image is identified and a series of 2D brain images is formed, considering the 3D brain volumes. As an example, the shape of the 4D image is 128, image converter creates 128 2D images from each volume. Three types of plotting functions epi, stat_map, and glass_brain were used to create three different types of 2D images from the raw fMRI images. A random slice from the sagittal, coronal, and axial direction was formed using the cut_coods parameter [26]. A total of 138 fMRI images in the proportion of 69 ASD

samples and 69 TD were converted to 20500 sample 2D images and saved in .png format.

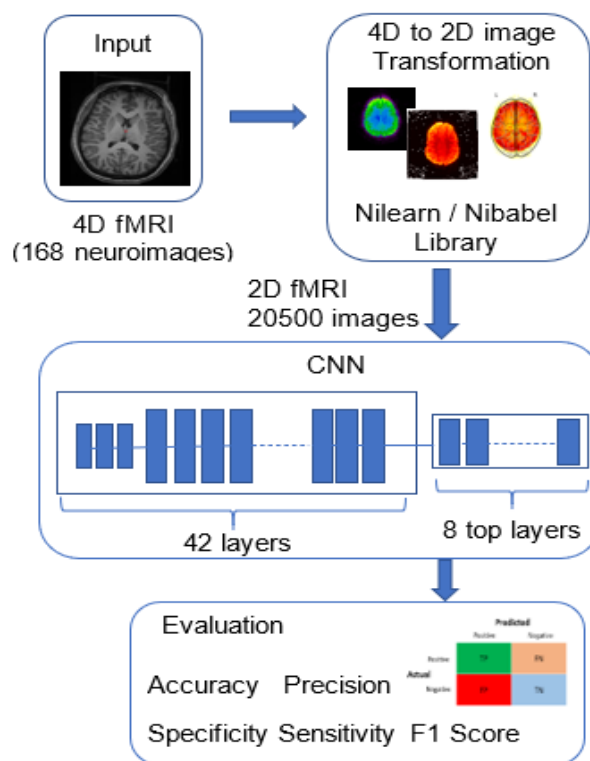


Fig. 1. ASD identification process

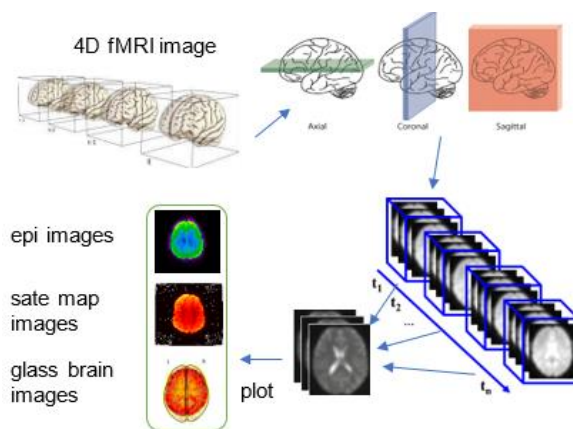


Fig. 2. Process of converting 4D fMRI image to 2D images

C. Transformed 2D image types

Three types of image plotting methods were used to train the neural network. Epi images are Echo-Planar Imaging. This is the type of sequence used to acquire functional or diffusion MRI data. Statistical images or stat maps plot cuts of a region of interest (ROI)/mask image of frontal, axial, and lateral. Epi images and stat images are 2D visualization images. The glass brain images represent the 3D view of the brain while plotting 2D projections of an ROI/mask image.

D. Augmentation

The training dataset was artificially increased using a data augmentation module, so that the training of the network was benefitted with a higher variation of input data.

For this purpose, Keras-based ImageDataGenerator was used by defining the image augmentation parameters such as batch size (32), rescale (1./255), transformations (shear, zoom, rotation). Here, the rescale parameter 1./255 transforms every pixel value from range [0, 255] -> [0,1], since 255 is the maximum pixel value. These augmentation methods increase the specificity and the sensitivity of medical images in the classification task. This may reduce network overfitting and support the model to generalize properly. Augmented epi images are shown in Fig. 3.

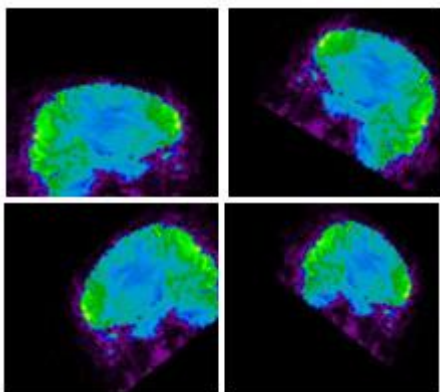


Fig. 3. Augmented epi images

E. Transfer learning settings with CNN

Inception v3 [32] model was selected as the CNN model to train the nine different image datasets using the TL approach. Inception v3 model pertained with ImageNet dataset (Natural images), which consist of 1.28 million training images, 100 k testing images, and 50 k validation images to classify 1000 classes. InceptionV3 model, layers are often connected in parallel instead of being stacked on top of one another and it is 42 layers deep. It comprises several inception modules that contain convolutions, average pooling, max pooling, dropouts, and fully connected layers. SoftMax is used to compute the loss. It employs techniques like regularizations, parallelized computations, dimension reduction, and factorized convolutions to optimize the network and enhance the model adaptation. The Inception v3 model was modified to adapt them to our classification task shown in Fig 4.

The InceptionV3 model deployed without the top layers and append new layers to the top layer. CNN was trained in two distinct ways and those are named MED1 & MED 2.

MED1: Initialize the Inception v3 by ImageNet weights and overlay with TP1 top layers

MED2: Initialize the Inception v3 by random weights and overlay with TP1 top layers

Fig. 4 (a) illustrates the Inception v3 model with TL settings and modifications. The classification was done by

setting the last dense layer to the softmax function. The models were trained by adjusting the hyperparameters. ADAM optimizer was selected with a 0.0001 learning rate. The dataset was randomly split into the sample ratio of 70:15:15. The implementation was done in python using Keras libraries

TP1 top layer block which is shown in Fig. 4 (b) is a combination of global average pooling layers, three dropout layers, three dense layers and one flatten layer. Here, GAP states the global average pooling layer, FL denotes the flatten layer, DEANs specifies the dense layer and DL states the dropout layer. Further, to reduce overfitting of the model L2 regularization was applied.

F. Statistical analysis

The CNN pre-trained model was evaluated using five statistical measurements, Accuracy (A), Recall (R) or Sensitivity, Precision (P), Specificity (S), and F1 score (F) [11] [15] [33]. The model identifies ASD subjects exactly as ASD is known as True Positive (TP) while the model which identifies TD subjects as TD is given as True Negative (TN). Further, the models which identify the ASD subjects as TD and TD subjects as ASD are referred to as False Negative (FN) and False Positive (FP), respectively.

Accuracy is defined as the closeness of a measured value to a known value, specified as the percentage of correctly classified samples. It can be calculated using the equation depicted in (1).

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \times 100\% \quad (1)$$

Precision describes how often the model provides an accurate prediction for positive class as shown in (2). That is the ratio of the correctly ASD positive labelled to all ASD positive labelled.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \times 100\% \quad (2)$$

Recall, also called sensitivity or true positive rate (TPR) is described as the percentage of correctly classified ASD subjects from all ASD subjects. The recall is calculated using (3).

$$\text{Recall (Sensitivity)} = \text{TP} / (\text{TP} + \text{FN}) \times 100\% \quad (3)$$

Equation (4) explains how the specificity or true negative rate is calculated as the number of correct negative predictions divided by the total number of negatives. It is the percentage of correctly classified control subjects from all control subjects:

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \times 100\% \quad (4)$$

The F1-score or balanced F-score is determined as the harmonic mean precision and recall. It focuses on the analysis of positive class. A high value for F1 suggests that the model performs better on the positive class. F1 score is calculated using Equation (5).

$$\text{F1 score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

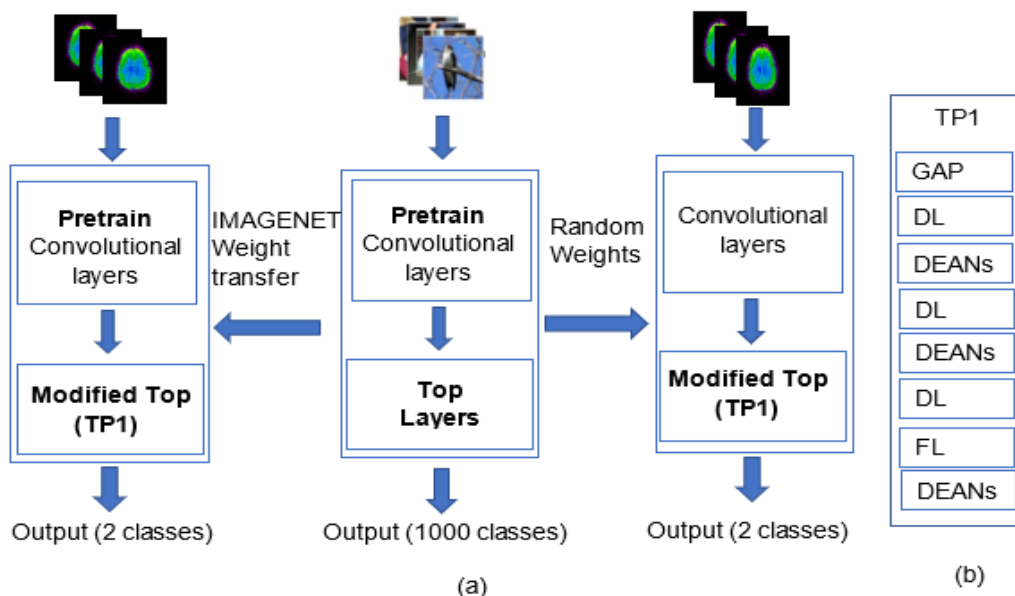


Fig. 4 (a) Transfer learning process of MED1 and MED2 (b) TP1 top layer

IV. MODEL EVALUATION

A. Experiment results

Inception v3 classification performance was calculated using the two methods MED1 and MED2. Each method, trained with nine different data sets, belongs to three different image plotting types. The performance of the model was measured with the test set, which is 15% of the data set.

All of the image types showed a similar and statistically significant performance in MED1 shown in Table I. The highest accuracy, sensitivity, specificity, precision, and F1 score was obtained by the axial view of glass brain images. On the other hand, in the accuracy metric, stat map Sagittal view obtained the lower result with 97.04%, followed by stat map Coronal view and stat map Axial view with 96.59%, 96.65% respectively.

The accuracy of all categories of MED2 was between 57% to 74%. The highest accuracy was observed in the Coronal view of epi images, which is 73.79%. The lowest accuracy value of 57.31% was observed in the Sagittal view of stat map images. There is a significant difference in Sensitivity (R), Specificity (S) in all types of image categories. In every category, the percentage of specificity is less than the percentage of sensitivity except in the glass brain sagittal view. That implies the fact that the percentage of correctly classified ASD is greater than the percentage of correctly classified Controls. The percentage value of the F1 score is around 74% in epi images and 70% in stat map images. The higher value of F1 shows that the MED2 model performs better on the positive class than the negative class.

There are only a few studies conducted on the effects of TL from ImageNet architectures. Most of the time this architecture does not provide the best performance on medical image datasets due to the lower capacity of data [23]. This is because when it comes to TL, two most important factors considered are the size of the new dataset (small or big) and its similarity to the original dataset.

In this research, the fMRI images lie in a different domain when compared to the ImageNet dataset. Not only that, when the size of the datasets are compared, the ImageNet dataset has a huge number of images than the fMRI image dataset. Nevertheless, it is observed in this study, that there is a clear impact on the ImageNet weights in ASD subject identification. The overall results demonstrate that the MED1 has a significantly higher performance than the MED2, in all image categories used to identify ASD subjects from controls. MED1 method, the Inception v3 model used the ImageNet weights as the initial weights for the training network. Even if ImageNet weights are trained using millions of Natural images, still it is possible to transfer those ImageNet weights to medical imaging, due to the properties of CNN, like gradual feature extraction in subsequent layers.

Lower layers identify basic features like lines and points, middle layers detect partials of objects, where top layers learn to recognize an entire object. Since any type of image consists of low-level features (points & lines) it is possible to start from a common low-level layer and then introduce specific higher-level layers according to the domain. Thus, the weights trained using ImageNet pre-trained dataset can be used as the initial weights to extract the basic feature of any type of image. Inception v3 model trained using these weights, achieved around 98% accuracy in epi images, 97% in stat map images, and 98% in glass brain images with equally high sensitivity, specificity, precision, and F1 score.

In contrast, in the MED2 the weights are initialized randomly, and the network starts learning from scratch by adjusting the weights. Inception v3 is a larger CNN with 42 convolutional layers, with 24 million parameters which need a large number of images to converge the network. When compared with the ImageNet, the size of the epi images, stat map images, and glass brain images were lesser, but still, the learning percentage of the network was between 58% to 70%, from the given data to identify ASD subjects from Control subjects.

TABLE I. PRECISION (P), RECALL OR SENSITIVITY(R), SPECIFICITY (S), F1 SCORE (F1), ACCURACY (A) OF MED1 AND MED2 METHODS

Image type	Display mode														
	Sagittal (x) %					Coronal (y) %					Axial (z) %				
	P	R	S	F1	A	P	R	S	F1	A	P	R	S	F1	A
MED1 (with ImageNet pre-trained weights)															
Epi images	97.82	97.31	97.87	97.56	97.69	96.96	98.23	96.97	97.59	98.59	98.25	99.27	99.26	98.89	98.76
Stat map	97.66	96.38	97.62	97.01	97.04	97.81	97.10	97.80	97.45	96.59	97.62	95.67	97.65	96.89	96.65
Glass brain	97.77	98.02	97.78	97.91	97.90	98.03	97.91	98.03	97.96	97.97	98.60	99.12	98.56	98.85	98.84
MED2 (with raw images)															
Epi images	59.40	97.57	34.40	73.84	65.79	71.55	78.25	69.39	74.50	73.79	65.62	84.02	56.44	73.60	70.18
Stat map	54.54	86.72	28.01	66.96	57.31	53.75	99.60	14.71	69.83	57.06	56.66	90.49	30.13	69.68	60.45
Glass brain	76.19	60.27	81.37	67.30	70.89	59.91	99.56	31.95	74.80	66.07	59.92	98.92	32.76	74.63	65.85

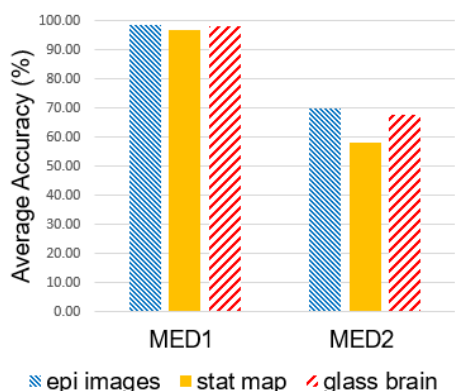


Fig. 5 Average accuracy of MED1 and MED2

Training a deep CNN network from scratch with random initialization of weights is a challenging task which consumes time. The accuracy can be increased using TL and pre-trained models, in a shorter period, compared to models trained from scratch.

The study examines three different types of unprocessed images, epi images, stat map images, and glass brain images. The epi images and stat map images represent the 2D visualization of Sagittal, Coronal and Axial view of 4D fMRI images while glass brain images represent the 3D visualization. According to Fig. 5, the highest average accuracy value of 98% was observed in epi images and glass brain images respectively from MED1. MED2 method also produced higher accuracy values, which were observed in epi images and glass brain images. Overall, epi image produced the best results, while stat map images yielded relatively poor results for both MED1 and MED2.

B. Comparison with the existing studies

The underlined research investigates methods to use TL to classify ASD utilizing unprocessed fMRI data by transforming the 4D image to a series of 2D images. Table II compares the proposed study with few other similar studies. Most studies carried out to identify ASD, have been conducted using natural imagery like facial images due to the domain similarity. But very few have investigated TL methodology with fine-tuning, using fMRI images to the target task. The study conducted by Husna et al. has achieved a higher accuracy of 87% using ResNet50, but the model suffers from overfitting [24]. It is a best practice to apply Regularization and data augmentation to avoid model overfitting [16].

Moreover, the epi image based InceptionResNetV2 model trained by Dominic et al. has obtained less accuracy due to a lesser number of sample images and model overfitting [25].

TABLE II. COMPARISON WITH RELATED STUDIES

Study	Features considered	Techniques	Accuracy %
[24]	2D images	VGG-16	63.40
		ResNet-50	87.00
[25]	epi images	InceptionResNetV2	57.75
[26]	stat map images, glass brain images	ensemble classifiers	82.70
[27]	DTI images	VGG19 based CNN	74.50
[28]	ROI correlation Matix	Combined model (MLP and ResNet-18)	74.00
Proposed study	epi images, stat map images, glass brain images	Inception v3	98.35
			96.76
			98.24

In contrast, this study has proposed a model with data augmentation as well as Regularization to avoid overfitting. Ahmed et al. have designed ensemble classifiers combining four different types of pre-trained networks. These include DenseNet, ResNet, Inception v3, Xception which are used to classify ASD from controls using various preprocessing pipelines. They have been able to achieve 82.7% accuracy with stat map and glass brain images [26]. In the context of this study, the images were created from raw fMRI images which imply the fact that preprocessing normalizes the images by reducing noise, together with fine features of the image. This study has gained better results compared to related studies.

C. Future research directions

The method was only applied to the KKI site of the ABDIE dataset. To implement a universal model to identify ASD subjects, it needs to experiment with all other sites of the ABDIE dataset. Combining the ABDIE site data may increase the number of sample points to train a deep CNN from scratch, which may benefit the creation of a universal set of weights for identifying ASD. Ultimately it is highly advantageous if the model can be enhanced to develop a universal set of weights to analyze and diagnose all ailments related to the brain. The underlined method opens up a new way of developing a computation model to identify ASD subjects using raw images. Furthermore, it decreases the computational cost compared to other studies which are

beneficial to develop an efficient computational model with necessary improvements.

V. CONCLUSION

Technology enhanced decision support systems facilitate the analysis of medical images. The study has introduced a transfer learning-based approach to identify ASD using fMRI images. We have shown that the ImageNet based pre-train models can be used to increase the performance in the medical image domain. The classification accuracy of the pre-train Inception v3 model with ImageNet weights was observed to be 98% in epi all image categories. This concludes that it can transfer the pre-trained weights with ImageNet to a highly diverse medical image domain with high accuracy, with a smaller number of sample data.

REFERENCES

- [1] D. A. Meedeniya, I. D. Rubasinghe, "A Review of Supportive Computational Approaches for Neurological Disorder Identification," in *T. Wadhera, D. Kakkar, (Eds.), Interdisciplinary Approaches to Altering Neurodevelopmental Disorders*, Chapter 16, pp. 271-302, IGI Global, 2020.
- [2] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, S. Jayarathna, "A Review of EEG-based Classification for Autism Spectrum Disorder." *Journal of Computer Science (JCS)*, vol.15, no.8, pp. 1161-1183, 2019.
- [3] S. De Silva, S. Dayarathna, G. Ariyaratne, D. Meedeniya, S. Jayarathna, "A Survey of Attention Deficit Hyperactivity Disorder Identification Using Psychophysiological Data." *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 15, no. 13, pp. 61-76, 2019.
- [4] World Health Organization, "International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)." Internet: <https://icd.who.int/browse10/2019/en#/F84.0>, 2019 [July 20, 2021].
- [5] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*. 5th Edition, American Psychiatric Association Publishing, 2019.
- [6] CDC, "Data & Statistics on Autism Spectrum Disorders," Internet: <https://www.cdc.gov/ncbddd/autism/data.html>, 2020 [Jul. 20, 2021].
- [7] H. Perera, K. Wijewardena, R. Aluthwelage, "Screening of 18–24-Month-Old Children for Autism in a Semi-Urban Community in Sri Lanka," *Journal of Tropical Pediatrics*, vol.55, no 6, pp. 402–405, 2009.
- [8] D. Haputhanthri, G. Brihadiswaran, S. Gunathilaka, D. Meedeniya, S. Jayarathna, M. Jaime, C. Harshaw, "Integration of Facial Thermography in EEG-based Classification of ASD," *International Journal of Automation and Computing (IJAC)*, vol.17, no. 6, pp. 837-854, 2020.
- [9] S. E. Schipul, T. A. Keller, M. A. Just. "Inter-regional brain communication and its disturbance in autism," *Frontiers in systems neuroscience*, vol 5, pp. 5-10, 2011.
- [10] G. Ariyaratne, S. De Silva, S. Dayarathna, D. Meedeniya, S. Jayarathna, "ADHD Identification using Convolutional Neural Network with Seed-based Approach for fMRI Data," in *Proc. 9th International Conference on Software and Computer Applications*, Langkawi, Malaysia, 2020, pp 31-35.
- [11] S. De Silva, S. Dayarathna, G. Ariyaratne, D. Meedeniya, S. Jayarathna, "fMRI Feature Extraction Model for ADHD Classification Using Convolutional Neural Network," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol 12:1, no. 6, pp. 81-105, IGI Global, 2021.
- [12] N. K. Logothetis, J. Pauls, M. Augath, T. Augath, A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, pp. 150–157, 2001.
- [13] M. Plitt, K. A. Barnes, A. Martin "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage: Clinical*, vol 7, pp. 359–366. 2015.
- [14] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *Neuroimage Clinical*, vol 17, pp. 16–23, 2018.
- [15] S. De Silva, S. Dayarathna, G. Ariyaratne, D. Meedeniya, S. Jayarathna, A. M. P. Michalek, "Computational Decision Support System for ADHD Identification," *International Journal of Automation and Computing (IJAC)*, vol.18 no.3, pp. 233–255, 2021.
- [16] H. G. Kim, Y. Choi, Y. M. Ro, "Modality-bridge transfer learning for medical image classification," in *Proc 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Shanghai, China, 2017, pp. 1-5.
- [17] W. Rawat, Z. Wang, "Deep convolutional neural networks for image classification: a comprehensive review neural computation," *Neural Computation*, vol 29, no. 9, pp. 2352 – 2449, 2017.
- [18] D. Rubasinghe, D. A. Meedeniya, "Automated Neuroscience Decision Support Framework," in *Deep Learning Techniques for Biomedical and Health Informatics*, B. Agarwal, Ed., Elsevier, Academic Press, Chapter 13, 2020, pp. 305-326.
- [19] H. Huang, X. Hu, Y. Zhao, M. Makkie, Q. Dong, S. Zhao, L. Guo, T. Liu, "Modeling task fMRI data via deep convolutional autoencoder," *IEEE transactions on medical imaging*, vol 37, no. 7, pp. 1551-1561, 2017.
- [20] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 770-778.
- [21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. "Going deeper with convolutions," in *Proc. IEEE conference on computer vision and pattern recognition*, Boston, USA, 2015, pp. 1–9.
- [22] S. De Silva, S. Dayarathna, D. Meedeniya, "Alzheimer's Disease Diagnosis using Functional and Structural Neuroimaging Modalities," in *Enabling Technology for Neurodevelopmental Disorders for Diagnosis to Rehabilitation*, T. Wadhera, D. Kakkar, Ed. Taylor & Francis CRS Press, USA, Ch. 11, 2021.
- [23] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Proc. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019, pp. 1-11.
- [24] R. N. S. Husna, A. R. Syafeeza, N. A. Hamid, A. R. Wong, R. A. Raihan, "Functional Magnetic Resonance Imaging For Autism Spectrum Disorder Detection Using Deep Learning," *Journal Teknologi (Science and Technology)*, vol. 83, no. 3, pp. 45-52, 2021.
- [25] N. Dominic, Daniel, T. W. Cenggoro, A. Budiarto, B. Pardamean, "Transfer Learning Using Inception-Resnet-V2 Model to The Augmented Neuroimages Data for Autism Spectrum Disorder Classification," *Communications in Mathematical. Biology and Neuroscience*, Article ID 39, pp. 1-21, 2021.
- [26] R. Ahmed, Y. Zhang, O. T. Inan, H. Liao, "Single Volume Image Generator and Deep Learning-based ASD Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3044–3054, 2020.
- [27] M. Chen, H. Li, J. Wang, W. Yuan, M. Altaye, N. A. Parikh, L. He, "Early Prediction of Cognitive Deficit in Very Preterm Infants Using Brain Structural Connectome with Transfer Learning Enhanced Deep Convolutional Neural Networks," *Frontiers in Neurosciences*. vol. 14, no. 858, pp. 1-11, 2020.
- [28] M. Tang, P. Kumar, H. Chen. A. Shrivastava, "Deep Multimodal Learning for the Diagnosis of Autism Spectrum Disorder," *Journal of Imaging*, vol. 6, no. 6:47, 2020.
- [29] D. Haputhanthri, G. Brihadiswaran, S. Gunathilaka, D. Meedeniya, S. Jayarathna, M. Jaime, Y. Jayawardena, "An EEG based Channel Optimized Classification Approach for Autism Spectrum Disorder," *Moratuwa Eng. Research Conference (MERCon)*, IEEE explorer, Sri Lanka, 2019, p. 123-128.
- [30] D. Martino, et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry*, vol. 19, no. 6, pp. 659-667, 2013.
- [31] D. Martino, et. al., "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," *Scientific Data*, vol. 4, no. 170010, 2017.
- [32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 2818-2826.
- [33] C. Bielza, P. Larrañaga *Data-Driven Computational Neuroscience: Machine Learning and Statistical Models*, New York, Cambridge University Press, 2020, pp. 205 – 206.

Smart technologies in tourism: a study using systematic review and grounded theory

Abdul Cader Mohamed Nafrees*
Office of the Dean
South Eastern University of Sri Lanka
nafrees@seu.ac.lk

F. H. A. Shibly
Department of Arabic Language
South Eastern University of Sri Lanka
shiblyfh@seu.ac.lk

Abstract - Tourism that uses smart technology and practices to boost resource management and sustainability while growing their businesses' overall competitiveness is known as smart tourism. Information and communication technologies (ICTs) have had a profound impact on the tourism industry, and they continue to be the key drivers of tourism innovation. ICTs have fundamentally changed the way tourism products are developed, presented, and offered, according to the literature. Any empirical studies or experiments must be focused on accepted or formed hypotheses. In this regard, grounded theory measures were used for interpretation, while a systematic review was performed to assess the research scope from current studies and works. The main goal of the study is to investigate and propose long-lasting and stable smart technologies for implementing smart tourism. Grounded theory is a concept that uses methodical rules to gather and dissect data in order to construct an unbiased theory. Fewer studies on smart technology in tourism have been conducted, with a majority of them concentrating on IoT, virtual and augmented reality, big data, cloud computing, and mobile applications. In either case, there is space for further investigation into this important field of study. As a result, this paper is a vital first step toward a clearer understanding of how smart technology can be applied to the tourism industry. The number of available research work on smart technologies in tourism were fewer from the selected journals and conference proceedings, which led to the accessibility of lesser data for analysis.

Keywords - IoT, smart technology, smart tourism, systematic review, tourism

I. INTRODUCTION

The tourism industry has been significantly affected by information and communication technologies (ICTs), and they continue to be the primary drivers of tourism innovation. Literature shows that information and communication technologies (ICTs) have radically changed the way tourism products are made, viewed, and offered [1]. The tourism industry's technical impact affects not only the manufacturers, but also the customers. The advancement of ICTs has explicitly denoted improvements in tourists' attitudes, which is central to the entire discipline of ICTs adoption in tourism. Clearly, ICTs' enormous popularity is shaping tourists' attitudes towards mobile apps, thus improving users' experiences [2]. Indeed, the broad reach of ICTs' involvement in tourism has sparked considerable debate among academics. It is also claimed that the internet has influenced the transformation of best operations and strategic practices in the tourism industry

[3]. Certainly, this is because the internet facilitates access to information to every corner of the globe. It is inevitable to admit that the application of ICTs in tourism is an important component in the supply chain [4].

Smart tourism is defined by a particular destination, attraction or tourist itself, depending on its technological abilities. Increased use of smart technology in their operations, from payment methods to interactive activities, is modernized in many destinations. Smart tourism ultimately aims to increase resource management efficiency and maximize competition [5]. Smart Tourism's European capital defines a clever destination as: "A destination that facilitates access to products, services, spaces and experiences from the tourism and hospitality sector via ICT instruments. It's a healthy social and cultural environment that is based on the social and human capital of the city. It also implements innovative, smart solutions and promotes the development and connectivity of companies."

It is explained in [2] in further detail: 'Smart Tourism Destinations take advantage of: (1) Technology embedded environments; (2) Responsive processes at micro and macro levels (3) End-user devices in multiple touch-points; and (4) Engaged stakeholders that use the platform dynamically as a neural system.'

Taking into account the available literature at the time of writing, researchers have provided their own definition of smart tourism as below.

'Smart tourism is the act of tourism agents utilizing smart technologies and practices to enhance resource management and sustainability, whilst increasing the businesses overall competitiveness'.

With various types of ICTs being created on a daily basis, the world continues to go digital. These ICTs use powerful operating systems like iOS12, Android, and others that are now common on modern mobile technologies. Indeed, getting access to mobile web or "apps" opens up a slew of new possibilities [6]. The notion that innovations are becoming smarter and use of wearable devices has recently emerged in academic discourse and within the tourism industry. Wearable technical technologies are expected to have a huge impact on people's interactions with their environments, despite their youth [7]. However, there is a scarcity of studies on the use of smart technologies in tourism in academic literature. As a result, this paper provides an interesting opportunity to further research on the use of smart technologies in tourism.

II. RESEARCH METHODOLOGY

Any research or scientific study must be conducted based on acceptable or formed theories. In that sense, grounded theory steps have been followed for the analysis

purposes; while the systematic review has been done to reach the research scope from the existing studies and works. The major objective of this study is to explore and recommend long-lasting and secure smart technologies to introduce to smart tourism. Grounded theory is a common philosophy with methodical rules for collecting and dissecting data to create an impartial theory [8]; while, a systematic review is a process of gathering and summarizing similar studies, which are conducted in the past to form a new conclusion and suggestions for the upcoming research work [9]. There were fewer researches conducted on smart technologies in tourism; those were mainly focused on IoT, virtual & augmented reality, big data, cloud computing, and smartphone applications.

The required pieces of information and data were collected from previous studies and qualitative research methods were used to analyze the gathered data. This study uses the grounded theory technique as the study uses qualitative data. Qualitative research methods use participant's experience, behaviors, and perception for data analyzing purposes [10]. A five-step process was introduced (Table I) to do a systematic review that used grounded theory for the content analysis [11].

TABLE I. FIVE STEP GROUNDED THEORY METHOD - SYSTEMATIC REVIEW

Steps	Process
Define	Defining inclusion/ exclusion criteria, field of research, select the source for the paper, and keywords for searching.
Search	Papers searched published after 2015, also previous published papers also included for the methodology part.
Select	Papers selected using the critical appraisal skills program (CASP) were used.
Analyze	Qualitative analysis performed
Present	Most suitable smart technologies found

According to the Table II, research papers and book chapters published after 2015 were searched in the initial stage. Furthermore, among the downloaded papers, only peer-reviewed journals and international conferences were included. Although, among these papers, research publications related to smart technologies in tourism were relatively very less. Therefore, some papers related to smart cities were also included in this study. Furthermore, papers related to ICT in tourism also were reviewed. In addition to that, search keywords were included; not only Scopus, emerald, IEEE, springer, and Google scholar to find the full articles published in a high indexed database; but also, smart tourism, smart technologies, IoT, cloud computing, and big data were included. According to the protocol of this study, research articles were shortlisted based on the paper's title and abstract, and Boolean operators were also used to get better results from the search. Finally, all the selected research works were validated based on the CASP tool, which uses the validity of the selected research articles [12]. At last 28 papers were selected based on these criteria and processes, which were closely related to Smart Technologies and Tourism.

Grounded theory was used to perform a different type of coding analysis, and qualitative analysis was performed using the Nvivo software tool [13].

TABLE II. NUMBER OF FINALIZED RESEARCH PAPERS

Database	No of Selected Papers
EBSCO	3
Google Scholar	12
Science Direct	6
Emerald	4
Springer	6
Tylor & Francis	3
IEEE	11



Fig. 1: Literature search overview

Thus, based on the 28 finalized papers, the most suitable smart technology was suggested for future smart tourism according to government support, data security, and cost-effectiveness.

III. EXISTING WORK

IoT enabled devices gather data from the tourist using sensors and store them in cloud storage, which then suggests to the tourist, in the future, about food preferences, near places, restaurants, and hotels; these services reduce extra hours spent by tourists for searching [14]. Meanwhile, collected data using IoT sensors changes to Big data which then can be used to predict tourist demand, enabling better decision-making, managing knowledge flows and interaction with customers, and providing the best service in a more efficient and effective way [15]. IoT and cloud computing are the essential core parts of developing Smart tourism, in the meantime, human capital, leadership, social capital, and innovation also support the Smart tourism destination [16]. [1-3] in line with the following work and further this study has used a network analysis approach to find how ICT supports smart tourism [17]. The researchers explored the fact that tourism focusing on smartphone technologies is the major sign that tourism industries expect from smart tourism; further, smart tourism promotes the implementation of IoT, cloud computing, and wireless communication technologies [18]. Furthermore, a team of researchers, proposed a tourism planner application with the help of IoT and big data, that not only helps the typical tourist but also persons with physical impairments [19]. Similarly, authors pointed out that tourists are expecting flexible and mobile-friendly tourism which can be easily provided by IoT [20]. Meanwhile, a team of researchers developed a system to transform Indian tourism digitally with the help of embedded systems and IoT; where this system assisted users to get to know the authentic history, heritage, culture and tradition of India via smartphone [21].

The authors pointed out that Big data which have been collected via IoT devices can be used to analyze tourism data. Data is collected in three different stages of tourism before, during the tour, and after the travel, based on the analysis results tourists can make their tour user friendly in real-time [22]. Furthermore, another research work pointed out that Smart tourism is not only about applying applications of different techniques but is also about easy and accurate accessibility of required tourism information before, during, and after the tour for the tourist use; but all these can be made possible with the help of Big data [23]. Meanwhile, a study about how smart technologies assist the marketing of tourism, shows how big data helps to track and forecast tourist flow and categorize tourists from the data of hotels and smart system management [24]. an IoT application based on smart city has been proposed, that confirmed the tourists save more than 50% of their time, while their satisfaction level is around 27% [25]. Closely, another study revealed that IoT in tourism can help enable automatic hotel check-ins and check-outs, locate travel destinations, and monitor tourist's health, which lead to cost reduction, better productivity, and traveler's satisfaction. But, there are challenges such as data security, investment cost, and technology infrastructure in implementing IoT on tourism [26].

A researcher pointed out that any internet connected wearable device can help the tourist by providing information, communication, sharing experiences, revealing setbacks encountered when traveling.; Furthermore, these devices can be accessible with voice command to avail help from tourist guides [27]. Similarly, a team has developed a prototype based on augmented reality (AR) using image processing and location data, which helps improve smart tourism by recommending scenic places, restaurants, hotels, and other important matters to the tourist [28].

Researchers stated that the Koran Tourism Organization, Tourism virtual reality (VR) mapping, and location-based tourist services provide required tourism information to the tourist, where these data can be collected from social media updates of tourists who have already visited those places; This information can help increase tourist visits by suggesting better places, food preferences, and hotel selections via web platforms or mobile applications [29]. In another study, Researchers proved that the websites, social media, and smartphone provide a huge support for tourism in terms of travel planning, which promotes both explorative and exploitative use but tourist's data security and privacy of data have a negative effect [30]. Likewise, Mobile technologies can help to implement VR in tourism, which is used to see the attractiveness of certain places in 3D shapes before tourists visit those places physically. Furthermore, they mentioned that the data privacy and security must be considered [31]. Meanwhile, the utilitarian and hedonic characteristics of mobile technologies are the main reasons for successful adoption of mobile technologies for travel; where these technologies provide greater assistance to the tourist before, during and after the travel for information accessibility [32].

The authors identified four-factors but ICT provision was not included, which doesn't mean that ICT is unimportant, but the knowledge deficiency of visitors to local conditions and characteristics caused a simple smart city structure that creates smart tourism [33]. But, in the

next research work, a group of authors mentioned that the internet penetration rates and the rate of use of Information and communication technology, existing smart city infrastructure and social networking create a way for the smart tourism destination. In the; meantime, policymakers must consider economic, social, environmental, and technological strategies to support smart tourism [13]. Rosanna Leung has conducted a survey among selected hoteliers in Taiwan, that confirmed the fact that hoteliers must be aware of the necessity of smart technology in terms of how social media and ICT promote hotel industries among tourists.; Furthermore, they believe technologies cannot replace employees, but that can help increase employee performance [34].

Authors state that technological implementation on tourism introduced smart tourism that supports tourists to make their travel easy throughout the entire tour where the Ambient Intelligence tourism is driven by a collection of disruptive technologies, and on the other hand, these technologies have many negative influences, especially on data privacy & security [35]. Similarly, a researcher mentions that the current tourism sector heavily depends on innovations like smart technologies, although the tourist's satisfaction is not only dependent on the technological factors that make the tourism accessible but also on services; some services can be provided only by humans [36]. But, an investigation revealed that the smart technologies play a major role in tourism to convert visited places into memorable ones via smart technologies tools and media [37].

Researchers proposed a model called Smart Tourism Destination (STD) based on the Delphi technique, which explored the fact that Smart Technologies alone are not enough to create smart tourism, but governance of STD is also needed [38]. Meanwhile, authors explain that ICT can frustrate tourists for authenticity, anxiety, addiction, narcissism, and mindlessness. On the other hand, it can help the tourist to avoid being alone during travel by providing virtual friends via smartphones [39].

Authors found that smart information systems, intelligent tourism management, smart sightseeing, ecommerce systems, smart safety, intelligent traffic, smart forecasting and virtual tourist attractions are tourists' key evaluation factors of smart tourism attractions, these factors help real-time data access, online booking, tourist flow forecast, better transport, and smart safety during the trip [40].

IV. DISCUSSION AND CONCLUSION

There were many research works conducted around the globe on the topic of tourism, and all the studies focused on finding and filling the gap in the tourism industry by providing ease, memorability, reduced costs, time management, and finding the places. In that sense, digital experts work on making smart tourism, especially on implementing smart technologies in tourism. Smart tourism has a difficult and dynamic environment where both physical and technological components are mixed and developed as a single object [41].

Nowadays, tourists expect to make their travel easy by finding high-rated restaurants to stay in, locate exact places to reach on time, cost-effective transport, fast and easy information access, secure information storage, and virtual travel to the tourism spots around the world before they

start their tour. Tourists prefer to visit a place if the accessibility of the required information is developed in a proper digitalized way [42]. On the other hand, tourism industries focus not only on profit but also on traveler's satisfaction by meeting traveler's expectations, and these expectations can be met easily by implementing smart technologies. The role of smart tourism is to provide a hedonic, noble, and significant experience [43].

From atheoretical perspective, this research work provides a meaningful contribution to tourism development using smart technologies. The main objective of this study is to review existing smart technologies in tourism and suggest the most suitable technology to improve the smart tourism industry for a better travel experience. A comprehensive systematic literature review was conducted to reveal various aspects of smart technologies in tourism and find the secure, quickest, and safest smart technologies to develop tourism using smart technologies. According to the review, this paper proposes a way to improve tourism using smart technologies by considering the facts selected from previous studies. Furthermore, this paper will help policymakers to make-up their concept of tourism in terms of smart technologies.

This study provides a solution to the gap that exists between smart technologies and tourism in terms of research areas. Findings of this study helps developers to use the suitable smart concept when designing new applications for tourism industries and tourists, which reduces the development time and cost. Furthermore, this work helps academics, researchers, and students to engage with better tourism studies in the future. The specialty of the used grounded theory strategy in the extraction of scientific classes suggest a helpful exploration technique for decision-makers. Likewise, it permits scientists to direct an examination that is interpretive and grounded in information.

Smart tourism is one of the most wanted research areas among academics and researchers and is the future of tourism. But there are only a few studies conducted so far, especially on smart technologies in tourism. Therefore, a detailed systematic review was conducted to earn the knowledge base study on smart technologies in tourism. The grounded theory method was used to analyze the data from the systematically reviewed articles and uncover the social processes. Smart tourism has to be developed more but the concept of smart development was developed as expected [44].

In various researches, authors suggest using IoT, big data and cloud computing technologies to implement smart tourism. Meanwhile, tourists expect user-friendly smartphone applications to access real-time information before, during, and after the tour at any time and from anywhere. But both tourism industrialists and travelers seriously consider data privacy and security, as all the collected data is stored in the cloud for analyzing purposes. Researchers suggested creating internet-connected wearable devices that can provide the required information from the cloud devices. Also, researchers mention that VR and AR devices could be developed to help show the scenic and tourist places before starting the tour. Further, mobile phone applications can be developed to access hotels and restaurants. On the other hand, smart technologies alone cannot make travel easy but human interaction must be mixed with them to develop a better smart tourism.

The findings of this study provide assistance to the academics, researchers, and industrialists to work on further effective implementations on research and development works on smart tourism as this article explored the existing smart technologies engaged with tourism. Based on this study, there are many technologies that can be adopted with tourism, for it to become smart tourism. Among these technologies, IoT was the most recommended and used smart concept to the tourism industry by many researchers and developers, which enabled automation. Travelers prefer real-time and trustworthy tourism information accessibility at any time and from anywhere. IoT can help monitor remotely, manage and control devices from anywhere anytime, and allow massive information access in real-time [45]. Therefore, cloud computing and big data are the most advanced technologies available today to securely store and analyze data collected through IoT devices. These information help tourists in better decision making; to plan their travel, but, the data privacy of users is questionable, although data hiding techniques such as encryption methods help keep user information secure.

There are many types of research conducted on smart tourism and smart technologies, but fewer researchers focused on smart technologies in tourism and fewer statistical analysis was done among the tourist and tourist industrialists about smart tourism, also fewer implementations were developed, but these researches do not completely express about the stakeholder's expectations. It is strongly recommended to conduct survey analysis from the perspective of the stakeholders of smart tourism, and then designers and developers can implement stakeholder's expectations either as a wearable device or as a smartphone application or both.

Other than the above, smart technologies only are not enough but other digital techniques and methods can be implemented with tourism industries such as STD. Also, it is very important to consider a tourist's mind for authenticity for anxiety, addiction, narcissism, and mindlessness while developing smart tourism.

V. RECOMMENDATIONS AND LIMITATIONS

The aim of this research was to look at how smart technology devices are used in the tourism industry. According to proof, the use of smart technology is revolutionizing the tourism industry, resulting in added value for both suppliers and customers. Smart Technology has moved the internet from mobile cyberspace to wearables on the body. Without participating in any physical activity, tourists can use this technology to obtain required information, communicate, share experiences, solve a variety of travel-related problems, and co-create their own value. According to the report, smart technology will turn tourists into explorers. Tourists will undoubtedly be inspired to re-construct their memories as a result of Smart Technology, which will enable them to add time, place, context, and personalization to their offers and experiences. This means that tourists can use only a voice command to program a series of events or actions for a specific period of time and at a specific venue, without the need for assistance from a tourism provider. The advent of smart technology has ushered in a new age of disintermediation, with visitors gaining influence over the entire service delivery process. As a result, the new face of

tourism will be focused on the optimization of "personalized reconstructed experiences" by customers. Smart products open up a wide variety of potential applications for the tourism industry, from both the supplier and customer viewpoints. Tourism providers, on the other hand, must take advantage of the interactivity, intimacy, and ubiquity of wearable devices by looking for ways to provide visitors with personalized, enhanced, automated, and novel experiences. The most critical considerations are privacy and protection of users. The use of smart technologies in tourism poses significant privacy and safety issues. The hotel's data portal, for example, is accessible to tourists who can "voice order" check-in and access up-to-date information about their accounts at the hotel. As a result, they could be enticed to gain unauthorized access to the establishment's data in order to satisfy their curiosity.

In terms of strategic advice, smart technologies are rapidly evolving technologies that will continue to have a direct effect on visitors and tourism organizations. With the Internet of Things, tourism providers should begin to streamline their existing business models and strategies in order to benchmark with rivals and address the challenges that will be faced in meeting the demands of visitors. While this paper makes a contribution by highlighting the use of smart technology in tourism and their future potential, it has some flaws. To begin with, the literature on smart technology and tourism is still in its infancy. As a result, the majority of the references in this paper about the use of smart technology in tourism were minimal academic sources. In either case, there is scope for further inquiry into this significant field of research. As a result, this paper is a significant first step toward a deeper understanding of how smart technology can be used in tourism. Further research into the value development of smart technology in tourism is recommended. It is also proposed that a thorough investigation be conducted to determine the economic implications of smart technology in this industry.

Further, the number of available research works on smart technologies in tourism were fewer from the selected journals and conference proceedings, which led to the accessibility of lesser data for analysis.

REFERENCES

- [1] Buhalis, eTourism: information technology for strategic tourism management. 2003.
- [2] Compuware, "Mobile apps: What consumers really need and want. A global study of consumer's expectations and experiences of mobile applications. Compuware, the Technology Performance Company.," 2012.
- [3] D. Buhalis and R. Law, "Progress in information technology and tourism management: 20 years on and 10 years after the Internet-The state of eTourism research," *Tour. Manag.*, vol. 29, no. 4, pp. 609–623, 2008.
- [4] Tourismembassy, "The use of new technologies in the tourism industry | Tourismembassy," [tourismembassy.com](https://tourismembassy.com/en/news/tourism-trends/the-use-of-new-technologies-in-the-tourism-industry), 2013. [Online]. Available: <https://tourismembassy.com/en/news/tourism-trends/the-use-of-new-technologies-in-the-tourism-industry>. [Accessed: 15-Mar-2021].
- [5] tourismteacher.com, "Smart tourism explained: What, why and where - Tourism Teacher," [tourismteacher.com](https://tourismteacher.com/smart-tourism/), 2019. [Online]. Available: <https://tourismteacher.com/smart-tourism/>. [Accessed: 15-Mar-2021].
- [6] R. Egger, "The impact of near field communication on tourism," *J. Hosp. Tour. Technol.*, vol. 4, no. 2, pp. 119–133, 2013.
- [7] A. Tate, "Google Glasses (Project Glass) : The Future of Human Computer Interaction? - Usability Geek," usabilitygeek.com. [Online]. Available: <http://usabilitygeek.com/google-glasses-project-glass-the-future-of-human-computer-interaction/>. [Accessed: 15-Mar-2021].
- [8] K. Charmaz and L. Liska, "Grounded Theory," *Blackwell Encycl. Sociol.*, 2015.
- [9] S. Reviews, "The Systematic Review: An Overview," *AJN, Am. J. Nurs.*, vol. 114, no. 3, pp. 53–58, 2014.
- [10] R. L. J. Ii, D. K. Drummond, S. Camara, R. L. J. Ii, D. K. Drummond, and S. Camara, "What Is Qualitative Research?," *Qual. Res. Reports Commun.*, vol. 8, no. 1, pp. 21–28, 2007.
- [11] J. F. Wolfswinkel, E. Furtmueller, and C. P. M. Wilderom, "Using grounded theory as a method for rigorously reviewing literature," *Eur. J. Inf. Syst.*, vol. 22, no. 1, pp. 45–55, 2011.
- [12] S. Nadelson and L. S. Nadelson, "Evidence-Based Practice Article Reviews Using CASP Tools: A Method for Teaching EBP," *Worldviews Evidence-Based Nurs.*, vol. 11, no. 5, pp. 344–346, 2014.
- [13] S. Shafiee, A. R. Ghatari, A. Hasanzadeh, and S. Jahanyan, "Developing a model for sustainable smart tourism destinations: A systematic review," *Tour. Manag. Perspect.*, vol. 31, no. May, pp. 287–300, 2019.
- [14] N. Wise and H. Heidari, "Developing smart tourism destinations with the internet of things," *Big Data Innov. Tour. Travel. Hosp. Manag. Approaches, Tech. Appl.*, pp. 21–29, 2019.
- [15] L. Ardito, R. Cerchione, P. Del Vecchio, and E. Raguseo, "Big data in smart tourism: challenges, issues and opportunities," *Curr. Issues Tour.*, vol. 22, no. 15, pp. 1805–1809, 2019.
- [16] K. Boes, D. Buhalis, and A. Inversini, "Conceptualising Smart Tourism Destination Dimensions," *Inf. Commun. Technol. Tour.* 2015, pp. 391–403, 2015.
- [17] G. Del Chiappa and R. Baggio, "Knowledge transfer in smart tourism destinations: Analyzing the effects of a network structure," *J. Destin. Mark. Manag.*, vol. 4, no. 3, pp. 145–150, 2015.
- [18] P. Liu and Y. Liu, "Smart Tourism via Smart Phone," in *International Conference on Communications, Information Management and Network Security (CIMNS 2016)*, 2016, pp. 129–132.
- [19] N. Michele et al., "Using IoT for Accessible Tourism in Smart Cities," in *Assistive Technologies in Smart Cities, IntechOpen*, 2018, pp. 31–49.
- [20] A. Verma and V. Shukla, "Analyzing the Influence of IoT in Tourism Industry," in *International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019)*, 2019, pp. 2083–2093.
- [21] V. Prusty, A. Rath, K. K. Rout, and Sivkumar Mishra, "Development of an IoT-Based Tourism Guide System," in *Advances in Intelligent Computing and Communication*, 2019, pp. 495–503.
- [22] D. Buhalis and A. Amaranggana, "Smart Tourism Destinations Enhancing Tourism Experience Through Personalisation of Services," *Inf. Commun. Technol. Tour.* 2015, pp. 377–389, 2015.
- [23] Y. Li, C. Hu, C. Huang, and L. Duan, "The concept of smart tourism in the context of tourism information services," *Tour. Manag.*, vol. 58, pp. 293–300, 2016.
- [24] V. Y. Moiseeva and V. A. Zolotovskiy, "Preconditions of Development and Perspectives of Use of Smart City Technologies for Regional Market of Tourism," in *State and Economy Smart Technologies for Society*, 155th ed., Springer Switzerland, 2021, pp. 85–91.

- [25] M. Nitti, V. Pilloni, D. Giusto, and V. Popescu, "IoT Architecture for a Sustainable Tourism Application in a Smart City Environment," *Mob. Inf. Syst.*, vol. 2017, p. 9, 2017.
- [26] T. Car, L. Pilepić Stifanich, and M. Šimunić, "Internet of Things (Iot) in Tourism and Hospitality: Opportunities and Challenges," *ToSEE – Tour. South. East. Eur.*, vol. 5, pp. 163–173, 2019.
- [27] R. Atembe, "The Use of Smart Technology in Tourism: Evidence From Wearable Devices *," *J. Hosp. Tour. Manag.*, vol. 3, no. 11–12, pp. 224–234, 2015.
- [28] Ö. F. Demir and E. Karaarslan, "Augmented reality application for smart tourism: GökovAR," in *Proceedings - 2018 6th International Istanbul Smart Grids and Cities Congress and Fair, ICSG 2018*, 2018, pp. 164–167.
- [29] C. Koo, S. Shin, K. Kim, C. Kim, and N. Chung, "SMART TOURISM OF THE KOREA : A CASE STUDY."
- [30] C. D. Huang, J. Goo, K. Nam, and C. W. Yoo, "Smart tourism technologies in travel planning: The role of exploration and exploitation," *Inf. Manag.*, vol. 54, no. 6, pp. 757–770, 2017.
- [31] J. Dorčić, J. Komsic, and S. Markovic, "Mobile technologies and applications towards smart tourism – state of the art," *Tour. Rev.*, vol. 74, no. 1, pp. 82–103, 2019.
- [32] R. Law, I. C. C. Chan, and L. Wang, "A comprehensive review of mobile technology use in hospitality and tourism," *J. Hosp. Mark. Manag.*, vol. 27, no. 6, pp. 626–648, 2018.
- [33] C. S. Chan, M. Peters, and B. Pikkemaat, "Investigating visitors' perception of smart city dimensions for city branding in Hong Kong," *Int. J. Tour. Cities*, vol. 5, no. 4, pp. 620–638, 2019.
- [34] R. Leung, "Smart hospitality: Taiwan hotel stakeholder perspectives," *Tour. Rev.*, vol. 74, no. 1, pp. 50–62, 2018.
- [35] D. Buhalis, "Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article," *Tour. Rev.*, vol. 75, no. 1, pp. 267–272, 2019.
- [36] D. Tüzünkan, "The Relationship between Innovation and Tourism: The Case of Smart Tourism," *Int. J. Appl. Eng. Res.*, vol. 12, no. 23, pp. 13861–13867, 2017.
- [37] S. Shen, M. Sotiriadis, and Y. Zhang, "The Influence of Smart Technologies on Customer Journey in Tourist Attractions within the Smart Tourism Management Framework," *Sustainability*, vol. 12, pp. 1–18, 2020.
- [38] J. A. Iyars-Baidal, M. A. Celdrán-Bernabeu, J. N. Mazón, and Á. F. Perles-Ivars, "Smart destinations and the evolution of ICTs: a new scenario for destination management?," *Curr. Issues Tour.*, vol. 22, no. 13, pp. 1581–1600, 2017.
- [39] J. Tribe and M. Mkono, "Not such smart tourism? The concept of e-lienation," *Ann. Tour. Res.*, vol. 66, pp. 105–115, 2017.
- [40] X. Wang, X. L. Robert, F. Zhen, and J. Zhang, "How smart is your tourist attraction?: Measuring tourist preferences of smart tourism attractions via a FCEM-AHP and IPA approach," *Tour. Manag.*, vol. 54, pp. 309–320, 2016.
- [41] R. Baggio, R. Micera, and G. Del Chiappa, "Smart tourism destinations: a critical reflection," *J. Hosp. Tour. Technol.*, vol. 11, no. 3, pp. 407–423, 2020.
- [42] N. Azis, M. Amin, S. Chan, and C. Aprilia, "How smart tourism technologies affect tourist destination loyalty," *J. Hosp. Tour. Technol.*, vol. 11, no. 4, pp. 603–625, 2020.
- [43] M. Kay Smith and A. Diekmann, "Tourism and wellbeing," *Ann. Tour. Res.*, vol. 66, pp. 1–13, 2017.
- [44] Alfonso Vargas-Sánchez, "Exploring the concept of smart tourist destination," *Enlightening Tour. A Pathmaking J.*, pp. 176–198, 2016.
- [45] T. hoon Kim, C. Ramos, and S. Mohammed, "Smart City and IoT," *Futur. Gener. Comput. Syst.*, vol. 76, no. July 2014, pp. 159–162, 2017.

Architectural framework for an interactive learning toolkit

Shakyani Jayasiriwardene*
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
shakyani@se.mrt.ac.lk

Dulani Meedeniya
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
dulanim@se.mrt.ac.lk

Abstract - At present, a significant demand has emerged for online educational tools that can be used as replacement for classroom education. Due to the ease of access, the preference of many users is focused on m-learning applications. This paper presents an architectural framework for an interactive mobile learning toolkit. This study explores different software design patterns and presents the implementation details of the prototype. As a case study, the application is applied for the primary education sector in Sri Lanka, as there is a lack of adaptive learning mobile toolkits that allow teachers and students to interact effectively. The study is concluded to be user-friendly, understandable, useful, and efficient through a System Usability Study.

Keywords - architectural framework, interactive learning, m-learning, primary education, usability

I. INTRODUCTION

Today, there is a highly increased demand for educational tools that promote online teaching and learning. Specifically, m-learning tools have become one of the most sought-after types of educational tools due to the high availability of personally owned mobile devices [1]. Although there are many existing applications, they lack the features discussed in this proposed solution. Moreover, the uprise of the COVID-19 pandemic and the sudden peak in the requirement for m-learning applications is challenging in the Sri-Lankan education sector [2]. Currently, the interactive online teaching and learning process is conducted via applications such as Zoom and WhatsApp. However, with a large number of class sizes, the primary students who have a low attention span and are likely to distract easily [3], it is challenging to impose effective interactivity and address each individual who is with different level of learning, within the allocated class time in online teaching [4].

Although several learning applications are available to address the virtual barrier between the teacher and students, the lack of technical knowledge in operating these applications has created a reluctance among Sri-Lankan teachers toward using them [2] [5]. Also, there is a lack of a platform for teachers to teach the lessons, rather than relying on pre-made lessons provided by the application itself. The available applications with similar functionality require some level of technical knowledge to operate.

In this light, m-learning that runs on smartphones has become a widely used method for teaching and learning. However, in some areas, there is limited access to internet connectivity and high-performing devices, which are essential to smoothly run such learning applications. Therefore, there is a need for a mobile learning application that can operate under constrained resources and both online and offline. In another point of view, it is important

to have a set of functionalities that allow teachers to manage their lesson videos and assess the understandability level of the students easily.

This paper proposes a software architectural framework for an interactive learning toolkit that can be used for the teaching and learning process. The main objectives of this study are to provide author video-based learning content and to interactively assess the student's skill level during the learning process under least device performance. Although this proposed toolkit can support the learning process in general, we have considered the primary education sector as a case study with more specific features. Therefore, this system provides a methodology to create interactive video lessons in a more effective manner which would be less costly in terms of performance and resource utilization. Also, it suggests the best-suited architecture to integrate functionalities to provide a user-friendly and efficient experience to the end-user.

The rest of the paper is organized as follows: Section II discusses the background studies and Section III presents the common architectural parameters considered for related applications. Section IV discusses the existing architectural patterns for mobile applications, while Section V describes the methodology. Section VI explains the implementation details and Section VII contains the evaluation for application usability. Section VIII discusses the contributions and Section IX concludes the paper.

II. BACKGROUND

E-learning focuses on creating an augmented learning environment in which technology can be utilized to provide a combination of different teaching and learning methods aiming to maximize the participation of students, rather than replacing the conventional learning techniques. m-learning is an extension of E-learning where it is capable to improve the productivity of students by allowing them to engage in learning without the restrictions of time and place with the utilization of handheld devices for the teaching and learning process. Several learning applications with different features are available in the literature that supports primary education. In another direction, few learning applications have considered the use of virtual environments for immersive learning [6]. This allows learner-centric education where the student can learn at their pace based on their skill levels. Thus, supports personalized learning.

Table I depicts the features and limitations of the above-mentioned applications. Most of the existing application has limitations such as limited and default content, lack of authoring ability for content, less incorporation of the

student's skill levels and issues in computational performances. Therefore, there is a requirement for a tool that allows any teacher to create their lesson content, author the lessons to enhance them as interactive lessons, under low-performance requirements so that the tool can be used in most, smart mobile devices.

TABLE I. EXISTING LEARNING APPLICATIONS

Tool name	Features	Limitations
Byjus learning app [7]	Syllabus-based videos with smart visualization, personal learning journey with knowledge graph, interactive and adaptive exercises, Real-time progress reports, individual guidance from mentors, and real-time tracking.	No authoring tools or access for teachers, Indian Syllabus, Paid subscription
Hapan – Kids' Learning App / Hapan 5 [8]	Hapan: Interactive game interface, practice exercises, self-explanatory UI, report cards for parents, mini-games Hapan 5: Revision app, follows the Syllabus, progress tracking, progress report generation	Performance issues, no authoring tools, redundant content, paid subscription Hapan 5: Limited to revision, no authoring tools
Kahoot! [9]	Attractive interface, verified educators, quiz-based learning	Learning is based on quizzes; no media authoring tools
Khan Academy Kids: [10]	Highly interactive, a library of content, tools for teachers, progress monitoring, adaptive learning path, playful characters to encourage, free	No authoring tools
Noon Academy – Student Learning App [11]	Online toolset for teachers, interactive classroom, online quizzes, breakout for group work, live chat with teacher and peers, test-preparation assistant	Limited for chosen tutors
ABC Kids – Tracing & Phonics [12]	Interactive, game-based, smart UI, Teacher Mode for progress reporting and activity toggling	No authoring tools
HOMER Learn & Grow [13]	Interactive, playful, personalized reading path, resources for parents, offline learning available	No authoring tools, redundant content
ABCmouse - Early Learning Academy [14]	Highly attractive and interactive, puzzles and quizzes, games, engaging characters, progress tracking for parents	No authoring tools, not free
Vedantu – Live Learning App [15]	Live interactive learning, in-class quizzes, leaderboard, test-preparation material, daily live interactive quizzes	No authoring tools
Udemy – Online Courses [16]	Offline learning available, customized learning reminders, note-taking and bookmarking, in-course quizzes, Q&A with instructors	Authoring tools limited to selected and paid instructors mostly paid subscription

III. ARCHITECTURAL ASPECTS IN LEARNING APPLICATIONS

A. Video streaming

To playback interactive videos, there must be a video streaming architecture that enables the preview of additional annotations attached to the video. The study by Meixner & Kosch [17], has introduced a set of requirements for Playback namely: Media interpretation,

Navigation, Media synchronization, Cache management, and Download management. Media Interpretation is the processing of the metadata files related to the video and transforming them into other internal data types. Navigation refers to different elements such as button panels and quizzes. Media synchronization supports viewing and hides annotations in the synchronized video timeline. Cache management manages deletion and cache occupation. Download management decides which elements and storylines to download, optimizing the downloaded quantity, and scheduling the download. Furthermore, Dellagiacomma et al. [18] and Gordillo et al. [19] also use the metadata interpretation approach for streaming interactive videos.

B. Content management

The learning apps contain a repository of learning materials which may include documents and media that are presented to the user. To manage these learning materials a weighted directed graph has been used in Alshalabi et al. [20]. In studies like Garcia-Cabot et al. [21] Multi-Agent systems have been used to manage the course content. Chen et al. [22] have used a Learning Object Repository to store the teacher-made learning objects, and a Learning Management System component to retrieve the courseware. In the study by Yarandi et al. [23], the software architecture contains a courseware knowledgebase to store the course content, and a Courseware Manager component to provide the User interface to manage the stored courseware. Tortorella & Graf [24] uses a course content database to store the material and a course content manager module is present to access this content.

C. Quality of service

To accept and use a certain application, the application must be able to satisfy the requirements and needs of the user [25]. Three types of quality factor frameworks have been proposed by Almaiah et al. [25], based on the DeLone and McLean's model (DL & ML) to ensure the quality of service in mobile learning systems: Information quality, system quality, and service quality. Furthermore, the ISO/IEC 25010: 2011 quality standard has introduced 2 quality models which include further characteristics and sub-characteristics [26]: Quality in Use and Product Quality.

D. Multiple access provision

In mobile learning applications, the main end-users are the Student/Learner and the Teacher/Instructor, whereas in some situations an Admin would also be made available to monitor and control the entire system. The user roles and access grants are usually provided through the application's interface where users may or may not be able to see different items of the app based on the user role. The study by Alshalabi et al. [20] contains three sub-modules connected to the System Interface Module, namely, the Admin Interface Module, Instructor Interface Module, and Student Interface Module. Moreover, Yarandi et al. [23] have presented separate modules for each user (Learner and Instructor), where the Learner and Instructor access the system through a User Interface Manager and a Courseware Manager, respectively. Using these separate manager modules, the access grants and permissions are defined for each user type. Further, when considering mobile

computing in a public cloud-based mobile application, the inherent long latency in data exchange may negatively affect the interactive nature of the application. To overcome this issue, Mobile-edge cloud computing has been introduced with Computation offloading [27] where the delay-sensitive and computation-intensive applications can offload the computation processes to nearby mobile-edge cloud servers, so that the application may function smoothly [28].

E. Usability

Usability measures the user feasibility and easiness to achieve the goal intended by a particular system [29]. Six categories of usability guidelines have been proposed by Kumar et al. [30] intending for mobile learning applications: Content organization, Navigation, Layout, Visual representation, Selection based, Consistency and standards, Help and feedback, Interaction, Customization, Learning experience, and Accessibility. A total of 121 usability guidelines have been introduced in the same study under the said categories. Moreover, Tahir & Arif [31] have introduced UI design criteria to consider when designing the user interface for mobile learning applications for children. This includes input/output, cognitive load, multimedia usage, customization, etc. The study has further categorized those design criteria based on the usability characteristics addressed by them: effectiveness, understandability, efficiency, learnability, operability, satisfaction, and attractiveness.

F. Interactivity

The interactivity parameter explains how the application enables user interaction by receiving input from the user to produce a specific output based on it. Meixner & Kosch [17] explains four main methods of interaction in interactive videos namely: Viewer to Video, Viewer to Annotation, Scene to Scene, Scene to Annotation. Here, Annotation refers to videos, animations, audio, images, text, etc. that are displayed in parallel with an interactive video story. Accordingly, Viewer to Video defines how the Viewer can interact intra-scene or inter-scene by different functions such as Play and Pause, and also by switching from one scene to another. Viewer to Annotation refers to a type of interaction where the user can click on hyperlinks on the video to display additional annotations related to the video. Scene to Scene explains how scenes interact through a predecessor-successor relationship, where the scenes will change from one to another based on different factors. Finally, Scene to Annotation defines how the annotation is derived by the scene itself. Here, the annotations are displayed and hidden based on time (time-based annotations). According to this study, interactivity from users can be enabled by the user interface such as by using a button panel. The other forms of interactivity are enabled by implementing specific logic.

G. Resource utilization

Mobile devices are restricted in their number of resources. Therefore, it is important to develop an application in a manner that would utilize those limited resources optimally to produce an efficient mobile application. To identify the resources to consider, Rawassizadeh et al. [32] have presented a resource classification. In the study, CPU, memory, battery, and disk

and network activities are identified as the five main resources that come along with mobile devices and are consumed by mobile applications, the battery being the most important. It also explains that each of these resources can affect one another. To efficiently utilize these resources to ensure smooth performance in a delay-sensitive and computation-intensive, multi-user, multi-task mobile application, a computation offloading mechanism is used in Mobile-edge cloud computing technology [33]. Initially, the users' computations tasks are sent to a Base Station. Afterward, they are executed in mobile-edge computing servers to send back the results to the mobile. This will assist in managing the limited battery life, communication resources, and computation resources of the system [34].

H. Authoring tools techniques

Authoring Tools are used to edit different media content such as videos, presentations, etc. In learning applications, they allow the Teacher's role to edit such media to create interactive learning content. There are two types of such videos, namely, non-linear videos and Annotated (Linear) Videos. To author interactive, non-linear videos four tools are required for video processing, video rearrangement, annotation editing, and export functions [17]. Another method to implement Authoring tools is by the use of directed graphs. This graph may contain nodes that are video clip placeholders and the edges as options for the user to choose the next placeholder [18]. It is also possible to create Learning objects using media obtained from various repositories including online sources such as YouTube and create a metadata file (JSON) for the authored Learning Object [19]. This method can be used to create Linear videos as well.

IV. EXISTING ARCHITECTURAL PATTERNS

A. Layered architecture

Layered architecture is defined as a design strategy that ensures the separation of responsibility across the objects of an application in an effective manner. Here, the system is separated into layers and each layer has a functionality specific to it. Each layer will be providing services to the layer above it. This architecture is suitable for instances where incremental development is preferred, as well as when the system development is handled by several teams (each team handling a specific layer), and when multilevel security is desired [35]. This architecture is beneficial to use when developing rich mobile applications [36] and to promote maintenance [35]. But, a complete separation of concerns may be challenging to achieve with this architecture, while it may be difficult to enable direct communication between non-adjacent layers. Also, the performance may be affected by the requirement to process requests at each layer [35].

B. MVC architecture

The MVC (Model-View-Controller) architecture is a design methodology where the presentation and interaction are separated from the system's data. The three components that compose the architecture are Model, View, and Controller. Each component is responsible for the system data, handling the presentation of the data to the end-user, and concerning with the user interactions and the integration of the interactions with the other two

components, respectively. This architecture is used when there are many ways to present and interact with the system data, and also when such future requirements are not clear. The MVC architecture is considered important in mobile application development, especially in iOS application development. However, this architecture may complicate simple application models [35].

C. Multi-Tier architecture

In the Multi-tier client-server architecture, different layers are executed in different processors as individual processes. Usually, the tiers consist of but are not limited to, the presentation tier, application processing tier, data management tier, and database tier. Multi-tier architectures can handle a large client base. This architecture may be useful when the data and the application are both volatile, when data from many sources are combined, and when it is required to scale in the future. However, when the system is large, there may be issues in identifying the sources of errors and identifying the responsibilities of teams working on developing each layer.

D. MVVM architecture

MVVM (Model-View-ViewModel) is specifically intended for modern UI development platforms where the View is handled by a designer. The View is the User Interface of the system. Model is the system data, and ViewModel manages the state of the view [37]. It will be beneficial when there is a need for good separation of concerns, and to reuse code. However, this architecture may complicate simple application structures, and also may cause considerable memory use due to data binding.

E. Client-server architecture

The Client-Server architecture contains different services, each provided by a server to be used by a client. This architecture is effective to be used when the system comprises a shared database that can be accessed from different locations by many clients. Further, when the system expects a varying load, the servers can be replicated, thus making this an effective architecture for such instances [35]. However, this has a single-point-of-failure, as well as the performance may depend on the network, making this architecture more vulnerable to failure.

F. VIPER architecture

VIPER (View-Interactor-Presenter-Entity-Router) is a Reference architecture that is popular in iOS application development. The VIPER architecture is intended for rich mobile applications, especially iOS applications [36]. Here, View handles the user interface items, as well as user inputs events and calls the Presenter. The Presenter is responsible to handle these calls and uses the Interactor to build the respective UI by retrieving the necessary data. The Entity components are used to retrieve domain objects and apply business logic on the data by the Presenter. Finally, the Router is accessed through the Presenter to handle navigation between the UI [38]. It provides better testability, loose-coupling, and better code structure which are suitable for medium to large-scale projects. Thus, it may not be much suitable for small-scale projects.

V. METHODOLOGY

A. Application overview

This System consists of a Mobile Application Authoring Tool to create interactive video-based Lessons by embedding pop-up activities in them. The System enhances learning by allowing teachers to create exercises based on educational videos related to the taught Syllabus that they can easily find online. As a case study, we have considered the video lessons provided by the National Institute of Education in Sri Lanka, with the YouTube channel NIE. Once a student enrolls in a Lesson and starts to play the video, the activities will pop-up at the defined timestamps and the video will pause. When the student enters an answer, the result will be recorded, and the video will continue to play. Also, the system focuses on the separation of concerns to enable scalability, reusability, manageability, and to lower the risk of failure. Further, it is concerned with providing optimal performance and resource use to enable smooth functionality and to increase the non-functional support to the end-user.

B. Design patterns

The main underlying architecture of the System is the Layered architecture based on which the system is divided into 4 layers: Presentation Layer, Application Layer, Business Layer, and Database Layer. Each layer will communicate only with its immediate layer(s).

1) *Presentation Layer*: The Presentation Layer consists of the User Interface which is the main interaction point with the end-user. All the .xml files that model the interfaces presented to the end-user groups, Student and Teacher, are included in this layer.

2) *Application Layer*: The Application server falls under the Application Layer which is the abstraction layer that functions to hide the Business logic from the presentation layer. There is only a single server to this system, and all the requests sent to the System through different events initiated at the Presentation Layer will pass through this layer to the Business Logic for processing. Similarly, the processed responses also move through this layer from the Business Logic Layer to the Presentation Layer to reach the end-user.

3) *Business Layer*: Business Layer comprises the Authentication Manager, Lesson Manager, Activity Manager, and the Metadata Manager that consists of other sub-modules and work together to address the business requirement of the system. The Authentication Manager interacts with the Application Layer to address the requests, and they do not communicate directly with the other components in the Business Layer. The other components except the Metadata Manager communicate with the application layer while interacting with the other components to fulfill different tasks.

4) *Database Layer*: Finally, the Database Layer consists of the database which is the single storage point of the system. All the resources that are being shared within the system are stored in the database layer.

Using the Layered architecture design pattern mainly contributes to the separation of concern which also isolates each layer enabling changeability in one layer without

affecting the others. It also contributes to testability in each layer separately, and maintainability by having a clear separation of code. Although the outcome of using this design pattern in this application is a monolithic architecture which may include overhead in future major changes made to the system, using a different architecture may pose a greater disadvantage based on development and performance [39]. Fig. 1 illustrates the high-level view of the Layered Architecture applied to the proposed system.

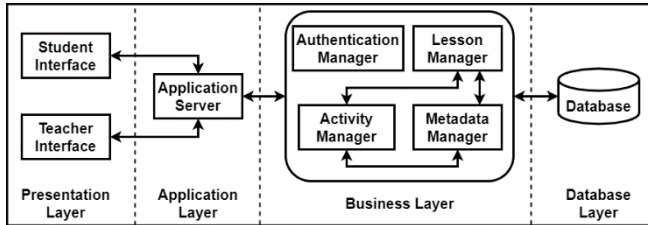


Fig. 1. A high-level view of the m-learning application.

C. Application architecture

The proposed software architecture supports the below architectural requirements for general learning environments [40] as follows.

- Provide information about the course
- Allow customization
- Automate the evaluation process
- Support the authoring of didactic material
- Enable the management of content by the instructor
- Enable learners' evaluation
- Support the delivery of didactic material
- Provide feedback mechanisms for evaluation

Fig. 2 depicts the architecture which enables the required functionalities.

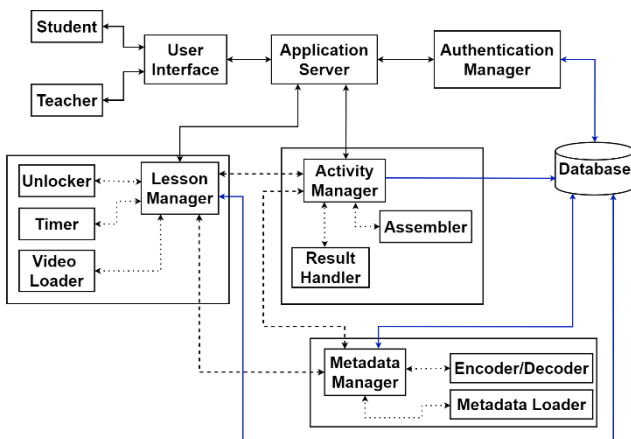


Fig. 2. Proposed architecture of the M-learning application.

The modules of the architecture can be listed as follows.

1) *User Interface*: The User Interface is the main interaction point of the Users with the System. The two main users who can interfere through the User Interface are

“Student” and “Teacher”. The Teacher is allowed to create, view, update, and delete Lessons, as well as to view students' progress [20]. The Student can enroll in Lessons, follow them and respond to activities, and finally view the results.

2) *Application Server*: The System comprises an Application Server, which is the central mediator between the User interface and the system's services. It is the entry point for the external user to the system's internal logic. As its main functionality, it will real-time manage the messages sent from the user interface by multiple users, to each of the other components in the system [41], and vice versa.

3) *Authentication Manager*: The Authentication Manager handles the registration and login attempts of the users to the system. A user can either be of the “Student” role or “Teacher” role. The registration attempts will be validated in the system based on the information input by the user at the time of registration. Once the registration is validated, the user will be provided access to the system with the related grants and permission based on the user role. Similarly, for a login attempt, the user's validity will be verified by checking for the availability of the username in the user profile, and the validity of the password [24] [42].

4) *Lesson Manager*: The Lesson Manager is the service or component of the system which manages all the existing and newly created Lessons or Learning Objects in the system. The “Learning Objects” refer to enriched videos [19] which consist of multiple activities to pop up at user-defined timestamps. This component is similar to a “Course Content Module/Repository” available in several related existing systems [20] [21] [22] where all the taught course materials are organized and managed.

a) *Video Loader*: This sub-module is responsible for loading the videos, which are the main course material of the system, from a video URL or by uploading from the filesystem [19].

b) *Timer*: The Timer's functionality is to listen to a Lesson during the time of video play for any activity timestamps, and to pass a message to the Activity Manager once such a timestamp is reached [19]. This sub-module ensures that the activities are retrieved at the correct point of time in the video.

5) *Activity Manager*: The Activity Manager handles the Activities related to each Lesson included in the Lesson repository of the System. Each Lesson will contain one or more than one activity per activity timestamp [19]. An Activity may be a pop-up quiz, based on text, images, etc.

a) *Assembler*: This sub-module functions to create an activity to be sent to the User Interface by using the decoded metadata. It will correctly identify which data to be used in which places in the activity code structure to setup up the final presentation for the user as a pop-up quiz.

b) *Result Handler*: This will store the results for each activity, based on the response received by the user [24], as a temporary file in the device file space. At the end of a Lesson, this sub-module will send the data in the temporary file to be stored in the database. This will minimize the

number of transactions required to store the user responses and other related parameters.

6) *Metadata Manager*: The Metadata Manager is focused on managing the Metadata files created for Lessons to define their structure and other parameters [18] [43]. Each time a Lesson is created/viewed/updated the Metadata manager will interfere.

a) *Metadata Loader*: This sub-module is responsible for loading a requested Lesson's metadata file from the database.

b) *Encoder/Decoder*: This sub-module performs the read/write functions of metadata files of the system. When a Lesson is created by a teacher and saved, a metadata file will be generated after the system processes the lesson data, and it will be stored in the database. This encoding function will encode all the data in the lesson into an XML format and save it. Once an existing Lesson is retrieved, the related metadata file in the database will be fetched and read or decoded to extract all the items included in the Lesson.

7) *Database*: The data will be stored in a NoSQL format, which is a non-relational database type, due to the requirement of better performance, flexibility, scalability, and due to the system consisting of different formats of data that will easily be stored and retrieved in the NoSQL format [44], [45].

VI. IMPLEMENTATION ASPECTS

To implement the creation, storage, and retrieval of the interactive learning videos, a mobile application with all the said features has been developed. Here, the teacher can upload a lesson video to his/her mobile device or use an existing video through a YouTube URL. Once uploaded, the video could be opened from the M-Learning application and the teacher could play the video.

To add interactive quizzes to the video, the teacher can choose desired time-points in the video where he/she requires the quiz to pop-up and then add a text-based, or image-based quiz from the palette provided in the User Interface, as the example quiz authoring structure depicted in Fig. 3 Several quiz structures will be provided, and their appearance will be set by default so that there will be minimum technical interference from the teacher's side. Everything required to develop the interactive video will be provided through a simple User Interface for easy understandability of the teacher. Further, the teacher can preview the video to test the output and make necessary changes. Once a quiz is added, the details of it will be stored temporarily as a .xml file in the device memory.

When the Teacher completes creating the interactive video and saves it, the temporarily stored information will be integrated to generate a .xml metadata file which includes the video details and quiz details that will be used to re-generate the video when retrieved.

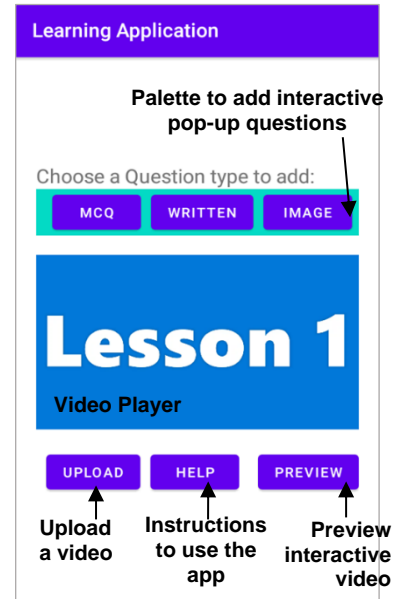


Fig. 3. The user interface for creating interactive video lessons

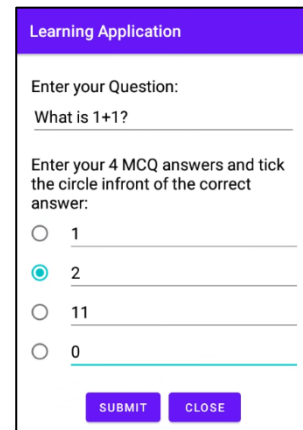


Fig. 4. Question creation interface

Then, the video will be added to the required lesson group, and students will be able to start accessing the interactive video lesson. Fig. 4 shows a sample GUI of a question creation. At the end of the video lesson, the students' answers and marks will be in the database to be stored.

VII. SYSTEM EVALUATION

The proposed mobile application was tested for its usability among a group of 22 users, 12 out of which were females and 10 males. The expertise/job profiles of the end-users ranged from Software Engineering to Education. The experience of the users in the education field concerning normal teaching ranged 2 – 20 years, whereas, with online teaching, the experience was limited to nearly a year. The mode of assessing the usability was through a System Usability Study [46] conducted through an online survey. The survey consisted of the ten questions of usability where the user could rate their positive and negative experiences on a scale of 1 to 5 where 1 represented "Strongly Disagree" and 5 representing "Strongly Agree". Then, the SUS score was calculated to evaluate how much the users have found the system to be usable. For this, the scores for each question were re-calculated based on the SUS score

calculation method, then the score was multiplied by 2.5, and finally, the average was obtained for 22 participants. Accordingly, the SUS score for this system was interpreted as approximately 80. Fig. 5 depicts the percentages of positive responses received for the usability aspect, whereas Fig. 6 depicts the negative responses of each participant.

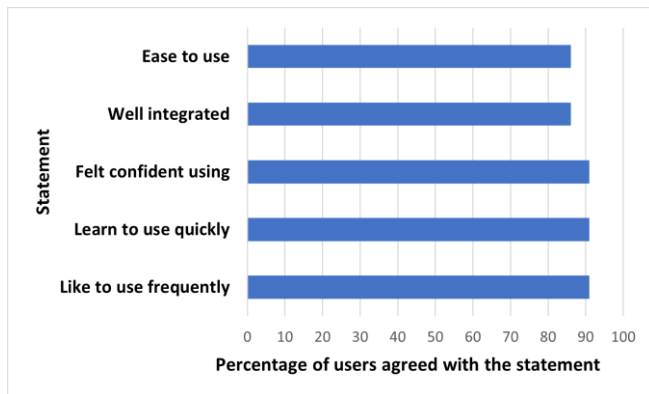


Fig. 5. Percentage of positive responses

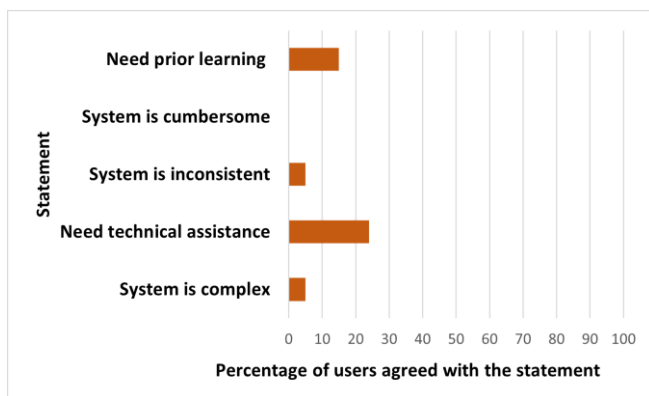


Fig. 6. Percentage of negative responses

This application can be further improved by including an adaptivity component where the authored assessments will be provided to the student based on his/her competency level, creating a personalized learning path. Also, a classroom management component can be added to form virtual classrooms where teachers can create lessons specific to different classrooms. In another point of view, by considering the development of technologies, interactive learning content such as virtual reality-based activities [47] can be incorporated to improve effective learning.

VIII. DISCUSSION

This study was conducted to develop an enhanced mobile learning toolkit that can provide lesson customization capabilities and an interactive learning experience to the end-users, along with a satisfactory level of usability and less cost in terms of performance and resource utilization. Using the proposed architecture and methodology, the said features can be achieved as follows: teachers can use the proposed authoring tools feature to customize the existing lesson videos or their lesson videos by adding pop-up quizzes and activities to the video at the chosen points of time. Then, students can view these lessons to learn, while actively participating in the learning

process by answering the pop-up activities added by the teacher, as explained previously. These answers and results are stored in the database for future use.

Moreover, the application interface is developed in a manner that is user-friendly and easy to understand by a novice user. Furthermore, the authoring tools feature is implemented in a manner that consumes comparatively fewer amounts of resources and memory of the mobile device, so that even a smartphone without many advanced features can run this application with the least performance issues.

This application addresses the shortcomings of the existing popular learning applications, mainly in the areas of authoring tools functionality and the related interactivity feature. The overall advantage of this application in terms of education is further to be studied in the future.

This system can be further improved with functionality to calculate the students' level of competency using the recorded assessment results, to provide adaptable learning content.

IX. CONCLUSION

This paper proposed an architectural framework for an interactive mobile learning toolkit that can be used to support primary education in Sri Lanka. The architecture caters to an m-learning application that can provide custom-made interactive video lesson content by the teacher to the student where the student will be tested for understandability during the video lesson. Also, a method has been proposed to store and retrieve these video lessons in a memory-efficient manner using metadata files. The survey conducted has shown a good system usability score for the proposed prototype solution.

REFERENCES

- [1] S. Criollo-C, S. Lujan-Mora, and A. Jaramillo-Alcazar, "Advantages and disadvantages of m-learning in current education," in *Proc. of the 2nd IEEE World Engineering Education Conference*, Buenos Aires, Argentina, 2018, pp. 1-6.
- [2] O. Ven Chandasiri, "The COVID-19: impact on education," *Int. J. Adv. Educ. Res.*, vol. 5, no. 3, pp. 13-14, 2020.
- [3] R. B. Hollis and C. A. Was, "Mind wandering, control failures, and social media distractions in online learning," *Learn. Instr.*, vol. 42, pp. 104-112, 2016.
- [4] D. A. Akuratiya and D. N. R. Meddage, "Students' Perception of Online Learning during COVID-19 Pandemic: A Survey Study of IT Students," *Int. J. Res. Innov. Soc. Sci.*, vol. 4, no. 9, pp. 755-758, 2020.
- [5] P. Chakraborty, P. Mittal, M. S. Gupta, S. Yadav, and A. Arora, "Opinion of students on online education during the COVID-19 pandemic," *Hum. Behav. Emerg. Technol.*, vol. 3, no. 3, pp. 357-365, 2020.
- [6] I. Perera, D. Meedeniya, C. Allison, A. Miller, "User Support for Managed Immersive Education: An Evaluation of in-World Training for OpenSim", *Journal of Universal Computer Science*, vol. 20, no. 12, pp.1690-1707, 2014.
- [7] Think and Learn Pvt Ltd., "BYJU'S - The Learning App," 2011. [Online]. Available: <https://play.google.com/store/apps/details?id=com.byjus.thelearningapp&hl=en&gl=US>. [Accessed: 11-Jul-2021].
- [8] Bhasha Lanka (Pvt) Ltd., "Hapan SuperKids," 2020. [Online]. Available: <https://play.google.com/store/apps/details?id=lk.hapan.hapan5&hl=en&gl=US>. [Accessed: 11-Jul-2021].
- [9] Kahoot!, "Kahoot! Play & Create Quizzes," 2012. [Online]. Available: <https://play.google.com/store/apps/details?id=no.mobitroll.kahoot.android&hl=en&gl=US>. [Accessed: 11-Jul-2021].

- [10] Khan Academy Inc, "Khan Academy Kids: Free educational games & books," 2018. [Online]. Available: <https://play.google.com/store/apps/details?id=org.khankids.android&hl=en&gl=US>. [Accessed: 11-Jul-2021].
- [11] Noon - The Social Learning Platform, "Noon Academy - Student Learning App," 2013. [Online]. Available: <https://play.google.com/store/apps/details?id=com.noonEdu.k12App&hl=en&gl=US>. [Accessed: 15-Jun-2021].
- [12] RV AppStudios LLC, "ABC Kids - Tracing & Phonics," 2019. [Online]. Available: https://play.google.com/store/apps/details?id=com.rvappstudios.abc_kids_toddler_tracing_phonics&hl=en&gl=US. [Accessed: 11-Jul-2021].
- [13] HOMER, "HOMER Learn & Grow," 2013. [Online]. Available: <https://play.google.com/store/apps/details?id=com.learnwithhomer.webapp&hl=en&gl=US>. [Accessed: 11-Jul-2021].
- [14] Age of Learning Inc., "ABCmouse.com Early Learning Academy," 2010. [Online]. Available: https://play.google.com/store/apps/details?id=mobi.abcmouse.academy_goo&hl=en&gl=US. [Accessed: 11-Jul-2021].
- [15] Vedantu Innovations Pvt. Ltd., "Vedantu: LIVE Learning App | Class 1-12, JEE, NEET," 2011. [Online]. Available: <https://play.google.com/store/apps/details?id=com.vedantu.app&hl=en&gl=US>. [Accessed: 11-Jul-2021].
- [16] Udemy Inc, "Udemy - Online Courses," 2013. [Online]. Available: <https://play.google.com/store/apps/details?id=com.udemy.android&hl=en&gl=US>. [Accessed: 11-Jul-2021].
- [17] B. Meixner and H. Kosch, "Creating and Presenting Interactive Non-linear Video Stories with the SIVA Suite," in Proc. of the 1st International Workshop on Interactive Content Consumption at EuroITV, Como, Italy, 2013, pp. 160–165.
- [18] D. Dellagiacom, P. Busetta, A. Gabbasiv, A. Perini, and A. Susi, "Authoring Interactive Videos for e-Learning: The ELEVATE Tool Suite," in Methodologies and Intelligent Systems for Technology Enhanced Learning, Proc. of the 10th International Conference, vol. 1241 AISC, L'Aquila, Italy: Springer, 2021, pp. 281–288.
- [19] A. Gordillo, E. Barra, and J. Quemada, "Facilitating the creation of interactive multi-device Learning Objects using an online authoring tool," in Proc. of the Frontiers in Education Conference, 2014, pp. 1–8.
- [20] I. A. Alshalabi, S. E. Hamada, K. Elleithy, I. Badara, and S. Moslehpour, "Automated adaptive mobile learning system using shortest path algorithm and learning style," Int. J. Interact. Mob. Technol., vol. 12, no. 5, pp. 4–27, 2018.
- [21] A. Garcia-Cabot, L. De-Marcos, and E. Garcia-Lopez, "An empirical study on m-learning adaptation: Learning performance and learning contexts," Comput. Educ., vol. 82, pp. 450–459, 2015.
- [22] C.-C. Chen, P.-S. Chiu, and Y.-M. Huang, "The Learning Style-Based Adaptive Learning System Architecture," Int. J. Online Pedagog. Course Des., vol. 5, no. 2, pp. 1–10, 2015.
- [23] M. Yarandi, H. Jahankhani, and A. Tawil, "An ontology-based adaptive mobile learning system based on learners' abilities," in Proc. of the IEEE Global Engineering Education Conference, Marrakech, Morocco, 2012, pp. 1–3.
- [24] R. A. W. Tortorella and S. Graf, "Considering learning styles and context-awareness for mobile adaptive learning," Educ. Inf. Technol., vol. 22, no. 1, pp. 297–315, 2017.
- [25] M. A. Almaiah, M. M. A. Jalil, and M. Man, "Empirical investigation to explore factors that achieve high quality of mobile learning system based on students' perspectives," Eng. Sci. Technol. an Int. J., vol. 19, no. 3, pp. 1314–1320, 2016.
- [26] J. Estdale and E. Georgiadou, Applying the ISO/IEC 25010 Quality Models to Software Product, in Proc. of the European Conference on Software Process Improvement, in Systems, Software and Services Process Improvement, X. Larrucea, I. Santamaria, R. O'Connor, R. Messnarz Ed., vol. 896. Springer, 2018, pp. 492–503.
- [27] X. Chen, L. Jiao, and W. Li, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," IEEE/ACM Trans. Netw., vol. 24, no. 5, pp. 2795–2808, 2016.
- [28] L. Tang and S. He, "Multi-User Computation Offloading in Mobile Edge Computing: A Behavioral Perspective," IEEE Netw., vol. 32, no. 1, pp. 48–53, 2018.
- [29] B. A. Kumar and P. Mohite, "Usability guideline for mobile learning apps: An empirical study," Int. J. Mob. Learn. Organ., vol. 10, no. 4, pp. 223–237, 2016.
- [30] B. A. Kumar, M. S. Goundar, and S. S. Chand, "Usability guideline for Mobile learning applications: an update," Educ. Inf. Technol., vol. 24, no. 6, pp. 3537–3553, 2019.
- [31] R. Tahir and F. Arif, "A Measurement Model Based on Usability Metrics for Mobile Learning User Interface for Children," Int. J. E-Learning Educ. Technol. Digit. Media, vol. 1, no. 1, pp. 16–31, 2014.
- [32] R. Rawassizadeh, A. Anjomshoaa, and M. Tjoa, "A Qualitative Resource Utilization Benchmarking for Mobile Applications," in Innovations in Mobile Multimedia Communications and Applications: New Technologies, I. Khalil, E. R. Weippl Ed., 2011, pp. 149–160.
- [33] B. Shebaro, O. Oluwatimi, and E. Bertino, "Context-Based Access Control Systems for Mobile Devices," IEEE Transactions on Dependable and Secure Computing., vol. 12, no. 2, pp. 150–163, 2015.
- [34] J. Guo, Z. Song, and Y. Cui, "Energy-Efficient Resource Allocation for Multi-User Mobile Edge Computing," in Proc. of the IEEE Global Communications Conference, Singapore, 2017, pp. 1–7.
- [35] I. Sommerville, Software engineering, 10th edition. London, UK: Pearson Education Limited, 2016, pp. 167–195.
- [36] F. J. A. Salazar and M. Brambilla, "Tailoring software architecture concepts and process for mobile application development," in Proc. of the 3rd International Workshop on Software Development Lifecycle for Mobile, Bergamo, Italy, 2015, pp. 21–24.
- [37] T. Lou, "A Comparison of Android Native App Architecture – MVC, MVP and MVVM," Master's thesis, Aalto University School, Espoo, Finland, 2016.
- [38] V. Humeniuk, "Android Architecture Comparison : MVP vs . VIPER," Master thesis, Linnaeus University, Växjö, Sweden, 2018.
- [39] J. D. Meier, A. Homer, D. Hill, J. Taylor, P. Bansode, L. Wall, R. Boucher, and A. Bogawat, Mobile Application Architecture Guide: Patterns & Practices, V1.1. Microsoft Corporation, 2008, pp. 1–138.
- [40] N. Freitas, D. Filho, L. Bortolini Fronza, and E. F. Barbosa, "Contributions for the Architectural Design of Mobile Learning Environments," IADIS Int. J. WWW/Internet, vol. 12, no. 1, pp. 94–112, 2014.
- [41] N. F. D. Filho and E. F. Barbosa, "A service-oriented reference architecture for mobile learning environments," in Proc. of the IEEE Frontiers in Education Conference, 2014, pp. 1–8, doi: 10.1109/FIE.2014.7044279.
- [42] M. A. Razek and H. J. Bardsi, "Towards Adaptive Mobile Learning System," in Proc. of the 11th International Conference on Hybrid Intelligent Systems, 2011, pp. 493–498, doi: 10.1109/HIS.2011.6122154.
- [43] R. Layona, B. Yulianto, and Y. Tunardi, "Authoring Tool for Interactive Video Content for Learning Programming," in Proc. of the 2nd International Conference on Computer Science and Computational Intelligence, Bali, Indonesia, 2017, pp. 37–44.
- [44] Y. Li and S. Manoharan, "A performance comparison of SQL and NoSQL databases," in Proc. of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, Canada, 2013, pp. 15–19.
- [45] V. Sharma and M. Dave, "SQL and NoSQL Databases," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 2, no. 8, pp. 2277–128, 2012.
- [46] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," Int. J. Hum. Comput. Interact., vol. 24, no. 6, pp. 574–594, 2008.
- [47] S. Hewawalpita, S. Herath, I. Perera and D. Meedeniya, "Effective Learning Content Offering in MOOCs with Virtual Reality – An Exploratory Study on Learner Experience", Journal of Universal Computer Science, vol. 24 , no. 2, pp. 129-148, 2018.

Temporal preferential attachment: Predicting new links in temporal social networks

Panchani Wickramarachchi*
Software Engineering Teaching Unit
Faculty of Science, University of Kelaniya, Sri Lanka
panchaniwickramarachchi@gmail.com

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

Abstract - Social networks have shown an exponential growth in the recent past. It has estimated that nearly 4 billion people are currently using social networks. The growth of social networks can be explained using different models. *Preferential Attachment (PA)* is a widely used model, which is often used to link prediction in social networks. PA tells that the social network users prefer to get linked with popular users in the network. However, the popularity of a node depends not only on the node's degree but also on the node's activeness which is reflected by the amount of active links the node has at present. Activeness of a link can be quantified using the timestamp of the link. The present work introduces a novel method called *Temporal Preferential Attachment (TPA)* which is defined on the activeness and strength of a node. Strength of a node is the sum of weights of links attached to the node. Here, the weights of the links are assigned according to their activeness. Thus, TPA captures the temporal behaviors of nodes, which is a vital factor for new link formation. The novel method uses *min - max* scaling to scale the time differences between current time and the timestamps of the links. Here, the *min* value is the earliest timestamp of the links in the given network and *max* value is the latest timestamp of the links. The scaled time difference of a link is considered as the *temporal weight* of the link, which reflects its activeness. TPA was evaluated in terms of its link prediction performance using well-known social network data sets. The results show that TPA performs well in link prediction compared to PA, and show a significant improvement in prediction accuracy.

Keywords - activeness of links, link prediction, social networks, TPA

I. INTRODUCTION

At present, around 4 billion users are using social networks, and still the number grows exponentially. Social networks serve different interests of the users. For example, social networks such as Facebook serve mainly as a friendship network which allow users to share their content and thoughts with their friends. In contrast, question and answering social networks such as Stackoverflow serve users to solve their programming problems by sharing them with other users of the social network. In addition, opinion posting social networks such as Reddit and Slashdot provide users a platform to post their opinions, thoughts, views and comments on various topics. Therefore, the growth of each social network depends on different facts and hence, predicting the growth of social networks has become a complex task [1], [2]. A plethora of researches have been carried out to devise novel models or alter the existing models to describe the growth of complex and heterogeneous social networks.

Social networks present a picture which has users connected via links. This picture of social networks can

further elaborate as a set of nodes connected via single or multiple edges (In network theory terminology, the users are referred to as nodes and the links referred to as edges). Here, the multiple edges represent the interactions that happen between the node pairs. For example, in Facebook, once a pair of users become friends, they interact with each other in multiple ways such as chatting, commenting, sharing posts, etc. All these interactions are considered as temporal edges and hence, the words edge and interaction use interchangeably to refer to the same entity. In network theory, the number of interactions between a node pair is referred to as the edge weight which reflects the closeness of the node pair. The total of the weights of edges attached to a node is said to be the strength of the node. In other words, the degree of the node is considered as the strength of a node. Here, the node degree is the count of all temporal edges attached to the node. The strength of a node reflects its popularity in the social network. The higher the strength, the higher the popularity. However, this is not always true due to the temporal behavior of nodes and edges. In other words, the strength of a node varies over time due to various factors. Therefore, the present research investigates the primary causes of temporal behavior of social networks. Although this study focuses on online social networks, it can be generalized to other types of social networks as well. The contribution of this paper can be summarized as follows.

- Provide an insight about the temporality of social networks.
- Discuss the limitations of existing static features used for link prediction in social networks.
- Introduce a non-parametric time-aware feature, *Temporal Preferential Attachment (TPA)* which captures the temporal behavior of nodes and edges.

The rest of the paper is organized as follows. Section II discusses the related research and provides a better insight about the importance of studying the temporality of social networks for link prediction. Section III presents the details of TPA, and link prediction performance of TPA. Section IV contains the experimental evaluation of the new method. Finally, section V concludes the paper with the summary of the research and future directions.

II. RELATED RESEARCH

Modeling modern social networks is a formidable task due to their complexity, heterogeneity and the size. Past researches have introduced various models to describe the growth of social networks [3], [4]. A growth model is a set of rules or a theory by which new nodes and edges are added to a social network. Among those growth models, the *Preferential Attachment (PA)* is a widely used method,

which is often used for link prediction in social networks. The intuition behind PA is that the nodes of social networks prefer to get linked with higher degree nodes or the popular nodes. PA quantifies this preference on popular nodes. Out of various PA based growth models, this section reviews some of the popular PA based growth models.

Barabási-Albert (BA) model [5] tells that the social networks grow according to the so-called power law (see Equation 7). The network starts with n nodes connected each other and grows by adding new nodes where each new node v randomly finds an existing node u to connect according to the probability proportional to the degree of u (see Equation 1).

$$\prod(d_u|v) = \frac{d_u}{\sum_{i \in N} d_i} \quad (1)$$

where N is the set of nodes in the network and d_u is the degree of node u . Although the BA model works well in modeling technological networks such as the Internet, it shows some limitations in modeling modern social networks such as friendship networks. The probability or the preference of choosing a node to connect does not depend only on the degree distribution of the nodes in the network but there are some other factors such as homophily, node attributes, and node activeness. Among them, homophily is described as the preference of new nodes to get linked with nodes which have similar interests. Considering this characteristic, homophily model [6] was introduced with homophily parameter δ which quantifies a certain property of a node. For any node pair u and v , the homophily parameters are defined as u_δ and v_δ . The difference $\Delta_{uv} = |u_\delta - v_\delta|$ tells the closeness of the node pair. Thus, the connection probability is defined as:

$$\prod(d_u|v) = \frac{(1-\Delta_{uv})d_u}{\sum_{i \in N} (1-\Delta_{iv})d_i} \quad (2)$$

Homophily model improves BA model by incorporating the similarity between node properties. Thus, the homophily model shows better performance in modelling modern social networks such as friendship networks. However, it still falls short in capturing temporality of nodes which is a key factor in deciding the connection probability. Therefore, an alternative model called Fitness model [7] was introduced to capture the short term node popularity. Fitness model is similar to BA model, but it includes an additional parameter called fitness parameter η ($0 \leq \eta \leq 1$) which captures the short term popularity of the node. The connection probability of Fitness model is defined as:

$$\prod(d_u|v) = \frac{\eta_u d_u}{\sum_{i \in N} \eta_i d_i} \quad (3)$$

Although the Fitness model captures the node temporality, it is still required to estimate the fitness parameter for each network. As a consequence, this model cannot generalise across different social networks. Also, parameter estimation is computationally intensive. Due to those limitations, researchers have introduced non-parametric link prediction methods. Non-parametric link prediction algorithm (NonParam) [8] uses a sequence of graph snapshots over time to capture the dynamic behavior

of nodes and edges. Compared to the baselines (Last time of linkage, Common neighbors, Adamic/Adar and Katz), NonParam algorithm performed well even in the presence of seasonal patterns. However, it can only predict pairs which are generated by 2-hop neighborhoods of last timesteps. Moreover, the non-parametric latent feature relational model is another link prediction method used to infer the latent binary features in relational entities [9]. This method has used feature-based methods to analyze the network data with the idea of Bayesian non-parametric approach. In capturing the subtle patterns of interactions, the latent relational model has performed better than class-based models.

Apart from that, researchers have introduced growth models which consider structural patterns such as motifs in temporal social networks [10], patterns and dynamics of users' behavior and interaction in social networks [11]. Inclusion of location information into PA based models have shown significant improvement in modeling the growth of various social networks [12]. This research has introduced a growth model which captures the growth of population in different geographic locations. It considers the account creation time and geographic information of each user. Although the above approaches have shown promising results in modeling the growth of modern social networks, still they have their own limitations.

III. LINK PREDICTION IN SOCIAL NETWORKS

Link prediction in social networks is a well-established research area. Social networks grow by adding new nodes as well as new links. Therefore, knowing the growth pattern of a social network is essential for link prediction in social networks. Link prediction problems can be classified into several sub-problems. For example, predicting new links, predicting missing links and hidden links are the popular link prediction tasks. This research focused on new link prediction, which can be defined as follows. For a given network at time t our task is to predict the potential links that can appear in time $t + 1$ [13]. Emergence of new links depends on various factors such as structural features, similarities between node and edge attributes. Common neighbors, Jaccard's coefficient, Adamic/Adar index, and PA are a set of popular neighbors based structural features used for link prediction [14]. Among them, PA quantifies this preference of getting linked with popular nodes. For example, preference of node pair ii and j getting linked can be quantified as shown in Equation 4.

$$PA_{ij} = degree_i \times degree_j \quad (4)$$

where $degree_i$ is the degree of node i . For example, in Figure 1, node A has degree 4 and node B has degree 3. Therefore, the $PA_{AB} = 12$. According to Equation 4, if the nodes have higher degrees their PA score takes a higher value. In case of link prediction, node pairs with higher PA are highly likely to get linked in future. Although PA looks like a promising method for link prediction based on the node popularity, the limitation of PA is it assumes that the popularity of a node solely depends on the node degree. In other words, the strength of the node, which assigns an equal weight (one) for each edge irrespective of its activeness. However, the popularity of a node depends not only on the node's degree but also on the activeness of the node which is reflected by the amount of active edges the

node has at present. In other words, the activeness of the node is reflected by the amount of recent interactions with its neighbors. The activeness of those edges is relatively higher than the old edges (old interactions). Thus, Activeness of an edge can be quantified using the timestamp of the edge. Based on the edge activeness, some of the recent researches have introduced alternative time-aware features which have shown their success in link prediction in social networks [15]– [17]. However, the inherent problems of most of these time-aware features are that they include parameters. Thus, it is required to optimize the parameters to obtain the optimal results. Parameter optimization is a tedious task as it consumes time and large amounts of computational power. As a consequence, some of those time-aware methods cannot generalise across different social networks. Those limitations motivated us to introduce a novel non-parametric time-aware feature which is an alternative to PA.

A. Temporal Preferential Attachment

The present work introduces a novel method called *Temporal Preferential Attachment (TPA)* which is defined on the strength or the weighted node degree where the weights of the edges are assigned according to the activeness of the links. Thus, TPA captures the temporal behaviors of nodes, which is a vital factor for new link formation. The novel method uses *min – max min – max* scaling to scale the time differences between current time and the timestamps of the links. Here, the *min* value is the earliest timestamp of the links in the given network and *max* value is the latest timestamp of the links. The scaled time difference of an edge is considered as the *temporal weights* (see Equation 5) of the link, which reflects its activeness.

$$\text{Temporal weight}_{ij} = \frac{T_{ij} - T_{\min}}{T_{\max} - T_{\min}} \quad (5)$$

where T_{ij} is the timestamp of the edge ij , T_{\max} is the latest timestamp in the network and T_{\min} is the earliest timestamp.

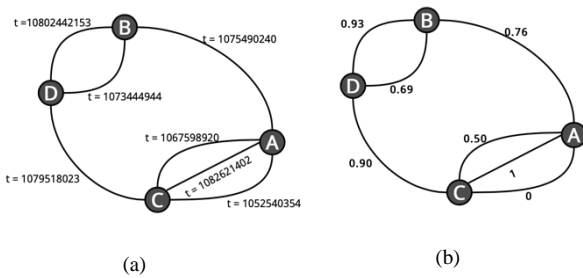


Fig. 1. A temporal social network. Figure (a): edges assigned with timestamps. Figure (b): after scaling the timestamps, each edge is assigned with a temporal weight.

According to Figure 1, older edges get lower weight and recent edges get higher weight. This is far better than assigning equal weights to all edges because the temporal weights reflects the activeness of the edges and hence, the activeness of the nodes they attached. Based on the temporal weights, TPA of nodes i and j calculate as shown in Equation 6.

$$TPA_{ij} = TS_i \times TS_j \quad (6)$$

where TS_i is the *temporal strength* of node i . Temporal strength of a node is defined as the total of temporal weights of the edges attached to the node. In Figure 1b, temporal strength of node A is 2.26 and temporal strength of node B is 2.38. Therefore, $TPA_{AB} = 5.38$ which is less than PA_{AB} but better captures the temporal strengths of the node pair. The effectiveness of novel method TPA was tested in terms of its link prediction performances on real-world social networks.

IV. EXPERIMENTAL ANALYSIS

The present study specifically focuses on link prediction in question and answering social networks and opinion posting social networks. In addition, one online friendship network was also used in the experiments to compare the effectiveness of TPA against PA in different settings. There are three types of interactions in question and answering networks: answers to the questions, comments to the questions, and comments to the answers. In this experimental analysis, we disregard the type of the interaction and consider each interaction as a temporal edge. TPA was evaluated in terms of its link predicting performances. The performance metric used to compare PA and TPA was area under curve (AUC) and ROC curves which give a better picture in model comparison.

The data analytics show that their degree distributions of the six networks follow the notion of power law (see Figure 2) which says that the fraction $P(k)$ of nodes in the network having degree k goes for large values of k according to the Equation 7.

$$P(k) = \lambda k^{-\gamma} \quad (7)$$

Here, γ is a parameter which typically takes values in between 2 and 3 for scale-free networks.

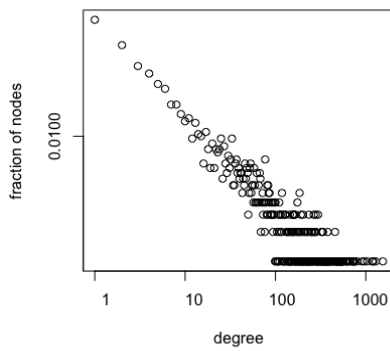
A. Data

Four question and answering social network data sets, one opinion posting social network data set and one online social network data set were used to test the effectiveness of TPA. Summary statistics of the data sets are shown in Table I. All data sets used in the experiment were taken from Stanford Large Network DataSet Collection (<https://snap.stanford.edu/data/>).

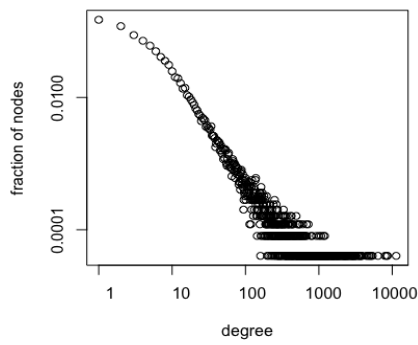
To create training sets and test sets, each data set was sorted in the ascending order of timestamps, and 80% of the sorted data set was taken as the training set and the rest 20% with latest timestamps were taken as the test set. In addition, all networks were assumed undirected. In each network, the largest connected subgraph was used to test the link prediction performance of PA and TPA. The training and test graphs were created in a way that the positive examples are the edges which are present in the test graph but not present in the training graph, and the negative examples are the non-edges which are common to training and test graphs. Also, all the nodes in the test graph are present in the training graph.

TABLE I. STATISTICS OF THE NETWORKS

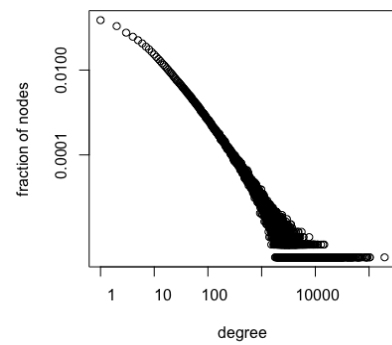
Network feature	CollegeMsg	Mathoverflow	Stackoverflow	Superuser	Askubuntu	Slashdot
Nodes	1899	24818	23977	53657	87485	51083
Edges	59835	506550	500000	500000	500000	140778
Time Span (days)	194	2305	201	1350	1875	13395
Nodes in Largest WCC	1893	24668	23906	52477	83497	51083
Edges in Largest WCC	59831	506395	499920	498942	496603	140778
Average clustering coefficient	0.11	0.31	0.08	0.12	0.1	0.02
Number of triangles	14319	1403919	849247	704332	371319	18937
Diameter (Longest shortest path)	8	10	10	13	13	17
Density	0.03	0.00164	0.00174	0.00035	0.00013	0.00011



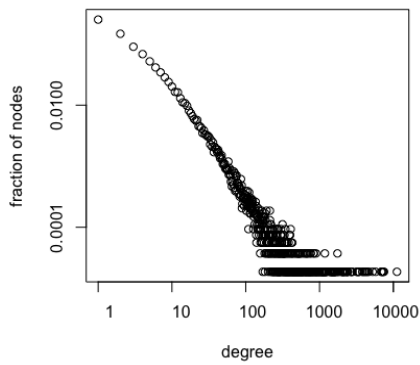
(a) CollegeMsg



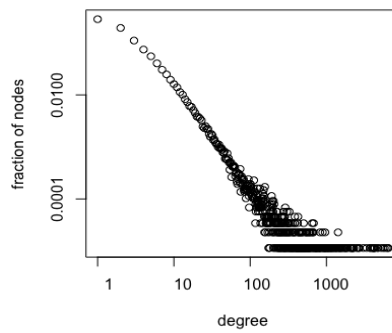
(b) Mathoverflow



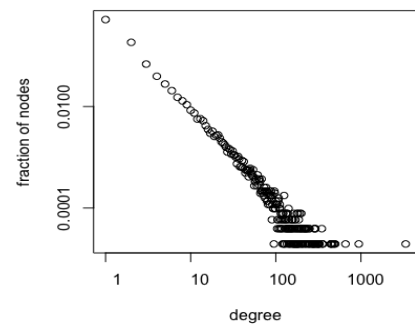
(c) Stackoverflow



(d) Superuser



(e) Askubuntu



(f) Slashdot

Fig. 2. Degree Distribution

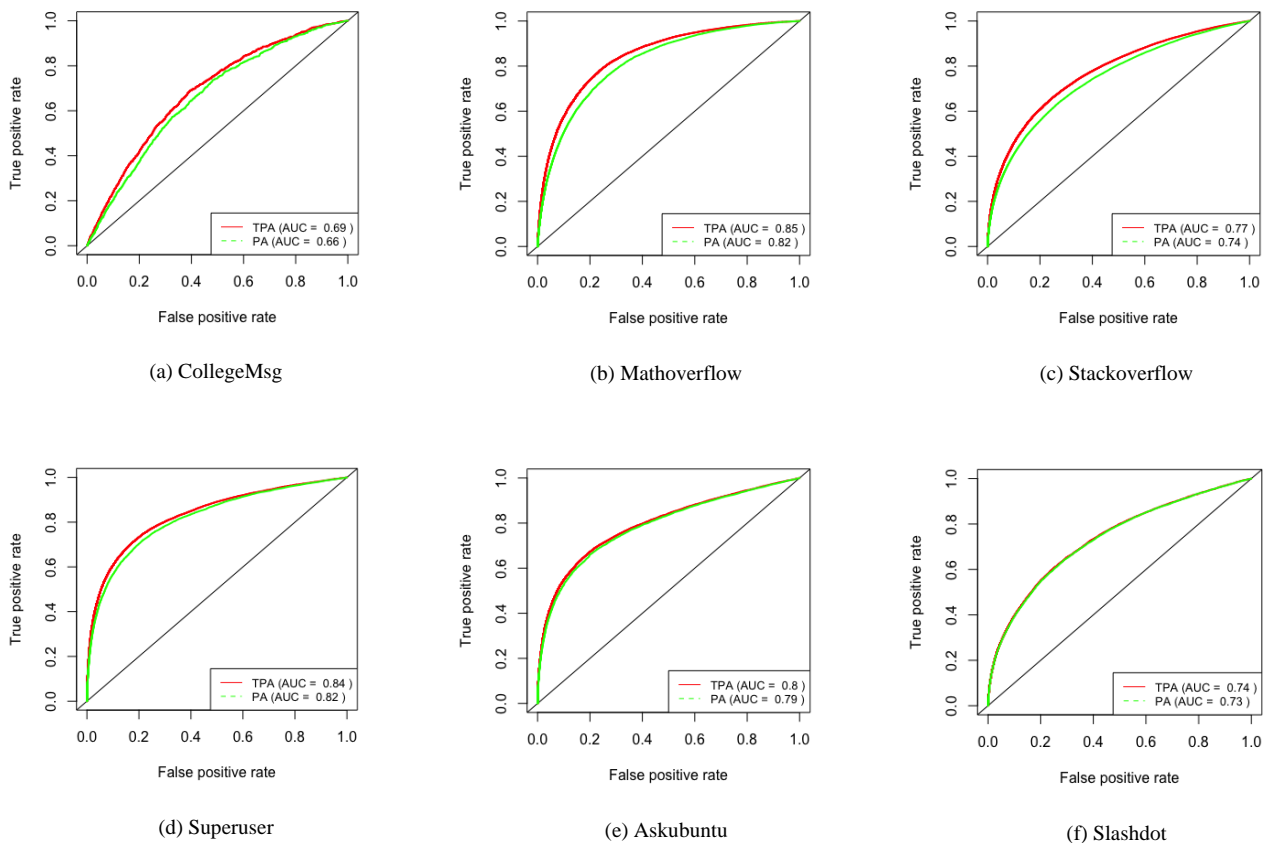


Fig. 3. Model comparison: ROC Curves of PA and TPA

TABLE II: LINK PREDICTION PERFORMANCE OF PA AND TPA. AUC COMPARISON OF PA AND TPA.

Network	AUC of TPA	AUC of PA
CollegeMsg	0.69	0.66
Mathoverflow	0.85	0.82
Stackoverflow	0.77	0.74
Superuser	0.84	0.82
Askubuntu	0.80	0.79
Slashdot	0.74	0.73

B. Results

The summary of the results of the experimental analysis is shown in Table II. It shows that TPA performs better than PA in link prediction in all six social networks. Among them, TPA shows 3% improvement in link prediction accuracy on Mathoverflow, Stackoverflow and CollegeMsg networks. TPA reports 2% improvement in link prediction accuracy on Superuser network. In Askubuntu and Slashdot networks, TPA reports 1% improvement in link prediction accuracy over PA. These results revealed that TPA performs well on most of the question and answering networks. The activeness of the nodes in question and answering networks stays for a short period of time. Once the question gets the right answer, all the interactions with that node stops, and the node becomes

inactive. Then the new links start to emerge around new questions rather than older ones. Owing to this nature, TPA performs better than PA in link prediction.

V. DISCUSSION AND CONCLUSION

Modelling the growth of social networks is a challenging task due to various factors. Among them, the temporality of nodes and edges is a key factor which influences the emergence of new edges. This research introduced a simple yet effective growth model TPA based on the node activeness. The underneath assumption of TPA is each node vv randomly finds an existing node u u to connect according to the probability proportional to the temporal strength of u (see Equation 8).

$$\prod(TS_u|v) = \frac{TS_u}{\sum_{i \in N} TS_i} \quad (8)$$

Here, TS_u is the temporal strength of node u . This growth model somewhat similar to the Fitness model [7]. The key difference is that the Fitness model includes a parameter but the TPA based growth model is non-parametric model. This growth model can be further improved by incorporating homophily and node attributes, which is the future direction of this research.

Although the novel growth model assumed that social networks obey the scale-free property, most of these real world networks do not follow the power law (see Equation 7). Among the social networks used in this study, degree

distributions of Superuser and Askubuntu follow the power law with the exponent of $\gamma = 2.1$. However, degree distributions of Mathoverflow and Slashdot follow the power law with the exponents less than two ($\gamma = 1.7$ and $\gamma = 1.9$). In Stackoverflow and CollegeMsg networks the power law exponents are 3.1 and 3.9 respectively. Typically, the γ of scale-free networks lies in between 2 and 3. The γ value of four above real-world networks stay outside the typical range, which mean that those networks are not typical scale-free networks. According to the Figure 2, the fraction of higher node degrees ($1000 \leq \text{degree}$) are much higher in Askubuntu, Math overflow, Stackoverflow and Superuser networks. It reflects the fact that those networks are growing around the higher degree nodes. Thus, the growth mechanisms of those networks might not fully explained by the power law assumption but still TPA growth model performs better than PA growth model.

Activeness of a node reflects by its interactions with its neighbors. Frequent and recent interactions make the node active. If the node is active then it should make two-way interactions with its neighbors. Otherwise, if the interactions are one-way, which means neighbors to node then the activeness of the node is questionable. In other words, the neighbors interact with the node but the node is not interacting with any of its neighbors. In this case, the node cannot be regarded as an active node. The present research considered both one-way and two-way interactions make the node active. However, it is required to investigate the one-way interactions and two-way interactions separately because in the one-way case only the edge is active but the node might not active. Therefore, it requires thorough investigation about different types of interactions to understand the insights of activeness.

Although TPA shows its own limitations, it shows better performance in link prediction compared to PA. Specially, TPA shows impressive performance over the temporal social networks. In fact, TPA is an effective non-parametric model which can be used to model the temporal social networks as well for link prediction.

REFERENCES

- [1] O. Mokryn, A. Wagner, M. Blattner, E. Ruppim, and Y. Shavitt, "The role of temporal trends in growing networks," *PLOS ONE*, vol. 11, p. e0156505, 08 2016. [Online]. Available: <http://www.cs.tau.ac.il/~ruppin/temporal.pdf>
- [2] F. Gao, K. Musial, C. Cooper, and S. Tsoka, "Link prediction methods and their accuracy for different social networks and network metrics," *Scientific Programming*, 2015.
- [3] G. G. Piva, F. L. Ribeiro, and A. S. Mata, "Networks with growth and preferential attachment: Modeling and applications," 2020.
- [4] M. Newman, "Newman, m.e.j.: Clustering and preferential attachment in growing networks. *phys. rev. e* 64, 025102," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 64, p. 025102, 09 2001.
- [5] A.-L. Barabasi and R. Albert, "Albert, r.: Emergence of scaling in random networks. *science* 286, 509-512," *Science (New York, N.Y.)*, vol. 286, pp. 509-12, 11 1999.
- [6] M. Almeida, G. Mendes, G. Madras, and L. Silva, "Scale-free homophilic network," *The European Physical Journal B*, vol. 86, 02 2013.
- [7] G. Bianconi and A.-L. Barabasi, "Competition and multiscaling in evolving networks," *EPL (Europhysics Letters)*, vol. 54, p. 436, 05 2001.
- [8] P. Sarkar, D. Chakrabarti, and M. Jordan, "Nonparametric link prediction in large scale dynamic networks," *Electronic Journal of Statistics*, vol. 8, 01 2014.
- [9] K. Miller, M. Jordan, and T. Griffiths, "Nonparametric latent feature models for link prediction," in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., vol. 22. Curran Associates, Inc., 2009.
- [10] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Feb 2017. [Online]. Available: <http://dx.doi.org/10.1145/3018661.3018731>
- [11] P. Panzarasa, T. Opsahl, and K. Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," *JASIST*, vol. 60, pp. 911-932, 05 2009.
- [12] K. Zhu, W. Li, and X. Fu, "Modeling population growth in online social networks," *Complex Adaptive Systems Modeling*, vol. 1, 12 2013.
- [13] D. Liben-nowell and J. Kleinberg, "The link prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, 01 2003.
- [14] L. Lu and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150-1170, 2011. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S037843711000991X>
- [15] L. Munasinghe and R. Ichise, "Time score: A new feature for link prediction in social networks," *IEICE Transactions on Information and Systems*, vol. E95.D, p. 821-828, 2012.
- [16] L. Munasinghe and R. Ichise, "Link prediction in social networks using information flow via active links," *IEICE Transactions on Information and Systems*, vol. E96.D, pp. 1495-1502, 2013.
- [17] E. Bu tu 'n, M. Kaya, and R. Alhaji, "Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks," *Information Sciences*, vol. 463-464, pp. 152-165, 2018.

Technology-enabled online aggregated market for smallholder farmers to obtain enhanced farm-gate prices

Malni Kumarathunga*

School of Computer, Data, and
Mathematical Sciences

Western Sydney University, Australia
m.kumarathunga@westernsydney.edu.au

Rodrigo Calheiros

School of Computer, Data, and
Mathematical Sciences

Western Sydney University, Australia
r.calheiros@westernsydney.edu.au

Athula Ginige

School of Computer, Data, and
Mathematical Sciences

Western Sydney University, Australia
a.ginige@westernsydney.edu.au

Abstract - Using scenario transformation methodology, we identified four scenarios that indicated a lack of trusted parties to sell harvest has forced smallholder farmers to sell the harvest to brokers who often collect the harvest at the farm gate at the lowest possible prices and sell in the market for large profits. As blockchain smart contracts provide a mechanism to reduce risk and establish trust between unknown trading partners, we transformed these into a scenario that establishes trust between farmer and unknown broker using smart contracts, generating a trust-enabled market. This scenario enables farmers to search for the optimum farm-gate price without relying on known brokers. The scenario is further enhanced to enable a Many-one-Many market linkage, facilitating automatic aggregated marketing. The paper presents the functional prototype of the scenario, explaining the functionality of the transformed system.

Keywords – aggregated market, blockchain, farmer linkage, smart contracts, trust

I. INTRODUCTION

Of the 570 million farms around the world, 90% of them are considered smallholder farms [1]. 1.5 billion people around the world depend on smallholder agriculture for their livelihood and 75% out of that are the world's poorest people who live in developing economies [2]. They receive only one-third to one-half of the final price for their produce [3] [4] [5]. Although there is a possibility of getting a better price, if the harvest is taken to distant markets, due to cost and lack of storage and transport facilities, rural farmers often sell their produce to a middle man who generates higher profits by procuring harvest at the lowest possible prices. Even though farmers manage to transport the produce to distance markets, they may not be able to compete with dominating larger traders and auction-based sales [6].

A survey carried out in a developing country, Sri Lanka, reveals that while some farmers sell their harvest directly in the market, where selling price changes vigorously, 90% of farmers depend on a middle person or a shopkeeper to sell their harvest [7]. Similarly, in India, fruit and vegetable farmers mainly rely on middlemen who control the market although do not add much value, to sell their produce. Middlemen receive 50% to 71% of the price difference between farm-gate price and resale price [3]. Fafchamps and Hill (2005) affirm that Ugandan farmers tend to sell their produce in the market particularly when the market is close or the quantity of harvest is high, despite the less lucrative farm-gate prices [8]. A survey from Turkey reports that farmers have less bargaining power

when it comes to selling the harvest due to the absence of apparent competition between commission agents [9]. Farmers from Perth, Australia have a major concern about the deductions done and margins received by the market agents [10]. Thus, the unavailability of organized markets and lack of buyers can be considered as some of the foremost reasons for less productive farm-gate prices, leading to poverty-stricken lives for smallholder farmers.

Muamba (2011) states that transformation of farmers' economic status from subsistent or semi-subsistent stage to specialized farmers who produce crops that have a comparative advantage, targeting their products to regional, national, and international markets, can be promoted by greater market participation [11]. Wealth stimulation can occur among farmers who have the potential to overcome the production constraints and the costs of market participation [12]. There are distinct types of markets associated with agriculture. The spot market is characterized by fewer barriers to entry, high transactions costs, and low returns. The contract productions to a known buyer for relatively undifferentiated crops are distinguished by potential barriers to entry, moderate risk of financial loss, and low transactions costs. The contract production to a known buyer for quality differentiated crops is similar to the former with a higher potential of financial returns as well as risks [12].

High marketing and transaction costs restrict smallholder farmers from market participation [3, 13]. Transaction costs can be classified into observable (pecuniary) and unobservable (non-pecuniary) transactions costs [14]. Observable transaction costs are visible when an economic exchange takes place such as transport, handling, packaging, storage, and spoilage. Unobservable transaction costs include information costs, negotiation costs, and monitoring costs [14]. Information Management Systems as an intervention approach have reported positive impacts in improving farmer's market participation and receiving higher farm-gate prices while lessening negative impacts [15] [16]. On the contrary, previous research reveals that there is no significant impact generated by the information intervention if markets are segmented [17] and the farmers have limited options to transport the harvest to the market [4] [17]. Thus, they are forced to sell to local middlemen. Research suggests encouraging farmers and new buyers into agribusiness because the limited competition for farmer's produce is the fundamental cause of lower farm-gate prices [4].

When new buyers enter into agribusiness, concomitant transaction costs arise in the form of information costs and negotiation costs from the farmer's perspective. While providing access to market information can result in reducing transaction costs, leading to higher market participation, facilitating the establishment of trust between farmers and buyers, targeting a trustworthy buyer-seller relationship can promote farmer's participation in markets [18]. Sako (1992) states that a smooth trading relationship requires contractual trust, expecting the promises to be kept, and competence trust, self-reliance in the trading partner's capability on carrying out the task [19]. Blockchain Technology, a distributed ledger platform that provides immutable, transparent, cheaper, faster, trustworthy, and secure transactions over a network with unknown users [20], together with smart contracts, executable code that facilitate execution and enforcement of the terms of an agreement between untrusted parties [21], has the potential of building trust between trading partners.

Thus, this research explores building trustworthy market linkages between farmers and buyers to obtain better farm-gate prices through enhanced market participation based on Blockchain smart contracts. Previous research claims that market linkages that support collective marketing have the potential of generating greater benefits for farmers [22] [23] [24]. Kumarathunga, et al (2020) analyses several online commodity market platforms, revealing most of them support one-to-one market linkages. Although some platforms provide many-to-one market linkages, this provision is implemented manually with the support of field partners who does the collection, limiting the scalability of the platforms [25]. Accessibility to markets depends on the extent of the production [26]. Thus, collectivization into cooperatives, self-help groups, or intermediary contracts is inspired due to the potential of reducing transaction costs for both farmers and the other trading party [13]. Therefore, in this paper, we present a functional prototype of a smart agricultural commodity market platform that supports aggregated marketing while enabling dynamic trust between farmers and buyers.

The remainder of the paper is organized as follows. In section II, we describe our research approach, leading to the functional prototype of the smart commodity market platform in section III and then the discussion in section IV. The conclusion is presented in section IV.

II. RESEARCH APPROACH

This research is carried out following Design Science Research (DSR) methodology, which is a method of addressing important unsolved problems in unique or innovative ways or solve problems in more effective or efficient ways [27]. A good starting point for DSR is identifying and representing opportunities and problems in an actual environment [28]. Improving the environment by introducing novel artifacts and the process of building these artifacts is the desire of design science research [29].

Thus, to understand the selling mechanisms practiced by smallholder farmers, we based our research on Sri Lanka, a developing country in the South Asian region. We selected the area of Nuwara Eliya, which has the major productions of upcountry vegetables such as carrot, beet,

leek, potato, and cabbage [30]. The distance between Nuwara Eliya and the Country's capital city, Colombo is 166.4 km. The manning market in Colombo is the wholesale market of fruits and vegetables grown across the country while Cargills is a supermarket network distributed across the country. Data for our research are gathered through discussions with about 30 smallholder farmers from different sub-areas: Palagolla, Kandapola, Kuda Oya, and Hawa Eliya. While the sub-areas are chosen randomly with the heuristic of representing the majority of the farming community, farmers are chosen according to the farm size, so the selected farmers are smallholders. The sample size of smallholder farmers is decided according to the Grounded Theory which emphasizes the flexibility of deciding the sample size as the research progresses. The researcher does the collection and analysis of data simultaneously, leading to real-time judgments on whether further data collection produces additional or novel contributions [31]. The sample size is decided when the researcher perceives that theoretical saturation is achieved [32]. Thus, the theories derived from the collected data are more likely to resemble reality [31]. According to DSR, the design cycle is the heart of any research project [28]. We chose the Scenario-based design method as the process of designing the artifact.

Scenario-based design is a family of techniques that uses to concretely describe how people will use a future system to accomplish tasks and activities at an early point in the development process rather than defining the system operations. A scenario is a story that describes actions and events that lead to a consequence. The goals, plans, and reactions of the people in the story are described as the actions and events [33]. Scenarios emphasize the people and their experiences, directing the user-appropriateness of the design ideas to the main focus. Design ideas can be refined from the feedback of the stakeholders about usage possibilities and concerns. Thus, the design will remain focused on users' needs and concerns since the scenario describes how the users will use the future system [33]. According to the discussions with farmers, we were able to develop 4 different scenarios on farmers' selling mechanisms as listed in Table I. The second step is analysing the scenarios to derive claims for each scenario, identifying the causal relationships. Next, each claim from each scenario is further analysed to derive positive and negative consequences [33]. The claims and consequences derived from the scenarios in Table I are listed in Table II. Deriving claims and their positive and negative consequences initiate originating some design moves with the heuristic of maintaining or even enhancing the positive consequences for the actors of the system while minimizing or eliminating the negative consequences [33]. Following this heuristic led us to perceive that farmers often choose a broker or buyer with pre-established trust, although they receive money later and the prices are low as illustrated in Table III. The level of trust reduces from top to bottom in the table. Thus, the process revealed the first design move of a future system.

- The system requires a mechanism to establish trust between farmers and unknown brokers to enable

TABLE I. SCENARIOS OF CURRENT SELLING MECHANISMS

Scenario 1
Bandara is a 45 years old farmer from Palagolla, Nuwara Eliya. He is a member of a farmers' society and has a farmer code given by the society. He grows carrot, leek, beets, and cabbage on his 2 acres' farm. He sells a certain amount of his harvest to Cargills supermarket who transfers the payable to a nominated bank account. He sells another certain amount of harvest to local brokers who pay within 2 or 3 weeks. Most of his harvest is sent to the manning market in Colombo in a truck. The truck driver (Sunil) comes to the farm. Bandara loads the harvest to the truck, writes a letter to the broker (Chinthaka) in Colombo, including the farmer code and quantities of each type of vegetable. Chinthaka decides the rates for each vegetable and the payable amount after deducting 2kg of vegetables for each 50kg bag as wastage. Chinthaka pays the transport charge to Sunil and reduces it from the payable amount. Then Chinthaka transfers the payable amount to a Bandara's nominated bank account after reducing a commission for selling the harvest from the payable amount.
Scenario 2
Nishantha is a 35 years old farmer from Kandapola, Nuwara Eliya. He grows carrots and leeks on a 1-acre farm. Nishantha sells his harvest to a local broker (Kamal) because Nishantha has trust in Kamal's paying back. In harvesting season, Kamal comes with a group of labours to help him with harvesting, but Nishantha does not have to pay for them. Kamal pays them. Nishantha and Kamal agree with a rate for the harvest, usually less than the rate in the Nuwara Eliya Dedicated Economic Centre. Nishantha does not know the rate Kamal sells. Usually, Kamal pays Nishantha within 2 or 3 weeks.
Scenario 3
Kalum is a 40 years old farmer from Kuda Oya, Nuwara Eliya. He grows leeks, carrots, and radishes on his 1/2 acres farm. He sells his harvest to a local broker (Namal) who pays Kalum within 2 or 3 weeks at an agreed rate. Sometimes he sells his harvest to an unknown broker for a lower rate because the unknown broker pays money on the spot.
Scenario 4
Ishan is a 50 years old farmer from Hawa Eliya, Nuwara Eliya. He grows carrots and potatoes on his 3/4 acres farm. In harvesting season, he makes a call to a broker (Nadun) from the Nuwara Eliya Dedicated Economic Centre, asks him to collect the harvest, and makes an agreement with the rate. Ishan harvests the potato and makes them ready for selling. But Nadun harvests carrots with the help of his labours. Nadun transports them to the centre. After 2 or 3 weeks, Nadun transfers the payable amount to Ishan's nominated account.

farmers to choose any broker who offers comparative rates without relying on known brokers.

Next, we realised that the quantities produced by these farmers are little due to the small extent of the farmlands, thus the cumulative of both observable and unobservable transaction costs can result in lower margins for marginal and small scale farmers. However, research has demonstrated that trading collectively has the potential of reducing transaction costs with better coordination [24], leading to higher revenues for farmers [23] from better bargaining positions [22]. Thus, the second design move is generated to facilitate aggregated marketing.

- The system requires a mechanism to support a market linkage that facilitates aggregated marketing for farmers to obtain better rates.

Both design moves are used to develop the transformed scenario. We presented the transformed scenario and the conceptual model for an online agricultural commodity market platform in a previous conference paper [25]. In this paper, we present the modified conceptual model to develop a functional prototype for a smart agricultural market platform. For simplicity, when explaining the

market platform, we have used buyer for both buyer and broker.

III. SMART AGRICULTURAL COMMODITY MARKET: THE FUNCTIONAL PROTOTYPE

The modified conceptual model of the smart agricultural commodity market platform is illustrated in Fig. 1. It has 5 major components.

A. Digital agribusiness ecosystem (DAE)

Digital Agribusiness Ecosystem, previously known as Digital Knowledge Agribusiness Ecosystem [34], consists of a database that has quasi-static information about crops, pests and diseases, land preparation, and growing and harvesting methods. It provides this information as actionable information to farmers through mobile apps. Two mobile apps called "Govi Nena" and "Gayankisan" are already being deployed and used by farmers in Sri Lanka and India respectively. When the farmer feeds what to grow and when to grow to the system through the mobile app, DAE provides a detailed cost of cultivation for each crop and crop calendar outlining essential tasks he should carry out to optimize yield as well as to manage pests and diseases better, leading to optimal output. DAE has the capability of predicting the expected harvest and expected harvesting date for each crop for each farmer according to the season and location [34].

B. Web site

Since DAE is capable of predicting the expected harvest for each farmer for each crop, the harvest can be aggregated based on geographical proximity, crop type, and expected harvesting date. Thus, many farmers can be clustered into one group according to the same parameters and made available to many buyers, forming Many-one-Many market linkages between them, enabling aggregated marketing. This market linkage is demonstrated in Fig.2. While the crops are still in the growing stage, the aggregated harvest according to the farmers' group is made available for buyers through the website in advance as displayed in Fig. 3. The harvest aggregation can be done according to administrative divisions in a country. For example, the administrative divisions in Sri Lanka are province, district, divisional secretariat division (DS Division), and Grama Niladhari division (GN Division – the lowest grass-root level division) [35]. Thus, for the buyers in Sri Lanka, aggregation can be carried out up to the GN division level. The buyers can fill in a bid form in the website as in Fig.4, entering the crop type he expects to buy, grade, the expected buying period, location, quantity, and the offered price.

C. Mobile app

Mobile App will be developed as an extension to existing apps in the ecosystem. When a buyer submits a bid, the bid is sent only to the mobile apps of a certain group of farmers as displayed in Fig.5. This filtration is executed against the geographical proximity, crop type, and expected harvesting date so that the buyer is facilitated with easy coordination and collection of the harvest during the harvesting period. When a farmer receives the bid, he has three options to correspond as displayed in Fig. 6.

TABLE II. ANALYSING THE FOUR SCENARIOS

	Claim	Consequences
1	has 2 acres farm	- produces small quantities of harvest
	sell the harvest to the Cargills supermarket.	+ has an agreed price and trust of paying - farmer has to do cleaning, grading, and packing
	sends the harvest to manning market in Colombo in a truck	+ gets his money transferred into his bank account + able to discharge his excess productions - does not know the rate which the buyer is going to sell his vegetables and the rate he will get - has to agree with any rate the seller decides because the harvest is already given - broker reduces 2kg for each 50kg as wastage. It is 4% of the total value. - transporting the vegetable-packed in a truck increases the wastage - broker reduces a commission for selling the vegetables. - farmer gets a little profit at the end when all deductions are made
2	Has 1 acre farm	- produces small quantities of harvest
	sells the harvest to the local buyer	+ no harvesting cost + no transporting cost + has developed mutual trust between farmer and buyer - receives the money within 2 or 3 weeks - rates are little less than in the economic centre
3	has 1/2 acres farm	- produces small quantities of harvest
	sells the harvest to a broker	+ gets money on the spot + no need to build trust between the farmer and the buyer - rates are low
4	has 3/4 acres farm	- produces small quantities of harvest
	local broker does the carrot harvesting and transports them to the economic centre	+ farmer does not have to bear a cost for harvesting carrot + farmer does not have to pay the transport charge + vegetable that goes to the market is fresh + farmer does not need storage for vegetable + harvesting labours may be experienced in harvesting, so the wastage is little
	sells the harvest to the local broker	+ has developed trust between farmer and broker + receives money to his bank account - receives the money within 2 or 3 weeks - rates are little less than the rates in the economic centre

1) *Accept the offer*

If a farmer is pleased with the price offered by the buyer, he can accept the bid by entering the amount of harvest he expects to sell at that price. The bid has an expiry date. Therefore, the farmer can accept it until the expiry date. However, if other farmers who received the same offer, accept the offer before him, the offer quantity can be saturated before the expiry date, supervening the expiration.

2) *Provide a counteroffer*

If the farmer is not content with the price, he is facilitated with the option of providing a counteroffer, entering a new price, and the amount expected to sell at that new price. Farmers can choose this option if the bid price is very low. In this case, the farmer is supposed to wait for the particular buyer's acceptance or rejection.

3) *Reject the offer*

The third option is to reject the offer if the price offered is not satisfactory enough. However, the farmer can anticipate more bids with different prices since the bids are for the expected harvest, not a ready lot.

When a farmer chooses one of the above three options, it is sent to the Contract Negotiator Module.

D. *Contract negotiator module (CNM)*

Contract Negotiator Module (CNM) is a server-side software module that maintains the coordination and communication between the farmer and buyer. CNM stores the bids offered by buyers and responses from the farmer in a database. Once an offer is saturated or expired, CNM analyses all the responses received from the farmers against the buyer's bid. This analysis can produce one of the following two results.

1) *The amount in total accepted offers = buyer's requirement*

TABLE III. FARMERS' CHOICE ORDER

Scenario 1	Scenario 2	Scenario 3	Scenario 4
Cargills Super Market (Agreement)	Local Broker	Local Broker	Broker from Nuwara Eliya Trade Center
Local Broker		Unknown Broker	
Manning Market in Colombo			

Since the buyer's requirement is fulfilled, CNM sends a notification to the buyer mentioning that his offer has been accepted by farmers, and requests his confirmation on whether he is intended to continue to the next step of establishing a contract.

2) *The amount in total accepted offers < buyer's requirement, but there are some offers from farmers with a higher price*

In this case, the CNM sends a notification to the buyer, stating that only a portion of his offer is accepted by farmers for the offered price. It also mentions that his requirement can be fulfilled at a higher price if he accepts the counter offers submitted by the farmers. If the buyer consents to the counteroffer price, that price is applicable for all the farmers who accepted that offer, not only for the farmer who submitted the counteroffer.

Once the buyer confirms his willingness to continue with the purchasing process, the next step is to establish a contract between the farmers and buyer. Thus, CNM asks each farmer to deposit 10% of the agreed total amount and the buyer to deposit 10% of the agreed total amount. These amounts are required as an honor to the contract that will be established between them. The buyer will be provided three options to pay the balance 90% of the total price according to the farmer's choice:

- deposit it in the system at the point of establishing the contract, so when the harvest is collected, the money is sent to the farmer, otherwise sent back to the buyer
- organize a cash payment at the time of collecting the harvest
- pay 3 days/ 1 week/ 2 weeks after collecting harvest (this depends on the buyer's rapport) – this can be done directly or through the system

The buyer and the farmers can do the deposit in the form of fiat money either via mobile money or e-banking. Once the deposits are done, the amount is converted into a unique type of cryptocurrency and sent into a blockchain network along with farmer's and buyer's data to establish a contract in the form of a smart contract. When the expected buying period approaches, the CNM requests confirmation from both parties whether the harvest delivery is performed, before sending an invoke message to the blockchain platform to execute the smart contract to transfer the money accordingly. When the smart contract is executed, the cryptocurrency is converted into fiat money and transferred to the relevant financial account: mobile money account or bank account.

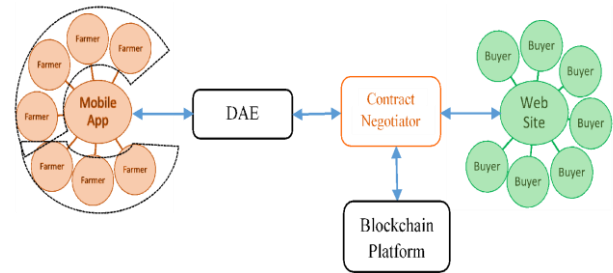


Fig. 1. Conceptual model of the proposed platform

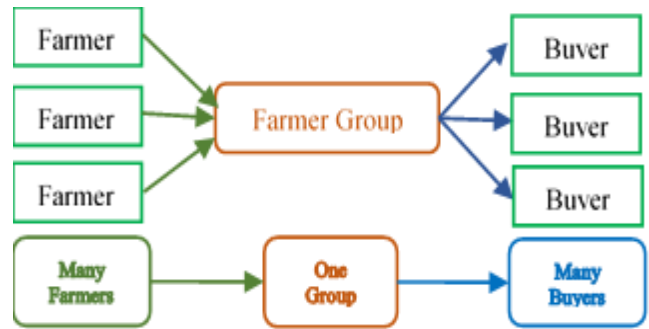


Fig. 2. Many-one-many market linkage

All the transactions are stored in a database to produce ratings and rankings for both farmers and buyers according to their behavior of honoring the contracts. The rank of the buyer or farmer will be calculated according to the number of successful transactions and the total number of contracts, while the rating is established according to the reviews received. When a farmer receives the offer, he can tap on the unique buyer id listed in the offer to see the buyer's rank and the ratings received from previous transactions. Similarly, when the buyer receives acceptance from a farmer, he can tap on the farmer's unique id to view the farmer's rank and ratings. This feature generates the possibility of establishing online trust between farmers and buyers.

E. Blockchain network

A Blockchain network is integrated into this platform to facilitate the process of contract establishment. When it receives a deploy message from CNM with the required data: farmer's data, buyer's data, the amount of cryptocurrency sent by both farmer and buyer, crop type, grade, expected harvesting period, agreed price, and amount of harvest for the particular crop, it deploys a new smart contract. Once it receives an invoke message from the CNM, it releases the cryptocurrency stored in the particular smart contract's account and let the CNM aware that the smart contract is executed.

IV. DISCUSSION

According to the scenarios derived from the discussions with farmers, we observed that farmers are in a trust bubble with a small number of brokers. They prefer selling the harvest to a known broker even at a lower price due to pre-established trust of getting paid although they receive money after 2/3 weeks. However, as farmers do not step out of their trust bubble, they miss the opportunity of

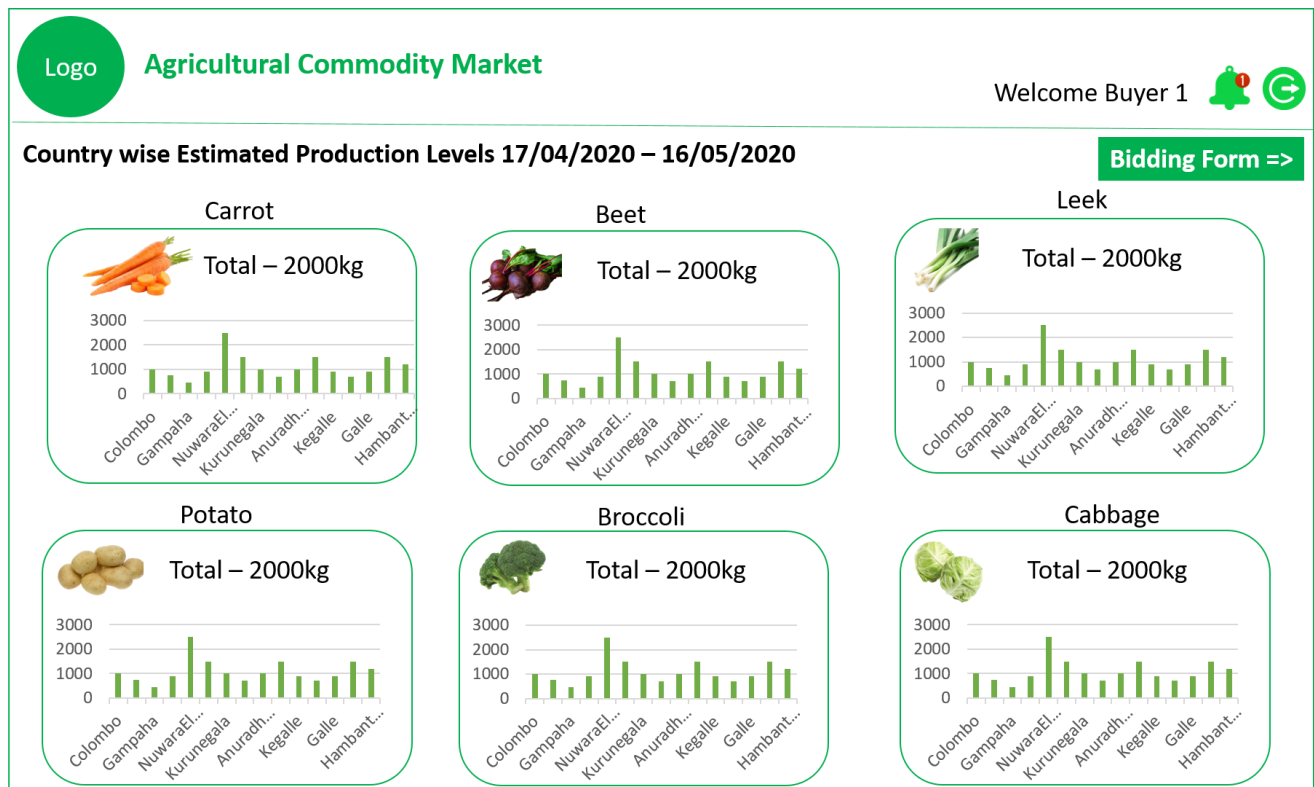


Fig 3. User interface for logged in buyers

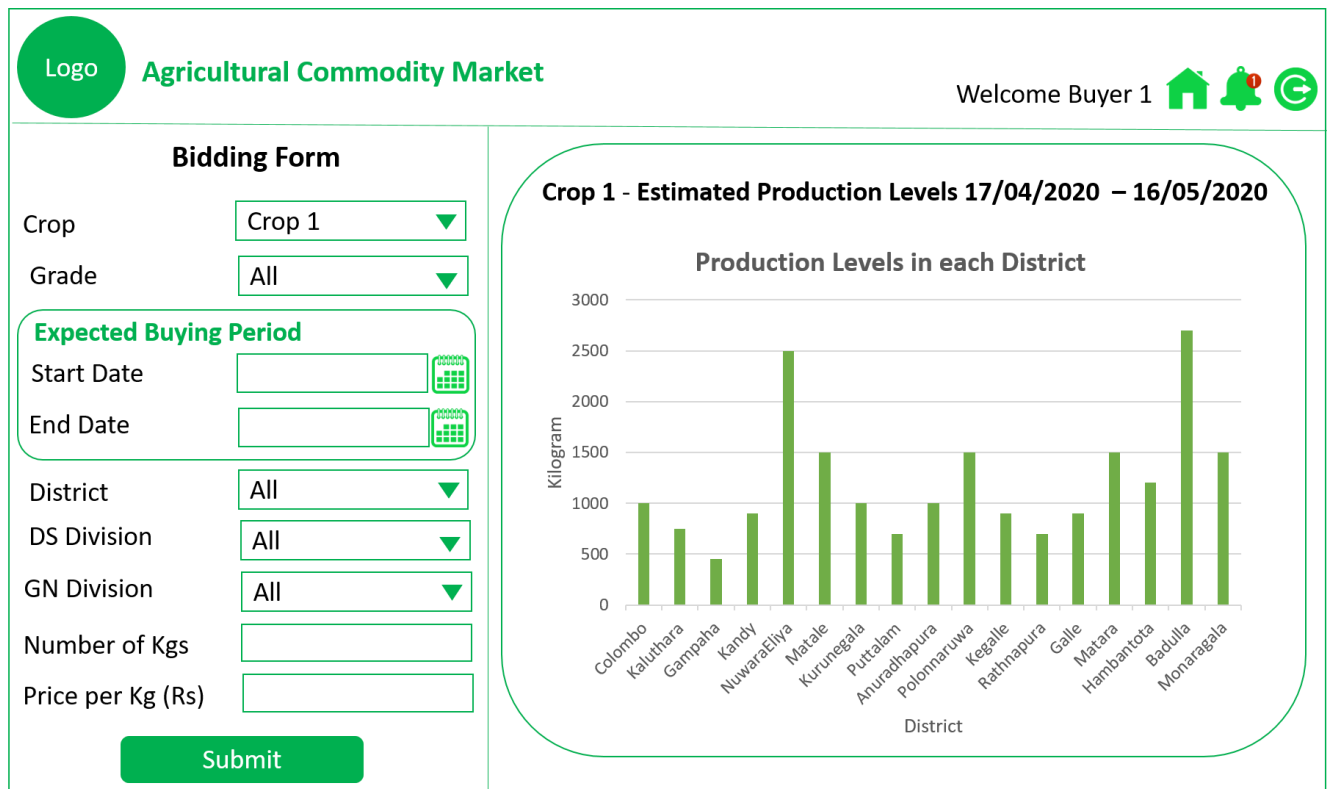


Fig 4. Bidding form

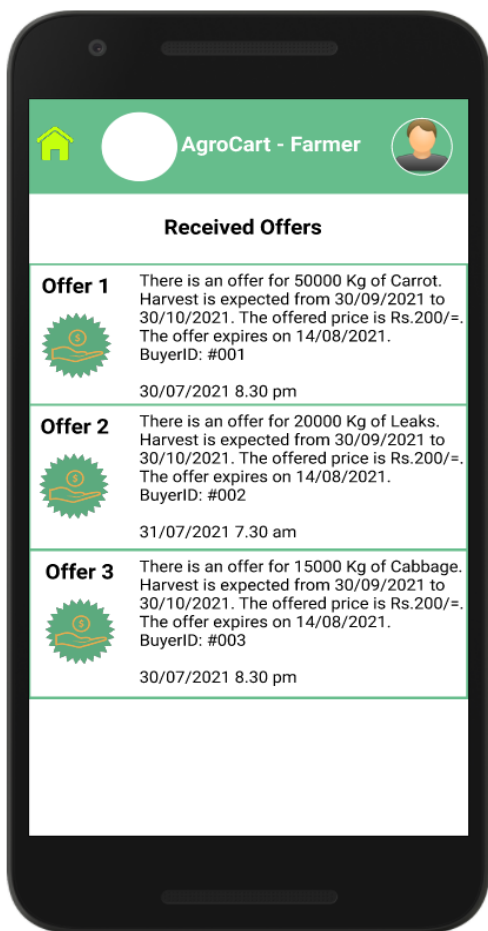


Fig. 5 (a). Received offers for the farmer

selling their harvest at a competitive price to an unknown broker. They do spot selling to unknown brokers only when they need instant money because on-the-spot buying brokers attempt to procure at the lowest possible price targeting higher margins. These findings correlate with research done by Batt (2003) among farmers in Perth, Australia. The researcher states that although farmers expected to transact with a market agent who offers the highest price, the highest price does not assure being paid. He further declares that farmers are paid after 14-21 days once the goods are received by the market agent [10].

The proposed smart agricultural commodity market platform provides a strategy for farmers to step out from their trust bubble for better price determination. While they receive a competitive price for their produce as a reward, they confront the risk of not being paid since the broker is now unknown, and there is no pre-established trust. To mitigate this risk and build trust, the proposed platform generates a Blockchain smart contract which executes by itself when the predefined terms are met. Thus, farmers can choose any broker who offers better rates. Once a broker agrees to buy harvest from the farmer at a specific rate, they can enter into a contract with agreed terms. The contract will ensure the payment is transferred to the farmer according to contract terms. Thus, this enables farmers to select any broker, guaranteeing an optimal price while assuring payments because the smart contract deployed on Blockchain is secure from vagaries from both farmer and the broker. The 10% deposit is proposed to compensate the

victim party if the other party did not follow the contract conditions.

Therefore, the static relationship between farmer and the known broker has transformed into a trust-enabled dynamic relationship between farmer and unknown broker

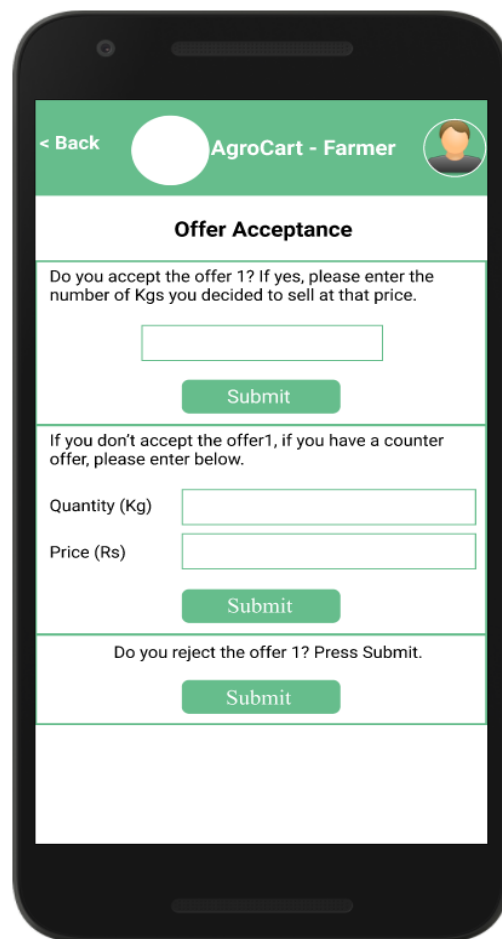


Fig. 5 (b). The three option farmer gets

since farmers are not bound to sell their harvest only to known brokers.

While eliminating the middleman and selling the harvest directly to buyers seems to be effective in reducing costs and getting better prices, transportation, storage cost, and wastage can negate these benefits. Besides, the middlemen can lose their source of revenue. Therefore, facilitating a dynamic trust-enabled relationship between farmer and broker is preferably more realistic for underprivileged farmers with no transportation or storage facilities, maintaining the existing nature of their agribusiness while increasing the number of brokers that a farmer can choose. While the existence of multiple brokers in the system can influence the farm-gate prices, it also eliminates the vulnerability of farmers, who have limited outside options, being abandoned by brokers. If brokers relinquish their business in some rural regions, farm-gate prices tend to decline dramatically as farmers have to adapt to available options. However, the exit of few brokers will not affect farmers since the platform facilitates farmers to choose any broker/buyer who offers comparative rates. Furthermore, the farmers will not have to rely only on brokers if they possess a transport advantage since the

trading can occur directly between farmer and buyer, eliminating the middleman. This feature also can lead to generating higher profits for farmers with better prices.

Following features are supported in the platform generating benefits not only for the farmer but also for the broker.

- forming automatic farmer groups enabling many-one-many market linkages, facilitating aggregated marketing

Thus, the farmer is enabled to achieve better prices with high bargaining power and low transaction costs, while provided access to bigger markets, enabling them to target regional, national, or international markets. Meanwhile, the buyer can collect the harvest with the least transaction costs due to better coordination between them.

- establishing contracts between farmer and buyer in advance in the form of smart contracts, reducing contract establishment costs

With an established contract, the farmer has an option to secure trade with a buyer who offers better rates, even the crop is still in the growing stage to reduce future market risks. Similarly, the buyer gets to secure a business opportunity. The pre-harvest and post-harvest wastage are minimized due to enhanced coordination with prior knowledge of buying period.

- enabling dynamic trust through blockchain smart contracts and rating and ranking system

The farmer is empowered to choose any buyer with comparative rates without relying on the known brokers from his trust bubble due to the dynamic trust enabled by the system. The rating and ranking system along with blockchain smart contracts contributes to building trust and reducing the risk of not getting paid. Since both parties deposit 10% of the total agreed amount as an assurance to honour the contract, in a case of breaching the contract, the victim is paid that deposit. Thus, the loss is minimized.

- Empowering both farmer and buyer to manage risks through disaggregation and aggregation

The farmer can disaggregate his production according to the grades and sell to different buyers at different prices. This process has the potential of reducing the overall risks by breaking down the risk into several parts since there is less probability for all the buyers to act unfaithfully at once. Similarly, from the buyer's perspective, he is enabled to aggregate the harvest from several farmers according to his requirement. Thus, risks are disaggregated in the cases of contract breaching from the farmer's side.

- facilitating buyers to pay the balance of 90% of the total agreed amount in three options

Since the buyers are getting three options to pay the balance, they can manage their finances according to their financial status.

Since a survey done in Sri Lanka reveals that 90% of farmers depend on brokers or shop keepers to sell their

harvest [7], we can reach an implication that the developed scenarios represent the majority of the farming community in Sri Lanka. According to MEAS 2014 report, the most accessible market for the majority of smallholder farmers in developing countries is the informal market where the price is discovered through arbitrary combinations of supply and demand, trader cartels, and customer loyalties for a particular buyer. However, 80-90% of agricultural products are traded in such informal markets, including farm gate sales, roadside sales, village markets, rural assembly markets, and urban wholesale and retail market sales [2]. All the harvest sales in the 4 scenarios we developed can be positioned in one of the above-mentioned informal markets. Thus, it enables the generalisation of the proposed commodity market platform for different types of crops in different areas, not only for Sri Lanka but also for other developing countries in the future. During this generalisation phase, there will be a step to identify the administrative divisions for the particular country to effectuate the farmer groups and production aggregation according to geographical proximity.

Although this is still in the functional prototype stage, we compared the proposed commodity market with existing blockchain-enabled markets for agricultural commodities with similar approaches. Liao, et al (2020) have presented an integrated market platform for contract production called BeIMP, targeting small-scale farmers [36]. One of the main differences between BeIMP and the proposed commodity market platform in this paper is the market linkages supported by both markets. While BeIMP supports one-to-one market linkage between farmers and buyers, the proposed market supports Many-one-Many market linkages, enabling aggregated marketing. A decentralized agricultural platform called KHET is being proposed to encapsulate the whole agricultural process, eliminating all the intermediaries from land renting to harvest selling. The markets in the KHET platform establish pre-contracts with farmers to buy farmer's produce [37]. Thus, KHET does not support aggregated marketing for farmers. A Community Supported Agriculture (CSA) model is proposed in the context of Vietnam, targeting small and tiny businesses. In this model, the end consumer directly pays the farmer in advance, sharing the risk with the farmer. However, this model has integrated blockchain for traceability option only and farmers do not have access to bigger markets through aggregated marketing [38]. Therefore, the proposed commodity market platform is distinguished from markets with similar approaches due to the aggregated marketing feature.

However, there is a possibility that farmers do not honour the contracts due to reasons beyond their control such as natural disasters and scarcity of Agri inputs. In such cases, farmers have to face the loss from both the harvest loss and the deposit loss due to the nature of the contract established. Thus, in the future, we expect to integrate the system with harvest insurance providers to ensure that the farmer is secured from such massive losses.

The initial proof-of-concept prototype of the market is developed as a website using HTML, CSS, and Typescript in frontend and node.js and MySQL in the backend. The prototype is evaluated to test the feasibility with the participation of experts in the Agri industry. According to DSR, generated design alternatives must evaluate against

the requirements until a satisfactory design is achieved [29]. Thus, based on the feedback from the experts from the Agri industry, the second prototype is decided to be developed as a mobile application, instead of a website. Furthermore, the feasibility of farmers paying 10% of the total agreed amount as an honour to the contract will be evaluated with the implementation of the second prototype.

V. CONCLUSION

High transaction costs, poor physical and institutional infrastructure, absence of market information, and insufficient markets inhibit smallholder farmers from market participation. We perceived that due to a lack of trusted buyers, farmers often choose the same brokers with pre-established trust although the rates they offer are low and receive money after 2/3 weeks. They sell the harvest to unknown brokers only if they receive money on the spot due to the risk of not getting paid and lack of trust. Thus, we present a functional prototype that supports a strategy to transform the static trust between farmers and known brokers into dynamic trust between farmers and unknown buyers. The prototype generates more options for farmers, enabling them to choose any buyer with comparative rates, generating competition among buyers that lead to better prices for farmers' harvest. Furthermore, supporting aggregated marketing through Many-one-Many market linkages results in reducing transaction costs for both farmer and buyer, facilitating farmers to generate higher profits with greater bargaining position. Thus, the proposed smart agricultural commodity market has the potential of uplifting the economic status of smallholder farmers, enabling them to receive better prices while reducing the transaction costs in market participation. Once the validity and feasibility is tested, the implementation of this platform will contribute to alleviating poverty among smallholder farmers and uplift their livelihoods.

REFERENCES

- [1] Food and Agriculture Organization. (19/06/2021). "Smallholder Family Farms". Available: <http://www.fao.org/economic/esa/esa-activities/smallholders/en/>
- [2] S. Ferris *et al.*, "Linking Smallholder Farmers to Markets and the Implications for Extension and Advisory Services," in "Modernizing Extension and Advisory Services," United States Agency for International Development 05/2014 2014, Available: https://www.agrilinks.org/sites/default/files/resource/files/MEAS%20Discussion%20Paper%204%20-%20Linking%20Farmers%20to%20Markets%20-%20May%202014_0.pdf, Accessed on: 19/06/2021.
- [3] I. Somashekhar, J. Raju, and H. Patil, "Agriculture Supply Chain Management: A Scenario in India," *Research Journal of Social Science and Management, RJSSM*, vol. 4, no. 07, pp. 89-99, 2014.
- [4] S. Mitra, D. Mookherjee, M. Torero, and S. Visaria, "Asymmetric Information and Middleman Margins: An Experiment with Indian Potato Farmers", *Review of Economics and Statistics*, vol. 100, no. 1, pp. 1-13, 2018.
- [5] (2020). AgStat. Available: <http://doa.gov.lk/SEPC/images/PDF/AgStat2020.pdf>
- [6] R. Ranjan, "Challenges to Farm Produce Marketing: A Model of Bargaining between Farmers and Middlemen under Risk", *Journal of Agricultural and Resource Economics*, vol. 42, no. 3, pp. 386-405, 2017.
- [7] P. Di Giovanni *et al.*, "User centered scenario based approach for developing mobile interfaces for Social Life Networks," ed, 2012, pp. 18-24.
- [8] M. Fafchamps and R. V. Hill, "Selling at the Farmgate or Traveling to Market," *American Journal of Agricultural Economics*, vol. 87, no. 3, pp. 717-734, 2005.
- [9] S. Lemeilleur and J.-M. Codron, "Marketing cooperative vs. commission agent: The Turkish dilemma on the modern fresh fruit and vegetable market," *Food Policy*, vol. 36, no. 2, pp. 272-279, 2011.
- [10] P. Batt, "Building trust between growers and market agents," *Supply Chain Management*, vol. 8, no. 1, pp. 65-78, 2003.
- [11] F. Muamba, "Selling at the farmgate or travelling to the market: A conditional farm-level model," *The Journal of Developing Areas*, vol. 44, no. 2, pp. 95-107, 2011.
- [12] D. Boughton *et al.*, "Market participation by rural households in a low-income country: An asset based approach applied to Mozambique," *Faith Economics*, vol. 50, pp. 64-101, 2007.
- [13] P. S. BIRTHAL, A. K. Jha, and H. Singh, "Linking farmers to markets for high-value agricultural commodities," *Agricultural Economics Research Review*, vol. 20, no. conf, pp. 425-439, 2007.
- [14] G. Holloway, C. Nicholson, C. Delgado, S. Staal, and S. Ehui, "Agroindustrialization through institutional innovation Transaction costs, cooperatives and milk-market development in the East-African highlands," *Agricultural economics*, vol. 23, no. 3, pp. 279-288, 2000.
- [15] P. Courtois and J. Subervie, "Farmer Bargaining Power and Market Information Services," *American Journal of Agricultural Economics*, vol. 97, no. 3, pp. 953-977, 2015.
- [16] E. Nakasone, "The Role of Price Information in Agricultural Markets: Experimental Evidence from Rural Peru," *IDEAS Working Paper Series from RePEc*, 2013.
- [17] M. Fafchamps and B. Minten, "Impact of SMS-Based Agricultural Information on Indian Farmers," *The World Bank Economic Review*, vol. 26, no. 3, pp. 383-414, 2012.
- [18] H. Lu, J. H. Trienekens, S. W. F. Omta, and S. Feng, "Influence of guanxi, trust and farmer-specific factors on participation in emerging vegetable markets in China," *NJAS - Wageningen Journal of Life Sciences*, vol. 56, no. 1, pp. 21-38, 2008.
- [19] M. Sako, Price, quality and trust: Inter-firm relations in Britain and Japan (no. 18). Cambridge University Press, 1992.
- [20] S. Underwood, "Blockchain beyond Bitcoin. (other applications of blockchain technology) (News)," *Communications of the ACM*, vol. 59, no. 11, p. 15, 2016.
- [21] M. Alharby and A. van Moorsel, "Blockchain-based Smart Contracts: A Systematic Mapping Study," *Fourth International Conference on Computer Science and Information Technology*, 2017.
- [22] D. Roy and A. Thorat, "Success in high-value horticultural export markets for the small farmers: The case of Mahagrapes in India," *World Development*, vol. 36, no. 10, pp. 1874-1890, 2008.
- [23] E. Fischer and M. Qaim, "Linking Smallholders to Markets: Determinants and Impacts of Farmer Collective Action in Kenya," *World Development*, vol. 40, no. 6, pp. 1255-1268, 2012/06/01/2012.
- [24] H. Markelova, R. Meinzen-Dick, J. Hellin, and S. Dohrn, "Collective action for smallholder market access," *Food policy*, vol. 34, no. 1, pp. 1-7, 2009.
- [25] M. Kumarathunga, R. Calheiros, and A. Ginige, "Towards Trust Enabled Commodity Market for Farmers with Blockchain Smart Contracts," in *Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference*, Nagoya, Japan, 2020, pp. 75-82.
- [26] A. Ali, A. Abdulai, and D. B. Rahut, "Farmers' Access to Markets: The Case of Cotton in Pakistan," *Asian Economic Journal*, vol. 31, no. 2, pp. 211-232, 2017.
- [27] A. Hevner and S. Chatterjee, "Design science research in information systems," in *Design research in information systems*: Springer, 2010, pp. 9-22.
- [28] A. R. Hevner, "A three cycle view of design science research," *Scandinavian journal of information systems*, vol. 19, no. 2, p. 4, 2007.
- [29] H. A. Simon, *The sciences of the artificial*. MIT press, 2019.
- [30] D. P. B. Dharmasena. (2017, 29/07/2019). High Input Vegetable Cultivation in Central Highlands of Sri Lanka. Available: https://www.academia.edu/34274054/High_Input_Vegetable_Cultivation_in_Central_Highlands_of_Sri_Lanka.
- [31] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Citeseer, 1998.
- [32] D. Silverman, *Doing qualitative research: A practical handbook*. Sage, 2013.
- [33] M. B. Rosson and J. M. Carroll, "Scenario-based design," in *Human-computer interaction*: CRC Press, 2009, pp. 161-180.

- [34] A. Ginige et al., "Digital Knowledge Ecosystem for Achieving Sustainable Agriculture Production: A Case Study from Sri Lanka," in 3rd IEEE International Conference on Data Science and Advanced Analytics, Montreal, QC, Canada, 2016, pp. 602-611.
- [35] Wikipedia. (25/08/2021). Administrative divisions of Sri Lanka. Available: https://en.wikipedia.org/wiki/Administrative_divisions_of_Sri_Lanka
- [36] C.-H. Liao, H.-E. Lin, and S.-M. Yuan, "Blockchain-Enabled Integrated Market Platform for Contract Production," IEEE Access, vol. 8, pp. 211007-211027, 2020.
- [37] S. Paul, J. I. Joy, S. Sarker, S. Ahmed, and A. K. Das, "An Unorthodox Way of Farming Without Intermediaries Through Blockchain," in 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 2019, pp. 1-6: IEEE.
- [38] D. H. Nguyen, N. H. Tuong, and H.-A. Pham, "Blockchain-based Farming Activities Tracker for Enhancing Trust in the Community Supported Agriculture Model," in 2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020, pp. 737-740: IEEE.

Automatic road traffic signs detection and recognition using ‘You Only Look Once’ version 4 (YOLOv4)

W. H. D. Fernando*
Department of Mathematics
Eastern University, Sri Lanka
harshadfernando@gmail.com

S. Sotheeswaran
Department of Mathematics
Eastern University, Sri Lanka
sotheeswarans@esn.ac.lk

Abstract - This paper presents an approach to detect traffic signs using You Only Look Once version 4 (YOLOv4) model. The traffic sign detection and recognition system (TSDR) play an essential role in the intelligent transportation system (ITS). TSDR can be utilized for driver assistance and, eventually, driverless cars to reduce accidents. When driving an automobile, the driver's attention is usually drawn to the road. On the other hand, most traffic signs are situated on the side of the road, which may have contributed to the collision. TSDR allows drivers to view traffic sign information without having to divert their attention. Due to the existence of a large background, clutter, fluctuating degrees of illumination, varying sizes of traffic signs, and changing weather conditions, TSDR is an important but difficult process in intelligent transport systems. Many efforts have been made to find answers to the major issues that they face. The objective of this study addresses road traffic sign detection and recognition using a technique that initially detects the bounding box of a traffic sign. Then the detected traffic sign will be recognized for usage in a speeded-up process. Since safe driving necessitates real-time traffic sign detection, the YOLOv4 network was employed in this research. YOLOv4 was evaluated on our dataset, which consisted of manual annotations to identify 43 distinctive traffic signs classes. It was able to achieve an average recognition accuracy of 84.7%. Overall, the work adds by presenting a basic yet effective model for real-time detection and recognition of traffic signs.

Keywords - Intelligent Transport systems, Traffic sign Detection, YOLOv4

I. INTRODUCTION

Traffic Sign Detection and Recognition (TSDR) is a critical work because detecting and accurately identifying traffic signs can alert drivers and pedestrians to the regulations they must observe, reducing the frequency of reckless accidents and, in some cases, deaths [1]. Due to factors such as different perspectives, degraded/damaged or discolored traffic signs, illumination on the traffic sign, and motion blur, traffic sign identification and recognition is a difficult process. The challenges of detection and classification of traffic signs are shown in Figure 1.



Fig. 1. Challenges of detecting traffic signs employing different lighting conditions, deformed signs, and variation of illumination

Traditional approaches including Bag of features methods and Regional Convolutional Neural Networks were used for the detection of traffic signs in the past but were discarded due to the poor performances produced by those approaches compared with the newer approaches.

In this paper, we used the You Only Look Once version 4 (YOLOv4) technique to detect and recognize traffic signs. YOLOv4 is a state-of-the-art approach for detecting visual objects in a real-time environment. A dense block, a dense net, and CSPDarknet53 form the backbone of YOLOv4. A YOLOv4 model's neck is made up of feature pyramid networks and a spatial pyramid pooling layer. Finally, the output is generated by the Dense prediction layer. YOLOv4 has dense prediction at layers 139, 150 and 161. These layers contribute directly to the ultimate output and their combined results are obtained [2].

The remainder of this article is laid out as follows. In section II, there are summaries of various methods used in previous works related to detection and recognition. The perspective on the terminologies used in this work is covered in section III. The fourth section is devoted to a detailed explanation of the proposed strategies. The experimental environment and testing results on traffic sign detection and recognition in section V. The suggested solution is discussed and concluded with future extensions in section VI.

II. PREVIOUS WORK

In [3], authors have used a YOLO network to detect and identify Vehicles, trucks, pedestrians, traffic signs, and traffic lights. Traffic signs were then submitted to a CNN, which further categorized them into one of 75 groups. The entire solution was built on a pre-trained YOLO v3 model for class detection, whereas a CNN was trained from scratch and excellent results were displayed on input images for the classification. Detected Traffic signs were cropped and fed into the CNN for classification. They have obtained a classification accuracy of 99.2% for detected traffic signs in various weather conditions. The Berkley Deep Drive Dataset was used to train the YOLO network. The Belgian TS Dataset and the German Traffic Sign Recognition Benchmark have been compiled into a single large dataset with over 120000 images of traffic signs which were then divided into 75 categories. Images were augmented by performing Gaussian Blur, Median Filter, Max Filter, Min Filter, and some simple image rotations. Filtering false expected bounding boxes with coefficients less than 0.5 was achieved using the non-max suppression algorithm. In this study, they used three CNNs. YOLO v3 for object detection and localization, another CNN for a

vehicle, truck, pedestrian, traffic sign, and traffic light classification, and a third CNN for traffic sign classification into 75 classes. Using three CNNs has led to the increase in computational cost while training the models and during realtime object detection.

In [4], authors have proposed a novel YOLOv3 architecture. On pictures, real-time detection with mean average precision (mAP) exceeding 88% has been demonstrated. The model was trained on a broad dataset of 200 different classes. The testing set consisted of 25% of images from the total number of images. As part of the image augmentation process, randomized placement of narrowly cropped traffic signs was done, as well as distortions such as changes to the shape, scale, luminance, and contrast on the training photos. YOLOv3 detection was based on a publicly accessible implementation based on the Darknet network. Weights that were pre-trained on the ImageNet database were used as the initial weights for the model. The number of filters in YOLOv3 or Tiny YOLO's final layer was increased to enable the detection of 200 classes. The learning rate was set at 0.001 and reduced every 15000 iterations, with the input picture size set to 608 608 pixels. After 10000 repetitions in Tiny YOLO, the learning rate was decreased. Both models were trained over 400 epochs on a machine with two 1080ti GPU. A predefined threshold of value 50 was used to calculate Intersection over Union. An accuracy(mAP) of 84.1% was achieved without using image augmentation and 88.1% mAP was obtained by using image augmentation on the YOLOv3 model, and an accuracy(mAP) of 72.1% was achieved without using image augmentation and 71.3% mAP with using image augmentation on the tiny Yolo model. Results have proven that when compared to Tiny YOLO, YOLOv3 was much more precise. Non-maxima-suppression algorithm was used to eliminate unwanted detections and double bounding boxes. YOLO v3 had a lesser number of hidden layers compared to yolo v4. Therefore, yolov4 had better detection accuracy. The time it took to train a YOLO v3 model was about two weeks. Although YOLO models provided greater accuracy and real-time performance, the training time complexity was significant.

In [5], authors have introduced a traffic sign recognition approach based on deep learning, with the primary goal of detecting and classifying circular signs. Initially, images were preprocessed to highlight key details to increase detection and classification accuracy. Image Enhancement, color space conversion from RGB (Red, Green, and Blue) to HSV (Hue, Saturation and value) image noise filtering using mean and median filters were included in Preprocessing stage. The hough transform and segmentation were used to detect and locate traffic sign regions. Morphological operation Opening was used to reduce the noise introduced by segmentation. Finally, deep learning was used to classify the detected road traffic signs. A basic CNN of lent-architecture was used with two convolutional layers with a kernel size of 5×5 , step one, and ReLU activation function which was able to learn complex features, two pooling layers with 2×2 kernel size, and two fully connected layers which contained 512 and 128 hidden nodes respectively. Finally, there were 43 hidden nodes in the output layer. The learning rate was set to 0.0001 at the beginning. German Traffic sign Recognition Benchmark (GTSRB) was used and the accuracy of detected circular

symbols was 98.2%. The entire dataset was split into two parts, a training set, and a testing set. 90% of the dataset was considered as the training set, while the remaining 10% was taken as the test set.

In [6], authors have proposed a method that addressed the problems of low detection and recognition accuracy of distant, small traffic signs and traffic signs which were affected by weather and illumination changes. YOLOv2 was used in real-time which had a fast-processing speed and few false detections to achieve the above goal. RGB images were obtained and used as input to the Yolo network. 22 convolutional layers and five pooling layers were used to build the YOLO v2 network. Each batch on the proposed system used a randomly selected image size from a selection of five. The YOLO network was used to estimate the bounding box and conditional class likelihood of each region in the input photos. Various image sizes were used to train a model that is resilient to scale shifts. A traffic sign dataset of 16 different types of traffic signs and 7160 annotations were created with an image size of 1093×615 pixels in JPEG format. The data volume was increased in this experiment by conducting high contrast, low contrast, noise, and flip horizontal data augmentations, which improved the generalization accuracy. Clear weather, night, and small objects were used in the test dataset which consisted of 123,241 and 140 images, respectively. During the test, 16 different kinds of traffic signs were discovered. An accuracy of 66.4 % and 60.0% was achieved as a result of data augmentation and training with various image sizes.

In [7], using cascade classifiers that were trained on HOG features authors have introduced a methodology to detect traffic signs. A CNN was used which ensured all traffic signs were identified. The CNN model was used to decide whether the candidate zone contained any traffic signs. The final decision was taken at the final stage of the cascade classifier. Image preprocessing was included in the HOG feature extraction to convert the image into a grayscale image. Then the gamma correction algorithm was used to normalize the grayscale image. Gradient components and ordinate coordinates were obtained separately by using Sobel and other edge detection filters with the original image. Cell segmentation and gradient histogram calculation was achieved by segmenting the image into several cells of the same size and counting the cell unit from the histogram. After that, several feature vectors were extracted, and cell units were grouped into a larger interval and feature vectors were superimposed to obtain the HOG features of the interval. Overlapping intervals were gathered and merged to get the final HOG features. Three convolutional layers, two max-pooling layers, and two fully connected layers were used to make the CNN model which was proposed in this study. 5×5 , 3×3 , and 3×3 filters kernels were used by each convolutional layer respectively. 300 and 42 nodes were present in each fully connected layer. A CNN was adopted to extract object features while HOG-CNN was trained to acquire candidate object regions. Weight sharing was not performed among the nine regression variables. Therefore, bounding boxes of multiple scales were predicted using HOG-CNN. The dataset was divided into a training set and a testing set where three-fourth of the images in the dataset

was used for the training purposes while the remaining was used for testing. An accuracy of 90.12% was achieved was the detection rate on video.

III. BACKGROUND

A. Object detection

Object detection is the process of locating objects which are present in an image and marking the detected object coordinates by using a bounding box. Object detection is a technique for determining the location of objects in an image [8].

B. Bounding Box prediction

A bounding box is a method of representing a specific part of an image, such as an object within a region of interest. A bounding box is a rectangular box that surrounds an object. It's usually expressed as an array of coordinate pairs, with the first pair corresponding to the x and y-axis coordinates in the upper-left corner and the second pair corresponding to the x and y axis coordinates in the lower-right corner [8].

C. Intersection over Union

Intersection over Union is a way of measuring the precision of an object detector on a given dataset [9]. The Intersection over Union is calculated using the ground-truth bounding boxes and the projected bounding boxes from the used model [9].

D. Non-maximum Suppression

To reduce redundant bounding boxes of an object, many object detection systems employ the non-maximum suppression processing approach. When non-maximum suppression is utilized, the number of detections in a frame is limited to the total number of objects [10].

E. You Only Look Once (YOLO)

“You only look once” (YOLO) is an object detection system that uses a deep neural network as its foundation and is designed to detect general objects quickly and accurately. The YOLO detector has excellent detection efficiency and a short detection time. At the same time, it generates various anchor boxes and confidence scores for those boxes [5]. During training, YOLO considers the whole image, allowing it to consider contextual details about objects. YOLO breaks the input image into square grids and then estimates how many bounding boxes each grid will have. A confidence level is calculated for each bounding box to determine the likelihood that it contains an object. The object's class is then estimated using a conditional class likelihood for each grid containing an object. During testing, conditional class probabilities and box confidences are combined to convey the chance of a class existing in the box as well as the accuracy with which the box fits the object [5]. There are multiple versions of YOLO.

- YOLOv1 [11] comprised two fully connected layers for likelihood prediction and 24 convolutional layers for extracting features.
- YOLOv2 [12] had the potential to train on large datasets and detect small objects with greater accuracy.
- YOLOv3 [13] architecture had 106 layers including residual blocks, skip connections, and

upsampling, which had a slower detection speed compared with the other versions. YOLOv3 detects at three distinct scales and from three separate network places, as well as a larger number of border boxes [5]. A simpler variant, known as Tiny YOLO, with a total of 22 layers, can be used for faster detection at the expense of lower detection accuracy [5]. Previous studies used YOLOv3 models, which had a lower detection rate, a larger computational cost, and a lower real-time performance.

- YOLOv4 [2] is an object detector that can be trained with a smaller mini-batch scale on a single GPU. This allows a single GPU to train an extremely fast and reliable object detector.

IV. METHODOLOGY

First, we manually labeled the dataset that was utilized to train the YOLOv4 detector for this study using the labeling image annotation tool and uploaded it to Google Drive. Next, a YOLOv4 model was trained on Google collaborators using the annotated dataset. The RGB images in our annotated dataset were not subjected to any form of preprocessing during model training. This model generates cropped photos of identified traffic signs, which are saved to Google Drive. The model was trained for 10000 epochs and achieved an average accuracy of 84.7%.

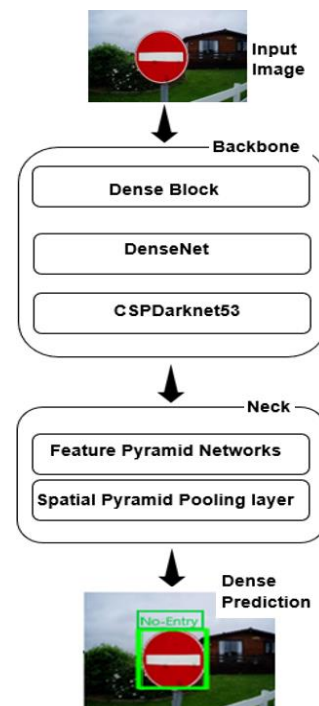


Fig. 2. The architecture of the YOLOv4 network

YOLOv4 considers the entire image during training, allowing it to consider contextual characteristics about objects. YOLOv4 divides the source image into rectangular grids and calculates the number of bounding boxes in each grid. For each bounding box, a confidence level is calculated to evaluate the possibility that it contains an item. For each grid containing an item, the object's class is then estimated using a conditional class probability.

Conditional class probabilities and box confidences are combined during testing to encode both the likelihood of a class being in the box and how well the box matches the object [5].

The overall framework of YOLOv4 is illustrated in Figure 2. The backbone architecture was used to describe the feature extraction architecture. YOLOv4's backbone was CSPDarknet53. A Dense block in the YOLOv4 backbone features multiple convolution layers, each of which has batch normalization, ReLU, and convolution. Dense Net is made up of many dense blocks connected by convolution and pooling layers in the middle. The Dense Block's input feature maps are separated into two parts by Cross-Stage-Partial connections (CSP), one of which will travel through a block of convolutions and the other will not. Following that, the outcomes are combined. This approach is used in the CSPDarknet53 backbone design.

FPN is a prominent methodology for producing object detection predictions at several scale levels. FPN up samples the preceding top-down stream and adds it with the adjoining layer of the bottom-up stream when producing predictions for a certain scale. Figure 3 shows how YOLOv4 uses feature pyramids to detect traffic signs at different scales. The output is passed through a 33% convolution filter to reduce upsampling artifacts and crevices [5]. Spatial Attention Module (SAM), Path Aggregation Network (PAN), and Spatial pyramid pooling layer (SPP) are implemented or replaced with the FPN approach in YOLOv4. Maximum and average pools are applied to input feature maps individually in SAM to produce two sets of feature maps. To produce spatial attention, the feature maps are sent into a convolution layer followed by a sigmoid function. This method is used to gather data and improve accuracy. The preceding layer's input is used by each subsequent layer.

V. EXPERIMENTAL SETUP

We examine the performance of our proposed YOLOv4 model for traffic sign detection and experimental findings using a set of calculated parameters and a dataset. The model was tested on 43 different traffic sign classes to gather all of the data

A. Dataset

The YOLOv4 model was trained and tested using our dataset [14], which was manually annotated. It was separated into a train set of 835 images with 1393 annotations and a test set of 133 images with 225 annotations, with a total of 968 images and 1618 annotations. Figure 4 illustrates several examples of our dataset's images.

B. Google Colab

Google Colaboratory is a cloud-based tool that mimics the functionality of Jupyter Notebooks. Colab requires no setup and offers unrestricted access to computing resources.

C. Darknet repository

The model is trained by using the Darknet framework from AlexeyAB's repository. Darknet is a C and CUDA-based open-source neural network framework. It is easy to

set up and supports both CPU and GPU computing. GPU backend was used to train the model.

D. Parameter calculation

Darknet repository was configured to match a batch size of 64 and 16 subdivisions. The learning rate was set to 0.001. The width and height of input images were set to 416x416. This YOLO v4 model consists of 161 layers which give detections at layers 139, 150, and 161. Max batches, Steps, and Filters used in this YOLOv4 model are given in equations (1), (2), and (3) respectively.

$$\text{Max batches} = \text{number of Classes} \times 2000 \quad (1)$$

$$\text{Steps} = \text{from (80\% of max batches) to (90\% of max batches)} \quad (2)$$

$$\text{Filters} = (\text{number of classes} + 5) \times 3 \quad (3)$$

E. Testing results

Precision, mean average precision and Intersection over Union were computed using the equations (4), (5), and (6).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4)$$

$$\text{Mean Average Precision} = \frac{1}{n} \sum_{k=1}^K (\text{Precision}_K) \quad (5)$$

$$\text{Intersection over Union} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (6)$$

In this study, the overall average accuracy of detection and recognition of the traffic sign over the test set for various situations was 84.7%. Detection and recognition accuracy achieved for each distinctive class is illustrated in Figure 5.

VI. CONCLUSION AND FUTURE EXTENSION

Because it was trained on Google Colab, the YOLOv4 model, which was used for traffic sign detection and recognition, was discovered to have a comparatively higher level of accuracy while saving a substantial amount of computing cost and time. The 161 layers in YOLOv4 contribute directly to the improved accuracy over prior YOLO versions. Higher results may have been obtained if the model had been trained on a larger number of epochs and images, as this results in a greater range of image contexts and image quality. 18 out of 43 classes got 100% accuracy and only two classes such as speed limit 80 and road work got less than 50% accuracy. This study was able to attain a mean average precision of 84.7 % for 10000 epochs when it came to concluding its conclusions. Overall, this study was able to confirm that YOLOv4 outperforms its predecessors in terms of traffic sign detection. It may be inferred that the detection works effectively in a range of situations, such as distorted input images and lighting fluctuations. In the future, the extended work of traffic sign recognition to be improved the performance by using skipped layer architecture and vocabulary voting technique [15].

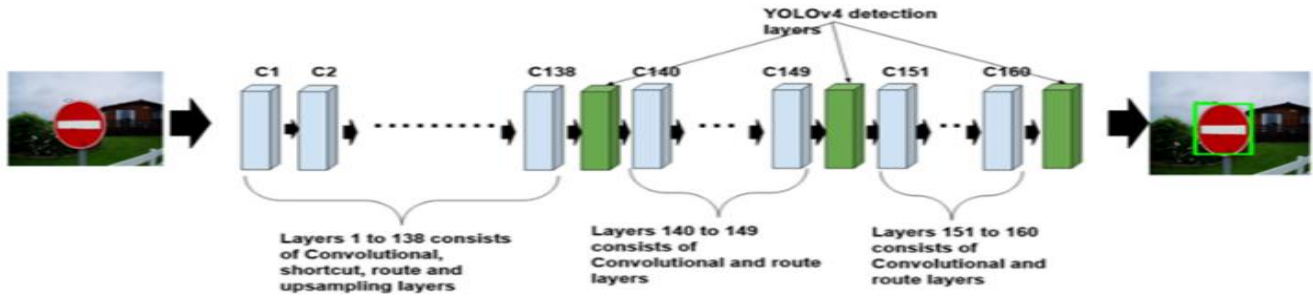


Fig. 3. How detections are found in feature pyramid network



Fig. 4. Some ample images of our dataset

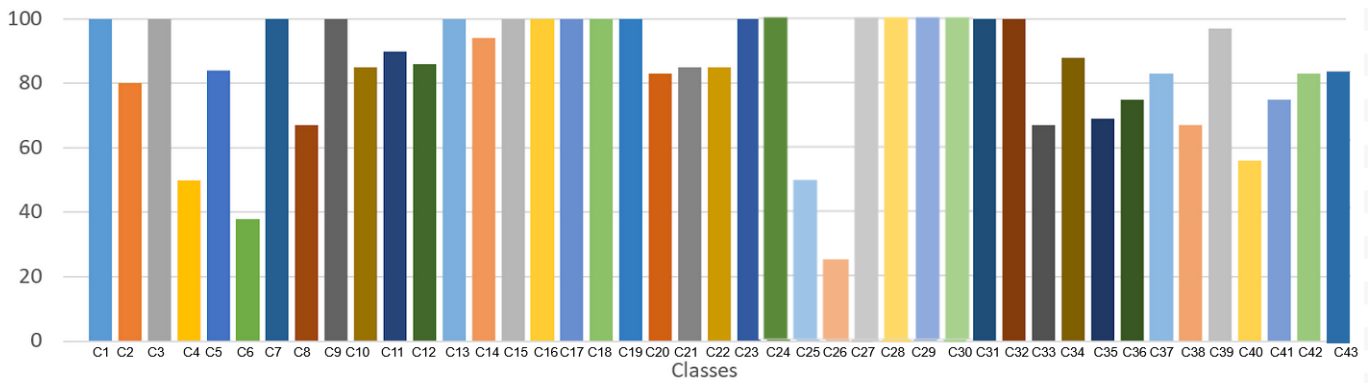


Fig. 5. Detection accuracy of the YOLOv4 model according to each class

- C1_Speed_limit(20km/h)
- C3_Speed_limit(50km/h)
- C5_Speed_limit(70km/h)
- C7_End_of_speed_limit(80km/h)
- C9_Speed_limit(120km/h)
- C11_No_passing_for_vehicles_over_3.5_metric_tons
- C13_Priority_road
- C15_Stop
- C17_Vehicles_over_3.5_metric_tons_prohibited
- C19_General_caution
- C21_Dangerous_curve_to_the_right
- C23_Bumpy_road
- C25_Road_narrows_on_the_right
- C27_Traffic_signals
- C29_Children_crossing
- C31_Beware_of_ice/snow
- C33_End_of_all_speed_and_passing_limits
- C35_Turn_Left_ahead
- C37_Go_straight_or_right
- C39_Keep_right
- C41_Roundabout_mandatory
- C43_End_of_no_passing_by_vehicles_over_3.5_metric_tons
- C2_Speed_limit(20km/h)
- C4_Speed_limit(60km/h)
- C6_Speed_limit(80km/h)
- C8_Speed_limit(100km/h)
- C10_No_passing
- C12_Right_of_way_at_the_next_intersection
- C14_Yield
- C16_No_vehicles
- C18_No_entry
- C20_Dangerous_curve_to_the_left
- C22_Double_curve
- C24_Slippery_road
- C26_Road_work
- C28_Pedestrians
- C30_Bicycles_crossing
- C32_Wild_animals_crossing
- C34_Turn_right_ahead
- C36_Ahead_only
- C38_Go_straight_or_left
- C40_Keep_left
- C42_End_of_no_passing

REFERENCES

- [1] M. Galvani, "History and future of driver assistance", IEEE Instrumentation Measurement Magazine, vol. 22, no. 1, pp. 11–16, 2019.
- [2] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao "YOLOv4: Optimal Speed and Accuracy of Object Detection", Google Scholar, pp. 1-17, 2020.
- [3] B. Novak, V. Ilić and B. Pavković, "YOLOv3 Algorithm with additional convolutional neural network trained for traffic sign recognition", Zooming Innovation in Consumer Technologies Conference (ZINC), pp.1-4, 2020.
- [4] A. Avramović, D. Tabernik and D. Skočaj, "Real-time Large Scale Traffic Sign Detection", 14th Symposium on Neural Networks and Applications (NEUREL), pp.1-4, 2018.
- [5] Y. Sun, P. Ge and D. Liu, "Traffic Sign Detection and Recognition Based on Convolutional Neural Network", Chinese Automation Congress (CAC), pp.1-4, 2020.
- [6] R. Hasegawa, Y. Iwamoto and Y. Wei Chen, "Robust Detection and Recognition of Japanese Traffic Sign in the Complex Scenes Based on Deep Learning", IEEE 8th Global Conference on Consumer Electronics (GCCE), pp.1-4, 2020.
- [7] L. Shangzheng, "A Traffic Sign Image Recognition and Classification Approach Based on Convolutional Neural Network", 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp.1-4, 2019.
- [8] Z. Zhong-Qiu "Object Detection with Deep Learning: A Review", IEEE Transactions on Neural networks and learning systems, pp. 1-21, 2019.
- [9] A. Kumar, Z.J. Zhang, and H. Lyu, "Object detection in real time based on improved single shot multi-box detector algorithm", EURASIP Journal on Wireless Communications and Networking, pp. 1-18, 2020.
- [10] S. Dasiopoulou, "Knowledge-assisted semantic video object detection", IEEE Transactions on Circuits and Systems for Video Technology, pp. 1210 - 1224, 2005.
- [11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, " You Only Look Once: Unified, Real-Time Object Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1 - 10, 2016.
- [12] X. Huang, X. Wang, and W. Lv, "YOLOv2: A Practical Object Detector", Cornell University Computer vision and pattern Recognition, pp. 1-7, 2017.
- [13] J. Redmon, and A. Farhadi, "YOLOv3: An Incremental Improvement Practical Object Detector", Cornell University Computer vision and pattern Recognition, pp. 1-6, 2018.
- [14] <https://www.fsc.esn.ac.lk/mathematics>.
- [15] S. Sotheeswaran and A. Ramanan, "Front-view car detection using vocabulary voting and mean-shift search," Fifteenth International Conference on Advances in ICT for Emerging

Regions (ICTer), 2015, pp. 16-20, doi: 10.1109/ICTER.2015.7377660.

Forecasting foreign exchange rate: Use of FbProphet

Fanoon Raheem*

Department of Information and Communication Technology
Faculty of Technology
South Eastern University of Sri Lanka, Sri Lanka
fanoonarfs@gmail.com

Nihla Iqbal

Department of Information and Communication Technology
Faculty of Technology
South Eastern University of Sri Lanka, Sri Lanka
mifnihla@gmail.com

Abstract - Foreign exchange rate prediction can be considered crucial in today's world. The exchange rate of a country plays a vital role in its economic growth. The Central Bank of a country holds the authority in managing the exchange rate and its policies. The study predicts the foreign exchange rate of American Dollar to Sri Lankan Rupee using FbProphet model; a time-series forecasting model developed and introduced by Facebook. The daily exchange rate values for USD/LKR were obtained and the values are predicted for another twenty-four months starting from November 2020. R Squared value is calculated to verify the fitting of the model and the value is 0.98, which indicates that the model for prediction very well fits for the data set used. And further, Mean Squared Error and Mean Absolute Error are calculated to measure the performance of the model. These metric measurements show that the model is appropriate for the data set which has been selected for the research study.

Keywords - exchange rate, FbProphet, forecasting, US Dollar

I. INTRODUCTION

In today's world, one of the most important liquid markets is the Foreign Exchange (FOREX) markets. The relative price between two different currencies is known as the exchange rate. It is the value of a money of a country's currency for undertaking international trade for goods, finance, and services, being the key to a country's monetary condition. The Central Banks are the monetary authorities of a nation which has been granted the power to manage the exchange rate as part of its monetary, financial, and economic development policies under relevant statutes. According to the perspective of macroeconomy, exchange rate policy is the key instrument for the mobilization of foreign capital and savings in order to fill the resource gaps in the domestic and also expand the investments [1].

The fluctuations in the exchange rate of a country have both favorable and unfavorable effects on the economic activities and standard of living of the people due to the trade being largely globalized and finance involving the exchange of currencies. Generally, appreciation in the currency of a country will have benefits, whereas depreciation will have the reverse impacts:

- Downfall in the domestic prices of products which are being imported because the import cost will be less if the domestic currency value is higher. This will result in a lower inflation depending on the volume of imports in local consumption and manufacturing activities.
- Reduction in the amount of outstanding foreign debt of a country which will lessen the burden of a nation's repayment of foreign debt.

- An imbalance in the trade of a country may be caused due to the increase in the imports as a result of lower cost in importing goods, which is unfavorable for the country.
- Another disadvantage is that there will be a downfall in the income of exporters which may discourage them in exporting products resulting in an adverse effect in the export industries. But, if a lower inflation prevails in the country, the demand for export products in the foreign countries will rise balancing the initial reduction in the exporter's income.

Sri Lanka maintains a healthy relationship with several foreign countries, as a result of which it receives more foreign exchanges. The American Dollar (US Dollar) is the common currency used by both the government and monetary policy makers of Sri Lanka. The transaction price of an US Dollar in the year 1970 was Rs. 5.95 Sri Lankan Rupees (LKR), which, after two decades, increased to Rs. 40 LKR [2]. Similarly, in the year 2020, the price has been elevated to Rs.180.76 LKR. In the economic point of view, the exchange rate is generally ascertained by the demand and the supply curve of the exchange rate which is much similar to the common commodity market system. The relative commodity price, inflation rate and interest rate are the main factors which influence on the exchange rate. It is also notable that the higher the exchange rate is the higher the promotion of economic growth of a nation.

This study attempts to forecast the exchange rates of USD/LKR for the next 24 months from November 2020 which would be useful for making economic decisions, using FbProphet model. The remainder of the paper is as follows. Section 2 of the paper is a literature review on the technologies adopted by researchers to predict the exchange rate of currency. Section 3 and 4 describes the methodology adopted to predict the USD/LKR exchange rate for this study and the results obtained from the model. Finally, section 5 concludes the study on time series forecasting of foreign exchange rate of USD/LKR.

II. LITERATURE REVIEW

According to the study conducted by [3] on designing and developing an algorithm to predict fluctuation of currency rates, the key purpose of the study was to compare the precision of three models: Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Networks (ANN) and Vector Support Machines (SVM). The import, export and USD currency exchange series for LKR data were chosen for training the data. It was possible to see that the SVM forecast performed better than other models after

training the data set and comparing each algorithm. Also, from the study it has been understood that the merging of SVM and SVR models has further strengthened the algorithm that can predict the fluctuations of the currency rates.

On another study by [4], the research is conducted using the Artificial Neural Network models to make multi-step forecasts of the Sri Lankan Rupee foreign exchange rate against three international currencies, to test the accuracy of these models and where present, to identify deficiencies. Basic Recurrent Neural Network, Multi-Layer Perceptron, Long-Term Memory, Gated Recurrent Unit and Convolutional Neural Network Architectures were the algorithms that are used for this study. With the exception of a few Gated Recurrent Unit models, many simulations have been able to forecast 10-day forward exchange rates with a greater degree of accuracy. The final output of the study showed that among the other algorithms, the Basic Recurrent Neural Networks with a single input layer, a hidden layer, a flattened layer, and an output layer is the best one to make the predictions.

A research had been conducted by [5] which aimed at comparing the forecast accuracy of the most widely used algorithms and to identify the more accurate one for forecasting Sri Lankan Rupees' daily exchange rates against the Euro and Yen. The NAR model (Nonlinear Auto Regressive Neural Network) with SCG learning and SVR model with Gaussian function were employed in the study conducted to make the forecasts. And the results of the study showed that SVR model outputted better predictions than ANN models.

Besides these, there is also a related work done by [6], which studied about the ways that United States US Dollar (USD) exchange rate can be predicted against Sri Lankan Rupees (LKR) using three different deep learning models, namely Long Short-Term Memory (LSTM), the Convolutional Neural Network (CNN) and Temporary Convolution Network (TCN). The findings of the research showed that the CNN model is superior to other models when it comes to financial time series prediction.

On another research by [7], the authors recommended a hybrid forecasting model for foreign exchange rate forecasting using EMD (Empirical Mode Decomposition) and FNN (Feedforward Neural Network) and the concert of the model is related with NAR and SVR (support vector regression) models. The methodology used EMD with several Intrinsic Mode Functions (IMFs) and one residual series to break down the original non-linear and non-stationary chain. In order to estimate the IMF exchange rate and the received residual inputs, the hybrid model is then used. The analytical results from the study proved that the Sri Lanka Rupee Euro and Yen daily exchange rate forecast was more accurate with EMD-FNN model.

The SCG algorithm trained Feedforward Neural Network (FFNN) performed better than BPR algorithms trained FFNN was put forward by [8] at a study conducted to discover a model that can foresee the US dollar with a better level of precision compared to Sri Lankan Rupee (USD/LKR) using existing neural network models.

In [9]'s research, GARCH model and the ANN model (FFNN model having Backpropagation algorithms) are used to compare the accuracy for the predictions of USD to LKR exchange rates. With both models, historical stagnated findings of the data and average of the other measures were

utilized as the response variable and the forecasting output were analyzed using a variety of popular statistical parameter. The findings revealed that ANN model performed better when compared to GARCH model.

Along with these, the research by Lingaraja and his co-authors [10] focused on long term volatility of Sri Lankan LKR against USD with nine other currencies that are considered to be emerging in Asian region, that would help in supporting financial decision making based on Asia. The study conducted used the GARCH model with correlation and the test was done based on Granger Causality test.

The subsequent analyses of USD/LKR exchange rate forecasts indicate that numerous Machine Learning models and algorithms have been used to forecast exchange rates. They include models ranging from various types of the Artificial Neural Network (ANN), ARIMA, Support Vector Machine (SVM), etc. The related study further shows that hybrid techniques have also been pursued in the design of the models. And each work offers a promising accuracy rating that has prompted this research to pursue a totally new paradigm that is distinct from all the other existing models, to come up with more successful predictions.

III. METHODOLOGY

The USD/LKR exchange rate prediction for the next twenty-four months starting from November 2020 is shown by the methodology adopted. For some important business decisions, such as whether to invest in USD to LKR currency pair or whether to purchase or sell USD/LKR pair, forecasting is known to be unavoidable.

A. Installation in Python

As the initial step, the library for FBProphet model need to be installed. FBProphet is available as an open-source library and based on the choice of programming language (Python or R) it can be used. To the study conducted, the Python3 is selected and therefore the python installation of the corresponding library was done.

B. Select and prepare data

Daily exchange rates of United States Dollar (USD) on Sri Lankan Rupee (LKR) were selected from the data repository of CurrencyConverter [11] for this study. The daily exchange rate from 2009-10-07 to 2020-11-22 were collected. The USD was selected as the currency to forecast the USD/LKR pair since, it is the widely used currency for trading and investments with LKR among the other currencies of world economy. Therefore, the input to the research is the exchange rate data from the timeline mentioned above (2009-10-07 to 2020-11-22) and provided the input in form of date and exchange rate in LKR, the study will forecast the subsequent 24 months. The dataset was prepared as a CSV file, having two columns, ds and y since the input data frame for FBProphet must be in a format that has ds and y columns indicating the date and the numeric values.

C. Exploratory visual analysis

Data visualization is important in order to understand the dataset used for the study. A visual representation of the data may, as always, be effective and informative [12]. A time series plot for the whole-time frame was generated by which the seasonal and abnormal deviations can be shown

if data were to be presented for such a prolonged period of time (2009 – 2020).

By plotting the data, the overview and the shape of the dataset used was visualized. Under this context, the ability to quickly dig into multiple timeline periods to better analyze the details and to find visual hints about possible patterns, intermittent and unexpected outcomes is understood and that is possible with one of the most valuable features provided by Plotly.

The Fig 1 shows how the data set is visualized in terms of years and values (exchange rates). In addition, the visualization shows that that the data is not fixed with a prominent increasing trend.

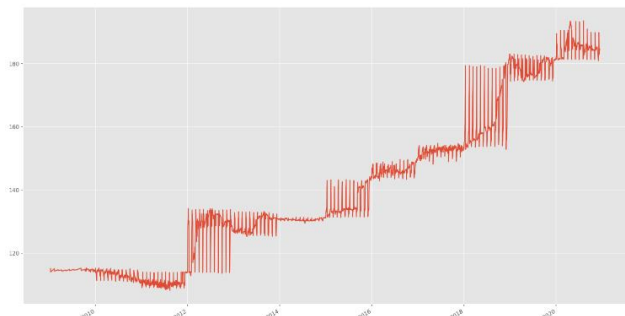


Fig. 1. Daily Exchange Rates (USD/LKR) from 2009 to 2020

D. Build the predictive model

The future predictions for the USD/LKR exchange rates are created by the predictive model built. The predictive model is developed by using FBProphet which is a time series forecasting model implemented by the data scientists of Facebook. Prophet is a technique based on an additive model for forecasting time series data where non-linear trends are consistent with yearly, weekly, and daily seasonality, plus holiday outcomes. For time series which have strong seasonal effects and a few seasons with chronological data, this works well [13]. The Prophet is responsive to missing values and generally it is capable of handling the outliers.

Sklearn Machine Learning Model is accompanied by FBProphet where the Prophet Class instance is generated to its fit and predict methods, as its syntax follows the Scikit learn’s train and predicting model. A data frame is used as the input to Prophet (consisting of ds and y columns). The Prophet object would then be constructed to match the model. The algorithm would be able to learn the data as a function of the model fit, which can be expanded later to a similar type of data.

Hence, the predictive model has been built by selecting the features such as dates and exchange rates. As stated already, the ds and y are considered to be the standard input format preferred by FBProphet. Here the ds have been assigned as the date whereas the y is the exchange rate corresponding to each date. Table I shows the input data frame that has been used in building the predictive model. The input data frame starts from 2009-10-07 and it goes up to 2020-11-22.

For the study conducted, the dataset has not been divided into training dataset and test dataset to build the prediction model instead the whole data has been used to fit the model, which has later given the predictions for the

exchange rates for future 24 months, i.e. the exchange rate up to November 2022.

TABLE I. USD / LKR EXCHANGE RATES (ds and y)

	ds	y
0	2009-07-10	114.8271
1	2009-08-10	114.7283
2	2009-09-10	115.2225
3	2009-10-10	114.8294
4	2009-11-10	114.8158

where,
 ds – datestamp, data type is date or datetime
 y – numeric value to predict

E. General model predictions

Once the model has been fit and instantiated, the predictions will be based on the data frame consisting of the future dates. In Prophet those future dates are known by the term, period. The USD/LKR exchange rate predictions are generated for the upcoming 24 months starting from December 2020. The methodology uses the frequency in terms of month (Freq = ‘M’) which implies the monthly data. Since the forecasted data covers 24 months, (Period = 24) the comparison can be made in between the actual and predicted values and it will be helpful in coming to a conclusion as to how well the model forecasts the exchange rates.

TABLE II. THE FUTURE DATES FOR FORECASTS (DS)

	ds
4084	2022-07-31
4085	2022-08-31
4086	2022-09-30
4087	2022-10-31
4088	2022-11-30

Table II, illustrates the future dates that are been selected by tail command to output the last part of the whole data frame. Based on those future dates, the predictions are generated.

Similarly, the Table III given below shows the future data frame (forecasts) for the USD/LKR exchange rate and the results from the full data frame show a quit a lot of data in various columns which includes the predictions based on trend, seasonality components as well the other additive terms. But for each future row, the focus has to be given to only few important columns including yhat, yhat_upper and yhat_lower.

yhat – stores the forecast values in this column

Therefore, such output data frame was generated using the appropriate Prophet function and it is shown below in Table IV. It consists of the forecasts that are tailed to last few months with each future row consisting of ds (date) and its resultant yhat, yhat_lower and yhat_upper values.

TABLE III. THE FORECASTS FOR USD/LKR EXCHANGE RATE

	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	additive_te	re_terms_lower	multiplicative_terms_upper	yhat
4084	2022-07-31	204.873181	181.133118	227.063420	183.767902	227.509829	-1.054975	•••	0.0	203.818206
4085	2022-08-31	205.639267	180.648363	228.610257	182.570091	230.564283	-1.637560	•••	0.0	204.001707
4086	2022-09-30	206.380640	178.974067	231.386060	181.536730	233.054414	-1.416293	•••	0.0	204.964347
4087	2022-10-31	207.146725	178.370946	233.421029	180.312570	235.170386	-1.603842	•••	0.0	205.542883
4088	2022-11-30	207.888098	178.961628	236.731663	179.630810	237.228222	-0.887813	•••	0.0	207.000285

TABLE IV. THE FORECASTS FOR USD/LKR EXCHANGE RATE WITH VARIABLE YHAT, YHAT_LOWER AND YHAT_UPPER

	ds	yhat	yhat_lower	yhat_upper
4084	2022-07-31	203.818206	181.133118	227.063420
4085	2022-08-31	204.001707	180.648363	228.610257
4086	2022-09-30	204.964347	178.974067	231.386060
4087	2022-10-31	205.542883	178.370946	233.421029
4088	2022-11-30	207.000285	178.961628	236.731663

The variable yhat characterizes the exact model predictions whereas the two variables yhat_lower and yhat_upper represents the lower limit and upper limit for the forecast. These two variables are used as measure to calculate the yhat values for future dates. Based on this, a conclusion can be realized that the forecasts will be stored into the yhat column.

F. Plot model predictions

The model predictions are plotted to clearly understand the actual values (original data), the predicted values (forecasted data) and forecast errors. The Fig 2 shows the plotting results where the actual values are drawn in black dots, the predicted values in blue lines and the blue shaded area showing the error of predictions. The plotting leads the way to quickly evaluate the results. The model predictions plot also generates a component plot in terms of individual components as shown in Fig 3.

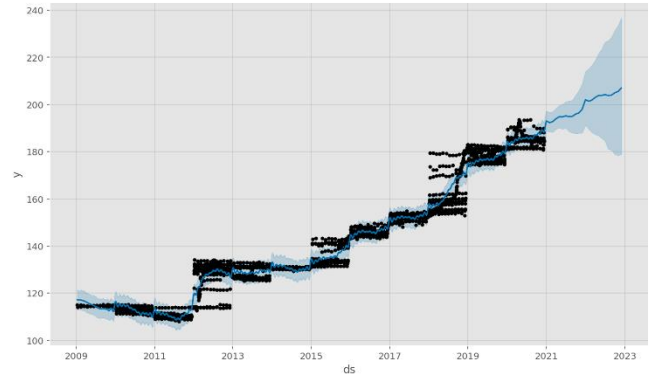


Fig. 2. The original and predicted values for USD/LKR exchange rate

The trend, weekly and yearly forecast components are plotted separately. The component plot is considered to be a vital one, as it better illustrates the factors of the forecast model.

From the individual component graph as shown in fig. 3 below, the conclusion can be made that for trend, Prophet has done a good job by showing the increasing pattern for USD/LKR exchange rates at the end of 2020. The weekly seasonality chart reveals that, the exchange rates are highest during the weekdays than that of the weekends. And during the annual holiday (December) seasonality the table shows a significant fall.

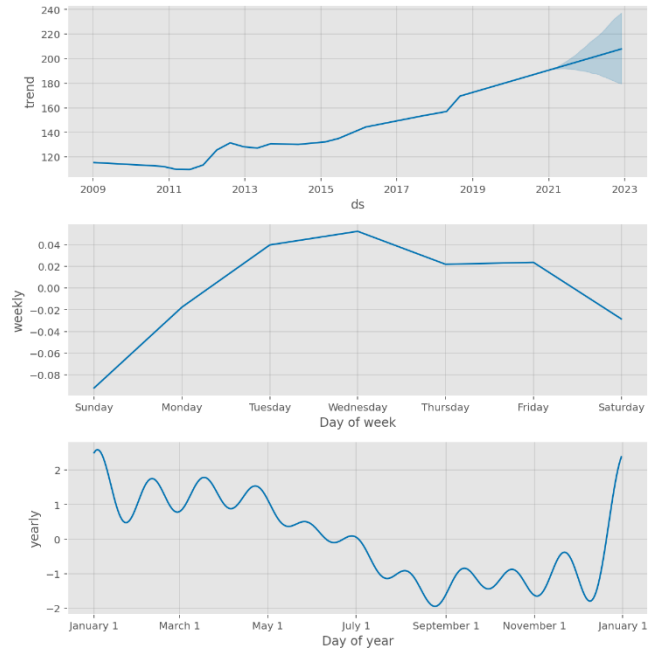


Fig. 3. Individual forecast model components for USD/LKR exchange rate

IV. RESULTS AND DISCUSSION

Forecasting foreign exchange rate is a complex task due to changings in the dynamics of its driving factors. It can be predicted by using various methods and this study uses FbProphet time series forecasting model. Daily values of USD/LKR exchange rates were used from 7th of October 2009 to 22nd of November 2020.

The performance of the FbProphet model is evaluated using the following metric measurements:

A. R squared Score

R squared is also known as the coefficient of determination which indicates how good a model fits for the given dataset. It also illustrates the closure of the regression line to the actual data value line. The R squared value ranges between 0 and 1 where 0 means that model is not appropriate for the given dataset whereas 1 denotes that the model perfectly fits with the given data set.

For the data set provided for the study in this research, the R squared value is 0.982 which means that the model fits for the exchange rate dataset.

B. Mean Squared Error (MSE)

MSE is the average of the square of the difference between the original and predicted values of the data. It is calculated using the formula given below.

$$\frac{1}{N} \sum_{i=0}^n (\text{actual values} - \text{predicted values})^2 \quad (1)$$

Where,

N - total number of observations per rows in the dataset.

\sum - difference between actual values and predicted values for each i value from 1 to n.

MSE is used to determine the performance of the regression model. The MSE value obtained for this study is 10.31 which means that the model is working efficiently with a 90% performance.

C. Mean Absolute Error (MAE)

MAE is the difference between the actual values and the predicted values. The result is obtained by getting the average of the error in each sample data set. The MAE value obtained for the dataset to predict foreign exchange rate is 2.1.

From the overall metric measurements taken, it can be determined that the model very well fits for the data set selected for the study and gives an efficient prediction on the foreign exchange rate values.

V. CONCLUSION

The research analyzed the USD/LKR exchange rate time series prediction for the next 24 months, starting in December 2020. Instead of making daily projections, the monthly estimates are made so that some other decisive variables such as volume swings, adjustment in prices, business cycles and market segments may also be subjected to adjustments. Therefore, it is hoped that such forecasts will help the decision-making of the financial quarters where reports are archived in a monthly manner. The forecasts are crucial factors in evaluating the currency pair's long-term future profits.

The methodology in section 3 reveals that the unique design of the real-life research will improve the predictability of USD/LKR currency pair that goes through heavy fluctuations during certain periods of the year in a way by using the enhanced time series forecasting algorithm – FBProphet. And in this study, the goal was to evaluate a highly accurate architectural model in USD/LKR currencies for the Machine Learning to predict the exchange rate.

The findings of section 4 of the analysis are promising since the model suits well with a strong r-squared value.

This shows that the model has a good impact scale. Related to the data collection comprising of the exchange rate of USD/LKR for a longer period, the utility of the model is improved.

The numerous methods of data mining for exchange rate forecasts are considered from the inspection of past studies. It has been shown that the predictive model that is been built from FBProphet is very useful in predicting USD/LKR exchange rate.

Further, the model could be compared with the other models such as ANN, ARIMA and SARIMA models. The comparison study would help in making decisions depending on the forecast values obtained from these models. Also, this would be useful in the economic growth of a nation.

REFERENCES

- [1] S. H. I. Rajakaruna, "An Investigation on Factors affecting Exchange Rate Fluctuations in Sri Lanka. Staff Studies", 47(1), pp 69, 2017. <https://doi.org/10.4038/ss.v47i1.4703>.
- [2] W. M. Madurapperuma, "Impact of Inflation on Economic Growth in Sri Lanka. Journal of World Economic Research", 5(1), 1, 2016. <https://doi.org/10.11648/j.wjer.20160501.11>.
- [3] N. Kuruwitaarachchi, M. K. M. Peiris, C. N. Madawala, K. M. A. R. Perera, & V. U. N. Perera, "Design and Development of an Algorithm to Predict Fluctuations of Currency Rates", 11th International Conference on Software, Knowledge, Information Management & Applications, At Colombo, 7, December 2017.
- [4] A. J. P. Samarawickrama, & T. G. I. Fernando, "Multi-Step-Ahead Prediction of Exchange Rates Using Artificial Neural Networks: A Study on Selected Sri Lankan Foreign Exchange Rates", 2019 IEEE 14th International Conference on Industrial and Information Systems: Engineering for Innovations for Industry 4.0, ICIIS 2019 - Proceedings, 2019, pp 488–493.
- [5] P. Nanthakumaran, & C. D. Tilakaratne, "A comparison of accuracy of forecasting models: A study on selected foreign exchange rates", 17th International Conference on Advances in ICT for Emerging Regions, ICTer 2017 - Proceedings, 2018-Janua, 2017, pp 324–331.
- [6] S. Aryal, D. Nadarajah, D. Kasthurirathna, L. Rupasinghe, & C. Jayawardena, "Comparative analysis of the application of Deep Learning techniques for Forex Rate prediction" 2019 International Conference on Advancements in Computing, ICAC 2019, 329(1), 2019, pp 329–333.
- [7] P. Nanthakumaran, & C. D. Tilakaratne, "Mode Decomposition and FNN: A Study on Selected Foreign Exchange Rates", 11, July 2018, pp 1–12.
- [8] C.D. Tilakaratne, "Forecasting Exchange Rates Volatilities Using Artificial Neural Networks", 2019. https://doi.org/10.1007/978-3-642-57652-2_4.
- [9] S. Nanayakkara, V. Chandrasekara, & D. Jayasundara, "Forecasting Exchange Rates using Time Series and Neural Network Approaches" European International Journal of Science and Technology, 3(2), 2014.
- [10] K. Lingaraja, C. J. B. Mohan, M. Selvam, M. Raja, & C. Kathiravan, "Exchange rate volatility and causality effect of Sri Lanka (LKR) with Asian emerging countries currency against USD" International Journal of Management, 11(2), 2020, pp 191–208. <https://doi.org/10.34218/IJM.11.2.2020.021>. USD LKR Historical Exchange Rate. (n.d.). Retrieved December 5, 2020, from <https://www.currency-converter.org.uk/currency-rates/historical/table/USD-LKR.html>.
- [11] Topic 9. Part 2. Time series with Facebook Prophet | Kaggle. (n.d.). Retrieved December 3, 2020, from <https://www.kaggle.com/kashnitsky/topic-9-part-2-time-series-with-facebook-prophet>.
- [12] Topic 9. Part 2. Time series with Facebook Prophet | Kaggle. (n.d.). Retrieved December 3, 2020, from <https://www.kaggle.com/kashnitsky/topic-9-part-2-time-series-with-facebook-prophet>.

Novel deep learning approaches for crop leaf disease classification: A review

E. M. T. Y. K. Ekanayake*
Postgraduate Institute of Science
University of Peradeniya, Sri Lanka
e.thissa@gmail.com

R. D. Nawarathna
Department of Statistics and Computer Science
University of Peradeniya, Sri Lanka
ruwand@pdn.ac.lk

Abstract - To encourage sustainable progress, it is suggested that in a world connected by virtual platforms, modern society should merge big data, artificial intelligence, machine learning, information and communication technology (ICT), as well as the “Internet of Things” (IoT). When real-life problems are considered, the above technology processes are essential in solving the issues. Food is an essential need of human beings. Food supply has become crucial, and it is very important to increase the adequate cultivation of plants for large populations due to huge population growth. At the same time, farmers are struggling with a variety of food plant diseases that significantly affect the harvesting and production in agricultural fields. Nevertheless, the agricultural productivity of rural areas is directly involved with the increase in the economic growth of developing countries such as Sri Lanka, India, Myanmar and Indonesia. Early identification of crop disease, using a well-established modern technique, is vital. It necessitates a number of processes observing large-scale agricultural fields as a disease can infect different parts of the plant such as leaf, roots, stem and fruit. Most diseases appear in plant leaves and have the potential to spread them all over the field within a very short time. This paper reviews several state-of-the-art methods that can be used for plant leaf disease recognition with a special reference to deep learning based methods.

Keywords - attention mechanism, Deep Learning, disease identification, image processing, Machine Learning

I. INTRODUCTION

Most Asian economies are based on agriculture. When people enhance food plant productivity, this often results in a degradation of agricultural fields due to being ignorant of the natural environmental impact on the plantation process. Because of crop plant pathogens such as fungus, organism, virus, bacterial infections, phytoplasmas, plant disease cannot be neglected. Therefore, identification of the crop plant disease is the main objective in the agricultural field. When a disease arises because of the above pathogens in any type of plant systems, it may infect all parts of the plant, including its leaves, roots, stems, crowns, tubes, flowers, fruits and seeds. Consequently, the identification and classification of the disease at an early stage is crucial. Direct observation of the field by crop experts is a common approach in the detection and identification of crop diseases, but this solution is an obsolete method. In addition, identifying the disease by monitoring the fields by experts will be extremely expensive in the large-scale farming industry. To take a better solution, we can analyse images of the crop plant leaf disease using image processing technology. This may include extracts of the feature of the diseased area in terms of colour, texture,

shape and other appearances from a measurable point of view in the plant area.

According to the level of expertise required, the cost of supervision will be high and time-consuming. A solution which uses an image processing technique, will assure more benefits in monitoring huge scale agricultural fields. Furthermore, this automatic identification of the crop disease, by analysing the symptoms of the related plant parts, makes the process both simple and economical. It will require computer vision to deliver an image-based programmed procedure control, the examination process, and the automation of robotic supervision.

Identifying crop diseases in a visual image is a difficult task and the accuracy of the identification can also become less valuable. This method can be used only in selected places. Using an automatic leaf disease identification technique will reduce the time and it will be more accurate, with less effort. When we consider the food plants, infected diseases are generally revealed by brown or yellow spots, early and late burn-patches, fungus, bacterial or virus diseases. The image processing technique is the way to measure the area affected by the disease, or determine the colour differences between a good location and the affected area. A few methods based on colour identification feature and K-means algorithm and threshold values are used for the segmentation process and identifying the disease.

The classification of a digital image process refers to the feature extraction information task from raster images. The resultant raster from the image classification process enables us to make a scale map. Supervised learning and unsupervised learning are the primary classification methods. Currently, there are a variety of ways to perform digital image classification interacting with thresholding methods. Most methods depend on colour identification, boundary detection, and the segmentation of digital images. Machine learning-based methods for crop disease identification and classification have become an important part of modern developments. Nowadays, most researchers tend to use new machine learning-based methods instead of traditional methods.

The main objective of this review is to suggest a better deep learning method for the identification and classification of plant leaf diseases at an early stage. In addition, it aims to compare and contrast the plant disease classification technologies with the latest deep learning methods, to verify the importance of the dataset of each method used, in order to assess the relevance of future enhancements for real world scenarios.

This paper is organized as follows. Section II presents a review of related literature. Section III summarizes the dataset, the proposed solutions' approach, and potential improvements. Section IV compares and contrasts the

average accuracy of the different methods in brief. Finally concluding remarks are given in Section V.

II. REVIEW OF TECHNIQUES FOR CROP LEAF DISEASE IDENTIFICATION

An advanced attention mechanism is suggested in the paper [1] that successfully operates the informative areas of an input image. Also, the method explains the usage of transfer learning to construct some fine-grained image classification model based on a developed attention mechanism. Close-grained detailed image classification is an exacting task due to the difficulty in recognizing distinguishing features. When the input image is a fully represented object, finding a suitable method is not an easy task. In this particular classification, the model considers visual disturbance such as overlapping and external light. To use this model for crop leaf disease identification, it should concentrate on the detailed regents of the input images.

The researchers have experimented with transfer learning with the convolutional neural network in the experiment [2]. The model modified a network layout to increase the learning ability of the plant disease characteristics. The MobileNet with the squeeze and extraction (SE) section was used in this experiment. To increase the qualities of both, the pre-trained MobileNet and SE section were embedded in the developed network called SE-MobileNet. The SE-MobileNet was the model used for the identification of paddy leaf diseases. The speciality of the model was the double usage of the transfer learning technique, which helped in gaining the optimal solution. There were two phases in this experiment. The first phase was training the SE-MobileNet for the extracted layers, and the end of the convolutional layers were stopped with the pre-trained shared weights on the ImageNet. The second phase was training the SE-MobileNet model using the target input dataset.

A classification and identification technique model constructed in [3] can be used in classifying crop leaf diseases. In this experiment, before the feature extraction process, pre-processes were completed. In the pre-processing section, all the RGB images were converted into grey level images to the next step, which was the feature extraction of the input image. The elementary morphological functions were applied as the second step on the input image. Then, the input image was converted into a binary level image. In the next stage, if the pixel value of the binary image was zero, the pixel was converted into a responsible RGB image value. Finally, using the Naïve-Bayesian classifier [4], the disease was identified.

Another novel approach [5] presents for the detection and classification of rice leaf viruses. It used K-means clustering, multiclass support vector machine (SVM) [6] and particle swarm optimization (PSO) [7]. Grey Level Co-occurrence Matrix (GLCM) was used for the feature extraction process. The virus classification was done using a Support Vector Machine (SVM) classifier, and the recognition of the virus accuracy was enhanced by optimizing the data with PSO. The paper [8] The performance of 13 CNN models for rice disease detection in transfer learning and deep features plus the SVM method is evaluated in this work. When compared to other models, the statistical analysis findings, deep characteristics of

resnet50 [9][10] and SVM classification model are superior. A comparison of all classification models based on CNN and conventional techniques was conducted.

An interesting model using RGB image acquisition is presented in the alternative experiment [11] to detect any type of plant disease affected by different agricultural crops. Converting the input RGB image format into Hue-Saturation-Intensity (HSI) format [12] and masking and removing the green pixels in the input image makes it accessible to the segmentation process, using Otsu's method [13]. Then, the texture features were calculated using the colour co-occurrence method and finally the disease was classified with the Genetic Algorithm [14].

The crop disease identification and classification process using a convolution neural network is presented in the paper [15]. This includes three convolution layers and three pooling layers followed by two fully connected layers. The results of the experiment clearly show the efficiency of the constructed model approach over the pre-trained models such as VGG16 [16], MobileNet and InceptionV3 [17].

The experiment [18] focused upon the leaf disease segmentation and classification of a few plants. Firstly, the disease area from the input images was segmented with an introduced superpixel cluster-based hybrid neural network. Texture, colour and shape were the main features whereby input images were classified under different classes. The experiment [19] tried to resolve the rough image dataset problem. The method initially limited the leaf area by applying the colour features of the input image. The classification process of the input leaf image depended on the structures of discriminatory characteristics. The property of the input image features showed a variety of patterns in the leaf area. Then, the researchers applied the feature discriminable characteristics with the Fisher vector in terms of different orders of the diversity of Gaussian distribution. In the paper [20], the EfficientNet [21] deep learning method experimented with in-crop leaf disease identification. The model performance was compared with several newly developed deep learning models. To train for the purpose, the researchers used the PlantVillage dataset in this experiment. The EfficientNet method and other deep learning models were trained with the transfer learning technique. In the transfer learning technique, each layer in the models was set up as trainable.

III. MATERIALS, METHODS AND PROPOSED ENHANCEMENT

The key components in the domain's reference study are the gathering of relevant material and the reviewing of the information with a competent analysis. In the first stage, the Google Scholar Web Science Indexing Facilities performed the keyword-based exploration for journal articles and conference papers. Two main search criteria were used to search the relevant articles. Those keywords were "plant disease classification " and " deep learning methods for plant disease identification" respectively. Initially, 10 articles were recognized. The selected articles were examined individually in the second stage. Key questions posed in analysis were: What was the dataset used? What were the disease categories in the dataset included? What methods were used and what was the level of average accuracy of the methodology they selected?

Table I shows a brief overview of the selected research papers on automatic crop disease identification, and their use of materials and methods. It summarizes the dataset, the methodology of the proposed solutions and future enhancement in the corresponding studies.

TABLE I. TABULAR LIST OF REFERENCE NUMBER OF REVIEWED PAPERS, THEIR METHODOLOGY AND FUTURE ENHANCEMENTS

Article Ref.	Dataset	Methodology	Future Enhancement
[1]	PlantVillage public dataset	Transfer learning method and the NASNetLarge fine-grained model based on attention mechanism.	Train and test the model with more extensive image datasets from various geographical regions, field conditions, image capture modes, and multiple sources.
[2]	PlantVillage dataset and Fujian Institute of Subtropical Botany dataset	Twice Transfer learning and a modified deep CNN approach used the "MobileNet" with "Squeeze and Excitation" (SE) block.	Researchers want to use it on mobile devices to track and diagnose a variety of plant diseases. The model applies to other similar fields such as online defect assessment, molecular cell recognition, and identification of location from disparate pictures.
[3]	Not specified	K-means clustering [22], Basic Morphological functions, "Naïve Bayesian" classifier, "Colour Co-Occurrence" method.	None
[5]	Not specified	K-means clustering Multiclass SVM and "Particle Swarm Optimization" (PSO) technique.	Developing combinations of more algorithms with fusion classification methods to improve the recognition rate of the classification process.
[8]	5932 field images	11 CNN models in transfer learning approach and deep feature plus support vector machine (SVM)	Testing for more varieties of rice diseases and a more fine-tuned "Convolution Neural Network" model with the expectation of better performance.
[11]	Not specified	RGB to HSI conversion and thresholding. Segment the components using "Otsu's method". "Colour Co-Occurrence" method and "Genetic Algorithm" as a classifier.	None
[15]	PlantVillage public dataset	CNN based model	Due to the testing accuracy is lower; modify the model using a larger number

			of pictures and a different crop and procedure to improve the same model on the same dataset.
[18]	Shri Mata Vaishno Devi University Dataset	Seven different machine learning algorithms (LR, LDA, KNN, CART, RF, NB, SVM) with Simple linear iterative clustering (SLIC) [23], "Adaptive Linear Neuron" (ADALINE) [24], "Scale-Invariant Feature Transform" (SIFT) [25]	Improving the learning rate to increase the segmentation performance and adopting a deep neural network for classification using some nature-inspired algorithms.
[19]	Selected categories of PlantVillage public dataset	MLP [26] and SVM classifier	Classifying different plant diseases and improving the classification accuracy.
[20]	PlantVillage public dataset	EfficientNet deep learning model	Improved models enabling plant pathologists and farmers to identify plant diseases rapidly in mobile contexts.

IV. Results

The tabular list is presented below in Table II, including the accuracy value and classification technology that have been covered to achieve that level of accuracy. In addition, figure 1 represents accuracy values of the paper reference number in this review paper.

TABLE II. LIST OF REVIEWED PAPERS WITH ACCURACY VALUES AND USED METHODS

Article Reference Number	Classification Technology	Average Accuracy (%)
1	NASNetLarge neural network model with Attention mechanism	93.05%
2	Twice Transfer learning and the SE-MobileNet model	99.33%
3	K-means clustering, basic morphological functions, Naïve Bayesian classifier, Colour Co-Occurrence method.	87%
5	Multiclass SVM and Particle Swarm Optimization Technique	97.91%
8	CNN based support vector machine (SVM)	97.62%
15	Convolution Neural Network	91.2%.
18	Computer Vision based approach	98.57%
19	MLP and SVM classifier	94.35%
20	EfficientNet deep learning model	99.91%

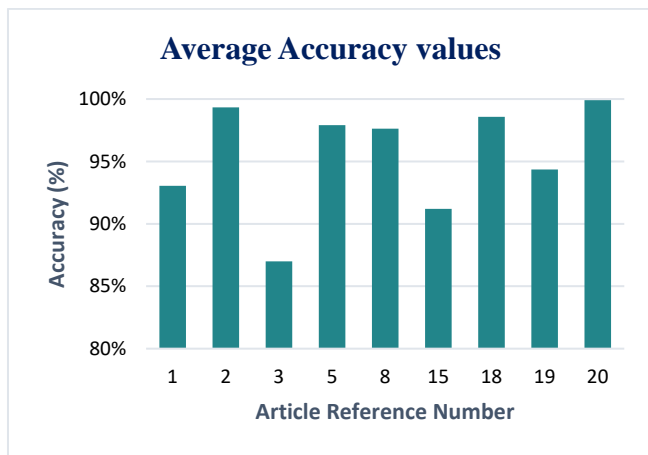


Fig. 1. Graph representation of accuracy values of reviewed papers

V. CONCLUSION

This paper provides a survey of different disease classification methods that can be used for crop leaf disease identification. An algorithm machine learning technique for automatic detection and classification of crop leaf diseases is described later. Most researchers used the PlantVillage public dataset for the algorithms and testing methods. Therefore, diseases related to these plants were taken for identification. With shallow computational efforts, the optimal result was gained, which also demonstrates the algorithm's efficiency in the identifying and classifying plant leaf diseases. Identifying the crop leaf diseases in the early-stage or initial stage is the main advantage of those methods. To maximise the recognition rate in the classification process Artificial Neural Network, Computer Vision-based approach, a deep learning model can also be used.

REFERENCES

[1] G. Yang, Y. He, Y. Yang, and B. Xu, "Fine-Grained Image Classification for Crop Disease Based on Attention Mechanism," *Front. Plant Sci.*, vol. 11, no. December, pp. 1–15, 2020, doi: 10.3389/fpls.2020.600854.

[2] J. Chen, D. Zhang, M. Suzaudola, Y. A. Nanekaran, and Y. Sun, "Identification of plant disease images via a squeeze-and-excitation MobileNet model and twice transfer learning," *IET Image Process.*, no. May, pp. 1–13, 2020, doi: 10.1049/ipr2.12090.

[3] D. Mondal, A. Chakraborty, D. K. Kole, and D. D. Majumder, "Detection and classification technique of Yellow Vein Mosaic Virus disease in okra leaf images using leaf vein extraction and Naive Bayesian classifier," *Int. Conf. Soft Comput. Tech. Implementations, ICSCCTI 2015*, pp. 166–171, 2016, doi: 10.1109/ICSCCTI.2015.7489626.

[4] N. Boyko and K. Boksho, "Application of the naive Bayesian classifier in work on sentimental analysis of medical data," *CEUR Workshop Proc.*, vol. 2753, pp. 230–239, 2020.

[5] Prabira Kumar Sethy, "Detection & Identification of Rice Leaf Diseases using Multiclass SVM and Particle Swarm Optimization Technique," no. 6, pp. 108–120, 2019.

[6] D. A. Pisner and D. M. Schnyer, *Support vector machine*. Elsevier Inc., 2019.

[7] F. Wang, H. Zhang, and A. Zhou, "A particle swarm optimization algorithm for mixed-variable optimization problems," *Swarm Evol. Comput.*, vol. 60, p. 100808, 2021, doi: 10.1016/j.swevo.2020.100808.

[8] P. K. Sethy, N. K. Barpanda, A. K. Rath, and S. K. Behera, "Deep feature based rice leaf disease identification using support vector machine," *Comput. Electron. Agric.*, vol. 175, no. May, p. 105527, 2020, doi: 10.1016/j.compag.2020.105527.

[9] I. Z. Mukti and D. Biswas, "Transfer Learning Based Plant Diseases Detection Using ResNet50," 2019 4th Int. Conf. Electr.

Inf. Commun. Technol. EICT 2019, no. December, pp. 1–6, 2019, doi: 10.1109/EICT48899.2019.9068805.

[10] M. O. Ramkumar, S. S. Catharin, V. Ramachandran, and A. Sakthikumar, "Cercospora identification in spinach leaves through resnet-50 based image processing," *J. Phys. Conf. Ser.*, vol. 1717, no. 1, 2021, doi: 10.1088/1742-6596/1717/1/012046.

[11] M. S. Arya, K. Anjali, and D. Unni, "Detection of unhealthy plant leaves using image processing and genetic algorithm with Arduino," *EPSCICON 2018 - 4th Int. Conf. Power, Signals, Control Comput.*, pp. 1–5, 2018, doi: 10.1109/EPSCICON.2018.8379584.

[12] W. Yi, Z. Jing, and G. Shuang, "Hue-saturation-intensity and texture feature-based cloud detection algorithm for unmanned aerial vehicle images," *Int. J. Adv. Robot. Syst.*, vol. 17, no. 3, pp. 1–8, 2020, doi: 10.1177/1729881420903532.

[13] P. Yang, "An improved Otsu threshold segmentation algorithm Wei Song *, Xiaobing Zhao and Rui Zheng Letu Qingge," vol. 22, no. 1, pp. 146–153, 2020.

[14] A. Tarafdar, B. P. Kaur, P. K. Nema, O. A. Babar, and D. Kumar, "Using a combined neural network – genetic algorithm approach for predicting the complex rheological characteristics of microfluidized sugarcane juice," *Lwt*, vol. 123, p. 109058, 2020, doi: 10.1016/j.lwt.2020.109058.

[15] M. Agarwal, A. Singh, S. Arjaria, A. Sinha, and S. Gupta, "ToLeD: Tomato Leaf Disease Detection using Convolution Neural Network," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 293–301, 2020, doi: 10.1016/j.procs.2020.03.225.

[16] A. Krishnaswamy Rangarajan and R. Purushothaman, "Disease Classification in Eggplant Using Pre-trained VGG16 and MSVM," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-59108-x.

[17] C. Wang et al., "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146533–146541, 2019, doi: 10.1109/ACCESS.2019.2946000.

[18] S. S. Chouhan, U. P. Singh, U. Sharma, and S. Jain, "Leaf disease segmentation and classification of *Jatropha Curcas L.* and *Pongamia Pinnata L.* biofuel plants using computer vision based approaches," *Meas. J. Int. Meas. Confed.*, vol. 171, p. 108796, 2021, doi: 10.1016/j.measurement.2020.108796.

[19] Y. Kurmi, S. Gangwar, D. Agrawal, S. Kumar, and H. S. Srivastava, "Leaf image analysis-based crop diseases classification," *Signal, Image Video Process.*, vol. 15, no. 3, pp. 589–597, 2021, doi: 10.1007/s11760-020-01780-7.

[20] Ü. Atila, M. Uçar, K. Akyol, and E. Uçar, "Plant leaf disease classification using EfficientNet deep learning model," *Ecol. Inform.*, vol. 61, no. June 2020, p. 101182, 2021, doi: 10.1016/j.ecoinf.2020.101182.

[21] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 10691–10700, 2019.

[22] P. Govender and V. Sivakumar, *Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)*, vol. 11, no. 1. Turkish National Committee for Air Pollution Research and Control, 2020.

[23] C. Wu et al., "Fuzzy SLIC: Fuzzy Simple Linear Iterative Clustering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2114–2124, 2021, doi: 10.1109/TCSVT.2020.3019109.

[24] M. Jannati, S. H. Hosseinian, B. Vahidi, and G. jie Li, "ADALINE (ADaptive Linear NEuron)-based coordinated control for wind power fluctuations smoothing with reduced BESS (battery energy storage system) capacity," *Energy*, vol. 101, pp. 1–8, 2016, doi: 10.1016/j.energy.2016.01.100.

[25] T. Lindeberg, "Scale Invariant Feature Transform," *Scholarpedia*, vol. 7, no. 5, p. 10491, 2012, doi: 10.4249/scholarpedia.10491.

[26] V. N. Ghate and S. V. Dudul, "Optimal MLP neural network classifier for fault detection of three phase induction motor," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3468–3481, 2010, doi: 10.1016/j.eswa.2009.10.041.

Thought identification through visual stimuli presentation from a commercially available EEG device

M. P. A. V. Gunawardhana*
Department of Physics and Electronics,
Faculty of Science,
University of Kelaniya, Sri Lanka
gunawardhana.mpav@gmail.com

C. A. N. W. K. Jayatissa
Department of Physics and Electronics,
Faculty of Science,
University of Kelaniya, Sri Lanka
jayatissa@kln.ac.lk

J. A. Seneviratne
Department of Physics and Electronics,
Faculty of Science,
University of Kelaniya, Sri Lanka
jehans@kln.ac.lk

Abstract - Thought identification has been the ultimate goal of brain-computer interface systems. However, due to the complex nature of brain signals, classification is difficult. But recent developments in deep learning have made the classification of multivariate time series data relatively easy. Studies have been carried out in the recent past to classify thoughts based on signals from medical-grade EEG devices. This study explores the possibility of thought identification using a commercially available EEG device using deep learning techniques. The crucial part of any EEG experiment is contamination-free data collection. Keeping the subject's mind concentrated only in the decided state is important, yet challenging. To address this issue, we have developed a graphical user interface (GUI) based program that allows stimulus controlling and data recording. With the use of the low-cost commercially available EEG device, accuracies up to 89% were achieved for the classification of high contrast signals. However, tests on complex thought identification did not produce statistically significant results over the chance accuracy.

Keywords - brain-computer-interface, classification, EEG, signal processing

I. INTRODUCTION

Electroencephalography (EEG) is the method of observing the electrical activity of the brain by the electrodes placed on the scalp. EEG is one of the most used brain imaging techniques in the medical field. Other than medical uses, EEG devices have found their way into the research field of Brain-Computer Interfaces (BCI). The ultimate goal of a BCI system is extracting thoughts directly from the brain. Studying this area often requires expensive research-grade EEG devices. But there are many advantages of using a low-cost device, mainly their accessibility. In recent years, there has been an increase in the availability of low-cost EEG devices in the consumer market. This study was conducted using one of these low-cost devices, the Emotiv Insight 5-channel EEG headset.

This study explores the feasibility of identifying thoughts by captured brainwave signals using a commercially available low-cost EEG headset. The focus of this study was to visually stimulate a subject's brain with stimuli of a limited number of stimulus classes and later identify the stimulus class from the recorded EEG data. This is not a simple task since EEG signals represent the electric potential changes on the scalp that correspond to the electrical activities in the brain that are received from the electrodes and are all higgledy-piggledy. Differentiating two EEG signals of two separate thoughts

is quite difficult with traditional methods. Therefore, the proposed method employs Deep Learning techniques. Proposed EEG experiments were all highly time-sensitive. The recording of the data needed to be done simultaneously with the presentation of the stimulus. This would not be possible without the use of an automated system to control the stimulus and capture data simultaneously. Therefore, a major contribution of this research is the development of the GUI. It allows seamless data capturing, managing, and saving. This makes the data-gathering stage effective and influences the overall outcome of the experiment.

If the classification is proven to be possible using a low-cost EEG headset, this technique can be extended to develop better low-cost BCIs. Another use case of this technique is that it can be used as the base for a communication platform that will assist differently-abled people with communication. This technique can also be used in game development to allow players to control certain actions based on what goes on in the player's mind. This will lead to mind-controlled gaming.

II. LITERATURE REVIEW

The Emotiv Insight EEG headset used in this study is a relatively low-cost commercially available device. Most of the published studies using this device have used the provided software by the MANUFACTURER. The study done by Stoelinga [1] has utilized raw EEG data from the headset. When using the manufacturer's software, it uses all the inbuilt sensors (Accelerometer, Gyroscope, Magnetometer, etc.) of the EEG headset to produce the output. Even though the use of the manufacturer's software could produce better results, it may not solely be based on EEG signals, since the signals picked up by the extra sensors could influence the outcome.

Experiments performed with EEG headsets vary widely from medical diagnosis [2]–[4], emotion recognition [5], to BCI applications [6], [7] all of which use some form of learning-based analysis for classification. All these studies used high contrast EEG data. Medical EEG data like Seizures [2], Epilepsy [3], or brain-dead and coma states [4] produce highly contrasting data. This is similar for emotion [5] and Motor-Imagery data [1], [6], [7]. Motor-Imagery is imagining moving a body part (e.g. raising an arm) without acting. Even though processing EEG signals to retrieve information is not new, classifying two distinct thoughts with low contrast data is still challenging.

Extracting thoughts from EEG data has been the primary goal of the BCI research. To that end, similar

studies have been conducted where one or multiple subjects were shown images of multiple classes and later tried to identify the thought of the class from the EEG data. In 2017 one study [8] proposed an automated visual classification. But a study published in 2020 [9], questioned the stimulus presentation method of the said previous study while proposing a randomized stimulus presentation. Both studies used raw data for the classification by Deep Learning techniques. Another study published in 2020 [10] used an Evoked Potential extraction on the EEG signal and achieved a higher classification accuracy. However, these studies have used EEG devices with higher electrode counts and higher sampling rates than the EMOTIV Insight headset used in this study.

Practical use of thought identification can be identified as a yes-no classification because the most fundamental linguistic response of human speech is answering a “yes-no” question. An EEG-based system that understands a simple yes-no thought of a subject is extremely useful for people who have speech and muscle control disabilities like Amyotrophic Lateral Sclerosis (ALS) patients. A study published in 2019 [11] used EEG data gathered from multiple subjects responding to self-referential questions on a screen. There were no visual stimuli attached with the questions. The questions were uniquely generated for each subject based on a questionnaire given to them. Similar to yes-no detection, lie detection was another area explored with EEG devices [12], [13].

BCI research study requirements are usually time-sensitive. Most of the studies which were focused on BCI research applications used their software tool for data collection. The tools were extremely specific for those studies and most often cannot be used by others. There were some studies [14] that designed EEG stimulus presentation software to use in other studies. But most of them either did not work with the used EEG headset of this study or did not include features critical for the experiments like having a darker background.

Through this literature review, it was made aware that there was not much research conducted in the area of differentiating the thoughts yes and no with EEG data. And it was made clear that the most efficient way of analyzing complex EEG data is by using a learning-based technique. It was also identified the need for a software tool of some kind to efficiently collect and manage the EEG data.

III. METHODOLOGY

The methodology of the study can be summarized as the flow diagram shown in Fig.1. To successfully execute the proposed procedure, the following factors were considered.

- Simultaneous image presentation while recording the broadcasted EEG signals.
- Tag the EEG stream with the class of the image shown on the screen.
- Having a fixed sample length for each stimulus
- A distraction-free data collection

When the subject is looking at an image for a set period, the class of the shown image needs to be saved (tagged) with the recorded EEG stream. If the EEG stream is to be manually tagged by the subject who is wearing the

EEG headset, the EEG data will get contaminated with the “thought of tagging”.

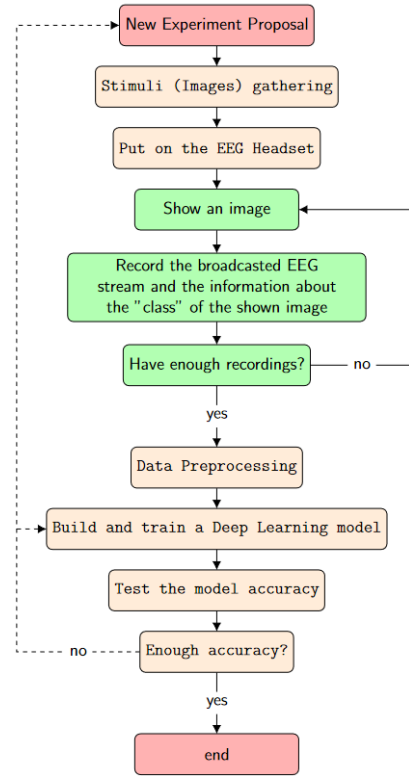


Fig. 1. Experiment procedure

Further, even if a third party was assigned for tagging with a mechanism similar to pressing a button every time the stimulus class changes, it will introduce human error into the experiment. A person performing the tagging will always introduce a random delay (error) between the time of stimulus change and the time of pressing the button.

Considering all these conditions, to make data collection consistent throughout the study, a program was developed to automate the proposed procedure.

A. Graphical User Interface (GUI)

The Graphical User Interface (GUI) was developed from scratch using the Python programming language. The main purpose of this GUI was to automate the tasks of capturing, saving, and managing the EEG data. Additionally, when building the GUI, special attention was given to the overall theme. A darker color pallet and low contrast fonts were used to keep the attention of the subject always on the area where the stimulus would be displayed on the screen when the software is used to gather data from stimuli. Since staring at a bright screen easily strains human eyes, using a darker background was found to be crucial for long recording sessions. Fig. 2 shows the main user interface of the GUI.

Here, the user can set the parameters of the experiment. Fig.3 shows what parameters are available to the user. Descriptions of the user-controllable parameters are as follows.

- 1) Interval – The period between two images.
- 2) Count – Number of images per one recording.

- 3) Subject Name – Name of the participant.
- 4) Project Name – Selection list of available image sets.
- 5) Start Recording – Button to click on to start the recording process

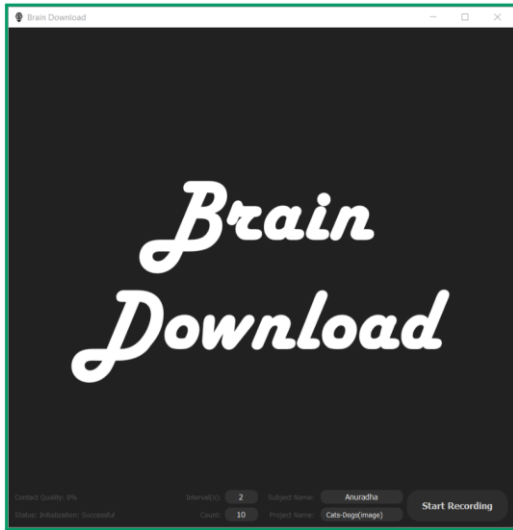


Fig. 2. Main interface

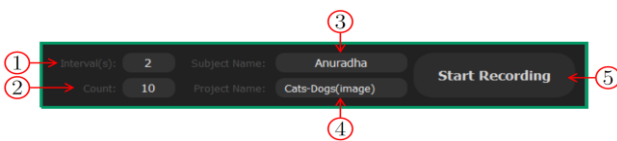


Fig. 3. User control parameters

B. GUI program flow

During the *image sequencing* process, which is in green color on Fig. 4, the GUI simultaneously records the EEG stream with some additional information.

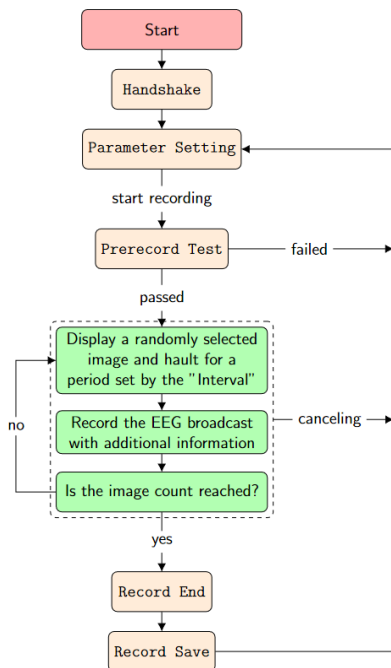


Fig. 4. Program flow of the GUI

A saved EEG recording contains data from 9 different variables,

- 5 – EEG channel (AF3, AF4, T7, T8, Pz)
- COUNT – A data packet counter.
- Contact Quality – Contact quality of the electrodes.
- TICK – Track the stimulus change.
- MARKERS – Track the class of the stimulus.

Data variables except TICK and MARKERS are directly captured from the broadcasted EEG stream. TICK and MARKER variables were added by the GUI to track the changes of the stimuli.

The TICK variable has two states, 0 and 1. Every time the GUI changes the image, the TICK changes its state. A record will contain several seconds long continuous stream of 5 EEG channel data. But to analyze the data, the stream needed to be separated into chunks depending on the stimulus shown period (interval). Since the variable TICK changes with every new image, it is used to identify the positions where the data stream needs to be split.

The MARKERS variable encapsulates the class of the image. When an image folder is selected, the GUI scans all the image files in the folder and identifies their unique classes. For example, for a folder that contains images of 4 types of vehicles [car, bus, train, bicycle], first, the program arranges the unique class names in the ascending order as [bicycle, bus, car, train] and assigns four index values starting from 0 as [0, 1, 2, 3]. When an image is shown on the screen, the value assigned to the image gets recorded as the MARKERS value throughout that image presentation period. For example, if an image of a car is shown on the computer screen, the value 2 will get recorded as the MARKERS until the next image is selected. After the stream is separated into chunks at the ‘preprocessing’ stage, they get labeled according to the values of the MARKERS variable.

When the GUI has shown a number of images specified by the researcher, the recording stops and the program saves the record in the computer hard disk as a .csv file.

Since the recording stage of this study stretched for several months, to save the records in a meaningful manner, the program uses the following naming convention when saving the recorded data.

[projectName][interval]sx[count][DATE]-[TIME].csv

The bracketed variables get replaced by the parameters set by the user. From this naming convention, all the necessary information about the record can be easily identified from the record name.

During communication, other than words humans often use body language and head movements to convey their inner thought to the other person. Used EEG headset can capture head movement data using the inbuilt sensors of accelerometer, gyroscope, and a magnetometer. When presented with a yes-no question, people unconsciously nod their head for the answer yes and move their head side-to-side for the answer no. Hence, head movements might give an extra edge with thought identification. Since the focus of the study is thought identification with EEG signals, the head movement data was not associated with the analysis.

C. Sample record

Each exported EEG record from the GUI contains EEG signals of watching several stimulus presentations. Fig. 5 shows a sample record of a subject watching 20 consecutive image presentations with 2-second image intervals.

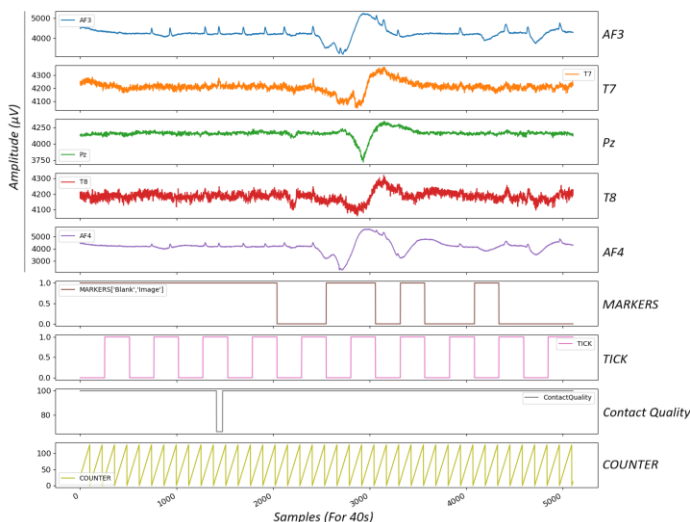


Fig. 5. Exported record from the GUI

D. Stimulus presentation methods

To visually stimulate the subject's brain, various methods were identified, in which the images of multiple classes can be presented.

One method of stimulus presentation is separating the whole image-set into subsets based on their class as suggested in [8] and continuously displaying images of one subset at a time, as shown in Fig.6. In Fig. 6, a two-class image set is separated into two sections (shown in two different colors for simplicity) and images of one set are displayed first before the images of the other set are displayed.

However, with this method, since all the images of a class are shown continuously, captured brain wave patterns are temporally correlated. This means EEG signals of each class will contain the patterns of the long-term mental state of the subject. For example, assume in this case (Fig. 6) the subject is looking at images of Class B first and then Class A during the experiment. In the beginning, the subject might be in an excited mood, and most of the EEG signals of Class B will capture that excited brain pattern. But at the end of the experiment, the subject might get bored, and those brain patterns will get captured in the EEG signals of Class A. When classifying these brainwaves, rather than detecting the thought of the presented stimuli, the brainwaves of excitement and boredom will get precedence.

Instead of separating the stimuli into individual classes and showing all images of one class before proceeding to the other classes, the image-set can be separated into smaller batches based on their class as shown in Fig. 7 and alternatively show each batch from separate classes. This method reduces the temporal correlation but not completely. If the length of the batch is too long, the error remains.

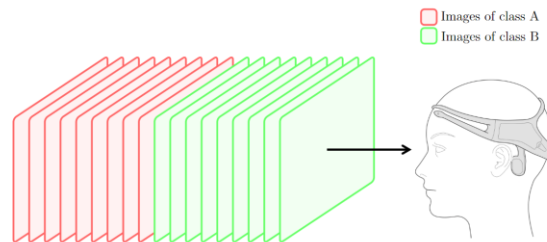


Fig. 6. Image-set was separate into classes

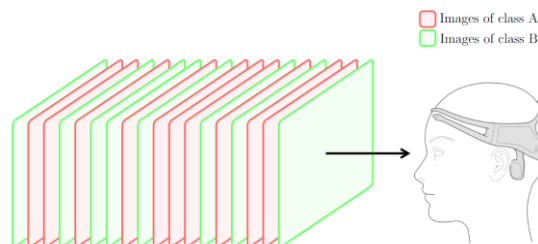


Fig. 7. Image-set was separated into multiple batches.

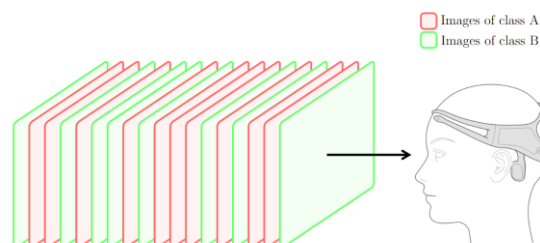


Fig. 8. The subject was presented with randomly selected images from the image-set

To eliminate the errors discussed above, it requires a randomized stimulus presentation [9] as shown in Fig. 8. If it is not randomized, and the presentation is similar to Fig. 7, the subject's brain will recognize the stimuli presentation pattern and will know what to expect in the next image. This can also be eliminated by using a randomized stimulus presentation. Hence, all conducted experiments in this study used a randomized stimulus presentation method.

E. Signal filtering and dataset conversions

At the preprocessing stage, the recorded long EEG signals were separated into smaller chunks of the subject watching one stimulus using the TICK variable. The MARKERS variable was used to label the separated chunks.

After the basic preprocessing, the obtained raw dataset was converted into 3 other forms to find out whether the classification accuracy of the deep learning model can be improved.

- The raw dataset was converted into the frequency domain using the Fast Fourier Transformation (FFT).
- Filtered the low-frequency blinking artifacts by adding a high pass filter at 12 Hz and filtered the 50 Hz electromagnetic interference by adding a notch filter.

- Using the Short-Time Fourier Transformation (STFT) each chunk in the raw dataset was converted into a stacked spectrogram.

To generate a stacked spectrogram, first, a chunk was selected from the raw dataset. Then each of the 5 EEG signals was converted into separate spectrograms using STFT (see Fig. 9). Then all the generated spectrograms were stacked on top of each other to generate a diagram similar to what is shown in Fig. 10.

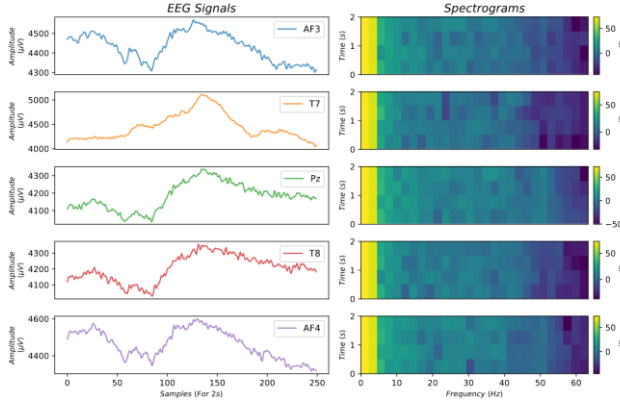


Fig. 9. 5 EEG signals converted separately into spectrograms

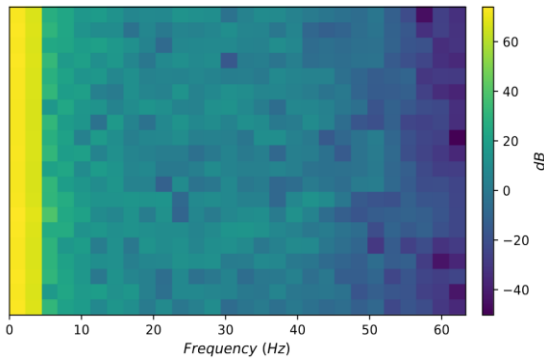


Fig. 10. Stacked spectrogram

F. Conducted experiments

To assess the feasibility of thought identification from the used low-cost EEG device, four experiments were conducted where the participant’s brain was visually stimulated by a presentation of image sequences. Only one subject was used for all the experiments conducted.

- *Experiment 1* – Simulated thinking “something” and “nothing” on the subject’s brain by randomly presenting images and blank screens to the subject.
- *Experiment 2* – Showed left and right directed arrows on the left and right edges of the screen respectively and the subject was instructed to directly look at them without moving the head. Since the image sequence is randomized, a reference mark was presented at the center of the screen after each arrow image.
- *Experiment 3* – Simultaneously displayed a yes-no question about the presented image and instructed the subject to think about the answer.

- *Experiment 4* – Showed images of cats and dogs, and the subject was instructed to identify the class of the image as a “cat” or a “dog”.

IV. RESULTS

In this study, we employed 2 deep learning models for the classification of the recorded EEG data. A one-dimensional Convolutional Neural Network (1D-CNN) was used for the classification of the multivariate time series data. The classification of the stacked spectrograms was done using a two-dimensional Convolutional Neural Network (2D-CNN) [15].

A. Classification results of the three experiments

TABLE II. CLASSIFICATION RESULTS

Experiment	Classes	Dataset	Classification accuracy (%)	
			1D-CNN	2D-CNN
Thinking “something” and “nothing”	Image, Blank	Raw	80	-
		FFT	79	-
		Filtered	80	-
		Spectrograms	-	74
Left-Right arrows	Center, Left, Right	Raw	68	-
		FFT	67	-
		Filtered	67	-
	Left, Right	Raw	89	-
		Filtered	91	-
		Spectrograms	-	69
Yes-No	Yes, No	Raw	45	-
		FFT	43	-
		Filtered	45	-
		Spectrograms	-	50
Cats-Dogs	Cat, Dog	Raw	52	-
		Spectrograms	-	54

Since experiment 2 presented a reference mark at the center of the screen, it contained EEG recordings of 3 separate classes of Center, Left, and Right. For three classes, the highest classification accuracy of 69% was achieved by the 2D-CNN model. For the classification of EEG signals of looking only Left and Right, the 1D-CNN reached an 89% accuracy.

It is important to notice that using spectrograms with a 2D-CNN model, there was no statistically significant improvement. All the results are fairly similar between both models. Also, the additional conversions of Fourier transformation and filtering done on the data did not increase the accuracies of the models.

B. Channel contributions

Individual datasets contain 5 separate EEG signals. Table I shows results of experiment 1 when all 5 EEG channels were concerned and maximum classification accuracy of 80% for the 1D-CNN model was achieved for

both the raw and filtered datasets. Table II shows the classification results of several EEG channel combinations of experiment 1. By selecting multiple combinations of EEG channels, the study tried to identify a channel or combination which contributes the most to the final accuracy. Only models that performed above the random chance accuracy of 50% are listed.

TABLE III. CLASSIFICATION RESULTS OF SELECTING MULTIPLE CHANNEL COMBINATIONS

Channel combinations	1D-CNN classification accuracy of experiment 1 (%)	
	Raw	Filtered
AF3, T7, Pz, T8, AF4	80	80
AF3, Pz	80	80
AF3, Pz, T8	80	78
AF4, Pz, T8	80	79
AF4, Pz	78	77
AF3	73	70
AF3, AF4	68	68
Pz	64	64
T8	54	56
T7	51	51

These results further clarify the fact that filtering done on the raw data did not affect the final accuracy of the model.

V. CONCLUSION

The automated data collection and tagging method implemented using the GUI were found to be crucial for acquiring contamination-free EEG samples. Since the GUI allowed effortless sample management, in a relatively short period we were able to gather EEG samples from multiple experiments.

Even though several studies have been published that converted the raw data into other formats such as spectrographs [4], [5] and scalograms [3], [6], [7], the analysis of this study suggests using only raw data for the classification is sufficient for the data gathered with Emotiv Insight 5-channel EEG headset, which is a low-cost EEG device.

Even though the classification of experiments 1 and 2 reached higher accuracies this might not be solely based on EEG signals. The coneo-retinal potential [16] might have played a major role in this. This is also confirmed by the results presented in Table II. When the frontal lobe channels of AF3 and AF4 are not selected, the accuracy drops considerably. Therefore, for complex thought identification tasks such as yes-no answer identification and distinct thought classification (thinking “cat” versus “dog”), we recommend using a device with a higher electrode count.

VI. LIMITATIONS

Since the GUI is built around one EEG headset (Emotiv-Insight) in mind, it cannot be used with EEG headsets of other manufacturers. But with minor changes,

the GUI can be made to work with other models of EEG devices of the same manufacturer (EMOTIV). However, the concept can be applied to any EEG device.

VII. RECOMMENDATION

All the experiments conducted for binary classification had a balanced stimuli presentation. Future research can be conducted to see the effects of unbalancing the image-set on the final accuracy. Also, for the analysis of this study the whole EEG signals of watching a 2-second stimulus was used. A study can be conducted to see the effect on the accuracy of the model when a shorter length is selected from the EEG signals.

REFERENCES

- [1] Stoeltinga, “Exploring the possibilities of the Emotiv Insight: discriminating between left- and right-handed responses Methods Participants,” no. 2013, pp. 1–11, 2016.
- [2] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, “Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals,” *Comput. Biol. Med.*, vol. 100, pp. 270–278, 2018.
- [3] Ö. Türk and M. S. Özerdem, “Epilepsy detection by using scalogram based convolutional neural network from eeg signals,” *Brain Sci.*, vol. 9, no. 5, pp. 1–16, 2019.
- [4] L. Yuan and J. Cao, “Patients’ EEG Data Analysis via Spectrogram Image with a Convolution Neural Network,” 2018.
- [5] F. Wang et al., “Emotion recognition with convolutional neural network and EEG-based EFDMs,” *Neuropsychologia*, vol. 146, no. June, p. 107506, 2020.
- [6] Y. R. Tabar and U. Halici, “A novel deep learning approach for classification of EEG motor imagery signals,” *J. Neural Eng.*, vol. 14, no. 1, p. 16003, 2017.
- [7] M. Dai, D. Zheng, R. Na, S. Wang, and S. Zhang, “EEG classification of motor imagery using a novel deep learning framework,” *Sensors (Switzerland)*, vol. 19, no. 3, pp. 1–16, 2019.
- [8] C. Spampinato, S. Palazzo, I. Kavassidis, D. Giordano, N. Souly, and M. Shah, “Deep learning human mind for automated visual classification,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4503–4511, 2017.
- [9] H. Ahmed, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, “Object classification from randomized EEG trials,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 3845–3854, 2020.
- [10] X. Zheng, Z. Cao, and Q. Bai, “An Evoked Potential-Guided Deep Learning Brain Representation For Visual Classification.”
- [11] J. W. Choi, K. H. Kim, and H. J. Baek, “Covert Intention to Answer ‘yes’ or ‘no’ Can Be Decoded from Single-Trial Electroencephalograms (EEGs),” *Comput. Intell. Neurosci.*, vol. 2019, 2019.
- [12] N. Baghel, D. Singh, M. K. Dutta, R. Burget, and V. Myska, “Truth Identification from EEG Signal by using Convolution neural network: Lie Detection,” 2020 43rd Int. Conf. Telecommun. Signal Process. TSP 2020, pp. 550–553, 2020.
- [13] J. Gao, H. Tian, Y. Yang, X. Yu, C. Li, and N. Rao, “A novel algorithm to enhance P300 in single trials: Application to lie detection using F-score and SVM,” *PLoS One*, vol. 9, no. 11, 2014.
- [14] M. Plong, K. Shen, M. Van Vliet, A. Robben, and M. Van Hulle, “Accurate Visual Stimulus Presentation Software for EEG Experiments,” pp. 1–4.
- [15] A. Simonyan, Karen and Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv Prepr. arXiv1409.1556*, 2014.
- [16] E. Marg, “Development of electro-oculography: Standing potential of the eye in registration of eye movement,” *AMA Arch. Ophthalmol.*, vol. 45, pp. 169--185, 1951.

LYZGen: A mechanism to generate leads from Generation Y and Z by analysing web and social media data

Janaka Senanayake*

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

Nadeeka Pathirana

Department of Information Technology
University of Sri Jayawardenapura, Sri Lanka
pathirana@sjp.ac.lk

Abstract - Identifying an appropriate target audience is essential to market a product or a service. A proper mechanism should be followed to generate these potential leads and target audiences. The majority of people who were born between 1981 and 2012 hold top positions in companies. These people are regular social media and website users, since they represent generations Y and Z. They usually keep digital footprints. Therefore, if an accurate method is followed, it is possible to identify potential contact points by analysing publicly available data. In this research, a novel lead generation mechanism based on analysing social media and web data has been proposed and named LYZGen (Leads of Y and Z Generations). The input to the LYZGen model was an optimised search query based on the user requirement. The model used web crawling, named entity recognition (NER), and pattern identification. The model found and analysed freely available data from social media and other websites. Initially, person name identification was performed. An extensive search was carried out to retrieve peoples' contact points such as email addresses, contact numbers, designations, based on the identified names. Cross verification of the analysed details was conducted as the next step. The results generator provided the final output, which contained the leads and details. Generated details were verified with responses captured via a survey and identified that the model could detect lead details with 87.3% average accuracy. The model used only the open data posted on the internet by the people. Therefore, it did not violate extensive privacy or security concerns. The generated results can be used, in several ways, including communicating promotional details to the potential target audience.

Keywords - lead generation, named entity recognition, web crawling, web data analysing

I. INTRODUCTION

There is a high number of instances of communicating about promotional details related to products and services. However, in most cases, these communications are conducted without identifying the potential audience. Resources and time of the advertisers or promotional campaign organisers might be wasted because of this. Therefore, identifying the potential leads should be the initial task of this whole process.

These potential leads can be generated by thoroughly analysing the web data [1]. Many young people who belong to Generation Y and Generation Z tend to keep digital footprints knowingly or unknowingly when they browse the internet and social media. That is the nature of Generation Y [2] and Generation Z [3].

Web crawling and web data analysis techniques can be applied to analyse the content of a web page [4], which is also known as web scraping. By using a spider, the analysis

can be performed in web scraping, and by using NER, person names can be identified [5] after analysing a textual input.

The use of web crawlers and web data analysis is not a novel area since various approaches were already proposed by academia. Their strengths and limitations are also discussed [6]. However, combining web crawlers to generate leads after identifying generation Y and Z behaviour in the digital space is not considered. The usage of websites and social media has increased rapidly, especially among the generations that were focused in this study. This increase is due to the travel restrictions imposed with the ongoing Covid-19 pandemic situation. In this paper, a model to detect leads and contact details of persons, using web crawling, web data analysis and named entity recognition, has been proposed. The generated data were validated again using web data analysis to determine the accuracy. In the model, all the steps in data collection and analysis were conducted on publicly available data on the web. Since the details were not extracted using any illegal approaches, there are no significant concerns of privacy violations [7].

Following research questions were answered in this research while building the LYZGen model.

- RQ1: What are the optimising strategies of web search queries?
- RQ2: How to apply web crawling and web data analysis to generate leads?
- RQ3: How to perform valid pattern recognition processes to identify lead-related attributes?
- RQ4: How to validate the accuracy of the contact details of the potential leads?

The generated details were re-evaluated for their accuracies by comparing them with survey results. This survey was conducted to record the name, details of designation, email address, and contact number from volunteers from academic, medical, financial and information technology fields. The survey results contain 179 records.

The rest of the paper is organized as follows: Section II contains related work. Section III gives an overview and the methodology of the LYZGen system to generate potential leads. Section IV presents the results and discussions related to the research. Finally, the conclusions and future work directions are discussed in Section V

II. RELATED WORK

There are various research studies conducted in the research areas of identifying leads, mechanisms of web crawling and web data analysing, NER methods, and user generations. However, to the best of our knowledge, there is no comprehensive research conducted after combining each of these individual areas to build a proper lead generation mechanism. In this section, related research studies in those mentioned areas are discussed.

People who live all around the world can be categorized based on different dimensions. Among all these dimensions, the “generation” has become one of the important societal categories introduced [8]. In the human context, a generation is defined as a group of people who were born and nurtured at a specific time. They have common characteristics and viewpoints which are affected by their growing time. It implies that there are characteristic discrepancies among generations.

In the current society, four to five generations are working side by side [9]. Among them, generation Z and generation Y are the latest generations who work in society nowadays. They deal with technology frequently. Generation Y is the first generation of people who came into the world of technology [2] when they were born. Generation Z is the first generation born with the technology; known as digital natives [3]. The new generation always tries to perform their tasks efficiently with the help of technology [10]. The research conducted in [11] identifies the fact that the leaders of using technology are the people from generation Y and Z. Compared to the other generations, they spend a significant amount of time surfing and browsing the internet for different purposes. Due to this behaviour, they tend to keep their digital footprints in cyberspace more than the other generations do.

One of the most important facts to consider when dealing with technology is security. Most people use their mobile devices to engage with the digital space actively. Several techniques used to detect and prevent attacks on the users, such as data theft, social engineering, and malware have been identified [12]. However, generation Y and Z people should consider their digital footprint to keep themselves safe since there are some ways for obtaining these footprints, which includes recording of footprints with or without the consent or acknowledgement of the users.

These digital footprints of generations Y and Z can be collected and analysed to generate leads. Lead generation is one of the most common marketing approaches used to identify potential customers. This method helps identify the target audience for a particular domain. Through identifying contact points, it is easy to reach the right people [13]. Various lead generation methods are being practised on several occasions [14].

To find the leads and related details, one way that can be used is to analyse the web page contents. An automated mechanism should be integrated to achieve that. Web crawling and web data analysis are the methods, which can be applied to this [15]. Web crawling is also known as web scraping. In web scraping, the feature known as spider visits websites and scrapes all the data after performing an analysis. In [4] and [16], many methods are proposed to conduct web scraping. Out of those methods, the spider-

based web scraping method was identified as the efficient method, and this is used in many web search engines.

The scrapped content should properly be analysed though an extensive web crawling method. If the content can be formed into a textual string, it is possible to apply several text mining methods to identify patterns [17]. NER is one of the commonly used methods to identify names with the help of text mining and natural language processing. NER can be categorised into three main categories as Hand-made Rule-based NER, Machine-Learning based NER and Hybrid NER. Mining names using human-made rules set is known as hand-made rule-based NER. Machine Learning-based NER can identify problems and classify problems, and then the System identifies patterns and relationships. After that, it makes a model using available statistical models and machine learning algorithms. Hybrid NER is the combination of rule-based and Machine Learning based NER approaches [18]. Based on the requirements, the types that need to be recognized could vary. Recognition can be done for a person, contact details, location, or other information related to a specific task.

Privacy and security of web data are also important to be considered. With technological enhancement, people tend to use online resources to do their day-to-day activities efficiently. When people use different web applications and mobile applications, they create social networks through digital platforms. Due to this, people make their details available in public, knowingly or unknowingly. These details may contain their experiences, opinions and knowledge. There can be private data such as name, contact information, gender, etc. [19] among those details. Sharing this type of information could have both. a negative and a positive effect. If a person shares sensitive information, a negative impact for that user can also be generated. For example, insurance companies can collect that information to identify users as risky clients [20].

III. METHODOLOGY

To overcome the identified problem by addressing the formulated research questions, the LYZGen model is proposed, as described in this section.

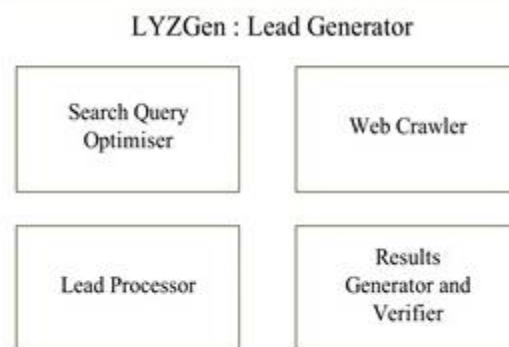


Fig 1. LYZGen architecture

The methodology of this work was distributed among four sub-systems. These subsystems were named Search Query Optimiser, Web Crawler, Lead Processor, and Results Generator and Verifier. The overview of the system is illustrated in Figure 1. Each of these subsystems were connected to generate verified results on potential leads.

An interface of the prototype system which performed those four subsystem processes is illustrated in Figure 2. The prototype was developed using the Java programming language.

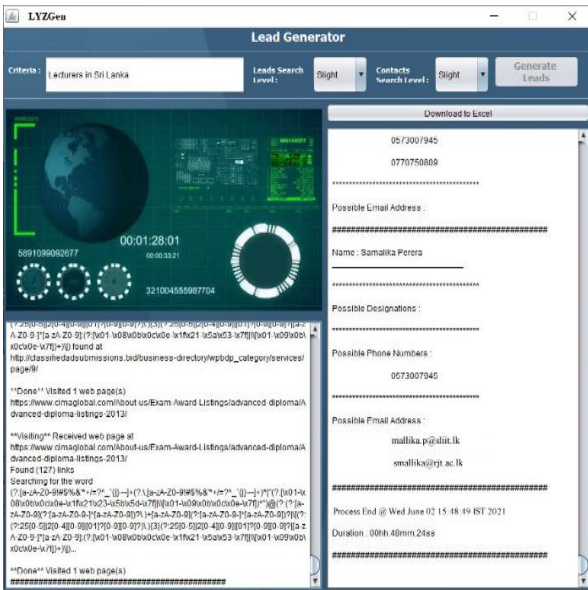


Fig 2. LYZGen prototype

A. Search query optimiser

The initial input to the model is the search criteria. There is a difference between web search queries entered by a person having good computer literacy and a regular person. But this model can be used by anyone. Therefore, search query optimisation should be performed as the first task to retrieve accurate search results [21]. The search criteria entered by the user was split into words, and a string array was created. Then, using “OR” and “AND” operators, the search query was optimised. The pre-stored article words were not taken into consideration initially when preparing the search string. Some examples for optimised search queries are listed in Table I.

TABLE I. SEARCH QUERY OPTIMISATION

User Input	Optimised Search Query
Lecturers in Sri Lanka	(“Lecturers”) AND (“Sri Lanka”) OR (“Sri Lanka”) AND (“Lecturers”) OR (“Lecturers Sri Lanka”) OR (“Sri Lanka Lecturers”) OR (“Lecturers in Sri Lanka”)
Cricket Players in Sri Lanka	(“Cricket”) AND (“Players”) AND (“Sri Lanka”) OR (“Cricket”) AND (“Sri Lanka”) AND (“Players”) OR (“Sri Lanka”) AND (“Cricket”) AND (“Players”) OR (“Sri Lanka”) AND (“Players”) AND (“Cricket”) OR (“Players”) AND (“Cricket”) AND (“Sri Lanka”) OR (“Players”) AND (“Sri Lanka”) AND (“Cricket”) OR (“Cricket”) AND (“Players”) OR (“Cricket”) AND (“Sri Lanka”) OR (“Players”) AND (“Cricket”) OR (“Players”) AND (“Sri Lanka”) OR (“Sri Lanka”) AND (“Cricket”) OR (“Sri Lanka”) OR (“Sri Lanka”) AND (“Cricket”) OR (“Sri Lanka”) AND (“Players”) OR (“Cricket Players in Sri Lanka”) OR (“Cricket Players Sri Lanka”)

B. Web crawler

Web crawler performs the searching and crawling process of the model. Once the search query was optimised, the web crawler was activated. The crawler can be customised with search depth (known as the Leads Search Level) as “Slight”, “Low”, “Moderate”, “Strong”, and “Extreme”. The number of outputs depends on the depth level. The time it takes to complete the search depends on the number of words the user input and the search depth. Then the search depth was converted into a numeric value. Values from 1 to 5 were assigned from Slight to Extreme. For example, if the user selects Strong (value is 3) as the search depth, the web crawler visits 30 (3×10) links and their sub-links in search engine results. The reason for multiplying by 10 is that one page of a search engine results contains ten results (links). The links visited by the web crawler were stored in a Java Collection to further process in the Lead Processor subsystem. Once the Lead Processor requests the crawl process to identify names and contact details, the Web Crawler subsystem crawled web pages while considering “Contact Search Level” as one parameter which defines the depth of the data analysis process of a given web link. The Contact Search Level also has five levels: “Slight”, “Low”, “Moderate”, “Strong” and “Extreme”. Similar to the Leads Search Level parameter, this also represents values from 1 to 5, and the value was multiplied by 10. “Mozilla/5.0 (compatible; Googlebot/2.1”, “Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.112 Safari/535.1”, and “Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US)” were used as the user agents when crawling the web pages [22]. The web pages can be either regular websites or social media sites.

C. Lead processor

The Lead Processor is the subsystem where most of the important steps happen in the overall model. Initially, this subsystem takes the input as the Java Collection (List) generated by the web crawler. As the first step of the lead processor, the names of the leads were identified using pattern recognition and NER. The lead processor sent a request to the Web Crawler subsystems with a list of links, and then the crawler visited each link. The pattern of the name was determined using “[A-Z]([a-z]+) [A-Z]([a-z]+)” regular expression. Identified possible names were stored in a Java Hash Set to avoid duplicates. The set was iterated through several NER classifiers to identify the person names using a similar process which was followed in [23]. This model used *english.nowiki.3 class*, *english.conll.4 class*, *english.all.3 class* and *english.muc.7 class* NER classifiers [24]. Java libraries developed using those classifiers were used in the LYZGen with the category information of “PERSON” [25]. Once the names are properly identified, the Lead Processor calls the Web Crawler subsystem to determine their contact numbers, email addresses, and designations. When calling the web crawler, the search queries were modified to receive accurate results based on the type of information (i.e. contact number, email, designation). For the email address pattern recognition, an advanced regular expression `(?:[a-zA-Z0-9!#$%&'*+/= ?^_{}~|]+(?:\.[a-zA-Z0-9!#$%&'*+/= ?^_{}~|]+)*)@(?!\|x01-||x08||x0b||x0c||x0e-||x0f||x21||x23-||x5b||x5d-`

name, designation, email address, and contact number. According to analysis in Table III, it is possible to say that the people from generations Y and Z are closely associated with cyberspace as the number of records in the survey were closely matched the LYZGen results.

TABLE II. COMPARISON OF MATCHING RECORDS

Category	# Records in the Survey	# Matching Records with LYZGen			
		Names	Designations	Email Add.	Contact Nos
Medical Officers	21	19	18	19	17
Lectures	29	27	26	26	24
Banking Officers	51	46	44	45	43
Software Engineers	78	71	68	67	63

TABLE III. COMPARISON OF ACCURACIES OF LYZGEN RESULTS

Category	Accuracy of LYZGen Results (%)			
	Names	Designations	Email Add.	Contact Nos
Medical Officers	90.48	85.71	90.48	80.95
Lectures	93.10	89.66	89.66	82.76
Banking Officers	90.20	86.27	88.24	84.31
Software Engineers	91.03	87.18	85.90	80.77

Fig. 5 shows the average accuracies of the LYZGen model when identifying attributes of leads.

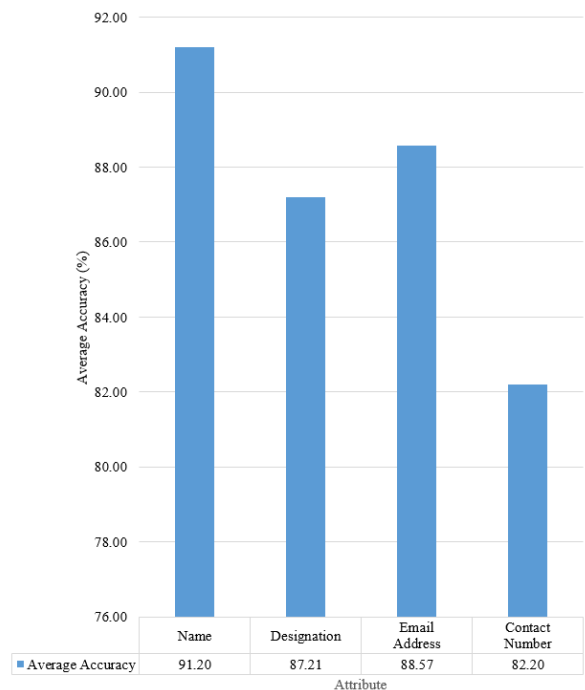


Fig 5. Average accuracies of lead details

Therefore, by analysing the above results, it is identified that the LYZGen model has a high accuracy of detecting names. The reason for that might be the attribute to be easily found when performing a web crawling process is the name. The email address is the second-highest

attribute, and designation is the third. The reason for that would be, the email addresses and employee designations are relatively easy to find in the publicly available web data. However, generating contact numbers is a difficult task. The reason for that is, they are hard to find in public sources. The accuracy of identifying contact numbers is somewhat low compared to the other attributes. The reason for that might be due to some pattern recognition issues. Overall, it is identified that the LYZGen model can identify leads and attributes with 87.3% average accuracy.

V. CONCLUSION AND FUTURE WORK

Having a proper lead generation mechanism is valuable in communicating promotional activities to the appropriate audience. Since generation Y and Z use technology and the internet more, it is possible to find digital footprints. In this paper, a novel lead generation mechanism was proposed, named LYZGen, to identify leads' details such as name, designation, email addresses, and contact numbers by analysing digital footprints and freely available data in websites and social media sites. There were four subsystems in the proposed model to perform lead generation with cross-validations. A survey was also conducted to validate the model. It was identified that the model can generate data with an average accuracy of 87.3%. The LYZGen model can be used by anyone who wants to generate leads from publicly available data without violating major privacy concerns. LYZGen can be used to generate leads to improve the strategies of marketing campaigns by identifying the most suitable target audience.

Though the generated results were conducted only in the Sri Lankan context, this model can generate results without limiting them to the context. The accuracy of generating results can be increased by improving some of the areas in the LYZGen model. We identified that the model sometimes detects incorrect person names not from the specific country due to the limitation of the NER classifier. That can be omitted if a context-based NER classifier is introduced. Currently, if the search level is selected as "Extreme", it will take a lot of time to generate the results since the crawler has to visit many web pages. The efficiency of the model can be further improved. Furthermore, the dataset generated from the current LYZGen model can be used in future research areas related to leads and contact details. Once a high number of data are collected, it will be possible to apply machine learning to improve accuracy.

REFERENCES

- [1] J. M. D. Senanayake and W. P. N. H. Pathirana, "Developing a Lead Generation Mechanism to Identify People's Contact Points Using Web Data Analytics," in Uva Wellassa University of Sri Lanka, Badulla, Sri Lanka, 2019.
- [2] S. Prasad, A. Garg and S. Prasad, "Purchase decision of generation Y in an online environment," *Marketing Intelligence & Planning*, vol. 37, no. 4, pp. 372-385, 2019.
- [3] W. P. N. H. Pathirana and D. N. Wickramaarachchi, "Software usability improvements for Generation Z oriented software application," in 2019 International research conference on smart computing and systems engineering (SCSE), Colombo, Sri Lanka, 2019.
- [4] Hernández, C. R. Rivero and D. Ruiz, "Deep Web crawling: a survey," *World Wide Web*, vol. 22, no. 4, pp. 1577-1610, 2019.

- [5] Goyal, V. Gupta and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," *Computer Science Review*, vol. 29, pp. 21-43, 2018.
- [6] M. Kumar, R. Bhatia and D. Rattan, "A survey of Web crawlers for information retrieval," *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 6, p. e1218, 2017.
- [7] S. Ribeiro-Navarrete, J. R. Saura and D. Palacios-Marqués, "Towards a new era of mass data collection: Assessing pandemic surveillance technologies to preserve user privacy," *Technological Forecasting and Social Change*, vol. 167, p. 120681, 2021.
- [8] L. Duxbury and C. Higgins, "An empirical assessment of generational differences in work-related values," *Human Resources Management Ressources Humaines*, p. 62, 2005.
- [9] J. Bejtkovsk'y, "The employees of baby boomers generation, generation X, generation Y and generation Z in selected Czech corporations as conceivers of development and competitiveness in their corporation," *Journal of Competitiveness*, 2016.
- [10] J. M. D. Senanayake and W. M. J. I. Wijayanayake, "Applicability of crowd sourcing to determine the best transportation method by analysing user mobility," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 4/5, pp. 27-36, September 2018.
- [11] T. Issa and P. Isaias, "Internet factors influencing generations Y and Z in Australia and Portugal: A practical study," *Information Processing & Management*, vol. 52, no. 4, pp. 592-617, 2016.
- [12] J. Senanayake, H. Kalutarage and M. O. Al-Kadri, "Android Mobile Malware Detection Using Machine Learning: A Systematic Review," *Electronics*, vol. 10, no. 13, p. 1606, 2021.
- [13] M. Rodriguez and R. M. Peterson, "The role of social CRM and its potential impact on lead generation in business-to-business marketing," *International Journal of Internet Marketing and Advertising*, vol. 7, no. 2, pp. 180-193, 2012.
- [14] Gupta and N. Nimkar, "Role of Content Marketing and it's Potential on Lead Generation," *Annals of Tropical Medicine and Public Health*, vol. 23, no. 17, 2020.
- [15] D. Shestakov, "Current challenges in web crawling," in *International Conference on Web Engineering*, 2013.
- [16] R. Janbandhu, P. Dahiwale and M. Raghuvanshi, "Analysis of web crawling algorithms," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 3, pp. 488-492, 2014.
- [17] T. Jo, "Text mining," *Studies in Big Data*, 2019.
- [18] N. e. r. approaches, "Mansouri, Alireza; Affendey, Lilly Suriani; Mamat, Ali," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339-344, 2008.
- [19] M. Taddicken, "The 'privacy paradox' in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure," *Journal of Computer-Mediated Communication*, vol. 19, no. 2, pp. 248-273, 2014.
- [20] L. Scism and M. Maremont, "Insurers test data profiles to identify risky clients," *The Wall Street Journal*, vol. 19, 2010.
- [21] D. Sharma, R. Shukla, A. K. Giri and S. Kumar, "A Brief Review on Search Engine Optimization," in *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2019.
- [22] T. Tanaka, H. Niibori, S. Li, S. Nomura, H. Kawashima and K. Tsuda, "Bot Detection Model using User Agent and User Behavior for Web Log Analysis," *Procedia Computer Science*, vol. 176, pp. 1621-1625, 2020.
- [23] S. Sulaiman and R. A. a. S. S. a. O. N. Wahid, "Using stanford NER and Illinois NER to detect malay named entity recognition," *Int. J. Comput. Theory Eng*, vol. 9, no. 2, pp. 147-150, 2017.
- [24] C. M. Costa, G. Veiga, A. Sousa and S. Nunes, "Evaluation of Stanford NER for extraction of assembly information from instruction manuals," in *2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2017.
- [25] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014.

A tree structure-based classification of diabetic retinopathy stages using convolutional neural network

M. S. H. Peiris*
Department of Mathematics
Eastern University, Sri Lanka, Sri Lanka
shashi.hiru15@gmail.com

S. Sotheeswaran
Department of Mathematics
Eastern University, Sri Lanka, Sri Lanka
sotheeswarans@esn.ac.lk

Abstract - Detection, and classification of medical images have become a trending field of study during the last few decades. There is a considerable amount of vital challenges to be overcome. Ample work has been carried out to provide proper solutions for those key challenges. This study was carried out to extend one such medical image classification process to classify the stages of Diabetic Retinopathy (DR) images from colour fundus images. The study proposes a novel Convolutional Neural Network (CNN) architecture which is considered to be one of the most trending and efficient forms of classification of DR stages. Initially, the pre-processing techniques were employed to the DR fundus images with Green channel extraction and Contrast Limited Adaptive Histogram Equalization (CLAHE). The data augmentation strategy was utilised to increase training images from the DR images. Finally, Feature extraction and classification were carried out by using the proposed CNN architecture. It consists of a 14 layered CNN model, which continues three main classifications. In this proposed classification, the images were classified into a tree structure based binary classification as No_DR and DR at the beginning, and then the DR images were again classified into two classes, namely Pre_Intermediate and Post_Intermediate. Moreover, those two classes were again separately classified into Mild, Moderate, and Proliferate_DR, Severe, respectively. The Kaggle is one of the benchmark dataset repositories which was used in this study. The proposed model was able to achieve accuracies of 81%, 96%, 84%, and 97% for the above-mentioned classifications, respectively.

Keywords - CLAHE, classification, CNN, diabetic retinopathy, green channel

I. INTRODUCTION

Detection and classification of medical images or medically-related objects in an image play an essential role as medical images are full of different characteristics which are absent in standard images. Preprocessing, segmentation, feature extraction, detection, classification, and prediction are some of the key challenges associated with medical image processing. Diabetic Retinopathy (DR) is the leading cause of vision loss and preventable blindness in grown-ups aged 20-74 years globally. The normal retina and diseased retina are shown in Figure 1.

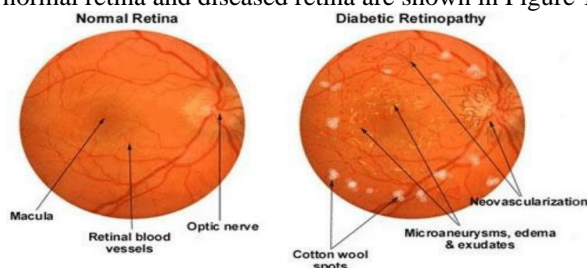


Fig. 1. Normal retina and DR

The main risk factor for the development of diabetic retinopathy is long-term diabetes which causes damage to blood vessels in the retina from high blood glucose levels [1]. DR can be classified into five stages as *No apparent retinopathy*, *Mild Non-Proliferative Diabetic Retinopathy (NPDR)*, *Moderate NPDR*, *Severe NPDR*, and *Proliferative Diabetic Retinopathy*. Visual loss can be prevented up to 90% with the proper management of DR [2].

Non-proliferative retinopathy (also named background retinopathy) emerges first and creates increased capillary permeability, microaneurysms, haemorrhages, exudates, macular ischemia, and macular edema (thickening of the retina resulted from fluid leakage from capillaries). Proliferative retinopathy progresses after non-proliferative retinopathy and is more critical; it may point to vitreous haemorrhage and traction retinal detachment [3].

The medical features of Diabetic retinopathy are as follows:

- Microaneurysms are the tiny swellings on the walls of blood vessels inside the retina that are caused due to absence of the Pericyte. These are the earliest clinically visible changes. Microaneurysms eventually rupture to form haemorrhages deep within the retina [4].
- Haemorrhages appear as large spots on the retina.
- Hard exudates form when protein drips from blood vessels, and they are wavy and yellow or white deposits of protein.
- Cotton wool spots form when leakage of blood vessels blocks the vessels. An eye with more than six cotton wool spots is generalised as a pre proliferative state [5].

Figure 2 depicts the sample pictures from each class of the DR stages mentioned above.

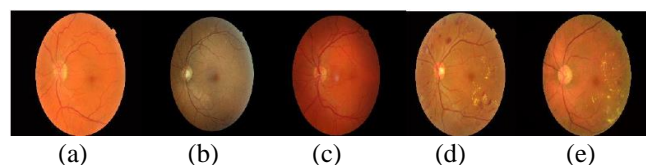


Fig. 2. Classification of DR stages

(a). No_DR (b). Mild (c). Moderate (d). Severe (e). Proliferate DR

When focusing on the detection of DR, there are several methods used. The existing architectures of CNN such as VGG16 [6], InceptionNetV3 [7], and AlexNet [8] can be cited as examples. Many Convolutional Neural Network [9 – 13] models are developed to achieve a successful classification. To begin the treatments for DR, it is crucial to diagnose and classify its stages. Therefore, it is a complex task for the Ophthalmologists to diagnose and classify DR as per the stages since the manual feature extraction is a time-consuming and less accurate process. Moreover, it requires expert skills. Thus detailing the fundus retinal images with computer-aided systems paves the way to an effective and accurate improved methodology rather than manual performance.

The objective of this study is to address the classification of the stages of Diabetic Retinopathy images with the use of a Hierarchical Convolutional Neural Network technique which initially classify the DR and No-DR images then the classified DR images will be classified as Pre_Intermediate and Post_Intermediate. Moreover, those two Pre_Intermediate and Post_Intermediate classes were again separately classified into Mild, Moderate, and Proliferate_DR, Severe, respectively.

The rest of the paper is ordered as follows. In Section II, different techniques that are related to DR classification are summarised. The background of this work is explained in section III. In Section IV, the proposed methodology is described in detail. Section V contains the experimental setup and the testing results obtained. Finally, Section VI is allocated for the conclusion and future extensions.

II. PREVIOUS WORK

In [9], an automated diagnosis system was developed to recognise retinal blood vessels, and a multi-class classification of DR was carried out. Green channel extraction and contrast limited adaptive histogram equalisation (CLAHE) were carried out as the preprocessing techniques. After preprocessing the images, feature selection was done followed by feature extraction. Finally, the images were classified using the Support Vector Machine (SVM) classifier. Two publically available datasets were used for this work. DIARETDB1 with 130 images where 42 mages for training and 88 images for testing and DIARETDB0 with 89 images where 28 images for training and 61 images for testing were used. The method proposed here obtained an accuracy of 93.6% and a sensitivity of 90.6% for all 219 images. It would be clearer if they could include the size of the used images in this paper.

In [10], detection of blood vessels, identification of the haemorrhages, and classification of DR into three classes were the main objectives taken into consideration. The images were classified as normal, moderate, and non-proliferate DR. 65 images of normal (30), moderate (23), and non-proliferate DR (NPDR) (12) were used from the STARE dataset with the dimension of 576×768. Green channel extraction and Adaptive histogram equalisation were used as the preprocessing techniques. A 3×3 median filter was operated to remove the noise. The matched filtered image was converted to binary equivalent with a global threshold value. Then binarization was carried out using a matrix. The images were then augmented. The classification was finally carried out using the Random

Forest technique based on the area and perimeter of the blood vessels and haemorrhages. The normal class with 20 training images and ten testing images achieved an accuracy of 90%. The moderate class with 15 training images and eight testing images achieved an accuracy of 87.5% and the severe NPDR class with four training images and eight testing images achieved an accuracy of 87.5%.

In [11], the authors have proposed a customised CNN architecture to classify diabetic retinopathy (DR) images. One thousand two hundred coloured fundus images were used from the Messidor dataset, where 840 are used for training images, and 360 are used for testing. Images were preprocessed by cropping to remove the black background and then resizing to 224×224, and the quality was adjusted using the histogram equalisation technique. Four CNN models were used where three were from pre-trained models such as AlexNet, VGG16, and SqueezeNet, and the remaining one was newly proposed. The performance of the classification of DR images of the newly proposed five-layered model was compared with the pre-trained models. In the proposed model, four separate kernels with size 3×3 were convolved in the first layer to extract features. Also, the image was zero-padded along by two. A pooling layer was also included in the first layer, and this layer reduces the calculations of the convolution layer and optimizes the time. The five-layered model produces a sensitivity of 98.94%, specificity of 97.87 %, and accuracy of 98.15%. It would be more effective if they could clarify the number of classes to which the images belonged and could use a higher number of images for testing and training.

In [12], the authors have considered the InceptionV3 architecture to classify Diabetic Retinopathy (DR). The dataset was taken from the famous Kaggle dataset which contains 35126 images. A five-class DR classification was done by splitting the dataset as 80% for training and 20% for testing with the input size as 299×299. Random scaling, resizing and centre cropping was done as preprocessing. The proposed model consisted of Inception V3 architecture and pre-trained on ImageNet as it can accelerate the process of training and also Inception V3 has a better performance on ImageNet. The architecture of the proposed model consists of five layers: Convolutional 2D layer, batch normalization layer, pooling layer, concatenate layer, and fully connected layer. Stochastic gradient descent (SGD) was used as the optimizer. Data augmentation was used with an early stop for 15 iterations to overcome the overfitting. Finally, the system was evaluated using 7023 test images. The system had achieved remarkable performance with an accuracy of 80% and a kappa score of 0.64.

In [13], the authors had employed a group of Convolutional Neural Networks (CNN) as a stage classification of Diabetic Retinopathy (DR). A fine-tuned three architectures; AlexNet, VGG16, and InceptionNet V3 were used to train the images. A total of 166 images from the Kaggle dataset were chosen to train the models. A five-class classification was done in this work. The images in the dataset were resized to pixels of 227×227, 224×224, and 299×299 for AlexNet, VGG16, and InceptionNet V3 respectively. The models AlexNet, VGG16, and InceptionNet V3 gained significant accuracies of 37.43%, 50.03%, and 63.23% for the dataset respectively. Higher rates of the accuracy of results have been achieved by the

InceptionNet V3 architecture. It would be effective if the authors could use a higher number of images to train and test these models.

III. BACKGROUND

A. Diabetic retinopathy

Diabetic Retinopathy is a related disease that is derived from Diabetes. The damage of the small blood vessels of the retina is the leading cause of it. Moreover, retinal blood vessels break down, leak or block. It affects the transportation of oxygen and nutrients inside the retina, causing vision loss over time. The presence of blockages, growth of abnormal blood vessels on the retinal surface increases the probability of bleeding leakages. These will result in vision blurring to vision loss over time.

B. Machine learning

Machine learning is a subfield of artificial intelligence where computers were made to learn from the data fed to them. It gives computers the ability to digest more data and reprogram themselves to execute a particular task with increasing precision. Then machines learn to perform a task more accurately through trials and errors. Machine learning usually uses several algorithms along with different tools to improve the prediction of desired outcomes [14]. Machine learning can be classified as supervised, unsupervised, and reinforced based on the algorithm it implements [15].

C. Convolutional Neural Network (CNN)

The neural network plays a major role in this report's work for the classification of Diabetic Retinopathy. Neural networks function similarly to the neurons in the human brain. It is important to note that all the neurons do not activate at once. Neurons are activated as per the signals received to carry out a particular task inside the body. This phenomenon is exactly used as neural networking in deep learning. CNN is formed of a set of layers that are stacked together. Each layer in the architecture owns a convolutional operator. Usually, a neural network inputs data process them with multiple neurons, and then outputs the results through an output layer [16]. Feature extraction and a fully connected layer are the two main parts of a basic CNN architecture. The convolution tool used to separate and identify the various features is known as the feature extraction, and the fully connected layer predicts the classes of the images using the features extracted in the previous layers.

IV. METHODOLOGY

The proposed tree structure-based binary classifications of DR are illustrated in Figure 3.

A. Preprocessing

Foremost in the experiment, the green channel was extracted from the procured images after centre cropping them to the size of 140×205. Then those images were subjected to Contrast Limited Adaptive Histogram Equalization (CLAHE). Figure 4 shows the preprocessed image samples.

B. Data augmentation

After obtaining the green channel and CLAHE processed images, those were subjected to data augmentation. The basic parameters used in this augmentation are flipping left, flipping right and rotation of 180° as shown in Figure 5. The other data augmenting parameters like shearing and zooming were not used since they did not have much impact on feature identification. The augmented images were saved separately and then were fed to the model. Data augmentation played a major role in extending the dataset to 49000 images. Datasets of images around 35000 were mostly found in the existing research works and that paved the way to derive an image set of 49000 images for this proposed work.

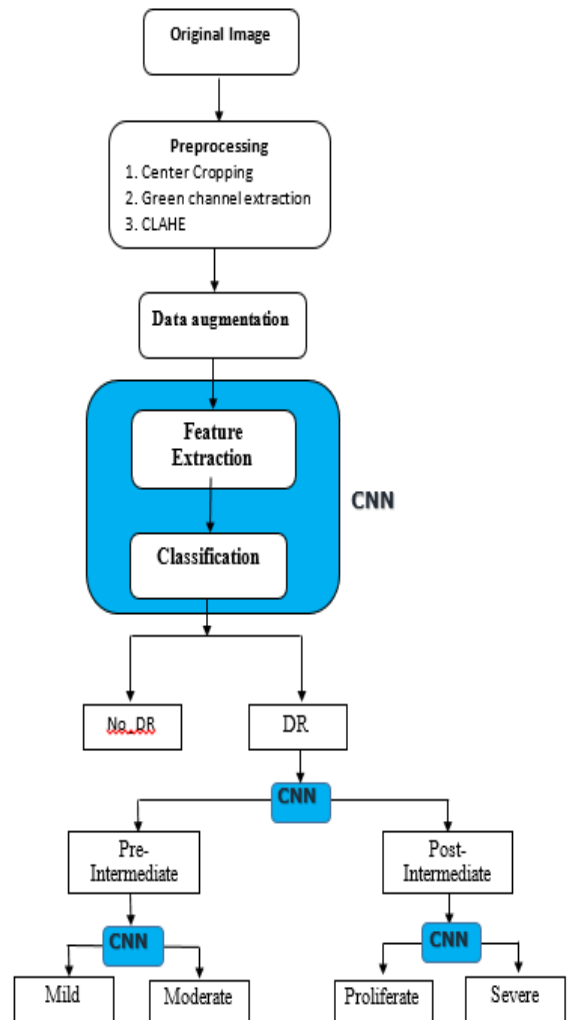


Fig. 3. Proposed methodology



Fig. 4. Sample of RGB and pre-processed images



Fig. 5. Samples of data augmentation

C. Proposed CNN Architecture

The proposed CNN model consists of a 14 layered architecture as shown in Figure 6. It contains four Convolution 2D layers of the same format, each followed by a max-pooling layer. Then a flatten layer is present. Next, there are two dense layers, followed by a dropout layer for each. The Softmax classification layer is present at last. The learning rate of 0.01 was used for each convolution layer due to the use of more epochs while training.

When moving deep inside the layers, the first two convolutional 2D layers are of kernel size (3,3) with a sum of 16 filters per layer. The padding 'same' is used here to receive the output with equal dimensions as the input. The ReLU activation function is used to overcome the gradient vanishing problem. The default stride (1,1) is used in addition to the above-mentioned. Each layer is followed by a max pooling layer with default values. The third Convolutional 2D layer is of kernel size (3,3) with 32 filters. The padding 'same' is used with the default stride (1,1) and the ReLU activation function is used to activate the neurons [17]. A max pooling layer is followed by this layer.

The fourth convolutional 2D layer is of kernel size (3,3) and 64 filters are available. The padding 'same' is used in

this layer as well with the ReLU activation function. A Max pooling layer is followed as stated above. Then a flatten layer is used to convert the data which comes from the above layers into a one-dimensional array for inputting it to the next layer. Next, there are two Dense hidden layers each followed by a Dropout layer (0.5). The two Dense hidden layers consist of 128 units per layer with a ReLU activation function. The final layer is the classification layer with the number of classes considered for the classification and Softmax as the activation function. The model was compiled with 50 epochs, a batch size of 32, a learning rate of 0.01, and "Adam" as the optimizer.

V. EXPERIMENTAL SETUP

This section provides a brief description of the training and testing images, and the experimental setup of Diabetic retinopathy classification with the obtained testing results.

A. Dataset

The dataset was used in this work from the Kaggle dataset repository [18] which was illustrated in Figure 7. There are a number of datasets available for diabetic retinopathy in Kaggle. The dataset which was used for this piece of work consists of 35126 fundus images. These images were of size 224×224 and were centre cropped to 140×205 to remove the black background. The objectives of cropping the images were to remove the black background as much as possible while preserving the majority of the retinal vessels. The number of images in the original dataset is given in Table I. Data augmentation is used to increase the number of images in each level of classification.

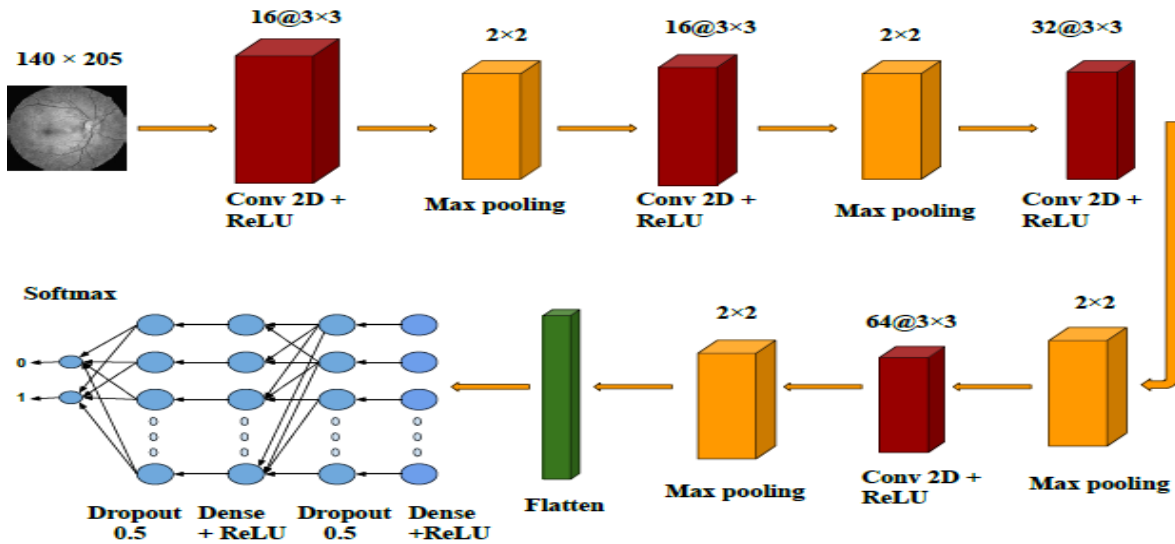


Fig. 6. Visualisation of the proposed CNN architecture

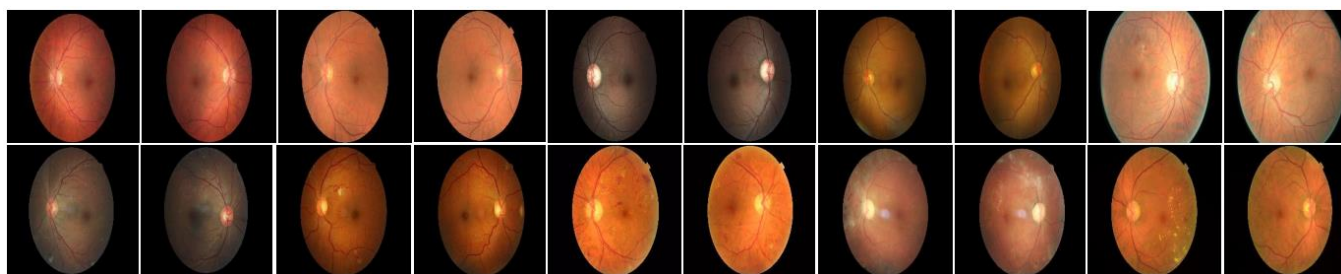


Fig. 7. Some sample images of the dataset

TABLE I. THE ORIGINAL DATASET IS IN DETAIL

Class ID	Class Name	Number of images
0	No_DR	25810
1	Mild	2443
2	Moderate	5292
3	Severe	873
4	Proliferate_DR	708

TABLE II. AVERAGE RESULTS OF THE CLASSIFICATION WITH 50 EPOCHS

Classification Level	Avg. Precision	Avg. Recall	Avg. F1-score	Train Accuracy	Test Accuracy
Level 1	0.65	0.64	0.63	0.9574	0.8111
Level 2	0.70	0.58	0.50	0.9896	0.9571
Level 3(A)	0.66	0.64	0.62	0.9764	0.8396
Level 3(B)	0.96	0.96	0.96	0.9957	0.9737

B. Tree based classification

Here, the results for the continued binary classifications were obtained. The Level 1 classification started with an image set of 49000 images and the Level 2 started with 24500 images per class. Finally, both Level 3(A) and Level 3(B) started with 12250 images per class. The order of the classification and results are displayed in Figure 8 and Figure 9, respectively.

C. Testing results

The model was trained and tested with images on the basis of 80% for training and 20% for testing. Accuracy, Precision, Recall and F1-score were also employed by obtaining the results in this work. We report the particular equations for the above parameters as follows:

$$\text{Accuracy} = (\text{TP} + \text{FP}) / \text{Total} \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{F1 score} = 2 \times (\text{Recall} \times \text{precision}) / (\text{Recall} + \text{Precision}) \quad (4)$$

Where, TP - true positive, FP - false positive, TN - true negative, FN - a false negative. The average results of the classification for 50 epochs are reported in Table II.

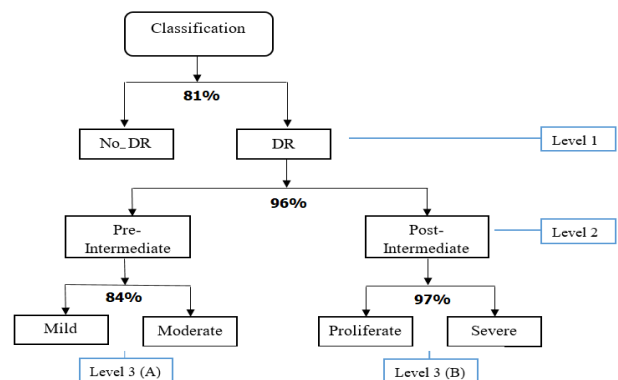
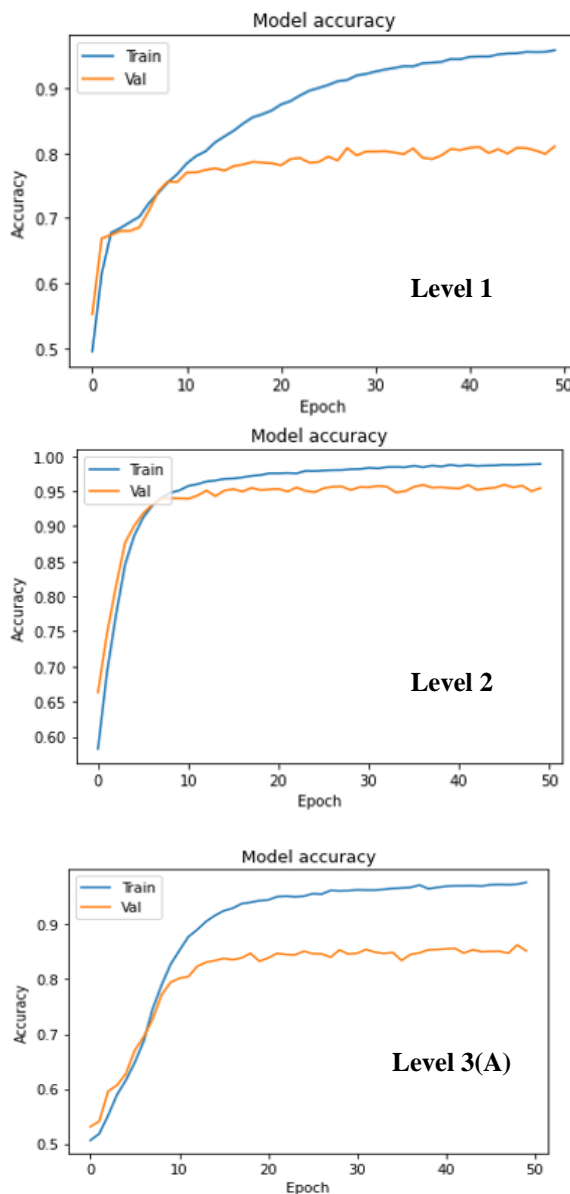


Fig. 8. Levels of the classification



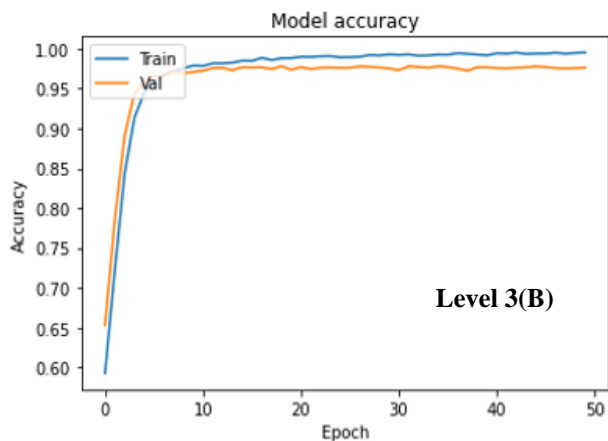


Fig. 9. Model accuracy for each level against epochs

All the experiments were carried out using a Virtual Machine (VM) from Microsoft Azure [19].

VI. CONCLUSION

In this piece of work, we have illustrated a proposed CNN architecture to classify Diabetic Retinopathy stages with a novel classification tree-based structure that continues with binary classifications. Moreover, the use of preprocessing techniques, Green channel extraction, CLAHE, and Data augmentation played a major role in achieving better accuracies. Centre cropping of all the images to the specified dimensions made it easy to remove the black background of the fundus images as much as possible. It was found out that the removal of the eye borders does not affect the feature extraction since a majority of the features are extracted from the retinal vessels present. The selection of the VM on training the models made a huge impact on gaining more accuracy. Hence, it can be concluded that this study which we proposed has been able to propose a model for the classification of Diabetic Retinopathy and has achieved worthy results for the novel classification approaches. While concluding the achieved results from this piece of work, it was able to achieve the particular accuracies of 81% for level 1, 96% for level 2, 84% for level 3(A), and 97% for level 3(B) on the proposed model. Deep learning approaches provide better results than geometrical approaches [20] of medical images. The expected future work of this particular study is to be stretched to enhance this model with a novel idea of classification and compare it with the bag-of-features approach.

REFERENCES

[1] S. Vujosevic et al., "Screening for diabetic retinopathy: new perspectives and challenges", *The Lancet Diabetes & Endocrinology*, vol. 8, no. 4, pp. 337-347, 2020.
 [2] L. Wu, "Classification of diabetic retinopathy and diabetic macular edema", *World Journal of Diabetes*, vol. 4, no. 6, p. 290, 2013.
 [3] W. Wang and A. Lo, "Diabetic Retinopathy: Pathophysiology and Treatments", *International Journal of Molecular Sciences*, vol. 19, no. 6, p. 1816, 2018.
 [4] V. Mayya, S. Kamath S. and U. Kulkarni, "Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A Comprehensive review", *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100013, 2021.
 [5] A. Mahdjoubi, Y. Bousnina, G. Barrande, F. Bensmaine, S. Chahed and A. Ghezzaz, "Features of cotton wool spots in diabetic retinopathy: a spectral-domain optical coherence

tomography angiography study", *International Ophthalmology*, vol. 40, no. 7, pp. 1625-1640, 2020.
 [6] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images", *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, p. 9420, 2019.
 [7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818-2826, 2016.
 [8] Z. Yuan and J. Zhang, "Feature extraction and image retrieval based on AlexNet", *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, 2016.
 [9] P. Adarsh and D. Jeyakumari, "Multiclass SVM-based automated diagnosis of diabetic retinopathy", *International Conference on Communication and Signal Processing*, 2013.
 [10] K. Verma, P. Deep and A. Ramakrishnan, "Detection and classification of diabetic retinopathy using retinal images", *Annual IEEE India Conference*, 2011.
 [11] Mobeen-ur-Rehman, S. Khan, Z. Abbas and S. Danish Rizvi, "Classification of Diabetic Retinopathy Images Based on Customised CNN Architecture", *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 244-248, 2019.
 [12] H. Chen, X. Zeng, Y. Luo and W. Ye, "Detection of Diabetic Retinopathy using Deep Neural Network", *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pp. 1-5, 2018.
 [13] A. Samanta, A. Saha, S. Satapathy, S. Fernandes and Y. Zhang, "Automated detection of diabetic retinopathy using convolutional neural networks on a small dataset", *Pattern Recognition Letters*, vol. 135, pp. 293-298, 2020.
 [14] O. Simeone, "A Very Brief Introduction to Machine Learning with Applications to Communication Systems", *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648-664, 2018.
 [15] M. Kang and N. Jameson, "Machine Learning: Fundamentals", *Prognostics and Health Management of Electronics*, pp. 85-109, 2018.
 [16] S. Albawi, T. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network", *International Conference on Engineering and Technology (ICET)*, 2017.
 [17] N. Gupta, P. Bedi and V. Jindal, "Effect of Activation Functions on the Performance of Deep Learning Algorithms for Network Intrusion Detection Systems", *Proceedings of ICETIT*, pp. 949-960, 2019.
 [18] Dataset: <https://www.kaggle.com/sovittrath/diabetic-retinopathy-2015-data-colored-resized>.
 [19] "Data Science Virtual Machines | Microsoft Azure", *Azure.microsoft.com*, 2021. [Online]. Available: <https://azure.microsoft.com/en-us/services/virtual-machines/data-science-virtual-machines/>. [Accessed: 25- Jun-2021].
 [20] D. V. D. S. Abeysinghe and S. Sotheeswaran, "Novel computational approaches for border irregularity prediction to detect melanoma in skin lesions," *International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 216-222, 2020, doi: 10.1109/SCSE49731.2020.9313042

Exploiting optimum acoustic features in COVID-19 individual's breathing sounds

M. G. Manisha Milani*
Faculty of Integrated Technologies
Universiti Brunei Darussalam, Brunei
manishamilani@gmail.com

Krishani Murugiah
Department of Biotechnology, Pavendar Bharathidasan College
of Engineering and Technology, India
krishanimurugiah@gmail.com

Murugaiya Ramashini
Department of Computer Science and Informatics
Uva Wellassa University, Sri Lanka
ramashini@uwu.c.lk

Lanka Geeganage Shamaan Chamal
School of Engineering
Sri Lanka Technological Campus, Sri Lanka
shamaang@sltc.ac.lk

Abstract - The world is facing an extreme crisis due to the COVID-19 pandemic. The COVID-19 virus interrupts the world's economy and social factors; thus, many countries fall into poverty. Also, they lack expertise in this field and could not make an effort to perform the necessary polymerase chain reaction (PCR) or other expensive laboratory tests. Therefore, it is important to find an alternative solution to the early prediction of COVID-19 infected persons with a low-cost method. The objective of this study is to detect COVID-19 infected individuals through their breathing sounds. To perform this task, twenty-two (22) acoustic features are extracted. The optimum features in each COVID-19 infected breathing sound is identified among these features through a feature engineering method. This proposed feature engineering method is a hybrid model that includes; statistical feature evaluation, PCA, and k-mean clustering techniques. The final results of this proposed Optimum Acoustic Feature Engineering (OAFE) model show that breathing sound signals' Kurtosis feature is more effective in distinguishing COVID-19 infected individuals from healthy individuals.

Keywords - acoustic features, COVID-19 breathing sounds, feature engineering, k-mean, PCA

I. INTRODUCTION

The word COVID-19 became familiar among every individual worldwide due to its adverse impact on daily routine life [1]. The first case is reported in a patient with severe respiratory syndrome with cough, fever and dizziness at Wuhan hospital in China [2]. The lung is the primary respiratory organ affected by this virus [3]. Lung auscultation is a method that plays a vital role in examining respiratory disorders by distinguishing normal respiratory sounds from abnormal sounds [4]. Abnormal breathing sounds are common in society, such as; bronchial breathing, stridor, wheeze, rhonchus, cackles, and pleural friction rub. The breathing sounds of patients with COVID-19 can be examined via lung auscultation methods [3].

The breathing sound waveforms of both COVID-19 infected individuals and healthy individuals are illustrated in Fig. 1. The normalised amplitudes of breathing sound signals are plotted against time. However, all characteristic differences and similarities may not be visualised via a waveform plot. Thus, further calibrations need to be done to identify significant signal characteristics to differentiate COVID-19 and healthy individual's breathing sounds.

The rest of the paper follows; Section II gives a background inspection and a literature review on audio signal processing applications to detect COVID-19 breathing sounds. Section III presents the proposed

methodology, while Section IV presents the obtained results. The conclusions of this study are presented at the end of the paper

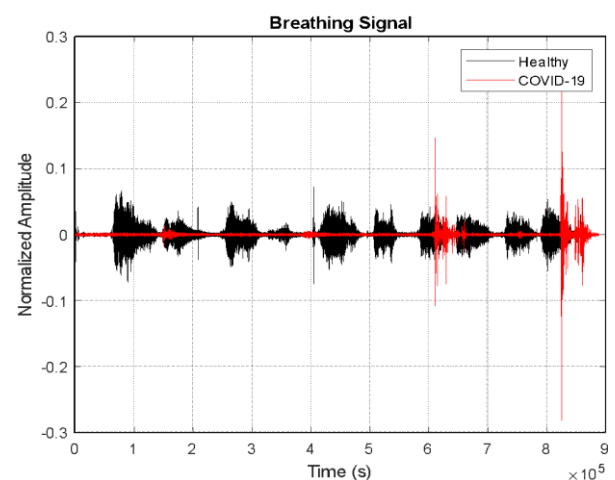


Fig. 1. COVID-19 and healthy breathing sounds in the time domain

II. LITERATURE REVIEW

Breathing is a chemical and mechanical process that includes inhaling and exhaling. In this process, Oxygen is inhaled in to the body, while Carbon Dioxide is exhaled [5]. Breathing is an essential process for all living creatures, including humans, because it impacts the whole body to regulate the functionalities of the organs. There are pathologies such as; asthma, pneumonia, and Chronic Obstructive Pulmonary Disease (COPD) that affect the breathing process [6]. Many of the pathologies undercover severe health problems that need proper treatment. Among many of these pathologies, the main problem facing the present society is detecting COVID-19 virus-infected persons. Thus, it is stated that the breathing process is the primary mode of transmission of the virus into the human respiratory system [7].

Many applications have already been presented for early diagnosis of various disorders that occur in different organs of the human body, mainly in the heart, brain, kidney, and lungs. Sound-based disorder identification techniques started to be experimented several years ago; thus, plenty of medical equipment was invented to hear and analyse these sounds of the human organs. The most significant sound analysis module is the stethoscope, which tends to listen to the inner sounds of hearts and lungs, including; murmurs, heart sounds, and breathing sounds. In

the modern world, Artificial Intelligence (AI) is a popular engineering concept; hence many of these equipments are developed to perform various applications, including; developments of smartphone apps, telemedicine, medical and surgery tools. Acoustic sounds would be critical data for future developments of these applications to identify COVID-19 patients.

Among many applications, audio-based smartphone applications are widespread in research studies to detect COVID-19 patients. For example, Stasak et al. [8] proposed a smartphone-based speech analysis application to detect pathological effects relevant to COVID-19 screening. Similarly, Imran et al. [9] proposed an AI-based smartphone app to detect COVID-19 infected people through their cough sounds. Breathing sounds are also integrated to screen COVID-19 infected people via smartphone applications. In their study, Faezipour et al. [10] proposed an idea to develop a smartphone-based breathing sound simulation app that can self-test a person's breathing patterns and identify his/her breathing complications. The idea of this app is specifically proposed to detect COVID-19 patients. Despite these smartphone-based applications, Huang et al. [3] recorded breathing sounds via an electronic stethoscope and sent these recordings to a computer-based signal analysis method. They used a time-frequency distribution of the waveforms of both COVID-19 virus-infected and healthy individuals to examine the characteristics in the signal patterns. Then these visualising results are compared with clinically proven data to differentiate COVID-19 and healthy people. Apart from identifying COVID-19 infected people through breathing sounds, a few more applications were developed to diagnose other breathing disorders. Yañez et al. [11] proposed a breathing rate monitoring system to use at home. This system allows early prediction of exacerbation of Chronic Obstructive Pulmonary Disease (COPD).

Audio processing is a fast-growing method in medical diagnosis to categorise the most effective acoustic features. Many studies are conducted to find the best feature selection of the audio signals generated by the human body. For example, Milani et al. [12] examined both frequency and time domain acoustic features to identify normal and abnormal heart sounds. Nagasubramanian et al. [13] analysed multivariate vocal sounds and acoustic features with deep learning techniques to predict Parkinson disease. Chambres et al. [14] used mel-frequency cepstral coefficients (MFCC) of lung sounds to detect individuals with respiratory diseases. However, many research studies are conducted at the present day with a scope of early diagnosis of COVID-19 virus-infected people; but, many of these studies are still at the proposal stage. Therefore, in the near future, there could be successful outcomes from them. Nevertheless, this study would focus on identifying the most effective acoustic features to detach the breathing sounds of COVID-19 individuals and healthy individuals. Therefore, the findings of this study shall be proposed to apply in the future and ongoing COVID-19 breathing sound analysis application to invent and develop technical solutions for the COVID-19 pandemic.

III. PROPOSED METHODOLOGY

An acoustic feature-based clustering method which shows in Fig. 2, is proposed in this study. This proposed methodology carries four (4) stages; data collection, signal

pre-processing, acoustic feature extraction, and optimum acoustic feature engineering. These four (4) stages conduct a particular task to obtain accurate outcomes in identifying COVID-19 and healthy individuals through their breathing sounds.

A. Data collection

The breathing sounds of COVID-19 and healthy individuals are collected from the Coswara open-access database [15], which contains various respiratory sounds, including; breath, cough, and voice [16]. However, only the breathing sounds are considered from the Coswara database to achieve the objective of this study. First, the sound quality is inspected manually before selecting the input sound recordings to the proposed methodology. All the sound recordings which are manually inspected (both visual and listening inspections) shows the sounds are incredibly in good condition. The sounds in the recordings are clear, and fewer background noises. A total number of forty (40) breathings sounds are taken for the training purpose, including twenty (20) sounds of each COVID-19 and healthy individuals. Then an additional four (4) sound recordings are selected to test the trained model. These four (4) recordings include; two (2) from COVID-19 and the remaining two (2) from healthy individuals, but they are considered as unknown in the testing process.

B. Signal pre-processing

The proposed signal pre-processing stage includes noise reduction and enveloping of the selected breathing sound signals. Breathing sounds can be considered as soft and low-pitched audio signals. A Finite Impulse Response (FIR) filter may be a better signal filtering solution to reduce the background noises and stabilise the signal [17]. These background noises may include the different sounds that are produced from internal organs of the body and other disturbances that occur during the sound recording process.

The filtered signal is then windowed with Hamming windowing method. The Hamming window has a fixed window function that can cancel the nearest side lobe of signals. Compared to other windowing methods, i.e. Hanning windowing, the performance speed and the noise cancellation is better in the Hamming windowing method [18].

In this study, each window of the filtered signal is designed for 30ms with a 10ms overlap. The proposed Hamming window is defined by:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right) \quad (1)$$

where ' n ' is the input sample number and ' N ' is the total number of input samples [19]. Hence, this windowed signal will address the discontinuity of the actual breathing sounds by giving a smooth and soft waveform to obtain more reliable information from its features.

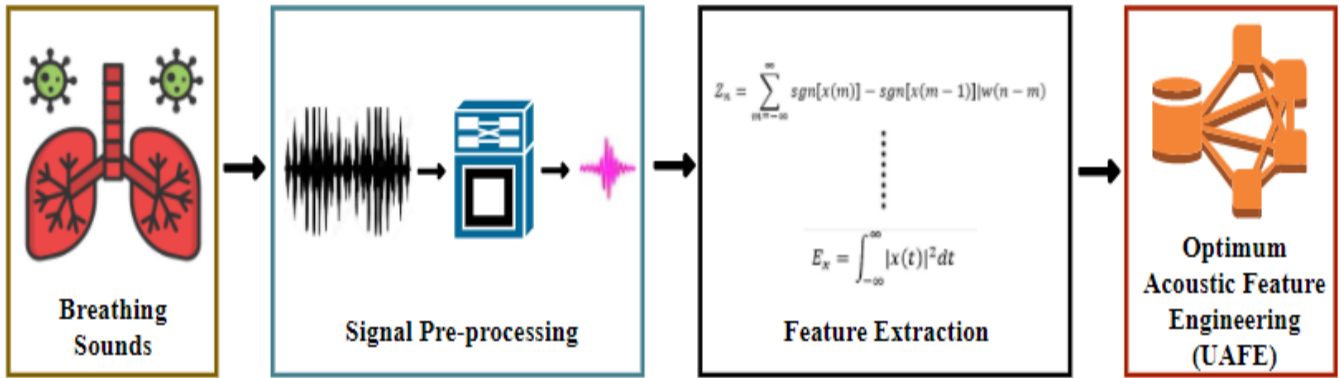


Fig. 2. Proposed methodology

C. Acoustic feature extraction

The information of each window of the breathing sound signals is extracted from the features of temporal, spectral, and frequency domains. These features are commonly used in different audio processing applications and provided acceptable results [20]. Twenty-two (22) multi-dimensional features are extracted from the selected three (3) domains. These extracted features may contain all key properties of the COVID-19 and healthy individual’s breathing sounds. A summary of extracted features is dispatched in Table I.

TABLE I. EXTRACTED FEATURES

Feature Domain	Name of the Features	Nor of Features Extracted
Temporal	Zero-Crossing Rate, Energy, Entropy of energy	3
Spectral	Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Roll-off, Chroma Vectors	17
Frequency	Harmonic Ratio, Fundamental Period	2
Total number of features extracted		22

D. Optimum Acoustic Feature Engineering

A novel feature engineering-based learning algorithm is proposed to achieve the stated objective of this study. The proposed Optimum Acoustic Feature Engineering (OAFE) method requires only the extracted features of each selected input class, such as; COVID-19 individuals and healthy individuals. This proposed OAFE method may directly influence the final data prediction; thus, it may provide better and most influential acoustic features from the extracted twenty-two (22) features. Hence, this method will be an effective solution to avoid misleading features.

The statistical features such as; mean, standard deviation, variance, Skewness and Kurtosis are considered as the inter-dependent properties of each extracted twenty-two (22) acoustic features. These statistical features may emphasise the inherent nature of the extracted features to achieve better clustering performance with higher accuracy.

However, these features are in a multi-dimensional space which may make the final clustering process uneasy. Therefore, a feature dimensional reduction is conducted via Principal Component Analysis (PCA) to remove redundant

features and keep the features in a low-dimensional space. Then, the final feature prediction is conducted through an unsupervised k-mean clustering algorithm to predict the class-based clustering to identify both COVID-19 and healthy individuals separately. The proposed OAFE process is illustrated in Fig. 3.

As of Fig. 3, the five (5) statistical features are calculated for each column of the input features matrix X . Thus, the columns represent twenty-two (22) extracted features, while rows of the matrix represent the number of input breathing sounds. Then the output feature matrix will become as $XStat$, which contains twenty-two (22) features as columns, while five (5) computed statistical features as rows. However, at the PCA dimensionality reduction stage, the feature matrix $XStat$ is turned $XPCA$ by having the first three (3) PCA values as columns and five (5) statistical features as rows. The reason for selecting only the first three (3) PCA values is because the PCA orders the eigenvectors in decreasing order, while the first three (3) PCAs may have a high impact on the feature clustering process.

To identify the optimum acoustic feature for the application of COVID-19 and healthy individual’s breathing sound identification, the feature matrix $XPCA$ is transposed and sent to the proposed k-mean clustering algorithm. Through the k-mean algorithm, the computed statistical features of a total of twenty-two (22) extracted acoustic features are ranked as highest influenced feature to lowest influence feature. Hence, the firmness of these features depends on their ranks. Therefore, the best influential features are considered as an optimum feature for the stated objective of this study.

IV. RESULTS AND DISCUSSION

The performance of the proposed OAFE method is evaluated using four (4) unknown breathing sound recordings. Before inputting these unknown breathing sounds, the training performance of the computed $XPCA$ feature matrix is assessed via computing its accuracy. Hence, the proposed k-mean clustering model provided 80% of overall training accuracy for all forty (40) input sounds. After the training is done, the PCA-based feature matrices of unknown four (4) breathing sounds are fed into the training model. The final two class clustering outcomes of these four (4) breathing sounds are illustrated in Fig. 4.

It can be seen that all selected features in the transposed feature matrix of $XPCA$ of all four (4) unknown breathing sounds distinguish two clusters; COVID-19

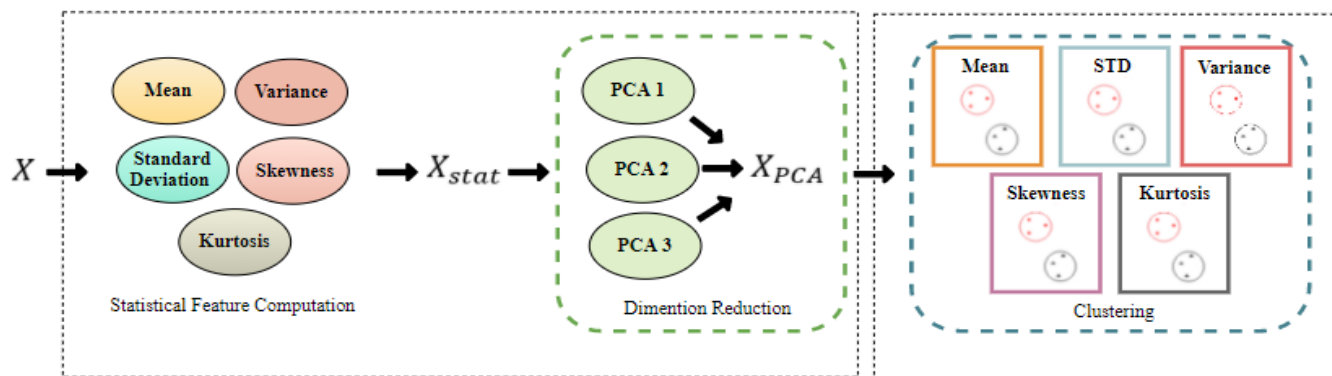
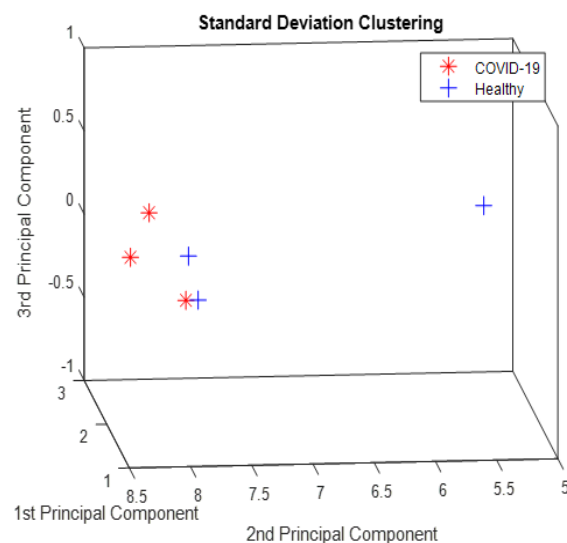


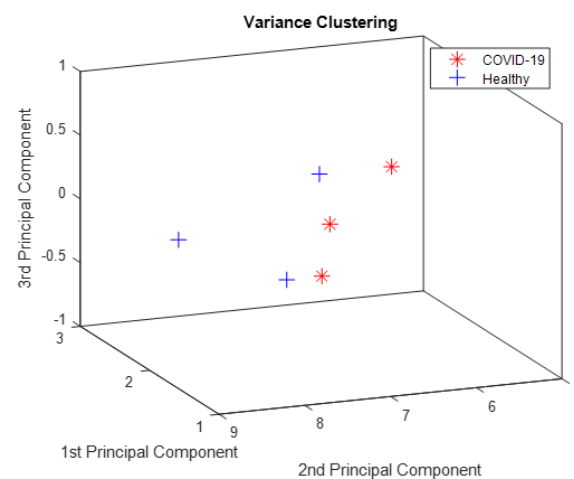
Fig. 3. Optimum acoustic feature engineering method

individuals and healthy individuals. However, PCA3 (third PCA value) in the Skewness predicted wrong. When further evaluating this wrongly predicted PCA, it is found that it belongs to a COVID-19 individual's breathing sound. Nevertheless, the overall performance of the executed statistical features of breathing sounds such as; mean, standard deviation, variance, Skewness, and Kurtosis indicated that these five (5) features extensively impact the stated purpose of this study.

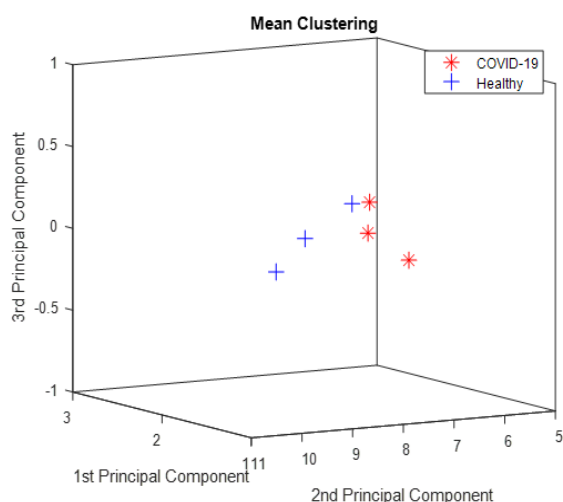
The traditional way of feature clustering for two or more classes is carried out by inputting a feature matrix (X) containing extracted features in high dimensional or low dimensional space with a number of input samples/signals. However, the novelty of the proposed OAFE method is to find the optimum feature or a set of features through the originally extracted features. Therefore, another set of features (in this study, five (5) statistical features) are computed from the original set of features to narrow down the most reliable information. Hence, the OAFE method does not contain the number of samples/signals as its input, yet it only contains features of the samples/signals in both rows and columns. In other words, this method considers each feature vector of the matrix X to calculate the five (5) statistical features. Thus, the combination of each feature vector creates a feature matrix containing; five (5) rows (statistical features) and twenty-two (22) columns (original features) before the dimensional reduction.



(b)



(c)



(a)

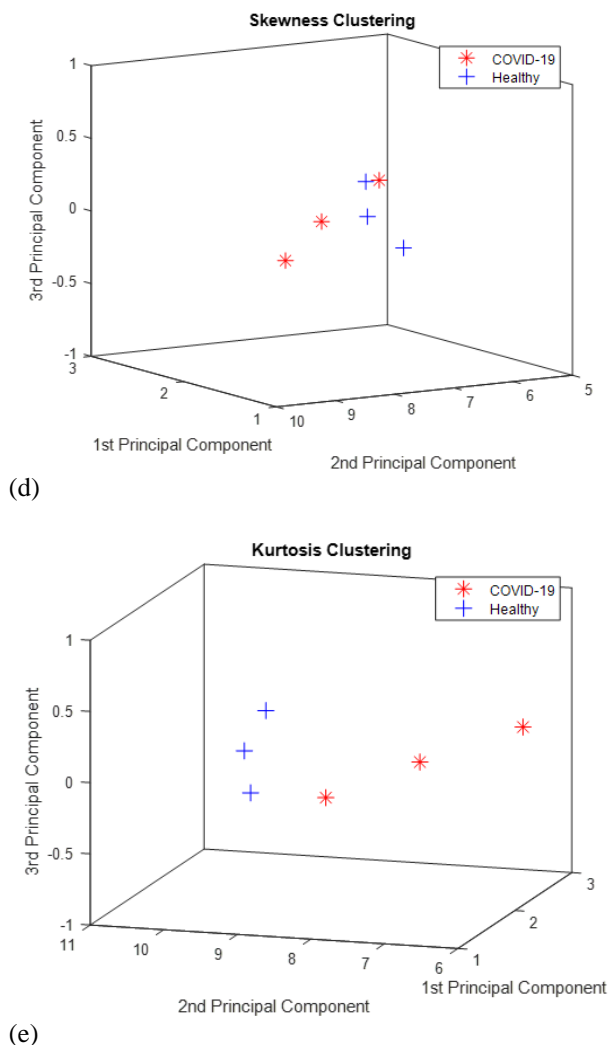


Fig. 2. Clustering results of both COVID-19 and healthy individual's breathing sounds: (a) Mean Clustering, (b) Standard Deviation Clustering, (c) Variance Clustering, (d) Skewness Clustering, (e) Kurtosis Clustering.

A cluster-based evaluation is implemented to find the most optimistic feature/features in the dimensionality reduced feature matrix *XPCA*. The results shown in Fig. 4 indicate that all five (5) statistical features effectively classify the breathing sounds of COVID-19 and healthy individuals.

TABLE II. OPTIMUM FEATURE RANKING

Ranking in Descending Order	PCA			Reason for Ranking
	1	2	3	
1) Kurtosis	✓	✓	✓	Distances between cluster points are longer.
2) Variance	✓	✓	✓	Distances between cluster points are less than Kurtosis.
3) Mean	✓	✓	✓	Distances between cluster points are close to each other, but the clusters can be clearly defined.
4) Standard Deviation	✓	✓	✓	Distances between PCA1 in both clusters are close to each other, yet the clustering is acceptable.
5) Skewness	✓	✓	✗	PCA3 of COVID-19 class clustered as a feature in healthy class. Thus, it is a wrong prediction.

The computed features are ranked in descending order and dispatched in Table II based on the Euclidean distance between the cluster points. The results indicate that the most relevant optimum feature vector is Kurtosis. The obtained test results are further verified via a mix and match method that mixed up all the breathing signals used in training and testing. Subsequently, the trained model is again tested for the PCA1 to PCA3 of *XPCA* matrix of randomly selected twenty (20) input breathing sounds. Remarkably, the final clustering observation of these features is identical to the results obtained for four (4) unknown breathing sound clustering results, which the Skewness predicted wrong for two (2) breathing sound signals of COVID-19 infected individuals. Hence, the results of Euclidean distance between the cluster points of this proposed mix and match method are identical to Table II. Therefore, it can be noted that the proposed method is well accurate to address the proposed issue of identifying the COVID-19 infected and healthy individual's breathing sounds.

V. CONCLUSION

Currently, the demand for an alternative PCR and other laboratory testing methods is higher to predict a COVID-19 positive individual in an early stage. This study displays a possible method to distinguish a COVID-19 individual from a healthy individual. The proposed method is based on the feature engineering technique examined via twenty-two (22) acoustic features. The proposed feature engineering model is a hybrid model that includes; model 1: computation of statistical features from original features and their dimension reduction, model 2: feature clustering. The novelty of this proposed feature engineering model is that it is altered from the traditional feature clustering method. The samples/signals are considered in the input feature matrix and the extracted features in the traditional engineering method. However, the proposed acoustic feature engineering method relies only on the features of each sample/signal.

The proposed feature engineering model examines; which feature is better to be used in any COVID-19 breathing sounds related application. The early stage of this hybrid feature engineering method computes five (5) statistical features such as; mean, standard deviation, variance, Skewness, and Kurtosis from all originally extracted twenty-two (22) acoustic features. This hybrid feature engineering model is named Optimum Acoustic Feature Engineering (OAFE), narrowing down the most effective statistical features of the original acoustic features. Among these five (5) statistical features, the most relevant feature/features are ranked in descending order. As of the obtained results, the most to the least compelling features are Kurtosis, variance, mean, standard deviation, and Skewness, respectively.

The proposed OAFE method with the signal pre-processing and feature extraction stages can be used in many practical applications such as; developing smartphone applications or hardware implementation to detect COVID-19 infected persons in real-time. However, this OAFE method will be further expanded by integrating more features like cepstral, wavelet and more to improve its performance. Also, the proposed clustering method can be made more robust by adding more training and testing data.

REFERENCES

- [1] Giovannini, H. Haick, and D. Garoli, "Detecting COVID-19 from Breath: A Game Changer for a Big Challenge," *ACS Sensors*, vol. 6, no. 4, pp. 1408–1417, 2021, doi: 10.1021/acssensors.1c00312.
- [2] Wu et al., "A new coronavirus associated with human respiratory disease in China," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020, doi: 10.1038/s41586-020-2008-3.
- [3] Y. Huang et al., "The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods," *medRxiv*, no. 628, 2020, doi: 10.1101/2020.04.07.20051060.
- [4] M. Sarkar, I. Madabhavi, N. Niranjana, and M. Dogra, "Auscultation of the respiratory system," *Annals of Thoracic Medicine*, vol. 10, no. 3. Medknow Publications, pp. 158–168, Jul. 01, 2015, doi: 10.4103/1817-1737.160831.
- [5] Sealy, "Every breath you take," *Materials Today*, vol. 7, no. 2. p. 1, 2004, doi: 10.1016/S1369-7021(04)00057-4.
- [6] Nazario, "Types of Lung Diseases & Their Causes," 2020. <https://www.webmd.com/lung/lung-diseases-overview> (accessed Jun. 20, 2021).
- [7] R. Ningthoujam, "COVID 19 can spread through breathing, talking, study estimates," *Curr. Med. Res. Pract.*, no. January, pp. 19–21, 2020, doi: 10.1016/j.cmrp.2020.05.003.
- [8] B. Stasak, Z. Huang, S. Razavi, D. Joachim, and J. Epps, "Automatic Detection of COVID-19 Based on Short-Duration Acoustic Smartphone Speech Analysis," *J. Healthc. Informatics Res.*, vol. 5, no. 2, pp. 201–217, 2021, doi: 10.1007/s41666-020-00090-4.
- [9] Imran et al., "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics Med. Unlocked*, vol. 20, p. 100378, 2020, doi: 10.1016/j.imu.2020.100378.
- [10] M. Faezipour and A. Abuzneid, "Smartphone-based self-testing of COVID-19 using breathing sounds," *Telemed. e-Health*, vol. 26, no. 10, pp. 1202–1205, 2020, doi: 10.1089/tmj.2020.0114.
- [11] M. Yañez et al., "Monitoring breathing rate at home allows early identification of COPD exacerbations," *Chest*, vol. 142, no. 6, pp. 1524–1529, 2012, doi: 10.1378/chest.11-2728.
- [12] M. G. M. Milani, P. E. Abas, L. C. De Silva, and N. D. Nanayakkara, "Abnormal heart sound classification using phonocardiography signals," *Smart Heal.*, vol. 21, p. 100194, 2021, doi: 10.1016/j.smhl.2021.100194.
- [13] Nagasubramanian and M. Sankayya, "Multi-Variate vocal data analysis for Detection of Parkinson disease using Deep Learning," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 4849–4864, 2021, doi: 10.1007/s00521-020-05233-7.
- [14] Chambres, P. Hanna, and M. Desainte-Catherine, "Automatic detection of patient with respiratory diseases using lung sound analysis," *Proc. - Int. Work. Content-Based Multimed. Index.*, vol. 2018-Sept., pp. 0–5, 2018, doi: 10.1109/CBMI.2018.8516489.
- [15] "Project Coswara | IISc." <https://coswara.iisc.ac.in/?locale=en-US> (accessed Jun. 20, 2021).
- [16] N. Sharma et al., "Coswara - A database of breathing, cough, and voice sounds for COVID-19 diagnosis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 4811–4815, 2020, doi: 10.21437/Interspeech.2020-2768.
- [17] M. G. M. Milani, P. E. Abas, and L. C. De Silva, "Identification of normal and abnormal heart sounds by prominent peak analysis," in *ACM International Conference Proceeding Series*, Sep. 2019, pp. 31–35, doi: 10.1145/3364908.3364924.
- [18] Pereles, "Window Functions in Spectrum Analyzers | Tektronix," Tektronix, 2013. <https://www.tek.com/blog/window-functions-spectrum-analyzers> (accessed Aug. 21, 2021).
- [19] Meddins, "The design of FIR filters," in *Introduction to Digital Signal Processing*, Elsevier, 2000, pp. 102–136.
- [20] M. Ramashini, P. E. Abas, U. Grafe, and L. C. De Silva, "Bird Sounds Classification Using Linear Discriminant Analysis," *ICRAIE 2019 - 4th Int. Conf. Work. Recent Adv. Innov. Eng. Thriving Technol.*, vol. 2019, no. November, pp. 27–29, 2019, doi: 10.1109/ICRAIE47735.2019.9037645.

A community-based hybrid blockchain architecture for the organic food supply chain

Thanushya Thanujan*
Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
thanushyal@esn.ac.lk

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

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

Abstract - This paper presents a novel blockchain architecture to incorporate community-level trust into the organic food supply chain by hybridizing Proof of Authority (PoA) and Federated Byzantine Agreement (FBA) consensus protocols. Community-level trust is an important aspect in the organic agriculture industry. Organic farming, in most parts of the world, happens in small scale farms where the farmers represent rural and less-privileged communities. Even though third-party certification systems exist for quality assurance in organic farming, due to many socio-economic reasons, participatory guarantee systems (PGS) have become a popular alternative among organic farmers and consumers. However, such participatory guarantee systems are still prone to frauds and have limitations in scalability as well. With the recent rise of blockchain technology, there is an emerging trend to adopt blockchain technology to enhance the credibility of organic food supply chains and mitigate the risk of fraudulent transactions. However, despite the popularity of participatory guarantee systems among organic farmer communities, the blockchain researchers have paid little attention to develop blockchain architectures by adopting the community-level trust into their consensus protocols. The hybrid consensus mechanism presented in this paper addresses that gap in existing blockchain research. Apart from discussing the details of the proposed blockchain architecture and the underlying consensus protocol, this paper also presents a qualitative analysis on the proposed architecture based on expert opinions.

Keywords - blockchain, community-level trust, Federated Byzantine Agreement, hybrid consensus mechanisms, proof of authority

I. INTRODUCTION

Consumer trust is an important aspect of organic farming. According to [1], consumer trust is a key prerequisite for establishing a market for credence goods, such as “green” products, especially when they are premium priced. Third-party certifications are commonly used to fulfill this need where a trusted organization accredits the quality of farming practices and the products of a particular farm. However, audits for such third-party certifications incur a significant cost for the farm being audited. Due to many reasons including the high cost of audit, third party certification is not a very trustworthy mechanism to ensure the credibility of organic food supply chains [2]. Alternatively, participatory guarantee systems (PGS) have become popular among organic farming communities, especially in rural areas since it helps avoid the entry barriers of third-party certification systems. According to [3], participatory guarantee systems are locally focused assurance systems that verify producers’ compliance to certain organic standards. PGS are based on active participation of stakeholders and are built on a

foundation of trust, social networks, and knowledge building and exchange [3].

According to [4], PGS are independent and decentralized systems of local communities that involve producers, consumers, students, professors, agronomists, etc. and the certification is based on a peer review conducted by the stakeholders through an annual visit to the farm. The key elements of this system are mentioned as participation, trust, transparency, learning process, horizontality, decentralization, formation of networks, local focus, and food security and sovereignty [4].

However, this community-based certification system has inherent limitations which hinders the market growth for organic products. According to [3], in practice, PGS are often run and administered by NGOs or farmer’s associations, with limited smallholder involvement, which could be seen as a major flaw in terms of trustworthiness. Moreover, whether this community-based certification system could grow beyond the local market while preserving its original characteristics remains doubtful in terms of scalability. As the organic food industry has a potential to grow beyond local markets, the question of how to ensure trust still remains largely unresolved. In other words, it is important to research on the ways and means of incorporating the stakeholder communities to the certification process while addressing the issues of trust and scalability when the market is growing beyond local boundaries.

In the recent past, many researchers have been interested in adopting blockchain technology to resolve the trust issue of food supply chains [5]. Blockchain refers to an emerging disruptive technology that enables the creation of decentralized information systems with immutable and trustworthy records of transactions. Blockchain-based systems in the domain of agriculture help provide a trustworthy link between farms and the external markets by keeping transaction records immutably in decentralized ledgers, thereby enabling the traceability of sequences of transactions pertaining to a particular lot of produce throughout its journey along the supply chain. Research on food supply chains mainly focus on ensuring the trustworthiness of the products, transparency of supply chain activities as well as the technicalities of the blockchain technology such as determining the most suitable architectures and consensus mechanisms which make the system scalable and secure [6]. Various traditional and hybrid consensus mechanisms have been proposed and tested in this context [7]. However, there is no evidence for a research that has attempted to incorporate the community’s consensus into the verification and

validation protocol (i.e., consensus mechanism or protocol) of a blockchain architecture in the organic food context.

This research addresses the issue of developing a highly scalable blockchain architecture for the organic food supply chain with a consensus mechanism that hybridizes the traditional Proof of Authority (PoA) protocol with the Federated Byzantine Agreement (FBA) protocol. The key hypothesis here is that the community, as in the case of PGS, is a powerful component in the process of ensuring the credibility of organic food supply chains and hence, needs to be incorporated into the verification and validation process. However, this needs to be done without bypassing the formal regulatory process of the territory where the supply chain is being operated. It is assumed that by hybridizing the PoA protocol with the FBA protocol it would be possible to create a consensus mechanism, which enables the incorporation of the community dimension into the verification and validation process, while adhering to the formal regulatory procedures imposed by the governing bodies. Hybridizing both these consensus aims to mitigate the scalability issues and enhance trustworthiness. PoA is proposed to empower the authorized persons to propose blocks. While the size of the network increases, FBA resolves the issues of scalability and latency. The hybrid blockchain architecture presented in this paper and its underlying consensus protocol is designed based on this assumption, after a thorough review of literature on existing consensus protocols as well as an interviewing process which involved different stakeholders of the organic food supply chain in Sri Lanka. The key objective of this paper is to present the details of the proposed blockchain architecture and also to have a discussion on the incorporation of community-level trust into the consensus mechanism pertaining to the organic food supply chain.

The rest of this paper is organized as follows. Section II presents a summary of the existing literature on blockchain-based systems in the organic food industry. Section III provides an overview of current consensus mechanisms. Section IV then introduces the proposed blockchain architecture and the hybrid consensus mechanism. Section V carries a concept review on the proposed architecture as a simple qualitative analysis and section VI provides conclusions and directions for future work.

II. LITERATURE REVIEW

Adoption of the blockchain technology in the organic and other agricultural supply chains has been a trending topic since the recent past. Such research pays attention to avoiding a range of issues in agricultural supply chains such as inefficiencies, safety concerns and scandals, using blockchain technology. In [8], a blockchain-based model for rice supply chain management (RSCM) is proposed for the Food Corporation of India, to avoid significant wastage of rice and enhance the operational efficiency. In [9], a framework is proposed to trace out the major issues in traditional rice supply chain management and deploy blockchain technology to resolve these issues. In another notable research, a blockchain-based architecture is proposed for the traceability and visibility in the soybean supply chain [10]. In that research, an Ethereum-based smart contract is implemented and tested to govern and ensure the proper interactions among key stakeholders in the soybean supply chain. To ensure a high level of

transparency and traceability, all the transactions are stored in the block chain's immutable ledger with links to a decentralized file system (IPFS). Another traceability intended blockchain-based application is presented in [11], which focuses on the berries supply chain, with evidence of the proof of concept with a pilot study. Moreover, a commercially important blockchain implementation is reported in 2017 where Walmart has successfully tested IBM's blockchain pilots for food provenance: pork in China and mangoes in America [12]. In that study, the challenges of implementing blockchain technology in the food supply chain and the opportunities for deploying blockchain solutions are also highlighted. Besides, an IoT-based blockchain architecture for enhanced transparency and traceability in food supply chains is proposed in [13].

Despite the undeniable benefits of blockchain, technical challenges and barriers to the adoption still remain. A study on the challenges and potential use of blockchain for assuring traceability and authenticity in the food supply chains is reported in [14] whereas another study on the challenges of adopting blockchain in food supply chains as well as a potential future direction by integrating blockchain with IoT is discussed in [15].

A few researches have been done on the adoption of blockchain technology in the organic food supply chains as well. [16] evaluates the application of blockchain technology to improve organic or fair-trade food traceability from "Farm to Fork" in light of European regulations with the intention of shedding light on the challenges in the organic food chain to overcome, the drivers for blockchain technology, and the challenges in current projects. The findings of the research highlights, among a few more, 1. optimizing chain partner collaboration and, 2. the selection of data to capture in the blockchain as key challenges. Furthermore, easy verification of certification data, accountability, improved risk management, insight into trade transactions, simplified data collection and exchange, and improved communication are highlighted as key benefits. Moreover, a prototype implementation of a blockchain-based system addressing the traceability issue in organic food supply chains is presented in [17].

III. OVERVIEW OF CONSENSUS MECHANISMS

Consensus mechanism or protocol plays a critical role in the implementation of a blockchain-based system. In other words, it can be considered as the backbone of blockchain technology. In literature, there are numerous consensus mechanisms reported, each with their own strengths and weaknesses [18]. As the applications' complexity grows, researchers have proposed hybrid consensus mechanisms where the features of traditional consensus mechanisms like Proof of Work (PoW) and Proof of Stake (PoS) are combined to have more advanced functionality. This section summarizes some existing literature on hybrid consensus protocols and introduces the two consensus protocols hybridized in this particular study to create a community-based blockchain architecture.

An improved hybrid consensus algorithm is proposed in [19], combining advantages of the Practical Byzantine Fault Tolerance (PBFT) algorithm and the POS algorithm. According to them, the proposed algorithm reduces the number of consensus nodes to a constant value by verifiable pseudorandom sortition and performs

transaction witness between nodes. The improved algorithm is tested and verified in terms of throughput, scalability, and latency. In [20], for incognito payments like tips, a hybrid consensus mechanism is proposed, which consists of a public and private blockchain. The public blockchain is based on the Federated Byzantine Agreement (FBA) consensus algorithm while BRAVO's private, incognito blockchain is based on an anonymizing Proof-of-Stake algorithm, which gives the end-users control on transaction speed, privacy, and cost. Furthermore, a hybrid consensus model (PSC-Bchain) composed of Proof of Credibility (PoC) and PoS consensus algorithms have been proposed in [21]. The PoS consensus is proposed as a means of saving energy. PoC is used to address the problem of coin collapse found in the PoS consensus method, and for credibility verification with the function of attack deterrence. Moreover, the model has combined a sharing mechanism with the proposed hybrid approach to emphasize security. The study has compared attack execution on both the classical blockchain and proposed hybrid blockchain, and also presented an attack analysis and security analysis. The experiment results have confirmed the enhanced scalability and performance of the blockchain-based e-voting system. Most of the existing studies on hybrid consensus mechanisms have focused on enhancing the security and scalability challenges. Notably, there is very little research in the agriculture domain, if not none, reported to have studied the adoptability of hybrid consensus mechanisms in their blockchain architectures. Given the nature of the problem being investigated, this study proposes to hybridize the PoA and FBA consensus protocols. The selection of these two protocols is based on a thorough desk review of existing consensus mechanisms [18] pertaining to the problem being investigated.

A. Proof of Authority (PoA)

The concept of Proof of Authority (PoA) was coined in 2015 by Gavin Wood, co-founder of Ethereum and Parity Technologies. Later in 2017, a solution to spam attacks on Ethereum's Ropsten test network using PoA was proposed [22]. Recently, the PoA protocol was adopted by commercial platforms such as Microsoft Azure, Ethereum Express, POA Network and VeChain [23]. PoA is considered a modified mechanism of Proof of Stake (PoS), which leverages the identity as a form of stake instead of a wealth (Ex. crypto tokens). Unlike Proof of Work (PoW), PoA eliminates the need of high computational power to validate a block. The core of this consensus is to empower the pre-authorized persons to create a new block of transactions by considering their individual identity as a stake. In other words, the block creator in PoA protocol puts his or her authority at stake when authorizing a transaction into the block. This acts as the key control mechanism to eliminate fraudulent transactions from the network.

Even though PoA is adopted by some public block chains, it still lacks the full decentralization. The validator should be an identifiable participant and selected among the pre-authorized nodes by the network, thus the potential validator group is often relatively small compared to the entire network [23]. Hence, it is more scalable while the group of validators are limited. Inherent features of PoA reveals that, though it sacrifices its decentralization, it

achieves high throughput and scalability, and it is well suitable for private blockchains [23].

B. Federated Byzantine Agreement (FBA)

Federated Byzantine Agreement is a consensus protocol stemming from the famous Byzantine Generals Problem [24], which explains a situation of avoiding complete failure of a decentralized peer-to-peer system while reaching a common consensus among majority, even though some of them are malicious. Other consensus protocols which belong to the same family includes the famous Proof of Work (Pow) protocol by Satoshi Nakamoto, the founder(s) of Bitcoin system as well as the protocols such as Practical Byzantine Fault Tolerance (PBFT) [25] and Delegated Byzantine Fault Tolerance (DBFT) [26]. PBFT is a promising consensus protocol, which is scalable when the group of nodes is small but becomes inefficient for large scale of networks [27]. DBFT is an advanced version of PBFT, which overcomes the scalability issue. FBA is the latest addition to the family, which ensures a robust decentralized system with the help of a concept called quorum slice [28], [29]. Several commercial blockchain systems such as Ripple and Stellar have adopted FBA successfully [30]. FBA is the most preferred protocol among the members of the BFT family because of its high throughput, network scalability, and low transaction costs [31].

As mentioned before, the novelty of FBA is its use of the concept of quorum slice to establish trust [29]. By definition, a quorum is a group of nodes that require to attain common agreement while communicating with each other. A quorum slice is a subset of a quorum, which is a small group of nodes in the system who have reached a consensus. In the FBA protocol, each participant node can choose which other nodes they trust, and their list of trusted nodes forms their quorum slice. Accordingly, it allows open-membership and forms decentralization. Quorum slices can be formed dynamically, thus an individual node can appear on multiple quorum slices called quorum intersection. This overlapping helps to achieve common consensus in a decentralized peer-to-peer network. Through the process of collective decision-making, it can surpass the impact of a faulty node's action.

Despite the promising advantages, the FBA has some shortcomings as well [32]. In this mechanism, each participant node can choose which other nodes they trust, and their list of trusted nodes forms their quorum slice. In such a situation, the nodes usually choose the nodes with a higher reputation. In other words, whether FBA actually reduces centrality is questionable and only a few studies have been done to elaborate on this. In [32] they have proposed a reputation mechanism to incentivize all the peers to be validators in a democratic way in order to be trusted by other peers in the network.

IV. PROPOSED FRAMEWORK

The proposed framework encompasses all the processes of the organic food supply chain, from the farm to the end customer. However, in order to reduce the complexity of the initial model, the processes are limited to those that involve the farmer and supermarket. The important component of transportation has been removed from this version of the framework but will be included in the future frameworks. As depicted by Fig. 1, at each of

these supply chain components, there is a set of actions that need to be recorded in the blockchain system. The role of the consensus protocol is to keep those actions securely (immutably) recorded in the system so they could be traced back to recall the history.

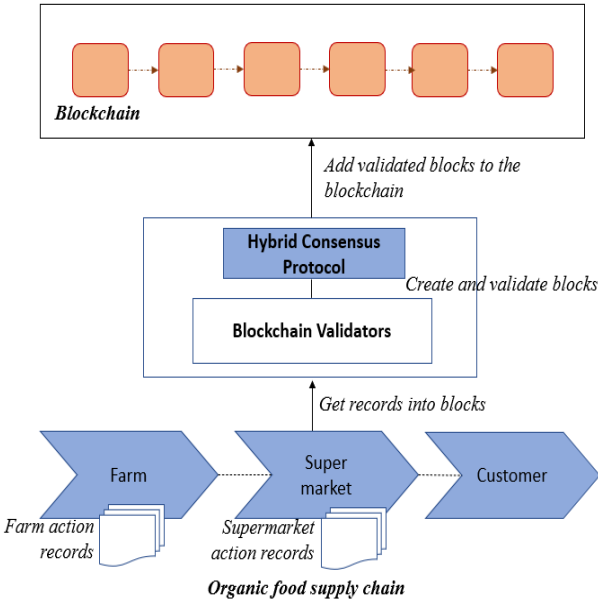


Fig. 1. Overview of the blockchain system

There are two possibilities with regards to a supply chain action. First, the action could be fraudulent. For example, it could be an action, which is not compatible with the concept of organic farming such as a farmer mixing synthetic fertilizers with organic fertilizers. Such actions should not be allowed to be recorded in the system. Second, a particular farmer or supermarket would attempt to alter an action, which is already stored in the system, maliciously. For example, one might attempt to change the recorded figures in a quality test report. Avoiding both of these possibilities is critical to ensure consumer trust on the organic food supply chain.

A blockchain system consists of a consortium of members known as nodes, who actively take part in the process of verification and validation of blocks. In the proposed architecture, there are two groups of members, namely, consortium members and community members. Consortium members are those who have a formal authority vested by the regulatory bodies to oversee, approve, and regulate actions in the organic food supply chain. For example, the agricultural inspector (AI) is a government appointed officer who has the authority to approve/verify some actions of farmers. Community members represent the communities of interest such as consumers, professionals, researchers, religious leaders, social activists, etc. These members do not have a formal authority but their participation in the verification and validation process of a particular action is very much influential to avoid fraudulent actions as well as alterations of records pertaining to past actions. Thus, in the proposed architecture, the involvement of the community members is considered vital in the process of validating a block.

A. Block creation

Block creation in the proposed architecture is done by the members of the first group (i.e., the group of members with authority). In the proposed architecture, block creation happens according to the PoA protocol. The member with the relevant authority pertaining to a particular action is given the chance to create a block and insert the record of the respective action into that block. However, the validation of the block (i.e., permitting the block to be added to the existing chain of blocks) is done based on a quantity defined as the stake of the member. The stake of a particular member is determined by the following formula.

$$S_i = A_i + R_i + T_i \quad (1)$$

Here, S_i is the stake of the i^{th} member of the system and A_i , R_i and T_i are the authority level, reputation and the duration served of that member. Authority is coming from the position the respective member holds and the duration served is computed using the period in service. Reputation is a value attributed to the block creator by the community (i.e., the second group of members). The reputation is computed by the following formula.

$$R_i = C_i + Q_i + P'_i + P''_i \quad (2)$$

Here,

C_i : Number of social connections of the i^{th} member

Q_i : Number of intersecting quorum slices of the i^{th} member

P'_i : The probability of creating a block

P''_i : Probability of success in validation

B. Community-level trust

Notably, the reputation (R) is a quantity related to the social recognition of the respective member. In other words, the community-level trust is incorporated into the system through this quantity of reputation (R). Thus, according to the equation (ii), the reputation is computed by involving the FBA protocol. To achieve a consensus, master node (i.e., node i) has to convince its own quorum slice rather than convincing a lot of nodes to trust. Accordingly, by the quorum intersection structure, the majority of the network nodes would be convinced, since each node trusts every other node on the network. Thus, by communicating with each other, if only the system-wide consensus is reached, that block is approved as a valid block and is appended to the existing chain of blocks.

C. Regulatory governance procedure – rewards and penalties

Participants' honesty and engagement can make the system stable or unstable. The system needs to have control mechanisms put in place to encourage transparent and legitimate actions while penalizing fraudulent actions. Thus, a reward and penalty mechanism is a necessity for the system. In the proposed system, this reward and penalty mechanism is driven by a quantity called trust index (I), which is defined by the following formula.

$$I_i = P'_i + P''_i \quad (3)$$

Here,

P_i' : The probability of creating a block

P_i'' : Probability of success in validation

The block creating node gets a reward for each successful validated block and the validating nodes in the block also get rewarded accordingly for the contribution to validate the block. This mechanism ensures the continuous engagement of the validators and helps sustain the blockchain system in the long run. There is also a penalty mechanism to remove a block creator from the consortium for any fraudulent activity after setting its trust index to zero.

D. System overview

The proposed blockchain system works as follows. The actors involved with the organic food supply chain do actions and transactions. When an action or a transaction is initiated, a member from the consortium members will become the master node, which is the member who has the highest stake to initiate a block. As mentioned earlier, the master node is a consortium member who has a formal authority to oversee, authorize and regulate actions and transactions of supply chain actors. As represented by equation (i), authority is a component of the stake of the consortium member. Thus, this part comes from the PoA component of the architecture.

Once a block is initialized, the master node attempts to reach a consensus in its own quorum slice. If a consensus is reached within that quorum slice, the members of that quorum slice communicate it to the other quorum slices they are involved in, through the quorum intersection structure. If a substantial percentage of the network reaches a consensus, the block is said to be validated and is added to the existing blockchain. This part of the consensus mechanism comes from the FBA component of the architecture.

As an example, if a farmer needs to record a seed certificate he just obtained in the blockchain, the agricultural officer is the formally authorized person to initiate the block when he signs the certificate. For the agricultural officer to initiate this block, he must have the required stake set by the system. In other words, the reputation of the agricultural inspector as well as the duration in the system will also affect the ability to initiate the block. After initiating the block, the agricultural inspector must convince the members of his quorum slice, who could also represent communities of interest such as the local head of police, the religious priests, other neighbouring farmers, professionals, etc. If a consensus was reached within the slice, the individual members can propagate the details of the block to their other quorum slices through quorum interactions. Through this mechanism, it is expected to reach deeper into communities of interest. If and only if a significant percentage (say 67%) of the network reaches consensus, the respective block carrying the record of the seed certificate would be added to the blockchain. The idea of quorum slices is depicted by Fig. 2.

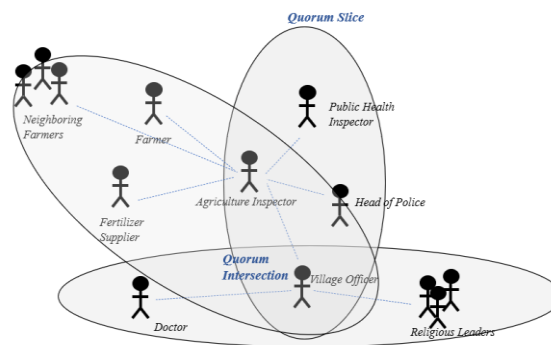


Fig. 2. An example of quorum slices having key officials in the intersections

V. CONCEPT REVIEW OF THE PROPOSED ARCHITECTURE

As the proposed framework is yet to be implemented and tested, as a first phase of validation, a concept level validation of the architecture was done with the involvement of experts in blockchain technology. A series of open-ended interviews were conducted with two academics with a sound track record in blockchain research as well as a practitioner from a leading software development company in Sri Lanka. The interviews were basically conducted focusing on the novelty and potential validity of the idea of adopting the community-level trust into a blockchain consensus protocol. According to the feedback of the experts, incorporation of community-level trust into the consensus protocol is a novel and a desired idea. Moreover, according to the experts' 1) incorporating the stakeholder communities to the certification process will strengthen the trust over the product 2) hybridizing the consensus protocol will mitigate the lapse of each and enhance the security and scalability of the system 3) a good incentive mechanism is required for the system to sustain 4) a solid reward mechanism and meticulous penalty mechanism should be defined to make the participants behave honestly. The experts' feedback further included some key limitations such as the difficulty of maintaining the credibility of the system while confronting the cultural barriers and social norms.

VI. CONCLUSION AND FUTURE WORK

The architecture presented in this paper is novel mainly due to the hybridization of two existing consensus protocols, namely, the Proof of Authority (PoA) and Federated Byzantine Agreement (FBA). Through this hybridization it is expected to obtain better consumer trust due to the incorporation of community-level trust into the consensus protocol as well as due to the enhanced transparency and scalability resulting from that. Besides, this is one of the very few hybrid blockchain architectures proposed aiming at the organic food supply chain. This paper explains the conceptual design of the proposed blockchain architecture in detail, giving insights into the basic components of the hybrid consensus mechanism. Furthermore, it presents a concept-level validation of the idea of incorporating community-level trust into the consensus protocol of the blockchain architecture, with the involvement of a few active researchers and practitioners.

However, this conceptual design needs to be tested to see its dynamic properties such as sustainability and scalability. After all, there is a highly significant social component due to the involvement of communities of interest in the block validation process. As this might bring lots of human-behaviour related dynamics into the actual behaviour of this blockchain system, the scalability and sustainability of this architecture is very much unpredictable. Hence, the testing of this system is thought to be done best in a simulation environment rather than in a real environment. There, the agent-based social simulation (ABSS) is looked at as a candidate approach in the testing process. As ABSS is acknowledged as the third way of doing science [33], mainly due to its ability to study emergent properties of complex social systems, it seems to be well suited to the testing of a complex system like this. Thus, future work of this research would be conducting experiments on the dynamic properties of the proposed blockchain architecture using the ABSS approach. Such experiments would reveal the potential limitations of the design and allow necessary corrective actions to be taken.

ACKNOWLEDGEMENT

Funding support from Accelerating Higher Education Expansion and Development Program (AHEAD) under the research grant AHEAD/RA3/DOR/KLN/SCI

REFERENCES

- [1]. K. Nuttavuthisit and J. Thøgersen, "The Importance of Consumer Trust for the Emergence of a Market for Green Products: The Case of Organic Food," *Journal of Business Ethics*, vol. 140, no. 2, pp. 323–337, May 2015, doi: 10.1007/s10551-015-2690-5.
- [2]. A. Zezza, F. Demaria, T. Laureti, and L. Secondi, "Supervising third-party control bodies for certification: the case of organic farming in Italy," *Agricultural and Food Economics*, vol. 8, no. 1, Nov. 2020, doi: 10.1186/s40100-020-00171-3.
- [3]. R. Home, H. Bouagnimbeck, R. Ugas, M. Arbenz, and M. Stolze, "Participatory guarantee systems: organic certification to empower farmers and strengthen communities," *Agroecology and Sustainable Food Systems*, vol. 41, no. 5, pp. 526–545, Jan. 2017, doi: 10.1080/21683565.2017.1279702.
- [4]. E. Nelson, L. Gómez Tovar, R. Schwentesius Rindermann, and M. Á. Gómez Cruz, "Participatory organic certification in Mexico: an alternative approach to maintaining the integrity of the organic label," *Agriculture and Human Values*, vol. 27, no. 2, pp. 227–237, Mar. 2009, doi: 10.1007/s10460-009-9205-x.
- [5]. J. Duan, C. Zhang, Y. Gong, S. Brown, and Z. Li, "A Content-Analysis Based Literature Review in Blockchain Adoption within Food Supply Chain," *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, p. 1784, Mar. 2020, doi: 10.3390/ijerph17051784.
- [6]. S. Saurabh and K. Dey, "Blockchain technology adoption, architecture, and sustainable agri-food supply chains," *Journal of Cleaner Production*, p. 124731, Oct. 2020, doi: 10.1016/j.jclepro.2020.124731.
- [7]. G.-T. Nguyen and K. Kim, "A Survey about Consensus Algorithms Used in Blockchain," *Journal of Information Processing Systems*, vol. 14, no. 1, pp. 101–128, Feb. 2018.
- [8]. A. Kakkar and Ruchi, "A Blockchain Technology Solution to Enhance Operational Efficiency of Rice Supply Chain for Food Corporation of India," in *Lecture Notes on Data Engineering and Communications Technologies*.
- [9]. M.V. Kumar and N. C. S. Iyengar, "A Framework for Blockchain Technology in Rice Supply Chain Management," in *Advanced Science and Technology Letters*, 2017, vol. 146, pp. 125–130.
- [10]. K. Salah, N. Nizamuddin, R. Jayaraman, and M. Omar, "Blockchain-Based Soybean Traceability in Agricultural Supply Chain," *IEEE Access*, vol. 7, pp. 73295–73305, 2019, doi: 10.1109/access.2019.2918000.
- [11]. J. D. Borrero, "Agri-food supply chain traceability for fruit and vegetable cooperatives using blockchain technology," *Revista de Economía Pública, Social y Cooperativa*, vol. 95, no. 0213-8093, pp. 71–94, doi: 10.7203/CIRIEC-E.95.13123.
- [12]. R. Kamath, "Food Traceability on Blockchain: Walmart's Pork and Mango Pilots with IBM," *The Journal of the British Blockchain Association*, vol. 1, no. 1, pp. 1–12, Jul. 2018, doi: 10.31585/jbba-1-1-(10)2018.
- [13]. S. Madumidha, P. S. Ranjani, S. S. Varsinee and P. S. Sundari, "Transparency and Traceability: In Food Supply Chain System using Blockchain Technology with Internet of Things," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 983-987, doi: 10.1109/ICOEI.2019.8862726.
- [14]. J. F. Galvez, J. C. Mejuto, and J. Simal-Gandara, "Future challenges on the use of blockchain for food traceability analysis," *TrAC Trends in Analytical Chemistry*, vol. 107, pp. 222–232, Oct. 2018, doi: 10.1016/j.trac.2018.08.011.
- [15]. H. F. Atlam, A. Alenezi, M. O. Alassafi, and G. B. Wills, "Blockchain with Internet of Things: Benefits, Challenges, and Future Directions," *International Journal of Intelligent Systems and Applications*, vol. 10, no. 6, pp. 40–48, Jun. 2018, doi: 10.5815/ijisa.2018.06.05.
- [16]. M. van Hilten, G. Ongena, and P. Ravesteijn, "Blockchain for Organic Food Traceability: Case Studies on Drivers and Challenges," *Frontiers in Blockchain*, vol. 3, Sep. 2020, doi: 10.3389/fbloc.2020.567175.
- [17]. B. M. A. L. Basnayake and C. Rajapakse, "A Blockchain-based decentralized system to ensure the transparency of organic food supply chain," 2019 International Research Conference on Smart Computing and Systems Engineering (SCSE), 2019, pp. 103-107, doi: 10.23919/SCSE.2019.8842690.
- [18]. T. Thanujan, C. Rajapakse, and D. Wickramaarachchi, "A Review of Blockchain Consensus Mechanisms: State of the Art and Performance Measures," in 13th International research conference holistic approach to national growth and security, pp. 315–326.
- [19]. Y. Wu, P. Song, and F. Wang, "Hybrid Consensus Algorithm Optimization: A Mathematical Method Based on POS and PBFT and Its Application in Blockchain," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–13, Apr. 2020, doi: 10.1155/2020/7270624.
- [20]. "Analysis between Dash, Zcash, Ripple (XRP) and BRAVO Pay," *Medium*, Oct. 08, 2018. [Online]. Available: <https://medium.com/@BRAVOPay/analysis-between-dash-zcash-ripple-xrp-and-bravo-pay-134dc925edf0>. [Accessed: 18-Jan-2021].
- [21]. Y. Abudris, R. Kumar, T. Yang, and J. Onginjo, "Secure large-scale E-voting system based on blockchain contract using a hybrid consensus model combined with sharding," *ETRI Journal*, Nov. 2020, doi: 10.4218/etrij.2019-0362.
- [22]. "poanetwork/wiki", GitHub, 2019. [Online]. Available: <https://github.com/poanetwork/wiki/wiki/POA-Network-Whitepaper>. [Accessed:10-May-2021].
- [23]. J. MAGAS, "Proof-of-Authority Algorithm Use Cases Grow: From Pharma to Games," *cointelegraph*, Nov. 16, 2019. [Online]. Available: <https://cointelegraph.com/news/proof-of-authority-algorithm-use-cases-grow-from-pharma-to-games>. [Accessed: 12-Apr-2021]
- [24]. L. Lamport, R. Shostak, and M. Pease, "The Byzantine Generals Problem," *ACM Transactions on Programming Languages and Systems*, vol. 4, no. 3, pp. 382–401, Jul. 1982, doi: 10.1145/357172.357176.
- [25]. M. Castro, *Practical byzantine fault tolerance*. Cambridge, Mass.: Institute Of Technolony, 2001.
- [26]. G. CHRISTOFI, "Study of consensus protocols and improvement of the Delegated Byzantine Fault Tolerance (DBFT) algorithm.," Master thesis, Faculty of the Escola Tcnica d'Enginyeria de Telecomunicaci de Barcelona Universitat Politcnica de Catalunya by.
- [27]. X. Zheng and W. Feng, "Research on Practical Byzantine Fault Tolerant Consensus Algorithm Based on Blockchain," *Journal of Physics: Conference Series*, vol. 1802, no. 3, p. 032022, Mar. 2021, doi: 10.1088/1742-6596/1802/3/032022.
- [28]. D. Mazières, "The Stellar Consensus Protocol: A Federated Model for Internet-level Consensus," Jul. 2017, [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.696.93&rep=rep1&type=pdf>.
- [29]. M. Kim, Y. Kwon and Y. Kim, "Is Stellar As Secure As You Think?," 2019 IEEE European Symposium on Security and

- Privacy Workshops (EuroS&PW), 2019, pp. 377-385, doi: 10.1109/EuroSPW.2019.00048.
- [30]. J. Innerbichler and V. Damjanovic-Behrendt, "Federated Byzantine Agreement to Ensure Trustworthiness of Digital Manufacturing Platforms," in *MobiSys '18: The 16th Annual International Conference on Mobile Systems, Applications, and Services*, Jun. 2018, [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3211933.3211953>.
- [31]. "Consensus Protocols That Meet Different Business Demands," Intellectsoft Blockchain Lab, Mar. 26, 2018. [Online]. Available: <https://blockchain.intellectsoft.net/blog/consensus-protocols-that-meet-different-business-demands>. [Accessed: 18- Jun-2021).
- [32]. A. Zoi, "Study of consensus protocols and improvement of the Federated Byzantine Agreement (FBA) algorithm," Master thesis, Faculty of the Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona Universitat Politècnica de Catalunya by.
- [33]. R. Axelrod, "Chapter 33 Agent-based Modeling as a Bridge Between Disciplines," in *Handbook of Computational Economics*, pp. 1565–1584.

Implementation of a personalized and healthy meal recommender system in aid to achieve user fitness goals

Chamodi Lokuge*
Faculty of Information Technology
University of Moratuwa, Sri Lanka
chamodi.16@itfac.mrt.ac.lk

Gamage Upeksha Ganegoda
Faculty of Information Technology
University of Moratuwa, Sri Lanka
upekshag@uom.lk

Abstract - Recent research implies that people's urge to stay healthy and fit has drastically improved and currently, many people are in need to maintain their physical fitness incorporating healthy food habits into their lives amidst hectic urban lifestyles. Thus, nutrition applications are mushrooming in the fitness domain to aid people to improve their dietary intake, track weight-related elements, and generate meal plans. Considering the applications that are typically built for meal planning, it was apparent that personalized nutrition incorporated with healthy meal suggestions is not well addressed, and hence the need for a personalized meal recommendation system that assists the users to achieve their fitness goals is identified. Learning users' food preferences and delivering food recommendations that plead to their taste and satisfy nutritional guidelines are challenging. Due to the lack of access to a proper meal planning application or without professional help most users follow ineffective, generic meal plans which hinder them from achieving their fitness goals and often cause long-term and short-term health complications. The proposed implementation aims to bridge the gap between the existing meal planning applications and the potential need for a personalized healthy meal plan. This paper succinctly presents the design and implementation of the proposed personalized and healthy meal recommendation system and further discusses the architecture and the evaluation of the design solution.

Keywords - automated meal planning, content-based filtering, personal nutrition, personalized meal planning, recommender system

I. INTRODUCTION

People's lifestyles have changed lately and they tend to consume more calories with less nutritional value, and these improper eating habits are extremely dangerous to one's health. It is indubitable that unhealthy eating habits can lead to deprivation of the right nutrition and eventually resulting in overweight, obesity, or malnutrition. As per the past literature, 80% of deaths referred to ten major ailments were related to improper eating habits [1]. Increased risk of strokes, diabetes, cardiac diseases, cancers, tooth decay, osteoporosis, depression symptoms, high cholesterol levels, high blood pressure are some remarkable short-term and lifelong ailments that could implicitly exhibit in individuals due to poor nutrition [2]. Global nutrition statistics demonstrated that people do not have adequate knowledge about the right nutrition which later results in macronutrient malnutrition [3]. Moreover, healthy meal planning requires a discerning knowledge about nutritional adequacy, gender, age, and level of physical activity which in most cases act as an obstacle to most individuals. Hence, even though healthy meal planning is starting to gain attention among people, these barriers discourage

individuals from adjusting their food habits to favor a healthier diet. The unavailability of finding healthy food alternatives that fit user tastes acts as one of the main barriers among individuals which hinders them from achieving their fitness goals. Learning users' meal preferences is a mandatory step in recommending healthy foods that users are more likely to find desirable. Despite the presence of personalized meal planning applications which have been specifically designed for the personalization of meal plans, many approaches still suffer from major limitations. PlateJoy [4], a personalized meal planning application, elicits users' meal preferences in the form of a questionnaire.

“(a) How often do you eat meat? No restrictions, No Red Meat, Pescatarian, Flexitarian, Vegetarian, Vegan

(b) Are there ingredients you prefer to avoid? Added sugar, Avocado, Beef, Bell pepper, Chicken”

Depending on the users' answers to the questions the application recommends a meal plan by avoiding distinctly unacceptable food choices made by the user, and thus only capable of recommending a meal plan of coarse-grained food preferences. Moreover, the application only focuses on delivering a personalized meal plan without embedding the nutritional guidelines.

Another main barrier that has been identified by the authors is the lack of meal planning approaches that take user's physiological data and plan their meals to meet the daily nutritional requirements by incorporating standard nutritional guidelines.

It is proven that the adoption of the right nutrition practices has been shown to be beneficial to prevent many non-communicable diseases [1] [5]. Drawbacks and limitations of previous meal planning approaches call the sheer lack of a meal planning system that correctly caters to meet the user's nutritional requirements and user's meal preferences. Hence, the proposed solution presents a meal planning approach that is focused on mitigating the issues in the meal planning domain and delivering the following features.

1. The delivery of a meal planning approach that delivers fine-grained user preferences by learning the user's meal preferences.
2. The delivery of a meal planning approach that integrates the nutritional guidelines to cater nutritional requirements of the user.

Personalization of meal planning is a lively research hitch focused on adding personalization capabilities in the meal recommendation domain. Recommender systems have been identified as the most successful tool which is capable of personalizing processes over several domains [6]. E-commerce [7], finance, marketing, tourism [8], and many other domains are using recommender systems to support users to deliver recommendations in an overloaded information context [9]–[13]. The proposed implementation contributes at developing and integrating a recommender system model that incorporates both user preferences and nutritional requirements in the food recommendation domain.

The proposed system will get users basic information (age, weight, height, gender) and user goals (weight loss/ weight gain/ maintain current weight) followed up by user meal preferences by asking a simple questionnaire. The level of physical activity (sedentary, lightly active, moderately active, very active) is taken as an input to the meal recommendation system as a parameter of the physical level of engagement of the user. The system then queries the Basal Metabolic Rate (BMR) and estimates the Daily Calorie Allowance for the user depending on the fitness goal based on various nutrition health measurements [14] [15]. The proposed system finally presents a weekly meal plan to achieve the user’s fitness goal that fulfills the nutritional requirements of the user after refining the meals to best match with the user’s taste. This paper is focused on developing a personalized meal recommendation system for healthy users that will eventually prevent the users from major chronic diseases related to unhealthy eating habits.

The remainder of the paper is organized as follows. Section II discusses the existing approaches in the meal planning domain and their corresponding gaps. Section III presents the design approach of the proposed implementation and section IV further discusses the system design architecture of the overall solution. Section V discusses the implementation of the proposed personalized and healthy meal planning system. Section VI presents the evaluation of the proposed system and section VII comprises the discussion. Section VIII finally concludes the paper.

II. RELATED WORK

Referring to the preceding literature, it is recognized that a multitude of studies have been conducted in the meal planning domain over the past years [16]–[23]. This section discusses the related work conducted on the food recommendation domain with correspondence to their gaps.

Eat This Much approach provides users with daily pre-defined meal plans fulfilling Calorie Intake (CI) level as stated by the user as a user input [24]. However, this approach has some limitations. This allows the user to select his meal preferences by distinctly avoiding certain food categories rather than allowing them to log their food preferences directly into the system which will finally result in suggesting coarse-grained food preferences. The approach does not deliver a meal plan adhering to the nutritional guidelines; hence the approach does not address the requirements of the user group who lacks adequate nutritional knowledge and thus fails in delivering a healthy meal recommendation.

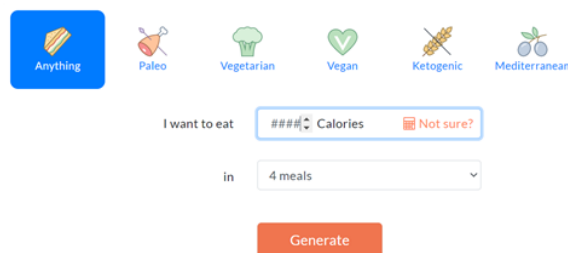


Fig.1. Screenshot of Eat-this-much application

MakeMyPlate approach lets the user restore an already existing recipe with another [25]. But the drawback of this approach is that the system doesn’t substantiate as to whether the replaced recipe is calorically equivalent with the initial recipe substituted by the user. Therefore, substituting meals as per the desire of the user might result in a caloric imbalance between the original meal and the replacement meal. Additionally, the approach does not deliver personalized meal recommendations to match the user’s taste.

Another existing meal planning approach is MyFitnessPal which takes in the user’s physiological information, desired weight, and outputs the daily calorie allowance for the user [26]. The approach does not display any intelligent behavior. It merely acts as a calorie counter for a particular user without even setting up meal plans.

The authors in [27] present the use of ingredient substitution on how ingredients can be fit well together as a means to get personalized recommendations. By observing the observations and the test results, authors in [27] have concluded that this approach can predict users’ preference for a recipe, but the whole list of ingredients is not taken into consideration. This research only focuses on predicting food recipes that adhere to user preferences and doesn’t take the fitness goal of the user into consideration.

Table I summarizes the existing approaches in the meal planning domain in relation to the tracking of calorie consumption, delivery of personalized meal recommendations, and adherence to the nutritional guidelines.

TABLE I. SUMMARY OF EXISTING APPROACHES IN MEAL RECOMMENDATION DOMAIN

Related work	Tracking of calories allowed	Delivery of personalized meal plans	Adherence to nutritional guidelines
Eat-This-Much [24]	✓		
Make My Plate [25]	✓		
MyFitnessPal [26]	✓		
LoseIt [28]	✓		
PlateJoy [4]	✓	✓	
Teng et al. [27]	✓		
Yang et al. [29]	✓	✓	✓
Nutrino [30]	✓	✓	
BNF’s Meal Plan [31]			✓

Following the existing approaches in the meal recommendation domain, it is identified that the taken approaches are not focused on delivering a healthy meal plan which is fine-grained to the user’s personal preferences.

The authors in [29] presented an approach to deliver a personalized and a healthy meal plan. However, their approach was limited to research on exploiting visual food features. Hence the approach followed in this paper will adhere to the delivery of a personalized and a healthy weekly meal plan to achieve the fitness goals of the users.

III. DESIGN APPROACH

This section describes the design approach of the proposed system with detailed explanations with relevance to the selection of the most appropriate technology in the context of use. The proposed implementation of the personalized and healthy meal recommender system is designed in a way by considering the user group opinions gathered from the initial survey conducted by the authors and by addressing the gaps of existing meal planning approaches in the domain and by incorporating nutritional measurements as depicted in Fig. 2.

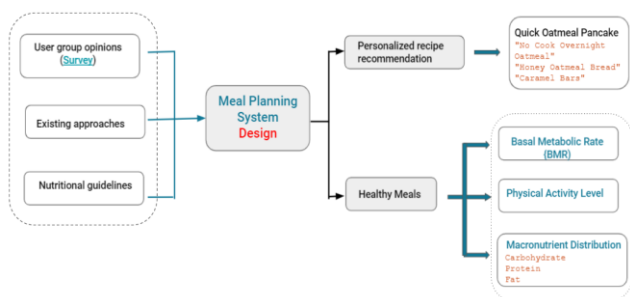


Fig.2. System design of proposed implementation

A. Initial survey

It was decided to conduct an initial survey to capture the sentiments of the individuals and to verify the perception held by individuals regarding the meal planning approaches. Additionally, the survey was aimed at understanding the barriers related to personalized and healthy meal planning. The survey was conducted targeting Sri Lankan individuals both residents and overseas Sri Lankans. The sample size of the survey was 103 participants. The participants were assessed based on their nutritional knowledge on meal preparation and asked to state their opinions on the need for a personalized and healthy meal plan to use in aid to achieve their fitness goals. A summary of the responses gathered is depicted in Fig. 3 and Fig. 4.

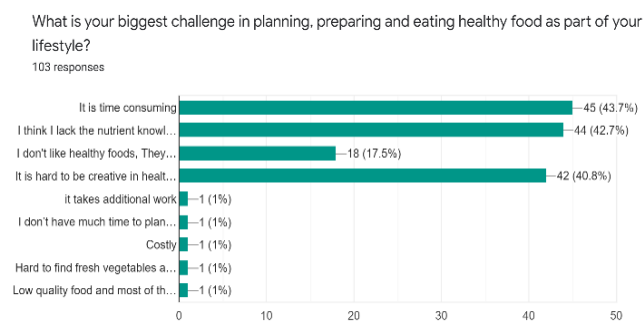


Fig.3 Challenges faced by individuals in healthy meal planning.

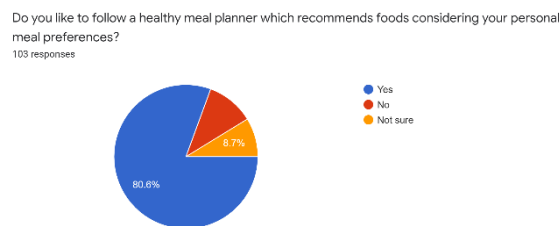


Fig. 4. Summary of responses for the need of the system.

The initial survey was additionally aimed at gathering the fitness goals of the general public, energy intake, physical activity level of individuals, other hindrances in healthy meal planning in order to deliver a more user-friendly meal planning approach. According to the nutritional survey statistics that have been conducted previously and as per the results obtained from the initial survey, only less than 5% of the participants could answer the knowledge about macronutrients (carbohydrate, protein, and fat) correctly. Moreover, 83 participants out of 103 participants, a percentage of 80.6% have stated the need for a personalized and healthy meal recommender system as in Fig.4 and hence the requirement for the proposed implementation was verified.

B. Selection of recommender system

The design for the recommender system in the proposed implementation has been conceived in an attempt to overcome the limitations faced by existing meal recommendation approaches. Hence, the most suited recommender system needs to be integrated into the system to deliver more fine-grained meal preferences by learning the taste of the user.

Gunawardana and Shani [32] identify two main tasks related to recommender systems as prediction task and recommendation task. In relation to the context of use and the working principles beneath the Recommendation Systems, RSs have been classified into some popular groups namely collaborative filtering, content-based filtering, and demographic filtering. Other categories are knowledge-based and constraint-based recommender systems [33]. Out of the aforementioned popular categories of RSs, content-based and collaborative filtering recommender systems are successful in the personalization process.

Authors in [34] have used collaborative filtering methods in the recommendation of food recipes and have concluded that content-based filtering strategies can be used to achieve more sensible accuracy and coverage. They have found only a marginal boost in the accuracy when collaborative filtering strategies are utilized [34]. Another major problem of the collaborative filtering approach is the method of combining and weighing the preferences of user neighbors. Knowledge-based recommender systems use users' preferences in the recommendation and the constraint-based recommender approach sets constraints like daily fat, carbohydrate, and protein intake limitations.

Content-based recommender systems rely on meta-data or features from individual items to recommend items that can be used in this context. Content-based filtering has been constructed to recommend similar item recommendations by analyzing the content of the user's

previous preferences [33]. Hence, the approach followed in this paper will adhere to the content-based filtering methodology with consideration to the context of use.

As authors in [35] addressed, embedding more rules and constraints in the recommender system will help in the improvement of the accuracy of the recommender system. The right balance between the nutritional needs of the user and the user's taste needs to be acknowledged rather than delivering recommendations in an isolated fashion. For instance, recommendations only based on user preferences may invigorate unhealthy eating patterns. Thus, the originality of this work also lies in coalescing more nutritional constraints in the system concerning user's physiological information and delivering fine-grained user-preferred meal plans.

IV. SYSTEM DESIGN ARCHITECTURE

The design architecture of the overall system is presented in this section with detailed designs and explanations, prioritized by the sequence of the design. To address the issue at hand, the authors have proposed a meal recommendation module consisting of a query module, recommender system, and a knowledge base (recipe data and nutritional information) as illustrated in Fig. 5.

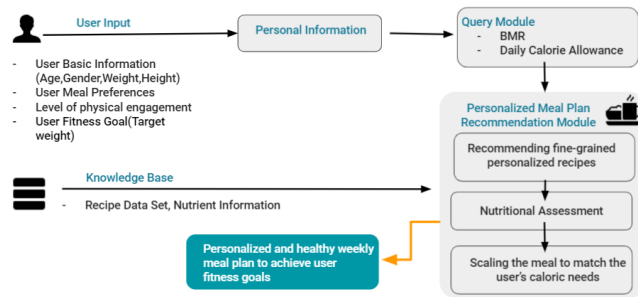


Fig.5. System design of the proposed implementation

The proposed implementation delivers an encompassment of a multitude of competence suited for recommending a personalized and healthy meal plan to achieve user fitness goals. The user's physiological information such as age, gender, current weight, height is taken as inputs to the system. Additionally, the goal weight of the user is taken as the fitness goal of the user. The user's fitness goal might be to lose weight, gain weight, or maintain the current weight. Therefore, meal recommendation is done in the order of the following steps.

1. Calculation of the user's caloric needs in correspondence to his BMR and the goal weight.
2. Delivery of fine-grained personalized meals to plead the user's taste.
3. Translation of daily calorie allowance into an actual meal plan and, optimizing and scaling the personalized meal recipes to meet the daily caloric allowance and the daily macronutrient requirement of the user.

The query module is specifically designed to compute the Basal Metabolic Rate (BMR) to determine the Daily Calorie Allowance of the user based on user inputs of age, gender, weight, height, and fitness goal. Harris-Benedict [36] equations and Mifflin St. Jeor [37] equations are the most adopted formulas used by nutritionists in the

calculations of Basal Metabolic Rate. The Mifflin St. Jeor equation is able to assess the weight more accurately with the changes in the lifestyle. In comparison to the Harris-Benedict formula, Mifflin St. Jeor's formula is having an improvement of 5% in the accuracy [38]. The following equations (Eq.1 and Eq.2) account for determining the BMR of males and females using the Mifflin St. Jeor formula.

$$BMR_{male} = 10 * weight + 6.25 * height - 5 * age + 5 \quad (1)$$

$$BMR_{female} = 10 * weight + 6.25 * height - 5 * age - 161 \quad (2)$$

To query the Total Energy Expenditure, BMR and the level of physical activity (PAL value) is taken into consideration. Energy expenditure and energy requirement are highly dependent on the Physical Activity Level (PAL). The level of physical activity is classified into 5 main categories by the 1981 FAO/WHO/UNU expert consultation (WHO, 1985) and given a range of PAL values based on the level of physical activity as stated in Table II [39]. Thus, Total Energy Expenditure can be calculated by multiplying the BMR and the corresponding PAL value given concerning the level of physical activity of the user.

TABLE II. CLASSIFICATION OF LIFESTYLE AS PER THE LEVEL OF PHYSICAL ACTIVITY

Category	PAL value
Sedentary (little or no exercise)	1.2
Lightly active (light exercise/sports 1-3 days/week)	1.375
Moderately active (moderate exercise/sports 3-5 days/week)	1.55
Very active (hard exercise/ sports 6-7 days a week)	1.725
Extremely active, hard daily exercise or physical job	1.9

Body mass change is associated with the daily caloric deficiency or a caloric surplus. A caloric deficiency results in weight loss while a caloric surplus results in weight gain. Likewise, a caloric balance between the caloric intake and the caloric expenditure results in maintaining the weight. As per the research conducted by the National Institute of Health in the USA, 3500 kcals per pound (0.45kg) rule can be used in achieving the fitness goals in the nutrition domain which states that cumulative energy deficiency of 3500 kcals is the equivalent of the loss of 1 pound per body weight [40]. The weekly steady rate of weight loss is considered to be one pound (0.45kg) i.e., 500 Kcal daily deficiency. Accordingly, a daily caloric surplus of 500kCals would result in a weight gain of 1 pound per week. Health Promotion guidelines state that Caloric Intake estimations for adult females and males range from 1600 to 2400 and 2000 to 3000 respectively based on their level of engagement of physical activity [41]. Moreover, females and males are not recommended to consume less than 1200 and 1500 kcals respectively [41]. Therefore, when recommending the daily calorie allowance to achieve user fitness goals, the aforementioned rules have been implemented in the proposed implementation of the personalized and healthy meal recommender system. The

proposed implementation is designed in a way to suggest the number of weeks (n) to reach the expected target weight (w') of the user (Eq. 3).

$$n = \frac{1}{7} \left(\frac{|w-w'|}{\frac{CI-TBEI}{500}} \right) \quad (3)$$

After querying the daily calorie allowance (CI), the personalized meal plan recommendation module aims at giving out personalized and healthy meal recommendations by translating the calculated caloric intake into an actual meal plan as sketched in Fig.5. This module uses a content-based recommender system to deliver fine-grained personal preferences. The content-based model fabricated in this research utilizes Latent Dirichlet Allocation (LDA) as a topic model to generate tags to group similar items of the recipes in the dataset in order to finally recommend personalized recipes based on the user's previous meal preferences. The similarity between the user preferred meal and all the recipe profiles in the dataset is obtained from cosine similarity. This is a semantic similarity measure that takes the cosine angle of two vectors to calculate the similarity as stated in Eq.4 [33].

$$sim(i, j) = \frac{r_i \cdot r_j}{\|r_i\|_2 \|r_j\|_2} = \frac{\sum_u r_{iu} r_{ju}}{\sqrt{\sum_u r_{iu}^2} \sqrt{\sum_u r_{ju}^2}} \quad (4)$$

In the proposed implementation, the user is given the chance to enter at least 3 user-preferred recipes via the application. During recommendation, the cosine similarity metrics are calculated from the recipes' feature vector and the user's preferred feature vector retrieved from user input. Hence the top 100 recipes are recommended in the descending order of similarity score to best match the recipes w.r.t user-preferred meals.

Subsequently, this initial set of personalized recipes is passed into the Nutritional Assessment Module. This module is designed to translate the daily caloric allowance into an actual meal plan by taking macronutrient distribution into consideration. The system will utilize the recommended daily protein requirement ($\geq 0.8\text{g/kg/day}$) as per the standard dietary guidelines and hence satisfy the daily protein need of the user [42]. Moreover, the system filters the fat percentage of the recommended recipes to be a minimum of 40% as recommended in guidelines in order to deliver a healthy meal recommendation [43]. After determining the appropriate macronutrient composition, the final phase of the personalized meal planning is to optimize the top-recommended recipes by the content-based recommender system. To do so, the top recipes recommended by the content-based recommender system are scaled to match the user's caloric need and filtered based on the rules implemented by the nutritional assessment module. The daily calorie allowance of the user is distributed equally among breakfast, lunch, and dinner. The proposed implementation finally outputs a weekly meal plan for breakfast, lunch, and dinner with the number of calories, portion size, link for the recipe, and a pie chart for macronutrient composition of the recommended recipe.

V. SYSTEM IMPLEMENTATION

This section describes the implementations carried out in each component of the system with regards to the methodologies and designs described in the previous sections.

A. Data set preparation

The recipe data set is scraped from allrecipes.com using python, selenium, and chrome web driver. Over 5000 recipes are scraped including the title of the recipe, ingredients, ratings, cook time, servings, calorie, protein, carbohydrate, cholesterol, fat, sodium, and ranking of the recipe. The recipe data with no nutritional information is eliminated and data types for cook time, calorie, protein, carbohydrate, fat, and rankings are changed to int and float data types. NLTK and Gensim libraries are used to clean and preprocess the dataset. Fig.6. illustrates a screenshot of the preprocessed dataset.

Recipe ID	URL	Calories	Protein	Carbohydrate	Fat	Sodium	Cholesterol	Ranking
4741	https://www.allrecipes.com/recipe/92899/my-canadian-friend-My Canadian Friends Bear	4.68	66	12	1ablsq	6	20	457
4742	https://www.allrecipes.com/recipe/93132/stovetop-granola/ Stovetop Granola	4.57	271	1	1ablsq	4	20	529
4743	https://www.allrecipes.com/recipe/95233/leftover-pot-pie/ Leftover Pot Pie	4.58	99	2	cup	6	0	506
4744	https://www.allrecipes.com/recipe/93337/citrus-broiled-duck-Citrus Broiled Alaska Salm	4.12	32	14	large	8	30	158
4745	https://www.allrecipes.com/recipe/9340/turkey-tetrazzini/ Turkey Tetrazzini	4.19	114	2	8	out	6	50
4746	https://www.allrecipes.com/recipe/9341/turkey-n-stuffing-b Turkey N Stuffing Bake	4.34	75	3	cup	5	0	615
4747	https://www.allrecipes.com/recipe/95597/authentic-and-easy Authentic And Easy Shrim	4.69	90	1	4	cup	1	20
4748	https://www.allrecipes.com/recipe/93623/shrimp-and-gravy Shrimp And Gravy	4.43	53	3	4	cup	4	40
4749	https://www.allrecipes.com/recipe/93747/venison-fajitas/ Venison Fajitas	4.71	81	1	Fajita	6	45	688
4750	https://www.allrecipes.com/recipe/94031/sweet-russian-cab Sweet Russian Cabbage Sc	4.65	401	1	1/2	poi	4	70
4751	https://www.allrecipes.com/recipe/9411/salmon-patties-ii/ Salmon Patties II	4.51	133	1	1/4	75	5	0
4752	https://www.allrecipes.com/recipe/93622/high-temperature High Temperature Eye Of	4.42	1000	1	lb	6	185	237
4753	https://www.allrecipes.com/recipe/94374/shrimp-leek-and-shrimp Leek And Spinach	4.16	51	2	1ablsq	4	60	714
4754	https://www.allrecipes.com/recipe/94570/absolutely-delicious Absolutely Delicious Grees	4.5	117	2	1/2	our	6	35
4755	https://www.allrecipes.com/recipe/94725/savannah-seafood Savannah Seafood Stuffin	4.83	67	1	2	cup	1	50
4756	https://www.allrecipes.com/recipe/93623/twice-baked-potato Twice Baked Potatoes II	4.49	152	1	large	6	0	722
4757	https://www.allrecipes.com/recipe/95465/easy-fried-spinach Easy Fried Spinach	4.21	138	1	4	cup	1	15
4758	https://www.allrecipes.com/recipe/95786/vegetarian-refried Vegetarian Refried Beans	4.44	103	1	3	pound	12	270
4759	https://www.allrecipes.com/recipe/95982/deep-fried-turkey Deep Fried Turkey Rub	4.93	47	2	5	bay	12	5
4760	https://www.allrecipes.com/recipe/9615/healthy-banana-cookies Healthy Banana Cookies	3.96	1000	1	lb	36	50	56
4761	https://www.allrecipes.com/recipe/9624/aunt-hazels-apple-Aunt Hazels Apple Oatme	4.44	120	2	3/4	cup	1	281
4762	https://www.allrecipes.com/recipe/9627/easy-oatmeal-cook Easy Oatmeal Cookies	3.6	108	1	3	cup	48	25

Fig.6. Screenshot of the preprocessed dataset

B. Content-based recommender system

In order to recommend fine-grained personalized recipes, it is important to provide labels to each recipe in the preprocessed dataset. For this topic-modeling purpose, authors have utilized the LDA model to have probability distribution across labeled topics as discussed in section IV. The LDA model is implemented after choosing the optimal number of topics and, by tuning the hyperparameters to improve the accuracy of the model as further discussed in section VI under evaluation of the recommender system. Fig.7 depicts the parameters used to build the optimized LDA model.

```
lda_model_final = LdaMulticore(corporus=corporus,
                               id2word=id2word,
                               num_topics=8,
                               random_state=100,
                               chunksize=100,
                               passes=10,
                               alpha=0.01,
                               eta=1)
```

Fig.7. Building the LDA model

Next, the LDA model is used to create an LDA matrix that holds the probability distribution for every recipe in the dataset as presented in Fig.8. Probability distribution

retrieved from the LDA matrix is utilized in the content-based recommender system to deliver personalized recipes based on the user's preferred recipes.

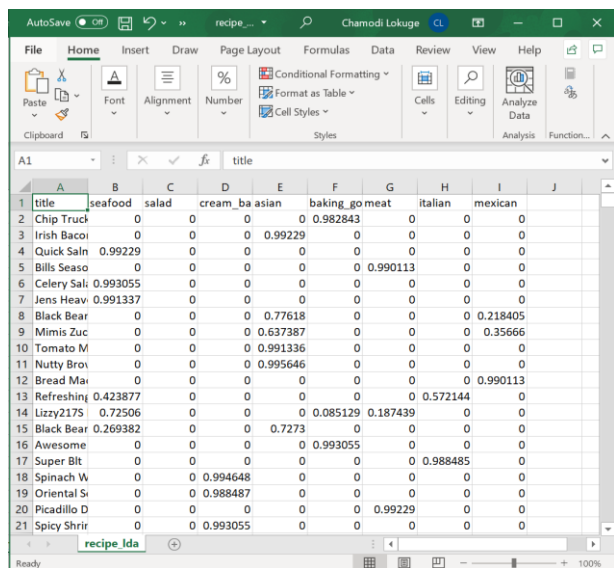


Fig.8. Screenshot of the probability distribution of topics in the dataset

C. Web application

The proposed system is implemented as a web application using python for the server-side development and the application is deployed in streamlit. The application takes in the user's personal information (age, gender, current weight, and height), user target weight, and user meal preferences via the user interface of the application. Fig.9 and Fig.10 demonstrate the UI implementation of the web application.

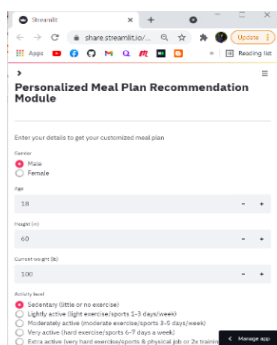


Fig.9. UI implementation of web application

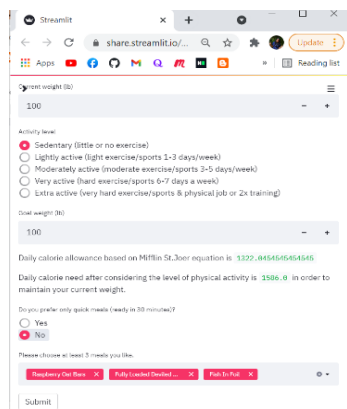


Fig.10. UI implementation of web application

The proposed implementation allows the user to add an optional filter for cook time to recommend recipes to prepare meals in less than 30 minutes. This was a user suggestion in the initial survey conducted by the authors at the initial phase of gathering user requirements. Upon the submission of the required information, the application outputs a weekly meal plan for breakfast, lunch, and dinner as demonstrated in Fig.11.

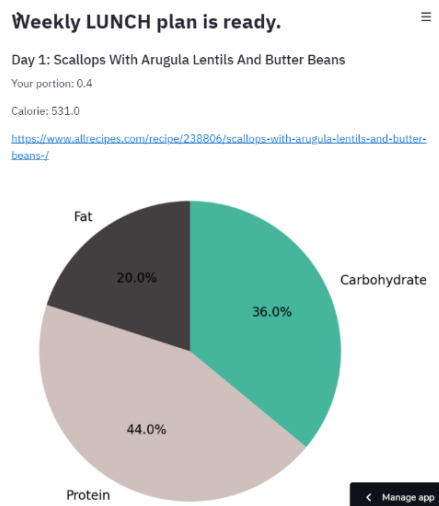


Fig.11. UI implementation of web application

VI. EVALUATION

The following section describes in detail how the proposed implementation of the personalized meal recommendation module is evaluated using different approaches, namely: (A) the validation and correctness of the recommender system, (B) evaluation by a real audience to determine the success at meeting the initially set objectives of the project.

A. Evaluation and validation of the recommender system

In order to test the quality of a recommendation system model, several evaluation metrics can be employed. The recommender system model incorporated in the proposed implementation is the Latent Dirichlet Allocation (LDA) model. This section describes a quantitative evaluation of the LDA model. Topic coherence and perplexity measures are some adopted intrinsic evaluation metrics that can be used to judge how good a given model is [44]. There were studies that argue the perplexity measure is sometimes not correlated with the human judgment of the model [45]. Thus, topic coherence is used to measure the semantic similarity between topics inferred by the model.

The LDA model is initially developed with 10 different topics where each topic is a mix of keywords and each keyword contributes a certain weightage to the topic. The baseline coherence score is 0.383 when the LDA model is built with default settings. The optimum number of topics needs to be determined in order to improve the baseline coherence score of the model. The graph in Fig. 12 presents the coherence score (c_v) over the number of topics (n). The highest coherence score is yielded when the number of topics is in the range of 7 to 8.

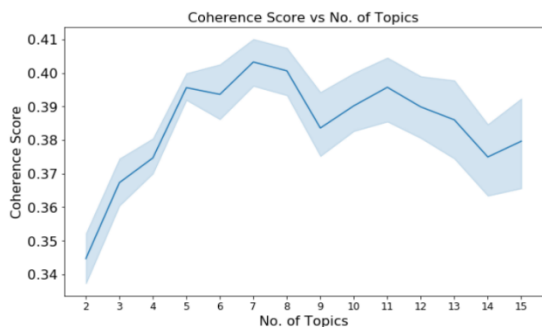


Fig.12. Coherence score over # of topics to determine the optimal # of topics

Additionally, optimal document-topic density (alpha) and word-topic density (beta) parameters need to be determined to improve the coherence score of the model. Using the LDA tuning results, it was observed that using a topic distribution of 8 and alpha of 0.01 and beta of 1, an improvement of 9.138% in coherence score over the baseline coherence value can be achieved.

Mean cosine similarity between content-based recommender system and raking-based recommender system for 1000 simulations is considered in order to validate the content-based recommender system. Ranking based recommender system is implemented to suggest recipes based on the 'ranking' of the recipes. Content-based recommender system is implemented to randomly pick 3 recipes to mimic the user behavior of choosing meal preferences via the web application. Both the systems were filtered based on the rules developed in the nutritional assessment module. Based on the results, the content-based recommender system scores a mean cosine similarity of 0.47 and the rank-based recommender system scores a mean similarity of 0.23 where the content-based recommender system scores remarkably a high mean similarity for 1000 simulations. The graph in Fig.13 illustrates the comparison between the mean similarity score of the two systems over 1000 simulations.

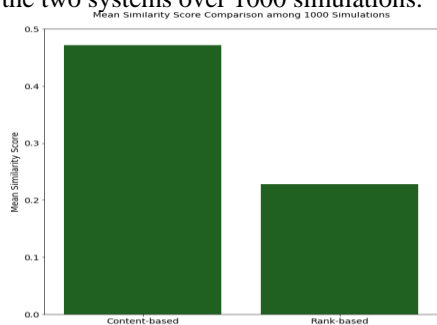


Fig.13. Graph of mean similarity score of content-based and rank-based systems

A. Evaluation by a real audience using the post-evaluation survey

Considering the initially set objectives in developing a personalized and healthy meal planning approach, it was decided to evaluate the system using a post-evaluation survey upon the completion of the relevant implementations. For evaluating the effectiveness of the proposed implementation, it was planned to conduct the experiment by allowing the participants to use the application deployed in streamlit over a period of one week. Phase 1 of the post-evaluation survey aimed at

targeting 10 individuals from Sri Lanka. Participants were given an introduction about how to use the application and asked to assess the application based on their user experience after the completion of the week. Following the completion of one week, all the responses of the 10 individuals were collected.

The majority of the participants rated the application positively as shown in Fig. 14. The country was in a locked-down state when the experiment was conducted thus people were not allowed to step out to prepare the meal plans suggested by the system. Hence, none of the participants have used the meal plans recommended by the system. Moreover, the recipe data set used is scraped from allrecipes.com which includes foreign recipes which was a drawback in the participants' point of view.

Please assess the Meal Recommender Application under the given criteria (for one time user)

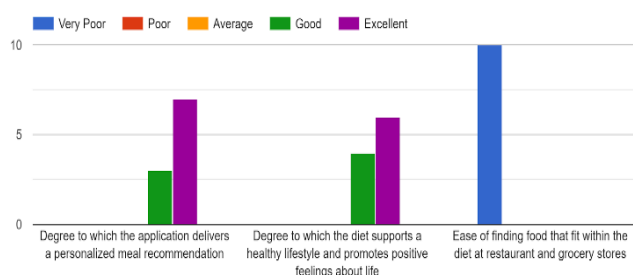


Fig.14. Summary of post-evaluation survey phase 01(one-time user)

Meals suggested by the application are mostly Malaysian, Japanese, and Australian cuisines. Therefore, it was decided to conduct Phase 02 of the post-evaluation survey targeting Sri Lankan participants currently living in Japan, Australia, and Malaysia.

As all of the participants are supposed to follow healthy meal plans, it was decided to choose individuals from a social media fitness group who are keen on planning their meals healthy. Among the individuals selected, 5 participants have followed the diet plan recommended by the application over a week. The majority of the participants rated the experience of the application from average to excellent. All of the participants have confirmed that the meals suggested by the application are personalized, healthy, and support the individuals in achieving their fitness goals. Fig.15 depicts the summary of ratings of post-evaluation survey phase 02 based on a one-week user experience.

Please assess the meal planning application based on your one-week experience.

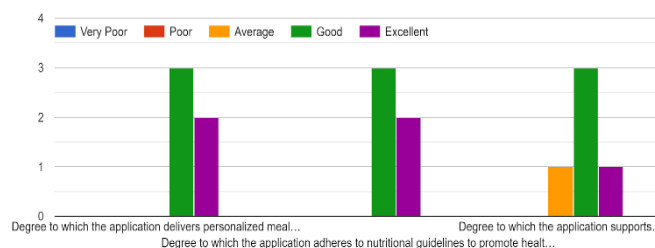


Fig.15. Summary of post-evaluation survey phase 02(one-week user experience)

VII. DISCUSSION

The existing studies in the meal planning domain focus solely on the meal plan generation task, while this paper proposes to provide a full-fledged solution for a more personalized and nutritional meal plan to achieve the fitness goals of the user. In general most of the food recommender systems play a better role in tracking the calorie consumption of the user, but do not adhere to provide the user the adequate nutritional needs or to help the user to achieve fitness goals [16], [17], [19]–[29], [34], [42], [46]–[59]. The primary objective of this paper is to understand the obstacles related to meal planning and thus, mitigate the shortcomings of delivery of a personalized and healthy meal plan.

The initial survey responses concluded that the majority of the individuals out of the 103 participants did not have adequate knowledge to plan their meals healthily. 80.6% of the participants were aware that unhealthy eating habits lead to major health diseases and over 80% of participants would likely to use a meal planner. It was evident that participants lack the adequate nutritional knowledge to plan their meals from the responses of the nutritional survey conducted along with the initial survey. It was mostly cumbersome to stick to a meal plan which did not go hand in hand with user taste. Based on the responses, participants have claimed the necessity of a meal of their choices which follows nutritional guidelines as presented below [60].

“I think many people lack the nutritional knowledge and do not know how to loose weight or gain weight by keeping track of their meals.”

“I do not tend to learn or keep track of all the nutritional values of the food I consume, so it's best to let a meal planner take care of it to me. But this again depend on how intrusive such an option in day to day life would be, for example having to consume food that do not align with my tastes is a negative.”

“I would always prefer to stick to a healthy and personalized meal plan. But since I lack the nutritional knowledge on how to prepare a meal plan on my own, I would surely use a meal planner that does the work for me.”

By analyzing the results obtained from the initial survey, it was determined that there exists a need for personalized and healthy meal plans in aid to achieve user fitness goals [60]. It was also identified that a combination of personalized and healthy meal planning approaches is favorable for many users.

Presenting a new web application and leaving a positive impression while engaging in the application is challenging. The authors of the proposed system ensure that the user interface (UI) design encompasses minimalistic UIs to make the application visually appealing to deliver a more aesthetically pleasing experience to motivate the users to follow along. The user requirements and perceptions about existing meal planning approaches are gathered during the initial phase of planning the system and thus the system is designed in a way to eliminate the complicated UIs and over flooding of information. The macronutrient distribution of the recommended recipes suggested by the system illustrates in pie charts to make

more sense to the user that later got positive comments in the post-evaluation survey.

The authors have conducted three surveys from the initial stage of planning the design, to the final phase of the proposed implementation. The results of the surveys are summarized below.

1. Out of 103 individuals who participated in the initial survey, a majority of them know that unhealthy eating habits lead to major ailments and hence in need of a personalized and healthy meal recommendation application.
2. The participants of the post-evaluation survey have concluded that the proposed implementation of the personalized meal planner application delivers healthy meal plans and supports in achieving their fitness goals.
3. Overall a positive perception was observed in the participants regarding the helpfulness of the implemented meal recommendation application for the users to achieve their fitness goals.

A. Limitations

One of the major limitations of the design methodology of the proposed implementation is currently the application is targeting healthy individuals with no medical complications. Due to the complexity of dealing with medical cases, and since it needs a lot of expert intervention, the current implementation of the proposed system aimed at delivering a healthy meal plan to a healthy user which will ensure that a user follows a healthy diet. It was evident that people eating unhealthy food choices and lacking the knowledge of nutrition may eventually lead to major chronic diseases which ultimately lead to premature death. Hence, the proposed implementation aims to deliver healthy meal recommendations which also pleads with their taste.

VIII. CONCLUSION AND FURTHER WORK

Following the inspection of existing meal planning approaches and their gaps, this paper presents a meal planning approach both personalized and healthy in aid to achieve user fitness goals. According to the past literature and the observations gathered during the various stages of design methodology, the following conclusions regarding the involvement of personalization and nutritional guidelines in the food recommendation can be identified.

1. Delivery of a meal recommendation application considering the user's meal preferences motivates the user to follow a healthy meal plan.
2. Delivery of a meal recommendation application considering nutritional constraints like macronutrient distribution has a positive impact in achieving the user fitness goals.
3. Delivery of a combination of both personalized and healthy meal recommendations is optimal for greater impact in achieving user fitness goals.

Further work of this paper will include an investigation of the possibility to integrate the capability of considering medical complications of the users in the meal recommendation.

REFERENCES

- [1] F. A. O. E. C. J. Who, "Diet, Nutrition and the Prevention of Report of a Joint WHO / FAO Expert Consultation," 2017.
- [2] F. Harmon Eyre, MD Richard Kahn, PhD Rose Marie RobertsonMD, "Preventing Cancer, Cardiovascular Disease, and Diabetes."
- [3] "Micronutrient-related malnutrition," World Health Organization, 2021. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/malnutrition>.
- [4] "PlateJoy." [Online]. Available: <https://www.platejoy.com/app/personalization>. [Accessed: 20-Jan-2021].
- [5] R. N. Walter C. Willett, Jeffrey P. Koplan, "Prevention of Chronic Disease by Means of Diet and Lifestyle Changes."
- [6] A. H. J. Bobadilla, F. Ortega, "Recommender systems survey," in Knowledge-Based Syst. Vol. 46., 2013.
- [7] and J. R. J. B. Schafer, J. A. Konstan, "E-commerce recommendation applications," in Data Mining Knowl. Discovery, 2001, pp. 115–153.
- [8] R. J. S. J. M. Noguera, M. J. Barranco, "A mobile 3D-GIS hybrid recommender system for tourism," in Inf. Sci, 2012, pp. 37–52.
- [9] D. Herzog and W. Wolfgang, "RouteMe: A Mobile Recommender System for Personalized, Multi-Modal Route Planning," pp. 67–75, 2017.
- [10] "Food Supplement Personal Assistant," 2019.
- [11] D. Bianchini, V. De Antonellis, and N. De Franceschi, "PREFer: a Prescription-based Food recommender system," pp. 1–37.
- [12] C. Anderson and W. W. International, "A s f r," no. Section 3, 2018.
- [13] M. Dascalu and S. Trausan-matu, "The Runner - Recommender system of workout and nutrition for runners," no. January, 2012.
- [14] 2 and Alice Ammerman Nasim S. Sabounchi, Ph.D., 1 Hazhir Rahmandad, Ph.D., "Best Fitting Prediction Equations for Basal Metabolic Rate: Informing Obesity Interventions in Diverse Populations," 2014.
- [15] "Basal metabolic rate studies in humans: Measurement and development of new equations," PubMed.
- [16] I. De et al., "A New mHealth App for Monitoring and Awareness of Healthy Eating: Development and User Evaluation by Spanish Users," 2017.
- [17] T. M. Garvin et al., "Cooking Matters Mobile Application: a meal planning and preparation tool for low-income parents," vol. 22, no. 12, pp. 2220–2227, 2019.
- [18] R. F. Id, R. Zennun, F. Id, F. Hwang, and J. A. Lovegrove, "Evaluation of the eNutri automated personalised nutrition advice by users and nutrition professionals in the UK," pp. 1–17, 2019.
- [19] A. Nezis, P. Jiskra, and M. Pontiki, "Towards a Fully Personalized Food Recommendation Tool," pp. 3–5.
- [20] C. Celis-morales, "Personalised Nutrition: paving a way to better population health A White Paper from the Food4Me project Written by the project partners," no. April 2020, 2015.
- [21] D. De Lenguajes and S. Aranda, "A rticle recommender system for the elderly," vol. 33, no. 2, pp. 201–210, 2016.
- [22] C. Ho and Y. Chang, "Design and Implementation of Intelligent Personalized Dietary Meal Recommendation System," no. Ccme, pp. 137–140, 2018.
- [23] J. Xie and Q. Wang, "Smart Health A personalized diet and exercise recommender system for type 1 diabetes self-management: An in silico study," Smart Heal., vol. 13, no. May, p. 100069, 2019.
- [24] "Eat This Much." [Online]. Available: <https://www.eatthismuch.com/>. [Accessed: 21-Jun-2021].
- [25] "MakeMyPlate." [Online]. Available: <http://www.makemyplate.co/>. [Accessed: 21-Jun-2021].
- [26] D. Evans, "MyFitnessPal", Br. J. Sport. Med., vol. 51, no. 14, pp. 1101–1102, 2017.
- [27] L. A. A. Chun-Yuen Teng, Yu-Ru Lin, "Recipe recommendation using ingredient networks," 2011.
- [28] "Lose It!" [Online]. Available: <https://www.loseit.com/>. [Accessed: 20-Jan-2006].
- [29] L. Yang et al., "Yum-Me: A Personalized Nutrient-Based Meal Recommender System," vol. 36, no. 1, 2017.
- [30] "Nutrino." [Online]. Available: <https://nutrino.co/>.
- [31] "BNF's 7 day meal plan." [Online]. Available: <https://www.nutrition.org.uk/healthyliving/helpingyoueatwell/7-day-meal-plan.html>.
- [32] G. S. Asela Gunawardana, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks," J. Mach. Learn. Res. 10, 2009.
- [33] K. Falk, "Practical Recommender System," 2019.
- [34] J. Freyne and S. Berkovsky, "Intelligent food planning: personalized recipe recommendation," 2010.
- [35] A. Felfernig and M. Stettinger, "An overview of recommender systems in the healthy food," 2017.
- [36] A. G. Z. Z. B. Moskowitz, "Harris-Benedict equation estimations of energy needs as compared to measured 24-h energy expenditure by indirect calorimetry in people with early to mid-stage Huntington's disease," PubMed, vol. Nutritiona.
- [37] "BMR Calculator (Basal Metabolic Rate, Mifflin St Jeor Equation)." [Online]. Available: <https://www.omnicalculator.com/health/bmr>.
- [38] "Which formula are recommended by nutritionists." [Online]. Available: https://www.researchgate.net/post/which_formula_are_recommended_by_nutritionists_and_scientists_to_measure_BASAL_METABOLIC_RATE.
- [39] J. F. E. Consultation, "Human energy requirements," 2001.
- [40] I. J. Obes, "What is the Required Energy Deficit per unit Weight Loss?," 2008.
- [41] "Office of Diseases prevention and Health Promotion." [Online]. Available: <https://health.gov/dietaryguidelines/2015/guidelines/>. [Accessed: 21-Jun-2021].
- [42] "2015-2020 Dietary Guidelines," U.S. Department of Health and Human Services. [Online]. Available: ion/previous-dietary-guidelines/2015.
- [43] "Dietary Reference Intakes (DRIs)." [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK56068/table/summarytables.t4/?report=objectonly>. [Accessed: 21-Jun-2021].
- [44] F. R. et AL, "Evaluating topic coherence measures."
- [45] "Evaluating Topic Models."
- [46] D. Ribeiro, J. Machado, J. Ribeiro, M. J. M. Vasconcelos, E. F. Vieira, and A. C. De Barros, "SousChef: Mobile Meal Recommender System for Older Adults," no. Ict4awe, pp. 36–45, 2017.
- [47] S. Chen, D. Chiang, T. Chen, Y. Chung, and F. Lai, "An Implementation of Interactive Healthy Eating Index and Healthcare System on Mobile Platform in College Student Samples," IEEE Access, vol. 6, pp. 71651–71661, 2018.
- [48] N. Suksom, M. Buranarach, Y. M. Thein, T. Supnithi, and P. Netisopakul, "A Knowledge-based Framework for Development of Personalized Food Recommender System," 2005.
- [49] M. Sadat, A. Tehrani, and J. Li, "Personalized Meal Planning for Diabetic Patients Using a Multi-Criteria Decision- Making Approach," 2019.
- [50] N. R. Lim-cheng, G. I. G. Fabia, M. E. G. Quebral, and M. T. Yu, "Shed: An Online Diet Counselling System," pp. 1–7, 2014.
- [51] S. Menal-puey and M. Mart, "Developing a Food Exchange System for Meal Planning in Vegan Children and Adolescents," pp. 1–14, 2019.
- [52] M. Harvey, "Automated Recommendation of Healthy, Personalised Meal Plans," pp. 327–328, 2015.
- [53] D. Elswailer and M. Harvey, "Towards Automatic Meal Plan Recommendations for Balanced Nutrition," pp. 313–316.
- [54] K. Nangung, T. Kim, and Y. Hong, "Menu Recommendation System Using Smart Plates for Well-balanced Diet Habits of Young Children," vol. 2019, 2019.
- [55] B. Ramzan et al., "An Intelligent Data Analysis for Recommendation Systems Using Machine Learning," vol. 2019, 2019.
- [56] M. Hane and A. Mashfiqui, "MyBehavior: Automatic Personalized Health Feedback from User Behaviors and Preferences using Smartphones," 2015.
- [57] R. Pop, M. Pop, G. Dogaru, and V. C. Bacarea, "A Web-based Nutritional Assessment Tool," vol. 22, no. 3, pp. 307–314, 2013.
- [58] J. Freyne and S. Berkovsky, "Intelligent Food Planning:

- Personalized Recipe Recommendation,” pp. 321–324, 2010.
- [59] M. Harvey, B. Ludwig, and D. Elswiler, “You Are What You Eat: Learning User Tastes for Rating Prediction,” pp. 153–164, 2013.
- [60] Summary of Initial Survey.” [Online]. Available: https://docs.google.com/spreadsheets/d/1HJEvJqpeU9YvNkl_4PJY1s5tiYA0AOIga39O5az93Uw/edit?usp=sharing.

Deep learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka

Siventhirarajah Sangeevan*
Software Engineering Teaching Unit
Faculty of Science, University of Kelaniya, Sri Lanka
sangeevan1995@outlook.com

Abstract - The study proposes a deep learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka. It is an intelligent system to get suitable pesticides prescriptions for plant leaf diseases. Home gardening has become popular and is rapid because of the current pandemic situation. However, plant diseases are a major problem in gardening activities, even in a home garden or in a commercial garden. Identifying and finding a solution for the plant disease is a big challenge for home gardeners rather than commercial farmers. The proposed system of deep learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka will be the best solution for identifying and finding a solution to the plant diseases. The system is using a trained model for prescribing pesticides. The model was built using the deep learning method and trained in the supervised learning process. The convolutional neural network algorithm was used in the model. Transfer learning with AlexNet pre-trained model was used to increase the performance in the proposed solution and the best accuracy of 88.64% was achieved in the experiments.

Keywords - convolutional neural network, leaf diseases, Machine Learning, pesticides

I. INTRODUCTION

Agriculture is one of the major livelihoods in Sri Lanka. People are engaging in cultivation in commercial gardens and also at a smaller level, in home gardens. While engaging in gardening, diseases to the crops are one of the major problems. Commercial farmers may have some knowledge of crop diseases and pesticides, but home gardeners do not have much knowledge of them. Even in some cases commercial farmers also fail to identify some diseases. So, in that situation, both must consult with some agricultural experts to find a solution. Home gardeners, however, don't have the luxury of time to spend consulting experts to find solutions to these diseases. So, they may search and find some unsuitable pesticides through the Internet or somewhere else and spend their money on it. In most cases, this may not work and thus demotivated, may even leave their home gardening activity. A smart solution to solve this problem may be feasible.

Deep learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka is a smart solution for this problem. A person without proper knowledge of crop diseases and pesticides also can use this system. Using this system, a user can simply input an image of a leaf that was affected by the disease and get the appropriate pesticide to cure that disease, as the output. Some diseases can't be easily identified by even an experienced farmer, so it will be a challenging thing for home gardeners. But this system can easily identify the disease and prescribe the pesticide as well. The system

mostly focuses on home garden crops, but whatever the crop in the home garden, it is also cultivated in commercial gardens. So, the system is not limited to a home garden, it can be used in a wide range like home gardens and larger gardens as well. Home gardeners would be most benefited by this proposed system.

The proposed system is using a trained model for prescribing pesticides. The model was built using the deep learning method and trained in the supervised learning process. The convolutional neural network algorithm was used in the model. The transfer learning method was used to increase the performance of the model. AlexNet was used as the pre-trained model for the transfer learning process. So, using this system the users can easily get the suitable and correct pesticide prescription for the leaf diseases.

II. LITERATURE REVIEW

In [1] authors evaluate the applicability of deep convolutional neural networks for the classification of plant diseases. They focused on two popular architectures, namely AlexNet and GoogLeNet. They analysed the performance of both these architectures on the PlantVillage dataset by training the model from scratch in one case, and then by adapting already trained models using transfer learning. In the case of transfer learning, they re-initialize the weights of layer fc8 in the case of AlexNet. They have achieved an accuracy level of 99.35%.

In [2] authors proposed a deep convolutional neural network model based on AlexNet and GoogLeNet to identify apple leaf diseases. The AlexNet gave a good recognition ability and obtains an average accuracy of 91.19%.

In [5] authors proposed a plant disease identification model framework based on deep learning. The RPN algorithm is used to train the leaf dataset in the complex environment, and the frame regression neural network and classification neural network is used to locate and retrieve the diseased leaves in the complex environment. The Chan-Vese algorithm is used to segment the image of diseased leaves. Resnet-101 was selected as the pretraining model, and the network is trained by using the dataset of disease leaves under a simple background. According to the comparison results, the average correct rate of their proposed method is 83.75%.

In [7] authors used the K-means clustering method for the segmentation of the image. They implemented their proposed methodology using Optimized Deep Neural network with Jaya algorithm in Python platform. The performance of their proposed method DNN-JOA is estimated and compared with the performance of existing classifiers such as ANN, DAE, and DNN. Using the

DNN_JOA classifier the highest accuracy is achieved for the blast affected leaf image which is 98.9%.

In [8] authors used K-means clustering, Support Vector Machine, and advance neural network for making an image classification model. K-Means algorithm is used to cluster the images, and then multiclass SVM is used for the classification process. The average accuracy of the classification of the proposed method is 95.83%.

III. METHODOLOGY

The system of Deep learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka has a trained model to prescribe pesticides for leaf diseases. So, in the proposed solution, the deep learning method was used, and the model was trained by a supervised learning approach. The convolutional neural network is a kind of deep neural network and it's commonly used to analysing visual imagery. It's a regularized version of multilayer perceptron. As the research is based on analysing the images, in the model training, Convolutional Neural Network has been used and the transfer learning technique also has been used to get the advantage of the AlexNet model. The high-level architecture diagram of the proposed system is shown in Fig. 1.

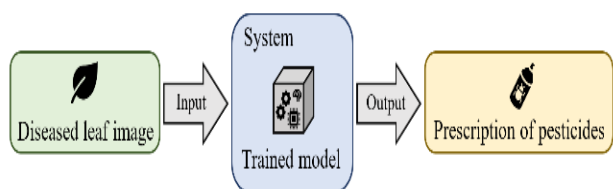


Fig. 1. The high-level architecture of the proposed system

Using transfer learning to train a model is more efficient than training a model from the scratch, and transfer learning has a higher start, higher slope, and higher asymptote. So, in the proposed system, the transfer learning technique has been used to increase the performance level and save time. The performance graph of the model with transfer learning and without transfer learning is shown in Fig. 2.

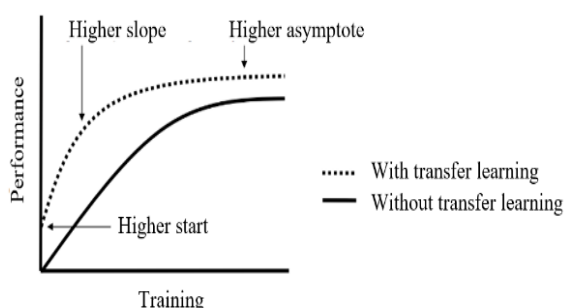


Fig. 2. Performance graph of learning types

The AlexNet model was used as the pre-trained model in the transfer learning process because the AlexNet is one of the best models which trained with a huge amount of data. The architecture of the AlexNet model is shown in Fig. 3. During the model training in the proposed system, the last layer of the AlexNet was reshaped and trained.

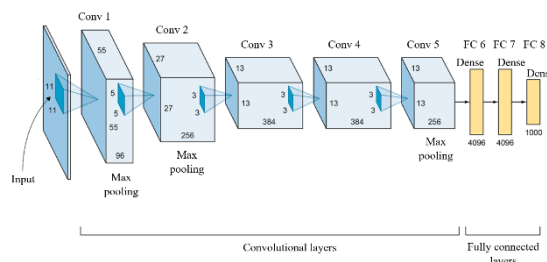


Fig. 3. The architecture of the AlexNet model

The dataset of images for training the model was collected from the Internet. As the research is focusing on Sri Lankan plants, it's difficult to find many plant types on the Internet. So, here only three types of plants and only twelve different diseases of those plants are used for the research. The images are in RGB colour format, and the size of the images is 256 x 256. Nineteen thousand and sixty-four leaf images were collected as the dataset. The dataset has twelve different diseases on three types of plants. Dataset also has healthy leaf images of those three plants. So, in the dataset, there are fifteen different types of classes available. Some sample images from the dataset are shown in Fig. 4.



Fig. 4. Sample images from the dataset

The data of pesticide details are also needed for model training to be used as the labels. Pesticides are called chemical control for plant diseases. Some diseases can't be cured by applying any chemicals. In this case, if the disease is severe, it must remove the affected plant from the garden. There will be a different chemical to control each leaf disease of crops. The chemical control methods of the selected twelve leaf diseases have been collected from the datasheets on the Internet. To control the leaf disease, the gardener must use a pesticide that contains the chemical which can control the specific disease. Then using the chemical control method details, the suitable pesticide details are also collected from the Internet. Since there are a lot of pesticide brands available, brands available in Sri Lanka should be identified. Thereafter, the prescribed pesticide can be locally bought by the customers. According to the findings in Table I, eleven diseases can be controlled by pesticides and one disease has no chemical control. The other three classes are healthy which do not need any usage of pesticides. So, now the dataset has nineteen thousand and sixty-four leaf images which fall under fifteen classes and has the pesticides data with the suitable chemical control methods of those diseases.

The experiment was done in the Jupyter Notebook editor using Python programming language. The model training was done using PyTorch open-source machine learning library. In addition to that, some python libraries also have been used.

Another important thing in the machine learning experiment is dataset preparation. The dataset has to be split for training, validation, and testing. So, the dataset was

split as 80% for training, 10% for validation, and 10% for testing.

TABLE I. DISEASE AND PESTICIDE DETAILS

Plant	Leaf Disease	Chemical Control	Pesticide
Bell pepper	Bacterial spot	Copper fungicide	Manar Maneb
Bell pepper	Healthy	No chemical needed	No pesticides needed
Potato	Early blight	Mancozeb	Hayleys Mancozeb
Potato	Healthy	No chemical needed	No pesticides needed
Potato	Late blight	Mancozeb	Hayleys Mancozeb
Tomato	Bacterial spot	Copper fungicide	Manar Maneb
Tomato	Early blight	Mancozeb	Hayleys Mancozeb
Tomato	Healthy	No chemical needed	No pesticides needed
Tomato	Late blight	Mancozeb	Hayleys Mancozeb
Tomato	Leaf mold	Chlorothalonil	Ronil Chlorothalonil
Tomato	Mosaic virus	No chemical control	No pesticides available
Tomato	Septoria leaf spot	Chlorothalonil	Antracol Propineb
Tomato	Target spot	Chlorothalonil	Antracol Propineb
Tomato	Two-spotted spider mite	Abamectin	Mig Abamectin
Tomato	Yellow leaf curl virus	Imidacloprid	Kobra Imidacloprid

The training dataset will be used for training the model. The validation dataset will be used for frequent unbiased evaluation of the model. This will be used to fine-tune the model's hyperparameters. The test dataset will be used to do the unbiased evaluation of the final trained model after the completion of training.

The model was trained in the system which has the CPU configuration of Intel(R) Core (TM) i7=4510U CPU @ 2.0GHz and the memory configuration of 8.0 GB DDR3. The dataset has a huge number of files, with nearly a thousand images per class. It is a very time-consuming task with the normal CPU. To minimize the training time, it must use a GPU. In the model training, the GPU of NVIDIA GeForce 840M was used in the system.

In the dataset, the images may not be of the same size. Most neural networks expect a fixed image size. So, it must transform the image into a specified size before loading the data to train the model.

In the proposed system, the transfer learning method was used and the AlexNet model was used as the pre-trained model for that. Therefore, it must reshape the last layer of the AlexNet before training. In the proposed system, currently, there are fifteen classes. According to

that, while initializing the model, it must reshape the number of neurons in the last layer to fifteen.

In neural network training, the optimizer is used to change the attributes like weight, biases, and learning rate to reduce the loss. It makes the training process fast.

The important thing in the experiment is the model training. Training the model means, learning the best values for the weights and bias from the examples. In supervised machine learning, the algorithm builds a model by examining many examples and try to find a model that minimizes loss. The model was trained with five hundred epochs. The learning algorithm will find the pattern in the training data that map the input data attributes to the target, and it outputs a model that captures these patterns.

IV. RESULTS AND DISCUSSION

The training process took six hundred and twenty-nine minutes and thirty-five seconds to finish the five hundred epochs. As a result of the experiment, the best accuracy of 88.64% was achieved during the training. The accuracy change over the number of the epoch is shown in Fig. 5, and the loss change over the number of the epoch is shown in Fig. 6.

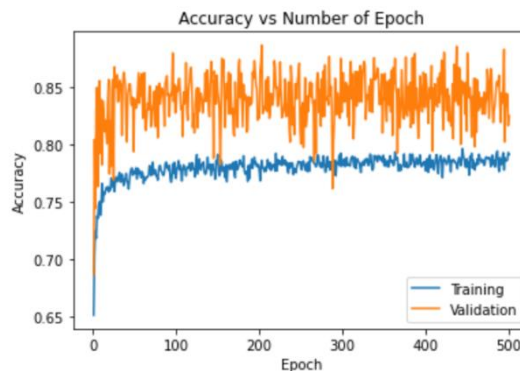


Fig. 5. Accuracy graph of training

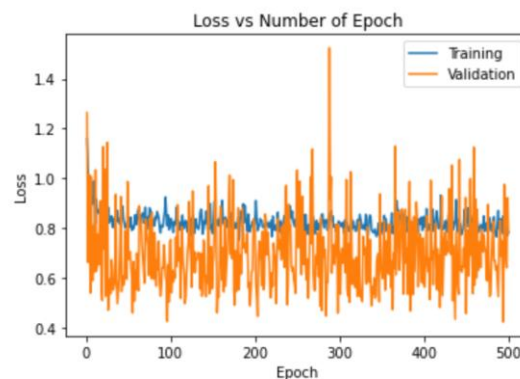


Fig. 6. Loss graph of training

The evaluation process is an important step in machine learning experiments. Through the evaluation process, one can find how well the trained model is performing. There are some evaluation metrics to measure the quality of the machine learning model. To evaluate the model, the test dataset can be used. This set of data is fully new and unseen data to the model. So, unbiased results of the evaluation can be gained from this. So, for the evaluation process, first, a confusion matrix is obtained as shown in Table II.

TABLE II. CONFUSION MATRIX FOR TEST RESULT

9	4	1	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	9	0	0	0	0	0	0	0	0	1	0	0	0
1	0	8	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	5	1	0	0	0	0	0	0	0	0	0
0	0	9	7	6	3	0	0	8	1	0	3	1	4	0
0	0	0	0	0	2	0	0	0	0	0	2	4	0	1
0	0	0	0	0	7	0	0	0	0	0	0	0	0	0
3	0	0	0	1	1	4	1	9	4	0	7	1	3	4
0	1	0	0	0	0	0	0	1	5	0	0	0	0	0
0	0	0	0	4	3	2	4	1	5	9	5	0	3	5
0	0	0	0	0	1	0	1	1	8	5	0	2	1	1
0	0	0	0	0	0	0	0	0	0	0	3	7	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	3	1	0	0	1	0	1	1	1	3	1	6	3	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	1	7	0	0	0	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	8	0	0	0	0	1	3	6
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Then to evaluate the model, the accuracy, precision, recall, and F1 score must be calculated.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total number of predictions}} \quad (1)$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (3)$$

$$\text{F1} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

According to the results shown in Table III, for each evaluation matrices, the performance of the model can be evaluated.

TABLE III. EVALUATION OF RESULTS

	Precision	Recall	F1 score
Class 0	0.92	0.95	0.94
Class 1	0.94	0.99	0.97
Class 2	0.90	0.98	0.94
Class 3	0.68	0.94	0.79
Class 4	0.91	0.64	0.75
Class 5	0.92	0.97	0.94
Class 6	0.96	0.48	0.64
Class 7	0.83	0.99	0.91
Class 8	0.89	0.83	0.86
Class 9	0.88	0.89	0.88
Class 10	0.93	0.97	0.95
Class 11	0.90	0.92	0.91
Class 12	0.73	0.78	0.75
Class 13	0.86	0.80	0.83
Class 14	0.89	0.99	0.94

Macro avg	0.88	0.87	0.87
Accuracy	0.88		

The experiment was carried out by using different methods to find a better solution for the dataset. During the experiments, the CNN model was trained from scratch by using the selected sample from the dataset with a selected number of epochs. The accuracy and loss graph of the CNN model training is shown in Fig. 7 and Fig. 8. The model using transfer learning with AlexNet was trained by using the same selected dataset sample with the same selected number of epochs. The accuracy and loss graph of the transfer learning with AlexNet model training is shown in Fig. 9 and Fig. 10. Using the same selected sample dataset, the experiment was also carried out by using the SVM classifier. The accuracy result of each experiment is shown in Table IV.

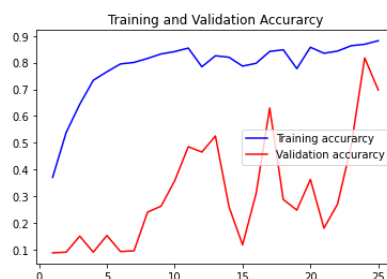


Fig. 7. Accuracy graph of CNN training

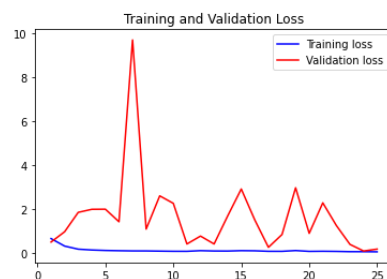


Fig. 8. Loss graph of CNN training

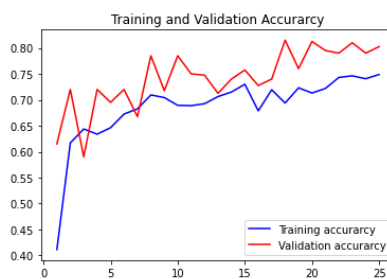


Fig. 9. Accuracy graph of AlexNet transfer learning

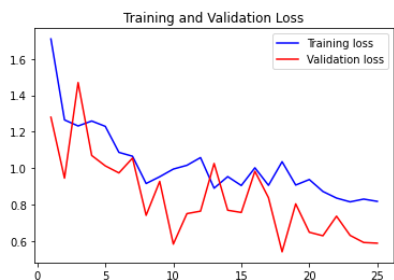


Fig. 10. Loss graph of AlexNet transfer learning

TABLE IV. ACCURACY OF EXPERIMENTAL RESULT

	Accuracy
CNN from scratch	69.75%
SVM	73.13%
Transfer learning with AlexNet	81.50%

According to the above results, transfer learning with AlexNet gave the best accuracy level. Therefore, this will be the best fitting method for the dataset. In the final implementation for the proposed system, the model training was carried out by using the method of, transfer learning with the AlexNet model. An example input of a diseased leaf image to the implemented system is shown in Fig. 11, and the pesticide prescription from the system for that input is shown in Fig. 12.

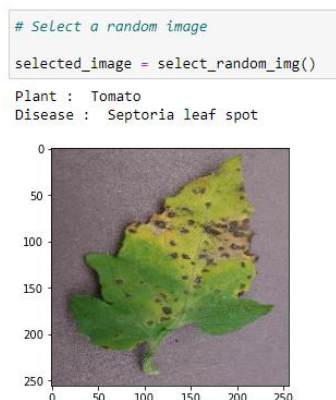


Fig. 11. The sample input image to the system

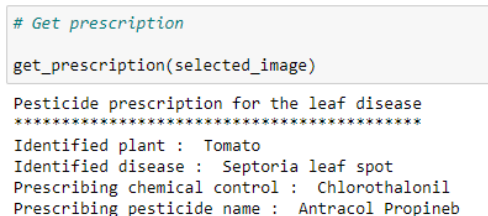


Fig. 12. A prescription from the system for the input

V. CONCLUSION

As Sri Lanka is an agricultural country, a solution for leaf diseases is an important thing. According to the current pandemic situation, the need for a smart solution emerged. To solve this issue using a computerized system, multiple machine learning models were trained and tested. According to the experimental results, the proposed system

is performing well in prescribing the most suitable pesticide for leaf diseases. There may be some existing systems to predict plant diseases, but the proposed system directly predicts suitable pesticides and it's a localized system for Sri Lanka. So, the proposed system, Deep learning-based pesticides prescription system for leaf diseases of home garden crops in Sri Lanka will be a great solution.

REFERENCES

- [1] Sharada P. Mohanty, David P. Hughes and Marcel Salathe, "Using Deep Learning for Image Based Plant Disease Detection", *Frontiers in Plant Science*, vol. 7, pp. 1419, September 2016.
- [2] Bin Liu, Yun Zhang, Dong Jian He and Yuxiang Li, "Identification of Apple Leaf Diseases Based on Deep Convolutional Neural Networks", *Symmetry*, vol. 10, issue 1, pp. 11, December 2017.
- [3] Justine Boulent, Samuel Foucher, Jerome Theau and Pierre Luc St Charles, "Convolutional Neural Networks for the Automatic Identification of Plant Diseases", *Frontiers in Plant Science*, vol. 10, pp. 941, July 2019.
- [4] Muammer Turkoglu and Davut Hanbay, "Plant Disease and Pest Detection Using Deep Learning Based Features", *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, issue 3, pp. 1636-1651, May 2019.
- [5] Yan Guo, Jin Zhang, Chengxin Yin, Xiaonan Hu, Yu Zou, Zhipeng Xue, and Wei Wang, "Plant Disease Identification Based on Deep Learning Algorithm in Smart Farming", *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 2479172, August 2020.
- [6] D. K. N. G. Pushpakumara, B. Marambe, G. L. L. P. Silva, J. Weerahewa and B. V. R. Punyawardena, "A Review of Research on Home Gardens in Sri Lanka: The Status, Importance and Future Perspective", *Tropical Agriculturist*, vol. 160, pp. 55-125, August 2012.
- [7] S. Ramesh and D. Vydeki, "Recognition and Classification of Paddy Leaf Diseases Using Optimized Deep Neural Network with Jaya Algorithm" *Information Processing in Agriculture*, vol. 7, issue 2, pp. 249-260, June 2020.
- [8] Nafees Akhter Farooqui and Ritika, "An Identification and Detection Process for Leaves Disease of Wheat Using Advance Machine Learning Techniques", *Bioscience Biotechnology Research Communications*, vol. 12, issue 4, pp. 1081-1091, December 2019.
- [9] Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, Dubravko Culibrk and Darko Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification", *Computational Intelligence and Neuroscience*, vol. 2016, pp. 3289801, June 2016.
- [10] Xiaoyue Xie, Yuan Ma, Bin Liu, Jinrong He, Shuqin Li and Hongyan Wang, "A Deep Learning Based Real Time Detector for Grape Leaf Diseases Using Improved Convolutional Neural Networks", *Frontiers in Plant Science*, vol. 11, pp. 751, June 2020.

What makes job satisfaction in the information technology industry?

Nimasha Arambepola*
Software Engineering Teaching Unit
Faculty of Science, University of Kelaniya, Sri Lanka
nimasha_2019@kln.ac.lk

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

Abstract - Having a rich human resource is critical for an organization to move towards success. Especially, for business organizations such as technology companies, the human resource is the driving factor of the company's growth which depends on employees' motivation, skills and quality of work. Employees often change their jobs when they are not satisfied with it. Different factors may cause a change in the level of job satisfaction of an employee. For example, the dynamic nature of the Information Technology (IT) industry is an impactful factor that determines the job satisfaction of IT professionals. Foreseeing the employees' job satisfaction makes it easy for a company to take swift actions to improve the job satisfaction of its employees. In this research, we analyzed the effectiveness of machine learning (ML) methods for predicting job satisfaction using employee job profiles. There are job-specific factors in each job domain, and those factors may influence job satisfaction levels. Therefore, this research focused on the following fundamental questions: 1) How do existing ML models perform when predicting job satisfaction of software developers? 2) Can the job satisfaction prediction models be generalized to the other job roles in the IT industry? This study compared the performance of classification models: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Neural Network (NN) in predicting the level of job satisfaction. Our experiments used two benchmark datasets: Stack Overflow developer survey and IBM HR analytics dataset. The experimental analysis shows that both employee-related factors and company-related factors contribute similarly to predicting job satisfaction. On average, the above ML models predict the job satisfaction of software developers with an accuracy of around 79%.

Keywords - classification models, data mining, job satisfaction, machine learning

I. INTRODUCTION

Human resource is the most important factor for the success of any organization. Therefore, most organizations and companies are seeking talented, knowledgeable and experienced candidates for their job openings. Due to the technological advancements and complex lifestyles of the modern people, the current job market shows rapid changes. New jobs are being created, and some of the existing jobs have been taken over by new technologies. For example, robots are serving at some of the airports to do certain tasks which were performed by human employees. With these drastic changes, employees are migrating to demanding jobs to discover their passion and to satisfy their life expectations. Job satisfaction is an important aspect due to the fact that it represents an overall summary of how an individual feel about a lifetime of work [1]. Therefore, job satisfaction can be described as a pleasurable or positive emotional state from the appraisal in any field of interest. Employees who are satisfied with their jobs have the enthusiasm to drive the company

towards success while improving themselves. Thus, employee job satisfaction is a vital factor that needs to be considered in the recruitment process. However, it is a challenging task to select the most suitable candidate from a plethora of applicants. There are popular filtering mechanisms used in human resource departments, which are mostly manual processes. For example, filtering candidates based on different factors in their resumes such as working experience and educational background. Owing to the new technologies and innovations, companies are moving towards novel techniques to make decisions regarding new recruits [2]. If the Human Resource (HR) managers can foresee the job satisfaction of a person, it will bring numerous benefits in terms of competitive advantage and efficiency in the recruitment process. On the other hand, it is beneficial for the employees to choose jobs with high job satisfaction. Different factors may influence the level of job satisfaction of an employee. For example, social, cultural and political factors such as employee salary, age, education level, and the complexity of the work to be done are some of the main influential factors for the level of job satisfaction. Nevertheless, the causes of employee job satisfaction or dissatisfaction mainly depend on the field that the employee works.

Various methods are available to predict job satisfaction. However, to the best of our knowledge, existing research is not focusing on predicting job satisfaction using machine learning (ML) techniques considering both employees' background data and company-related factors. In this research, we analyzed the performance of several ML models based on two case studies namely Stack Overflow developer surveys [12][26][27] and IBM HR analytics [25]. Different features extracted from the Stack Overflow developer survey were used to predict the job satisfaction of software developers. Then the study was extended to generalize the prediction model for predicting job satisfaction of other job roles in the IT industry (generalized model). Features extracted from IBM HR analytic dataset were used for the generalized model. We considered four different classifiers namely Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Neural Network (NN) for the prediction models. The objectives of this research are as follows:

- Studying the available approaches for predicting job satisfaction.
- Identifying the main influential factors of job satisfaction of the IT professionals.
- Exploring the possibility of generalizing job satisfaction prediction models to other job roles in the IT industry.

The rest of the paper is organized as follows. In the following section, we discuss a selected set of existing research studies related to our topic. In section 3, we present our empirical analysis of prediction models and the performance of each model. Then the findings are discussed with the results of the experimental analysis. Finally, we conclude this paper with the future directions of the research.

II. RELATED WORK

The advancement of internet technologies has allowed acquiring insights for accurate decision making through analytical formulas and data processing techniques. Employee related decision making such as employee turnover, attrition and job involvement is a crucial task as optimistic employees are the key success factors of a company. Therefore, recent researchers have focused on exploring the applicability of ML models in employee-related decision making [2]-[6]. Job satisfaction data mining has been widely used to extract meaningful knowledge about employee satisfaction. This approach is applicable for various domains and contexts in predicting the satisfaction level, identifying the most affecting factors of job satisfaction and taking remedial actions to improve the performance of employees [7] [8]. Even though both job satisfaction and career satisfaction are related to global life satisfaction, these two are independent of each other. For instance, while career satisfaction is related to turnover intention and leaving in the IT field, job satisfaction of IT professionals is highly related to employee turnover, which is a persistent problem in the IT industry [1]. It has shown that the level of job satisfaction strongly affects the turnover intention of software developers [3]. Employee job satisfaction is based on both objective and subjective data [9]. For example, a research study has been carried out to find the impact of family factors and the role of work in predicting career satisfaction. It was evaluated by collecting data from 344 participants through an online survey. In there, hypothesis testing has confirmed that there is a significant relationship between job satisfaction and work-family balance in improving the level of job satisfaction [10]. Many companies collect and keep employee records and data to study their job satisfaction. However, influential factors of job satisfaction may differ based on the industry, job role as well as the country and the region. For example, recent research has shown that personal development opportunities, relationship with the supervisor, and adherence to the duty roster are the most important factors for job satisfaction in the hospitality industry in the Alpine region [28].

A considerable number of research studies have been conducted using the data extracted from Stack Overflow as it is a world popular Q&A platform for software developers. However, the majority of them are related to the questions and answers [11],[12] posted in the Stack Overflow website rather than the Stack Overflow developer survey responses. Most of the existing research studies on predicting career/job satisfaction in different disciplines have used mostly statistical analysis methods rather than using sophisticated ML techniques. Therefore, it is worth exploring the potential of ML methods in predicting job satisfaction. ML is a branch of Artificial Intelligence (AI) that learns and improves automatically through experience. In there, classification is a supervised

ML approach that uses label data to train the model which used to predict the labels of unknown examples. ML algorithms have been used for predicting, classifying and clustering various kinds of data in different domains and industries such as healthcare, financial and marketing. Most of the previous ML-based forecasting have been conducted as empirical analysis by comparing the performance of existing ML models. For example, a research study has examined the performance of the five existing classification algorithms when predicting the likelihood of hospital readmission [13]. According to their comparison, SVM has shown the best performance among the chosen algorithms, while the results of LR and Naïve Bayesian (NB) are lower than the other classifiers. Besides job satisfaction, satisfaction level prediction is another area of forecasting that has applied in different domains. For example, the customer satisfaction level prediction is used to improve products and services. Since companies are not only relying on product quality but even more on a service quality level, there is a significant need for identifying the customer satisfaction level. Thus, a research study of predicting customer dissatisfaction has been carried out using five existing classification models [14].

Ensemble ML algorithms such as RF are widely used for both classification and regression problems due to their excellent accuracy, ease of use and robustness. This is because the method of combining multiple independent learning algorithms increases the predictive performance that could be obtained from any of the single learners alone. To reduce the learning time and the computational cost, the fast algorithms such as decision trees are widely used in ensemble methods [15]. Binary classification is the most commonly used classification type where the target variable has only two classes. Researchers have shown that decision trees and NN perform well in binary classification through several studies [16] [17]. In addition, SVM, Decision Tree, RF and NB can be used as multi-class classifiers. For instance, student academic performance prediction using their academic progress, personal characteristics and behaviors relating to learning activities [18] [19] are two case studies which have used multi-class classification. Therefore, classification models are ideal for predicting the level of job satisfaction.

III. JOB SATISFACTION OF IT PROFESSIONALS

According to the recent analysis, healthcare and information technology(IT) related jobs are the top-rated jobs in the world. As a result, there has been tremendous growth in the software and IT industry over the last few years. Software development ranked as a top demanding job and software engineering has been rated as one of the rapidly expanding sectors in the world. Although the demand for software developers is nothing new, it has seen a significant rise in the last couple of years. According to the predictions, employment of software developers will increase by 22% from 2019 to 2029, which is much faster than the average of all other occupations [20]. Therefore, more employees are moving into the software development industry. However, IT related job specific factors may influence the level of job satisfaction of IT employees. For example, since software development is often a deadline-oriented process, the level of stress among software developers tends to be high. This is especially common among the less experienced developers. Moreover,

adapting to rapidly changing cutting edge technologies is one of the most challenging tasks for software developers. Even though the idea of flourishing happiness among developers is often promoted by software companies, because of the above reasons, the IT industry has become the industry with the highest turnover rate in 2018 [21]. Therefore, the present work analyzed the factors which influence the job satisfaction of IT employees. This experimental analysis consists of three tasks which are shown in Fig. 1.

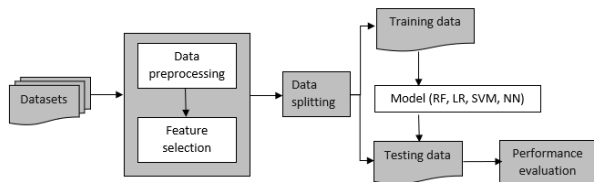


Fig. 1. Proposed methodology

In the first task, we retrieved data from the data sources and preprocessed to remove the noise. The second task was feature engineering and selecting the most discriminative features for training ML models. Finally, the prediction performances of the trained ML models were tested.

A. Data

The ever-increasing volumes of data and information shared on social media and collaborative sites have become a rich and valuable source of knowledge for a wide spectrum of research needs. When there is a need to learn about a new topic or to answer a particular query, people look for fast access to relevant information sources that would help them address that need. In the IT industry, software developers often visit online question and answering (Q&A) sites to find answers for their coding problems. Stack overflow is a well-known free Q&A website for IT professionals and enthusiastic software developers. Each year, Stack overflow collects data from the software developer community and makes the anonymized data available for researchers and other interested parties. This is named as the "Stack Overflow developer survey" which provides highly accurate data about software developers all around the world. Hence, we choose Stack Overflow developer survey datasets (dataset1) which have been released recently in three consecutive years: 2018, 2019 and 2020 [12] [26] [27]. The dataset1 was used for training the job satisfaction prediction model for software developers. It is mainly composed of categorical data such as Country, Developer Type, Gender, etc. Researchers use this public dataset for retrieving insights of the behavior of IT employees [22]. In 2018, they published their Annual Developer Survey results for the eighth consecutive year with the largest number of respondents yet [23]. Responses have been collected in January 2018 and nearly 100,000 developers have responded to this 30-minutes survey. Apart from the Stack overflow dataset, International Business Machines (IBM) HR analytic dataset (dataset2) [25] was used to analyze job satisfaction of both IT and non-IT employees in the IT industry. This dataset consists of job-related features common for employees in many industries such as age, job role, monthly income, education, etc. The volume/size and the number of features of each dataset before the preprocessing stage are shown in Table I.

TABLE I. COMPOSITION OF EACH DATASET

Dataset	Size	Features
StackOverflow developer survey 2018	98,855	129
StackOverflow developer survey 2019	88,883	85
StackOverflow developer survey 2020	64,461	61
IBM HR analytic	1,500	35

B. Experimental design

When datasets become bigger in both volume and variety with a large number of features, it is necessary to apply ML techniques to extract patterns and knowledge from the data. Effective data preprocessing and feature engineering techniques are vital for better performance of ML models. Hence, this study used two-dimensionality reduction techniques to select features from the dataset. First, unique identifiers such as "response_id" were removed from datasets as they do not hold any significant importance to the analysis. Then the features which have more than 50% of missing values were removed. Considering the RF feature importance, 53 features were selected from dataset1 to train ML models for predicting the job satisfaction of software developers. Only 33 features were considered among 35 features in dataset2 to train the generalized model to predict job satisfaction of other job roles in the IT industry. Since the majority of the selected features were categorical, missing values in the selected features were replaced with the mode. Even though the chosen ML algorithms are robust to the overfitting problem, we removed the classes with fewer frequencies in some features such as gender. For example, we considered the users whose gender is either male or female and removed the other gender categories which have very few examples in the dataset1. Since most of the ML algorithms accept only numerical data, categorical data were converted into numerical values using Label encoding. The label is job satisfaction in both scenarios. It has seven classes in the dataset1 namely, Extremely Satisfied, Moderately satisfied, Slightly satisfied, Neither satisfied nor dissatisfied, Slightly dissatisfied, Moderately dissatisfied and Extremely dissatisfied. However, we made it into three classes for better performance by grouping the first three classes into one class called 'Satisfied' and the last three classes into one class called 'Dissatisfied' and remaining the class 'Neither satisfied nor dissatisfied' as it is. The label has four classes in the dataset2, but were made into three classes. After the preprocessing and feature engineering stages, datasets were split into training and testing data such as 80% of the dataset for training the classifiers and 20% for testing.

In this research, we used supervised classification methods. Since the output variable has more than one class in both scenarios, the multi-class classification technique was used to classify job satisfaction. We compared four different predictive models namely, Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Neural Network (NN) to see the difference of the performance in predicting job satisfaction of software developers. These four algorithms have been selected due to their flexibility in handling a range of classification problems with a large feature space [15].

IV. RESULTS AND DISCUSSION

This section discusses the results of the experiments and the limitations of the study with future directions. After removing the noisy data in the preprocessing stage, we considered 97,869 records, 87,740 records and 63761 records for this study from Stack Overflow developer survey 2018, 2019 and 2020 respectively. The total number of 1472 records were considered from the IBM HR analytic dataset to train and test the generalized model for predicting job satisfaction of both IT and non-IT employees in the IT industry. The variable or the feature importance provides the statistical significance of the variables in the dataset. This is very important when using the multi-class classification methods to make predictions as it can be used to identify whether the selected features contribute or do nothing in classification with the chosen ML models. In this experiment, a total number of 53 features were selected as the most important features from Stack Overflow developer survey datasets for predicting the job satisfaction of software developers. A total number of 33 features were selected as the most discriminate features for predicting the job satisfaction of other job roles in the IT industry

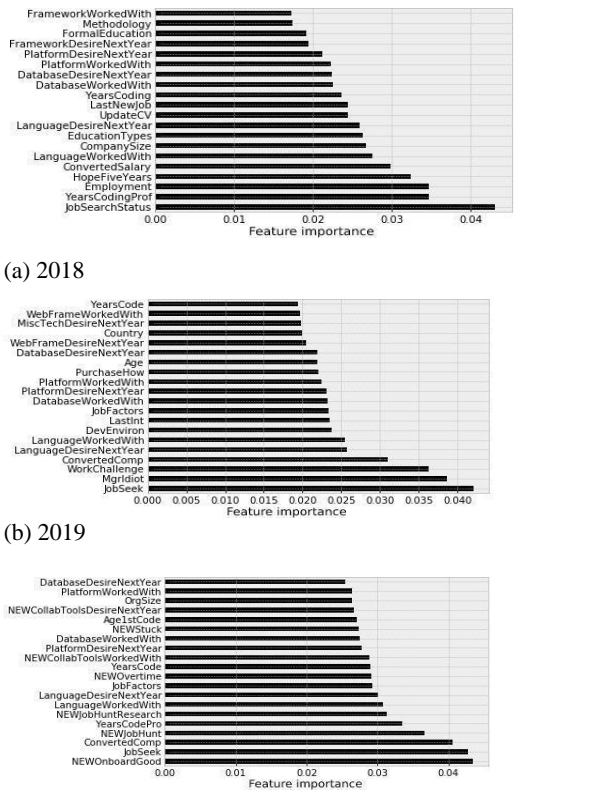


Fig. 2. Feature importance of Stack

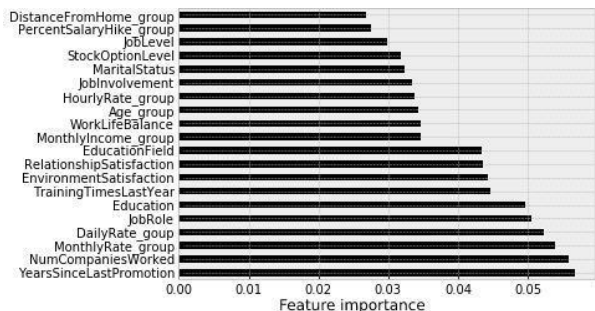


Fig. 3. Feature importance of IBM HR analytic dataset

According to the descriptive statistical analysis and the graphs of feature importance (figure 2 & figure 3), some variables in the dataset are less significant than some other variables for predicting the level of job satisfaction. For example, it shows that the contribution of the feature, “jobSeek” is one of the most significant features. In addition, graphs shown in Fig. 2 show the variations of job satisfaction influential factors for software developers in past consecutive years. Overall, common main influential factors for deciding the level of job satisfaction of software developers are as follows:

- Availability of training and managerial support
- Monthly income
- Years of coding experience
- Company size
- Challenges in workplace
- Programming languages work with
- Platforms work with

Following are the key factors to decide the level of job satisfaction of both IT and non-IT job roles in the IT industry as shown in figure 3.

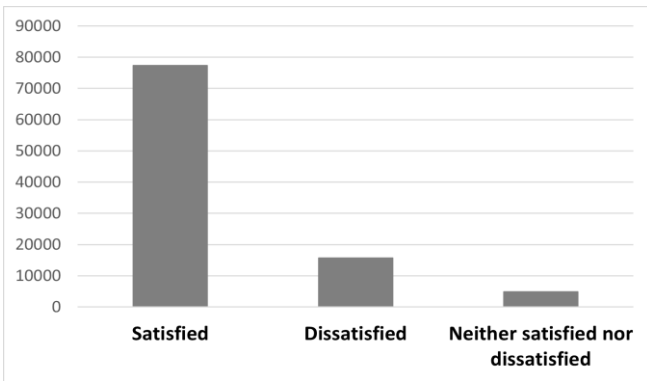
- Promotions
- Number of companies worked with
- Monthly income
- Job role
- Education
- Training
- Environmental satisfaction
- Relationship satisfaction

Feature important graphs shown in figure 2 & figure 3 show that both employee-related factors and company-related factors are contributing similarly when deciding the level of job satisfaction of IT employees. For example, providing training opportunities, monthly income, promotions and company environment are a few of the factors that HR managers and companies could directly involve to seed a high level of job satisfaction among employees

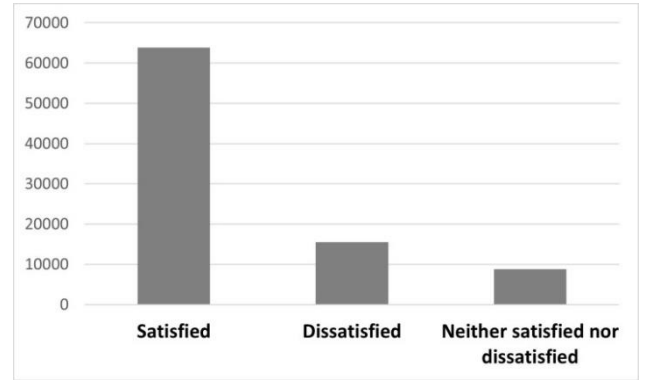
The problem of class imbalances causes a decrease in the accuracy of the predictive models. There are minority classes in the labels of all datasets. Thus, we reduced the classes into three by aggregating similar classes. Although the reduction of the number of classes increased the data in each class, the class imbalance still presents as shown in Fig. 4. Therefore, Synthetic Minority Oversampling Technique (SMOTE) [24] was used to synthesize new examples for the minority classes. After reducing the number of classes and removing the class imbalance, the accuracy of each model increased. For example, RF model performance comparison with 7 classes with the class imbalance, and with 3 classes after applying SMOTE is shown in Table II. Accuracy is not a good indicator of model performance in this study due to class imbalances. Because it is biased as the rare classes can be masked by

the majority classes. Thus, we used four performance measures namely accuracy, precision, recall and f1-score as the evaluation criteria for this study. Moreover, hyperparameter tuning was used to improve the performance of each model. For instance, hyperparameters in RF are (1) maximum depth: the maximum depth of the tree (2) maximum features: the maximum number of features Random Forest is allowed to try in an individual tree and (3) number of estimators: the number of trees in the forest. A grid search was performed over the specified parameter values using the cross-validation technique to assess model performance and to find the best set of parameters. The best parameter value of maximum depth is 12, maximum features are 50, and the number of estimators is 25 for RF. Then the SVM hyperparameters were tuned with Radial Basis Function (RBF) kernel function and the best parameter value for both C and gamma is 1.

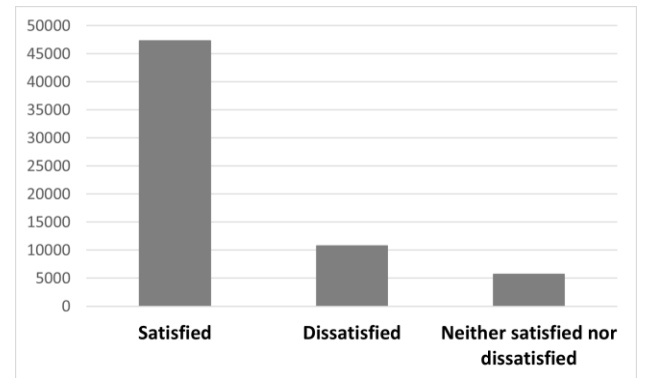
A feed-forward NN model was built using Keras and TensorFlow. We created a fully connected network with two hidden layers. Because of the advantages of computational efficiency and non-linearity, we used the “relu” activation function for the input layer and the hidden layers. Since this NN model is for multi-class classification, the “softmax” activation function is used for the output layer. Finally, the network used the efficient Adam gradient descent optimization algorithm and logarithmic loss function, “sparse_categorical_crossentropy” for compilation. With these parameters, RF shows the highest accuracy, precision, recall and f1-score among the chosen classifiers for all the datasets when predicting job satisfaction of software developers.



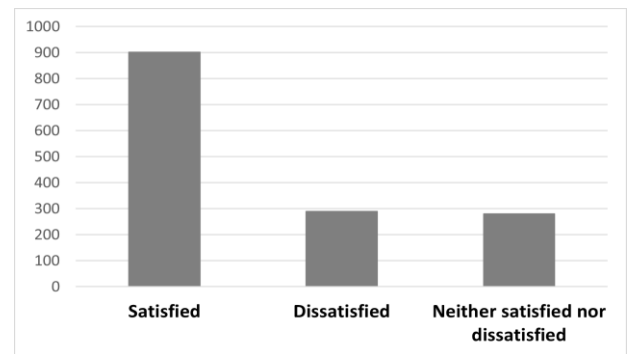
(a) 2018



(b) 2019



(c) 2020



(d) 2020 IBM HR analytic

Fig. 4. Class distribution of labels in Stack Overflow developer survey (a)2018, (b)2019 & (c)2020 and (d)IBM HR analytic dataset

TABLE II. RF MODEL PERFORMANCE WITH 7 CLASS LABEL VS 3 CLASS LABEL

	Label with 7 classes				Label with 3 classes			
	Accuracy	Precision	Recall	f1-score	Accuracy	Precision	Recall	f1-score
RF	0.68	0.62	0.68	0.62	0.80	0.74	0.80	0.75

This is because RF is an ensemble algorithm that consists of a group of decision trees. Table III shows that the RF shows 80% accuracy of job satisfaction of software developers while others show around 79% accuracy.

TABLE III. EVALUATION METRICS OF CLASSIFICATION MODELS

Dataset	ML model	Accuracy	Precision	Recall	f1-score
Stack Overflow 2018	RF	0.80	0.74	0.80	0.75
	SVM	0.79	0.62	0.79	0.69
	LR	0.79	0.69	0.79	0.70
	NN	0.79	0.62	0.79	0.69
Stack Overflow 2019	RF	0.76	0.68	0.76	0.71
	SVM	0.73	0.53	0.73	0.62
	LR	0.74	0.68	0.74	0.67
	NN	0.73	0.53	0.73	0.61
Stack Overflow 2020	RF	0.76	0.69	0.76	0.70
	SVM	0.74	0.55	0.74	0.63
	LR	0.75	0.65	0.75	0.67
	NN	0.74	0.55	0.74	0.63
IBM HR analytic	RF	0.33	0.31	0.33	0.31
	SVM	0.38	0.33	0.38	0.28
	LR	0.37	0.35	0.37	0.34
	NN	0.35	0.35	0.35	0.35

According to the results in the latter section in table III, it shows that the classifiers RF, SVM, LR and NN are not performing well with the IBM HR analytic dataset. Therefore, the above classifiers cannot predict employees' job satisfaction as a generalized model for predicting job satisfaction of other jobs in the IT industry. These results reveal that job-specific factors have a high contribution in deciding the level of job satisfaction. For example, above mentioned models performed well with predicting job satisfaction of software developers, and years of coding experience, programming languages work with and platforms work with are some of the most significant job-specific factors apart from the salary, training and workplace challenges. These factors are based on a globally collected dataset, and this experiment can be further extended to see the applicability of the above factors for the local IT industry by collecting local datasets.

The present study is a first step towards forecasting the level of job satisfaction of software developers using ML models. However, this can influence the software developer's survival in the software industry and further aid in the recruitment process. The findings of this study help HR managers to improve the identified company-related factors which caused a change in the level of job satisfaction of employees. It will increase the company reputation directly and indirectly through positive behavior and commitment of employees. For example, establishing training programs for newly recruited employees, giving promotions and career development opportunities, and building good relationships are worth the success of a company through highly satisfied employees. In addition,

the results of this study are beneficial for software developers to choose job opportunities where they can gain high job satisfaction considering their background data and company-related factors. As future works, three directions can be followed as follows. (1) The accuracy of the model can be increased by including more training data which is collected from different sources other than the Stack Overflow developer survey. (2) This work can be extended by implementing a NN model that finds the best weights of each factor for job satisfaction. (3) The effect of the "work from home" approach can be analyzed to change the job satisfaction level among employees in the IT industry.

V. CONCLUSION

In this research, we investigated supervised ML models for predicting the job satisfaction of IT employees. Prediction performance of multi-class classifiers namely RF, SVM, LR and NN were compared using two benchmark datasets. Accuracy, precision, recall and f1-score were used as the performance metrics to evaluate and compare the classifiers. The experimental results show that the above ML models can predict the level of job satisfaction of software developers using their background data and company-related data with an accuracy of around 79%. Further, we investigated the performance of the aforementioned classifiers when predicting job satisfaction of both IT and non-IT employees in the IT industry. It reveals that the above classifiers cannot be utilized as generalized models to predict the job satisfaction of IT-related employees who are not software developers. In addition, seven (07) factors were identified as the most influential factors of job satisfaction of software developers. In summary, the findings of this study are beneficial for several parties such as IT employees, IT related companies and researchers in this domain.

REFERENCES

- [1] J. Lounsbury, L. Moffitt, L. Gibson, A. Drost, and M. Stevens, "An investigation of personality traits in relation to job and career satisfaction of information technology professionals," *JIT*, vol. 22, pp. 174–183, 03 2007.
- [2] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting employee attrition using machine learning techniques", vol. 9, no. 4, 2020.
- [3] V. Wickramasinghe, "Impact of time demands of work on job satisfaction and turnover intention: Software developers in offshore outsourced software development firms in sri lanka," *Strategic Outsourcing: An International Journal*, vol. 3, pp. 246–255, 11 2010.
- [4] P. Rohit and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, 10 2016.
- [5] Y. Choi and J. Choi, "A study of job involvement prediction using machine learning technique," *International Journal of Organizational Analysis*, vol. ahead-of-print, 08 2020.
- [6] A. A. A. Khaled Alshehhi, Safeya Bin Zawbaa and M. U. Tariq, "Employee retention prediction in corporate organizations using machine learning methods," *Academy of Entrepreneurship Journal*, vol. 27, 08 2021.
- [7] M. Murawski, N. Payakachat, and C. Koh-Knox, "Factors affecting job and career satisfaction among community pharmacists: A structural equation modeling approach," *Journal of the American Pharmacists Association: JAPhA*, vol. 48, pp. 610–20, 09 2008.
- [8] A. Domagala, J.-N. Pena-SA nchez, and K. Dubas-Jakobczyk, "Career satisfaction of polish physicians - evidence from a survey study," *European Journal of Public Health*, vol. 29, 11 2019.

- [9] A. Altamimi, "Literature on the relationships between organizational performance and employee job satisfaction," *Archives of Business Research*, vol. 7, 2019.
- [10] N. Gopalan and M. Pattusamy, "Role of work and family factors in predicting career satisfaction and life success," *International Journal of Environmental Research and Public Health*, vol. 17, p. 5096, 07 2020.
- [11] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in stackoverflow," 03 2013, pp. 1019–1024.
- [12] A. Joorabchi, M. English, and A. Mahdi, "Text mining stackoverflow: Towards an insight into challenges and subject-related difficulties faced by computer science learners," *Journal of Enterprise Information Management*, vol. 29, pp. 255–275, 03 2016.
- [13] S. Alajmani and H. Elazhary, "Hospital readmission prediction using machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 10, 01 2019.
- [14] S. Meinzer, A. Thamm, U. Jensen, J. Hornegger, and B. Eskofier, "Can machine learning techniques predict customer dissatisfaction? a feasibility study for the automotive industry," *Journal of Artificial Intelligence Research*, vol. 6, pp. 80–90, 01 2017.
- [15] S. Ahamed and E. Daub, "Machine learning approach to earthquake rupture dynamics," 06 2019.
- [16] Y. Alejandro and L. Palafox, "Gentrification Prediction Using Machine Learning," 10 2019, pp. 187–199.
- [17] G. Deepali, A. Brar, and P. Sandhu, "Modeling of fault prediction using machine learning techniques," 08 2020.
- [18] Thi, H. Dinh, T. Pham, C. Loan, G. Nguyen, N. Thi, and N. Thi Lien Huong, "An empirical study for student academic performance prediction using machine learning techniques," *International Journal of Computer Science and Information Security*, vol. 18, 04 2020.
- [19] M. Asim and Z. Khan, "Mobile price class prediction using machine learning techniques," *International Journal of Computer Applications*, vol. 179, pp. 6–11, 03 2018.
- [20] Bureau of Labor Statistics, U.S. Department of Labor, *Occupational Outlook Handbook, Software Developers, 2019* (Accessed on: June 13, 2021). [Online]. Available: <https://www.bls.gov/ooh/computer-and-information-technology/software-developers.htm>
- [21] P. Petrone, "See The Industries With the Highest Turnover (And Why It's So High), 2018" (Accessed on: August 02, 2020). [Online]. Available: <https://www.linkedin.com/business/learning/blog/learner-engagement/see-the-industries-with-the-highest-turnover-and-why-it-s-so-hi>
- [22] T. Ahmed and A. Srivastava, "Understanding and evaluating the behavior of technical users. a study of developer interaction at stackoverflow," *Human-centric Computing and Information Sciences*, vol. 7, 12 2017.
- [23] StackOverflow, "Stack overflow 2018 developer survey," 2018 (Accessed on: January 27, 2021). [Online]. Available: <https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey>
- [24] S. Uyun and E. Sulistyowati, "Feature selection for multiple water quality status: Integrated bootstrapping and smote approach in imbalance classes," *International Journal of Electrical and Computer Engineering*, vol. 10, pp. 4331–4339, 08 2020.
- [25] pavansubhash, "IBM HR Analytics Employee Attrition & Performance," *Kaggle.com*, 2017. <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>.
- [26] M. Chirico, "Stack Overflow Developer Survey Results 2019," *kaggle.com*, 2019. https://www.kaggle.com/mchirico/stack-overflow-developer-survey-results-2019?select=survey_results_public.csv (accessed Aug. 22, 2021).
- [27] A. Khan, "Stack Overflow Developer Survey 2020," *www.kaggle.com*, 2020. <https://www.kaggle.com/aitzaz/stack-overflow-developer-survey-2020> (accessed Aug. 22, 2021).
- [28] P. Heimerl, M. Haid, L. Benedikt, and U. Scholl-Grissemann, "Factors Influencing Job Satisfaction in Hospitality Industry," *SAGE Open*, vol. 10, no. 4, p. 215824402098299, Oct. 2020.

Feature selection in automobile price prediction: An integrated approach

Sobana Selvaratnam*

Department of Physical Science

Vavuniya Campus of the University of Jaffna, Sri Lanka
sobanaselvaratnam@gmail.com

T. Jeyamugan

Department of Physical Science

Vavuniya Campus of the University of Jaffna, Sri Lanka
tjeyamugan@vau.jfn.ac.lk

B. Yogarajah

Department of Physical Science

Vavuniya Campus of the University of Jaffna, Sri Lanka
yoganbala@yahoo.com

Nagulan Ratnarajah

Department of Physical Science

Vavuniya Campus of the University of Jaffna, Sri Lanka
rnagulan@univ.jfn.ac.lk

Abstract - Machine learning models for predictions enable researchers to make effective decisions based on historical data. Automobile price prediction studies have been a most interesting research area in machine learning nowadays. The independent variables to model the price and the price predictions are equally important for automobile consumers and manufacturers. Automobile consulting companies determine how prices vary in relation to the independent variables and they can then adjust the automobile's design, commercial strategy, and other factors to fulfill specified price targets. Furthermore, the model will assist management in comprehending a company's pricing patterns. The ability of machine learning systems to predict outcomes is entirely dependent on the effective selection of features. In this paper, we determine the influencing features on automobile price using an integrated approach of LASSO and stepwise selection regression algorithms. We use multiple linear regression to build the model using the selected features. From the experimental results using the automobile dataset from the UCI machine learning repository, the influencing features on automobile price are width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, and drive wheels. Training data accuracy for predicting price was found to be 92%, and testing data accuracy was found to be 87%. The proposed approach supports selecting the most important characteristics of predicting the price of automobiles efficiently and effectively. This research will aid in the development of a model that uses the selected attributes to predict the price of automobiles using machine learning technologies.

Keywords - automobile price prediction, feature selection, LASSO, stepwise selection

I. INTRODUCTION

One of the greatest and most important innovations in human history is the automobile. In 2020, almost 78 million automobiles were produced worldwide [1]. The price of an automobile is determined by a number of distinct features and elements and the accurate car price prediction necessitates specialist expertise. Customers who purchase a new car may assure their investment to be worthy. The automobile consulting companies must comprehend the aspects that influence automobile pricing. The manufacturers always have attention to the elements which are important in estimating the price of automobiles and interest on how well those variables accurately predict an automobile's price. An automobile price prediction system is, therefore, needed to accurately estimate the automobile's price based on a range of factors.

In the field of computer science, machine learning approaches have revolutionized the discipline. Automobile price prediction studies using machine learning approaches [2-6] guide better decisions and take smart actions for high accuracy predictions in real-time. Feature selection is one of the initial steps for the machine learning model assessment to reduce model complexity and increase model performance when it comes to generalization, model fit, and prediction exactness [7]. The problem of feature selection has been extensively researched in the literature [8-9]. Wrapper methods, filter methods, and embedding techniques are the most common feature selection approaches [10]. However, predicting an automobile's pricing and selecting the optimal features are complex tasks since automobiles have many properties but some of the factors only can describe the automobile price.

In supervised machine learning algorithms, when the response variable is a real or continuous value, it is a regression problem. The relationship between one continuous dependent variable and two or more independent variables is explained by multiple linear regression [11], a simple machine learning approach. The goal of this study is to find an appropriate technique for choosing optimal features for the price prediction of automobiles. The technique of selecting the smallest number of effective explanatory variables can more properly characterize a response variable. Stepwise selection [11], a wrapper method, and the LASSO regression methods [12], an embedded method, are the better feature selection methods, which provide a high prediction accuracy, supports to improve the interpretability of the model by removing extraneous variables that aren't related to the response variable, and prevents overfitting. In this study, the LASSO and stepwise selection methods were used in a hybrid way to build an appropriate model for the dataset to select the optimal features. The LASSO method [12] has been used for selecting the optimal features from the numerical variables and removing the multicollinearity of the variables. The stepwise selection method has been used to find the optimal features from the categorical variables. The stepwise selection method is applied again for the selected features from the numerical and the categorical variables to tune the final optimal feature set since the feature set chosen does not contain any multicollinearity. We proved the effectiveness of this integrated approach feature selection method for predicting automobile prices using the multiple linear regression approach with the selected features.

II. RELATED WORK

Supervised learning is used in the vast majority of actual machine learning applications. Supervised machine learning techniques were utilized in the literature to predict the price of automobiles such as linear regression analysis [2,4,6], k-nearest neighbours [4,6], naïve Bayes [6], artificial neural network [5], support vector machine [5], and random-forest [2-5] and decision tree [4,6]. However, most of the research studies are highly interested in used automobile datasets [2-6]. For the feature extraction, different strategies were used by these studies, such as descriptive statistics [2], the correlation between variables [3,4,6], and data pre-processing [5]. There were no specific methods, only the heuristic, and basic statistical methods, used in these studies for selecting the optimal features. These research studies utilized different datasets and filter out the different sets of features such as (price, kilometre, vehicle type, and brand) [3], (number of doors, colour, mechanical and cosmetic reconditioning time, used to new ratio and appraisal to trade ratio) [4], and (brand, model, car condition, fuel, age, kilowatts, transmission, miles, colour, doors, drive, leather seats, navigation, alarm, aluminum rims, AC and more) [5]. Moreover, the main weakness of these studies is the low number of records that have been used [4,6].

Various approaches for solving the feature selection problem have been proposed in the literature. Wrapper methods [13], which use the output of an estimator or model in the selection process, and filter methods, which use heuristics to choose an ideal subset, are the standard strategies of feature selection. Popular regression methods have been used to extract the features for various prediction problems such as LASSO, OLS regression, ridge regression for Diabetes [14], LASSO for Diabetes [15], and LASSO for heart disease [16]. Muthukrishnan et al [14] proved, by decreasing the coefficients to zero, LASSO outperforms the other approaches. Valeria Fonti et al [17] showed the LASSO approach aids in the selection of a model with the most important properties, reduces the overfitting, increases the model interpretability, and has a very good prediction accuracy. New correlation matrices have been introduced in recent years that may have greater expressive capacity when measuring correlations between variables and feature selection [18-19]. However, these new correlation methods focus only on non-linear relationships rather than linear relationships.

Many of the unique algorithms have been constructed using only one form of selection strategies, such as a filter, wrapper, or embedded optimal feature collection procedure. Ensemble methods recently developed strategies [20] to choose influenced variables for machine learning purposes. In an ensemble method, multiple types of feature selection approaches are not taken into account. Furthermore, the use of ensemble feature selection is associated with automobile problems has not been investigated. Recently, optimal feature subsets formed by hybrid approaches combining filters, wrapper, and embedded feature selection approaches in medical datasets [21] and Gene expression data [22], which were performed well for feature selection. Hybrid filter-wrapper cluster-based feature selection method was applied for software defect prediction [23], short-term load forecasting [24], and intrusion detection systems [25]. Best of our

knowledge, this study is the first attempt to select the optimal features to efficiently predict the price of a new automobile using a hybrid wrapper-embedded method in a unique approach.

III. METHODOLOGY

A. Dataset

The primary dataset was gathered from the automobile dataset from the UCI machine learning repository [26]. Each column in our collection represents a feature of the automobile, and each row represents one automobile. The dataset consists of 26 parameters, as listed in Table I, and the details of 205 automobiles. The outcome of the prediction on the automobile dataset is the price which is a continuous variable and predictors with both numerical and categorical values.

TABLE I. DESCRIPTION OF THE ATTRIBUTES AND THE DATATYPE OF THE AUTOMOBILE DATASET.

Attribute	Attribute Range
Symboling	3, -2, -1, 0, 1, 2, 3
normalized-losses	continuous from 65 to 256
Make	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, Volvo
fuel-type	diesel, gas
Aspiration	std, turbo
num-of-doors	four, two
body-style	hardtop, wagon, sedan, hatchback, convertible
drive-wheels	4wd, fwd, rwd
engine-location	front, rear
wheel-base	continuous from 86.6 to 120.9
Length	continuous from 141.1 to 208.1
Width	continuous from 60.3 to 72.3
Height	continuous from 47.8 to 59.8
curb-weight	continuous from 1488 to 4066
engine-type	dohc, dohcvt, l, ohc, ohcvt, ohcvt, rotor
num-of-cylinders	eight, five, four, six, three, twelve, two
engine-size	continuous from 61 to 326
fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi
Bore	continuous from 2.54 to 3.94
Stroke	continuous from 2.07 to 4.17
compression-ratio	continuous from 7 to 23
Horsepower	continuous from 48 to 288
peak-rpm	continuous from 4150 to 6600
city-mpg	continuous from 13 to 49
highway-mpg	continuous from 16 to 54
Price	continuous from 5118 to 45400

B. Mathematical background

Multiple Linear Regression Model: Multiple Linear Regression is a statistical approach that predicts the outcome of a response variable by combining numerous explanatory variables. Multiple Linear Regression models can be described as below:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon_i \quad (1)$$

where dependent variable y_i , explanatory variables x_i , regression coefficients $\beta_0, \beta_1, \dots, \beta_k$, a number of explanatory variables k , and error term ϵ_i .

LASSO Estimator [12]: The LASSO estimator can be defined by the solution to the l_1 optimization problem,

$$(2) \quad \text{Minimize } \left(\frac{\|Y - X\beta\|_2^2}{n} \right) \text{ subject to } \sum_{j=1}^k \|\beta_j\|_1 < t$$

where t is the upper bound for the sum of coefficients.

This optimization problem is equivalent to the parameter estimation that pursues,

$$(3) \quad \hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right)$$

where $\|Y - X\beta\|_2^2 = \sum_{i=0}^n (Y_i - (X\beta)_i)^2$,

$\|\beta\|_1 = \sum_{j=1}^k |\beta_j|$ and $\lambda \geq 0$ is that the parameter that controls the strength of the penalty, the high value of λ , the greater amount of shrinkage.

Stepwise Selection [11]: Forward and backward selections are combined in a stepwise selection. It starts with no predictors and then adds the most significant predictors one by one (like forwarding selection). Remove any predictors that no longer improve the model fit after each new predictor is included (like backward selection).

C. Integrated approach for feature selection

Predictive model training and deployment pipelines often include data pre-processing techniques, exploratory analysis, and feature engineering. In practice, cleansing data sets before feeding them to a learning algorithm is typical to increase model predictive performance and generalization potential. The pre-processing of the automobile dataset included removing inconsistent and noisy data and managing missing values. The detail of the pre-processing is described in the Feature selection section. A correlation matrix was used to investigate the dependency between variables and detect multicollinearity. A multiple linear regression model was initially built with all the 26 features in the dataset to check the r-squared value and the most significant features for the model.

When we applied the LASSO [12] and stepwise approaches [11] separately to the automobile dataset, they did not perform well. Many numerical factors were substantially correlated with price in the correlation analysis; however, this was not the case for categorical variables. Furthermore, in the automobile dataset, numerical parameters were more strongly influenced by pricing than category factors. The LASSO and stepwise selection methods were therefore used in an integrated way to build an appropriate model for the dataset to select the optimal features and get predictions. The approach is further described with the intermediate results in the Feature Selection section.

The pre-processed data split into a training dataset and testing dataset with ratios of 70 and 30 respectively. The training dataset has been used for model fitting and feature selection and the test data has been used for evaluating the prediction accuracy. An integrated approach using LASSO and stepwise methods was used to select the appropriate feature for predicting the price of automobiles. Data preparation and model building are processed by using the R programming language in Rstudio. We implemented the LASSO method making use of the glmnet package and the plotmo package in R.

D. Price prediction process

The selected optimal features from the integrated approach were used to build a model using multiple linear regression for predicting the price. The training set was evaluated first using the accuracy as r-squared. The test set was evaluated using the same subset of features and computed the accuracy. The model performance indicator for regression issues is based on the coefficient of determination, r-squared, and the percentage of r-squared of the price variation is explained by the variation in the optimum selected independent variables.

IV. FEATURE SELECTION

A. Data preprocessing

The data pre-processing step consists of removing the inconsistent and noisy data. The missing data were also removed if any variable has missing values above 50%. The imputation process was performed for other predictors with a small percentage of missing values. In our dataset, we first find out the variables with missing values, and then it regresses on other variables. The missing values of that variable were replaced by predicted values. Moreover, influencing outliers are revealed, and taken action to remove non-influencing outliers.

The price and log (price) histograms are shown in Fig.1. While the price range varies widely with a lengthy tail, log (price) appears to follow a normal distribution. As a result, the outcome of the model development and evaluation procedure will be log (price).

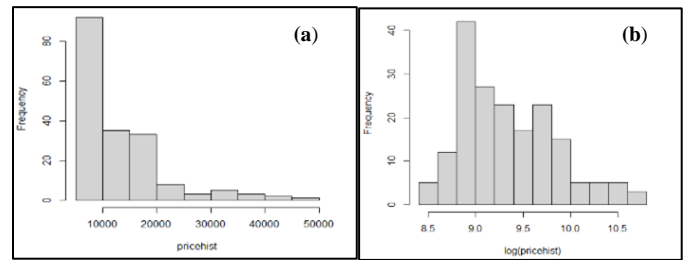


Fig. 1. (a) Price and (b) log(price) histograms

B. Data visualization and exploration

We evaluated the variable visually using matrix linear plots and bar plots. The wheelbase, length, width, curb weight, bore, and horsepower variables have a positive linear relationship with price than height, compression ratio, and peak rpm. City mpg and highway mpg have a negative linear relationship with price.

The correlation matrix of numerical attributes is visualized in Fig. 2. From the correlation matrix we can deduce that the response variable price is highly correlated with horsepower, bore, engine size, curb weight, width, and length. The price is also negatively correlated with highway mpg and city mpg. Some independent variables are highly correlated with each other such as wheelbase, length, width, height, curb weight, engine size, and bore. As a result, the vast majority of numerical variables are multi-correlated covariance variables.

The correlation matrix of the categorical variables was created using Kendall's Tau-b, which is visualized in Fig.3. From the correlation matrix, we can deduce that that price is highly correlated with the number of cylinders and drive

wheels. The price is also negatively correlated with the body style, make, and fuel type.

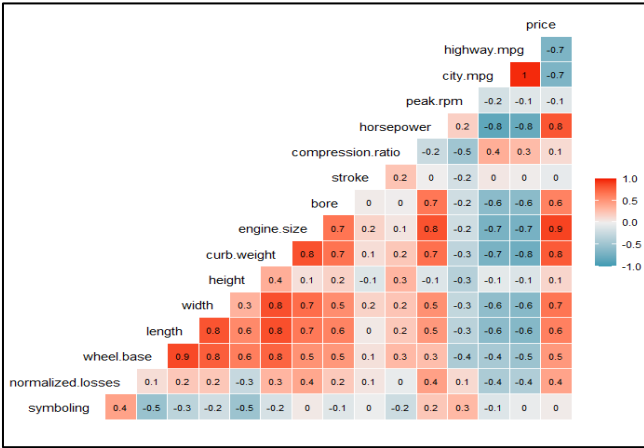


Fig.2. Correlation matrix for numerical variables

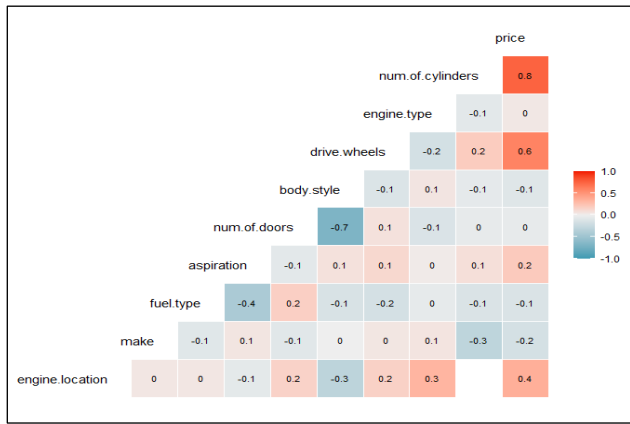


Fig.3. Correlation matrix for categorical variables

C. Variable analysis using multiple linear regression

Using all of the 26 variables in the dataset, a multiple linear regression model was created. We find out the most significant variables for the model using analysis variance (ANOVA). Table II presents the results of ANOVA.

TABLE II: ANOVA FOR THE AUTOMOBILE DATASET

Response Variable: log(price)				
Predictors	Df	F value	Pr(>F)	
symboling	1	41.8091	1.793e-09	***
normalized.losses	1	1102.8527	< 2.2e-16	***
make	15	140.0633	< 2.2e-16	***
fuel.type	1	0.6761	0.41243	-
aspiration	1	148.3762	< 2.2e-16	***
num.of.doors	1	52.6042	3.102e-11	***
body.style	4	14.2010	1.149e-09	***
drive.wheels	2	67.7883	< 2.2e-16	***
engine.location	1	4.5218	0.03533	*
wheel.base	1	169.4448	< 2.2e-16	***
length	1	90.6747	< 2.2e-16	***
width	1	44.9671	5.324e-10	***
height	1	2.2505	0.13596	-
curb.weight	1	114.3238	< 2.2e-16	***
engine.type	2	0.5032	0.60576	-
num.of.cylinders	3	1.6966	0.17086	-
engine.size	1	2.5527	0.11250	-
fuel.system	3	3.0498	0.03094	*

bore	1	1.6475	0.20155	-
stroke	1	2.0016	0.15948	-
compression.ratio	1	4.7318	0.03139	*
horsepower	1	0.7001	0.40427	-
peak.rpm	1	0.0857	0.77019	-
city.mpg	1	2.7325	0.10070	-
highway.mpg	1	3.8492	0.05187	-
Residuals	132	-	-	-

Based on the p-values of Table II, symboling, normalized losses, make, aspiration, num.of.doors, body style, drive wheels, engine.location, wheel.base, length, width, curb.weight, fuel.system, and compression.ratio are the most significant variables and other variables are not significant in the model. Thus, we can concern these significant variables for the final model.

D. LASSO implementation

We create a model to predict the price for the automobile dataset and to find out which explanatory variables to include in the final model using the LASSO regression method (In glmnet, alpha = 1 for the LASSO regression and alpha = 0 for the Ridge regularization). Glmnet generates a series of various models based on the tuning parameter λ . To determine the influencing features, we first utilized the function on all of the numerical explanatory factors in the automobile dataset. The analyses' findings are depicted in Fig. 4 and Fig.5. We can see when each variable entered the model and how much it changed the response variable using these charts. From Fig. 4, lasso included only 10 predictors out of 15 predictors which removed the following predictors such as normalized losses, length, curb weight, horsepower, peak rpm.

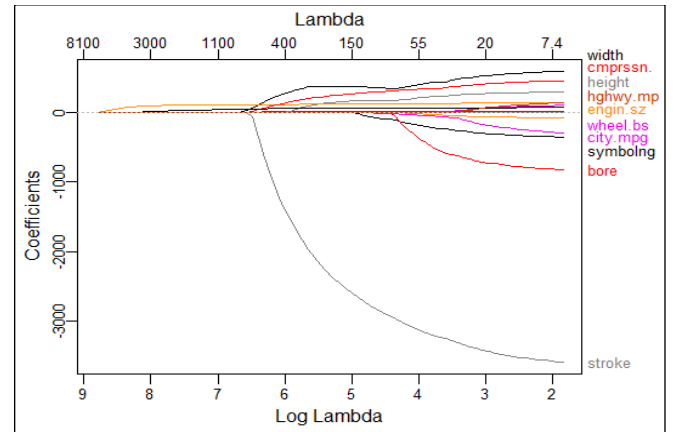


Fig.4. Glmnet graph for the numerical variables

A Correlation matrix was computed for the removed predictors by LASSO. Curb weight and horsepower are highly correlated (0.7326893) with price but they are highly correlated with each other. LASSO handles, therefore, the multicollinearity problem efficiently.

Fig.5 shows the top nine influencing predictors of automobile price. width, compression ratio, highway mpg, engine size have positively affected the model, and city mpg, symboling, bore, and stroke has negatively affected the model. To determine the value of λ , use k-fold cross-validation to find the λ value that generates the lowest test mean squared error (MSE).

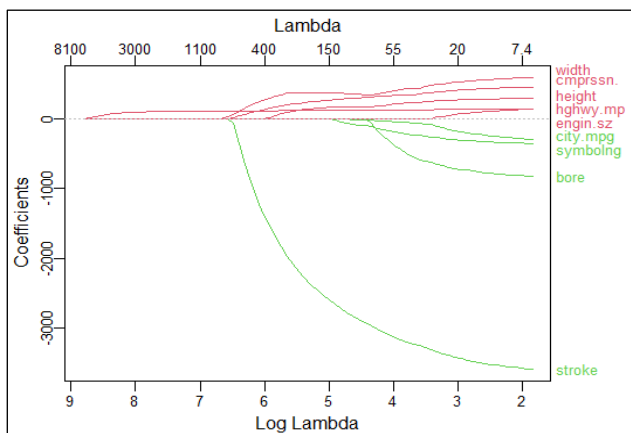


Fig.5. Lasso graph for influencing explanatory variables

The LASSO approach extracts different values for λ to determine the best acceptable value for, such as λ -min (first vertical line in Fig.6), which offers the minimum mean cross-validated error, and λ -1se (second vertical line in Fig.6), which produces a model with error within one standard error of the minimum. At this point, we can select the value for λ that is most appropriate for the problem.

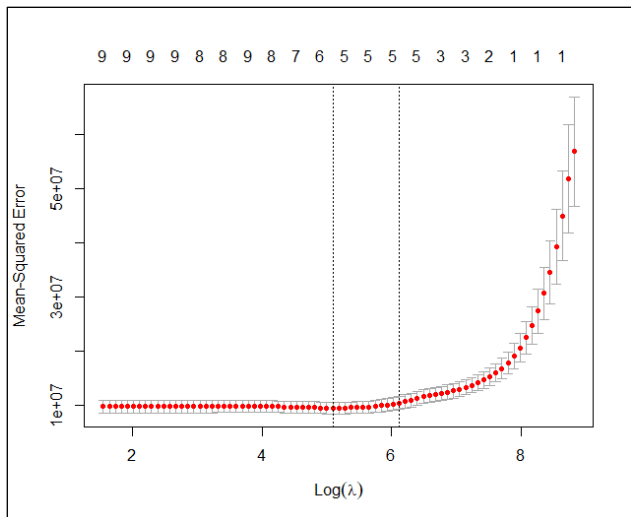


Fig. 6. Cross-validation

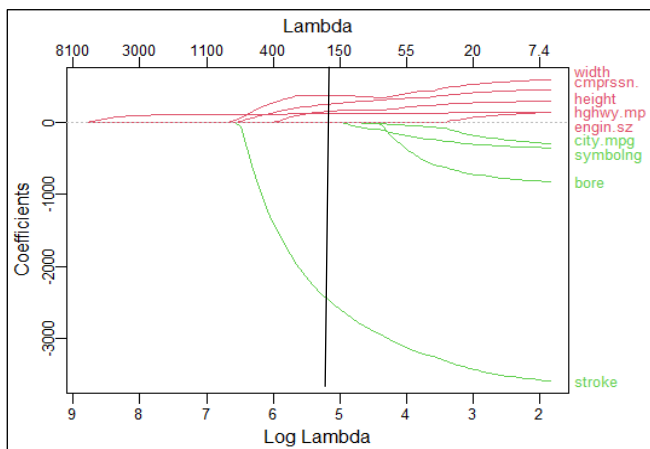


Fig. 7. Most important features

Because the aforementioned Fig. 6 plot exhibits an exponential trend, λ -min is not obvious in our analysis. So,

we compute those two λ values (λ -min value =162.7058 and λ -1se value= 543.3253)

TABLE III: 10 X 1 SPARSE MATRIX OF CLASS

Predictors	Coefficient
(Intercept)	1525.35581
symboling	-
width	94.73858
height	-
engine.size	147.21568
bore	-
stroke	-2760.41510
compression.ratio	268.16365
city.mpg	-277.45107
highway.mpg	-

From Table III, no coefficient is shown for the predictors symboling, height, bore, and highway mpg because as a result of the lasso regression, the coefficient was reduced to zero. This means it was deleted entirely from the model because it did not influence it. By combining the plots in Fig. 7 and Table III, we can conclude. The most significant numerical variables for the price prediction from the automobile dataset are Width, compression ratio, engine size, city mpg, and stroke, which have been selected according to the λ -min value.

E. Stepwise selection implementation

The stepwise selection method was utilized for the categorical variables in the automobile dataset to find out the influencing features. The results of the stepwise selection regression method are shown in Table IV.

TABLE IV: STEPWISE SELECTION METHOD'S OUTCOME OF THE CATEGORICAL VARIABLES

	Df	Sum of Sq	RSS	AIC
log(price) ~ make + aspiration + num.of.doors + body.style + drive.wheels + num.of.cylinders + fuel.system				
+ engine.location	1	0.00903	2.3681	-436.75
+ engine.type	3	0.01134	2.3658	-432.87
- body.style	4	0.29426	2.6714	-431.56
- aspiration	1	0.28457	2.6617	-426.02
- num.of.doors	1	0.51683	2.8940	-415.48
- fuel.system	4	0.72231	3.0994	-412.84
- num.of.cylinders	3	0.71775	3.0949	-411.02
- drive.wheels	2	0.78058	3.1577	-406.49
- make	15	2.88326	5.2604	-368.19

The above results (Table IV) present the final step of the stepwise selection for categorical predictors from the automobile dataset. From this method, seven influencing predictors are filtered such as make, aspiration, number of doors, body style, drive wheels, number of cylinders, fuel system. After that, we have applied the stepwise selection method to selected numerical and categorical predictors selected from lasso and stepwise selection.

The results (Table V) present, width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, drive wheels, number of cylinders, fuel system are filtered out by stepwise selection method from the dataset. Predictors selected by the stepwise method are analysed by ANOVA. From the ANOVA (Table VI), the number of

cylinders and fuel systems are not significant. So, we removed them from the model.

TABLE V: STEPWISE SELECTION METHOD'S OUTCOME OF THE SELECTED NUMERICAL AND CATEGORICAL VARIABLES

	Df	Sum of Sq	RSS	AIC
log(price) ~ width + engine.size + city.mpg + stroke + make + aspiration + num.of.doors + body.style + drive.wheels + num.of.cylinders + fuel.system				
- num.of.cylinders	3	0.07271	1.4078	-502.28
+ compression.ratio	1	0.00588	1.3292	-501.52
- city.mpg	1	0.05763	1.3927	-499.63
- fuel.system	4	0.13305	1.4681	-498.99
- width	1	0.06693	1.4020	-498.80
- stroke	1	0.10576	1.4408	-495.35
- num.of.doors	1	0.12324	1.4583	-493.83
- drive.wheels	2	0.15806	1.4931	-492.86
- aspiration	1	0.23138	1.5665	-484.82
- body.style	4	0.36160	1.6967	-480.76
- engine.size	1	0.34923	1.6843	-475.68
- make	15	1.44497	2.7801	-440.54

V. PRICE PREDICTION

The integrated approach's best attributes (width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, and drive wheels) were used to create a model that employed multiple linear regression to forecast the price. We obtained 92% accuracy for the price prediction using the training set. We evaluated the final model using the testing dataset and we obtained 87% testing accuracy. The high r-squared values show that the selected independent variables with the hybrid feature selection method truly determine the price of automobiles. To reduce the overfitting problem and improve interpretation capabilities, the number of features in the chosen algorithms was maintained as minimal as possible.

The experimental results suggest that a hybrid approach integrating LASSO (embedded method) and Stepwise (wrapper method) regression techniques provides a high level of prediction accuracy and a reasonable rate of feature reduction. For the proper comparison with other approaches in the literature, no study in the literature uses the UCI machine learning [26] repository data to predict automobile prices. Some researchers used this automobile dataset for various purposes such as data-guided approach to generate multi-dimensional schema for targeted knowledge discovery [27], mapping nominal values to numbers for effective visualization [28], and attribute identification and predictive customization [29].

VI. CONCLUSIONS

The study presented a hybrid approach to select the optimal features to build an efficient model for the price prediction of automobiles. First, the dataset is analysed and pre-processed for the model building and then split the dataset into train and test datasets. Next, the feature selection was conducted using the training dataset using lasso and stepwise selection regression methods in an integrated way. The most relevant features for the prediction of automobiles are width, engine size, city mpg, stroke, make, aspiration, number of doors, body style, and drive wheels. These optimal features were evaluated in the

multiple linear regression model with training dataset accuracy of 92% and testing dataset accuracy of 87% respectively. The findings show that combining embedded and wrapper feature selection to build a hybrid form of feature selection yields better outcomes.

REFERENCES

- [1] Agencija za statistiku BiH. (n.d.), retrieved from: <http://www.bhas.ba>. [accessed January, 2021.]
- [2] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buaya and P. Boonpou, 2018, "Prediction of prices for used car by using regression models", 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 115-119.
- [3] N. Pal, P. Arora, S.Sumanth Palakurthy, D.Sundararaman, P.Kohli, 2017, How much is my car worth? "A methodology for predicting used cars prices using Random Forest", CoRR, abs/1711.06970
- [4] P. Gajera, A. Gondaliya, J.Kavathiya, 2021, "Old Car Price Prediction With Machine Learning", International Research Journal of Modernization in Engineering Technology and Science, Volume:03, Issue:03, pp.284-290.
- [5] E. Gegic, B. Isakovic, D.Keco, Z.Masetic, J.Kevric, 2019, "Car Price Prediction using Machine Learning Techniques", TEM Journal. Volume 8, Issue 1, pp. 113-118.
- [6] S. Pudaruth, 2014, "Predicting the Price of Used Cars using Machine Learning Techniques, International Journal of Information & Computation Technology, Volume 4, Number 7 , pp. 753-764.
- [7] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, 2007, "Data preprocessing for supervised learning", International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 1, pp. 4104-4109.
- [8] A. L. Blum and P. Langley, 1997, "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol. 97, no. 1, pp. 245 - 271.
- [9] H. Motoda and H. Liu, 2002, "Feature selection, extraction and construction", Communication of IICM (Institute of Information and Computing Machinery, Taiwan), vol. 5, pp. 67-72.
- [10] Guyon, I., Elisseeff, A, 2003, "An introduction to variable and feature selection". Journal of machine learning research, pp.1157-1182.
- [11] M.A. Efronymson, 1960, "Multiple regression analysis - Mathematical Methods for Digital Computers", Ralston A. and Wilf.H. S., (eds.), Wiley, New York.
- [12] R. Tibshirani, 1996, "Regression shrinkage and selection via the lasso". J. R. Stat. Soc. Ser. B (Methodological), 58, pp. 267-288.
- [13] A. Y. Ng, 1998, "On feature selection: Learning with exponentially many irrelevant features as training examples", in Proceedings of the Fifteenth International Conference on Machine Learning, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 404-412.
- [14] R. Muthukrishnan and R. Rohini, 2016, "LASSO: A feature selection technique in predictive modeling for machine learning", IEEE International Conference on Advances in Computer Applications (ICACA), pp. 18-20.
- [15] P. M. Kumarage, B. Yogarajah and N. Ratnarajah, 2019, "Efficient Feature Selection for Prediction of Diabetic Using LASSO", 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 2019, pp. 1-7.
- [16] P. Ghosh et al., 2021, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques", in IEEE Access, vol. 9, pp. 19304-19326.
- [17] V Fonti, E Belitser, 2017, "Feature Selection using LASSO", VU Amsterdam Research Paper in Business Analytics, Volume 30, pp. 1-25.
- [18] D. N. Reshef, Y. A. Reshef, M. Mitzenmacher, and P. C. Sabeti, 2013, Equitability analysis of the maximal information coefficient, with comparisons, CoRR, abs/1301.6314.
- [19] A. Luedtke and L. Tran, 2013, "The generalized mean information coefficient", arXiv: Machine Learning", [Online]. Available: <https://arxiv.org/abs/1308.5712>
- [20] D.Guan, W. Yuan, Y. Lee, K. Najeebullah, and M.K. Rasel, 2014. "A review of ensemble learning based feature selection". IETE Technical Review, 31(3), 190-198.
- [21] C.W.Chen, Y.H.Tsai, F.R.Chang, W.C.Lin, 2020, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results". Expert Systems; e12553.
- [22] S. Shilan Hameed, O. O.Petinrin, A.Osman Hashi and Faisal Saeed, 2018, "Filter-Wrapper Combination and Embedded Feature

- Selection for Gene Expression Data”, *Int. J. Advance Soft Compu. Appl*, Vol. 10, No. 1.
- [23] F. Wang, J. Ai and Z. Zou, “A Cluster-Based Hybrid Feature Selection Method for Defect Prediction”, 2019, IEEE 19th International Conference on Software Quality, Reliability and Security (QRS), pp. 1-9.
- [24] Z. Hu, Y. Bao, T.Xiong, R.Chiong, 2015, “Hybrid filter–wrapper feature selection for short-term load forecasting”, *Engineering Applications of Artificial Intelligence*, Volume 40, pp. 17-27.
- [25] M.Kamarudin, C. Maple and T. Watson, 2019, “Hybrid feature selection technique for intrusion detection system”, *Int. J. High Performance Computing and Networking*, Vol. 13, No. 2, pp.232 – 240.
- [26] Automobile dataset from UCI machine learning repository, <https://archive.ics.uci.edu/ml/datasets/automobile>.
- [27] R.L. Pears, M. Usman, A. Fong, 2012, “Data Guided Approach to Generate Multidimensional Schema for Targeted Knowledge Discovery”, 10th Australasian Data Mining Conference (AusDM 2012), Sydney, Australia.
- [28] G. E. Rosario, E. A. Rundensteiner, D. C. Brown and M. O. Ward, “Mapping nominal values to numbers for effective visualization”, *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714)*, 2003, pp. 113-120.
- [29] A. A. F. Saldivar, C. Goh, Y. Li, H. Yu and Y. Chen, "Attribute identification and predictive customisation using fuzzy clustering and genetic search for Industry 4.0 environments," 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), 2016, pp. 79-86.

Estimation of the incubation period of COVID-19 using boosted random forest algorithm

P. P. P. M. T. D. Rathnayake*
Department of Industrial Management
University of Kelaniya, Sri Lanka
thidasala.demintha@gmail.com

Janaka Senanayake
Department of Industrial Management
University of Kelaniya, Sri Lanka
janakas@kln.ac.lk

Dilani Wickramaarachchi
Department of Industrial Management
University of Kelaniya, Sri Lanka
dilani@kln.ac.lk

Abstract - Coronavirus disease was first discovered in December 2019. As of July 2021, within nineteen months since this infectious disease started, more than one hundred and eighty million cases have been reported. The incubation period of the virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), can be defined as the period between exposure to the virus and symptom onset. Most of the affected cases are asymptomatic during this period, but they can transmit the virus to others. The incubation period is an important factor in deciding quarantine or isolation periods. According to current studies, the incubation period of SARS-CoV-2 ranges from 2 to 14 days. Since there is a range, it is difficult to identify a specific incubation period for suspected cases. Therefore, all suspected cases should undergo an isolation period of 14 days, and it may lead to unnecessarily allocation of resources. The main objective of this research is to develop a classification model to classify the incubation period using machine learning techniques after identifying the factors affecting the incubation period. Patient records within the age group 5-80 years were used in this study. The dataset consists of 500 patient records from various countries such as China, Japan, South Korea and the USA. This study identified that the patients' age, immunocompetent state, gender, direct/indirect contact with the affected patients and the residing location affect the incubation period. Several supervised learning classification algorithms were compared in this study to find the best performing algorithm to classify the incubation classes. The weighted average of each incubation class was used to evaluate the overall model performance. The random forest algorithm outperformed other algorithms achieving 0.78 precision, 0.84 recall, and 0.80 F1-score in classifying the incubation classes. To fine-tune the model AdaBoost algorithm was used.

Keywords - AdaBoost, boosted Random Forest, COVID-19, incubation period

I. INTRODUCTION

The Coronavirus disease 2019 (COVID-19) is one of the disastrous infectious diseases identified in late 2019 from a seafood wholesale market in China. Some of the common symptoms of COVID-19 include fever, dry cough, difficulty in breathing, muscle pain, sputum production, diarrhea, and sore throat [1]. While the majority of cases display mild symptoms, some progress to pneumonia and multi-organ failures. As for current findings, the death rate per diagnosed case is 4.4 percent; however, it could range between 0.2%-15% based on the age group and other health problems [2]. The virus typically spreads from one person to another via respiratory droplets released mostly during coughing and sneezing. As of July 2021, the virus has spread over 222 countries and territories resulting in 188,404,542 cases and 4,059,223 deaths [16]. Due to the high rate of diagnosed cases and deaths, the World Health Organization (WHO) has

declared the COVID-19 disease as a pandemic on 11th March 2020.

Incubation period of COVID-19 can be defined as the time range a person spends between exposure to the virus and symptom onset. During the incubation period, most of the patients do not show any symptoms of being infected, but they are capable of transmitting the virus to others [17]. It is very important to isolate the suspected cases during this period to avoid virus transmission. Since the incubation period greatly varies among individuals, it is very important to identify the incubation period accurately in order to decide quarantine periods and to allocate limited resources effectively towards controlling the pandemic.

WHO has declared a time range of 2 to 14 days as the incubation period of COVID-19 patients [19]. Since there is a range to the incubation period, every suspected case should undergo a quarantine period of 14 days. During the quarantine period, active monitoring and resource allocation for the suspected cases are mandatory. Although all the suspected cases are quarantined for 14 days, some may have lesser incubation periods than others, because incubation period greatly varies depending on patients' gender, age, chronic disease history, direct/indirect contact with the affected persons, and the residing country. If there is a mechanism to identify the incubation period of each individual based on their characteristics, it will help prevent unnecessary resource allocation for quarantine/active monitoring, and effectively use the limited resources towards controlling the pandemic. The main purpose of this study is to develop a predictive model to classify the incubation period of the COVID-19 suspected cases based on their characteristics.

Section-wise organization of the paper is as follows. Section - II discusses related work. Section - III describes the methodology of the system. Results are discussed in detail in Section -IV. Finally, section - V presents the conclusion and future work directions.

II. RELATED WORK

A. Findings on incubation period

There are a number of studies to calculate the mean incubation period for the selected populations. One study has calculated the incubation period using 181 cases. This study has referred to patients' residing country, exposure date and time, dates of symptom onset, fever onset and hospitalization and calculated the median incubation period as 5.1 days [3]. The study states that 97.5% of the cases develop symptoms around 11.5 days. Another early analysis has referred to 158 cases outside the Chinese regions and estimated the median incubation period as 5 days which ranges from 2 to 14 days [4]. Authors have estimated the incubation period using lognormal

distribution. This study specifies that the median time from illness onset to hospital admission was 3-4 days and the median delay between illness onset to death is 17 days. Another analysis based on 10 confirmed cases in China estimates the mean incubation period as 5.2 days (ranges from 4 to 7 days) [2]. This study specifies that children are less likely to be infected and may show milder symptoms. They have identified that age is one of the crucial factors that decide the incubation period. Their studies specify that 27% of the patients are hospitalized after two days of symptom onset which implies that time available to seek medical attention is generally short. Another analysis on 88 affected cases in Chinese regions outside Wuhan, specifies a mean incubation period of 6.4 days which ranges between 2.1 to 11.1 days [5]. They have obtained the possible values for the incubation period by considering the number of days the person has stayed in Wuhan and the date of symptom onset and fitted three parametric forms for the incubation period: The Weibull distribution, the gamma distribution, and the lognormal distribution.

B. Factors affecting to the incubation period

Studies about factors affecting the incubation period of COVID-19 patients are limited. One study has identified that age is directly related to the incubation period. This study was based on 136 patients who had travelled to Hubei, China, and identified the median incubation period as 8.3 days for all patients, 7.6 days for younger adults, and 11.2 days for older adults. This study specifies that elderly patients have a longer incubation period [6]. A study conducted by referring to Chinese COVID-19 patients specify that men's cases tend to be more serious than women's cases [7]. Using a public dataset of 37 cases, Authors have identified that the number of male deaths from COVID-19 is 2.4 times the number of female deaths. Further, they have identified that the percentage of males were higher in the deceased group than in the survived group. There is strong evidence which suggests that men may have a larger concentration of ACE2 (angiotensin-converting enzyme 2) receptors in their body, which helps coronavirus to latch on and spread inside the body. This is one of the primary reasons why COVID-19 seems to affect men seriously, when compared to women [8]. Centre of disease control and prevention in the United States has identified that the people who have cancer, chronic kidney disease, COPD immunocompromised state (weakened immune system) due to solid organ transplant, obesity, BMI of 30 or higher), serious heart conditions such as heart failure, coronary artery disease or cardiomyopathies, sickle cell disease, type 2 diabetes mellitus have a higher risk of getting severely ill from COVID-19 [9]. Since chronic diseases directly affect the immune system of patients, the incubation period can differ from the immunocompetent people. Studies regarding the factors affecting the incubation period of COVID-19 patients are limited. Out of those studies one study has identified that the age is directly related to the incubation period. Authors have identified that the median incubation period for a set of COVID-19 patients who had traveled to Hubei, China was 8.3 days, and for the younger adults the incubation period was 7.6 days, and for older adults, 11.2 days. This study specifies that elderly patients have a longer incubation period than the younger adults [6]. A study conducted on two populations of COVID-19 patients from two

geographic locations to identify the deviation of incubation period across residing location, has proved that there is a deviation of incubation period across two regions. Out of the 181 patients used for the study, 108 patients were diagnosed outside of mainland China with a median incubation period of 5.5 days and 73 patients diagnosed inside China with a median incubation period of 4.8 days [3]. The above literature specifies that the patients Age, Gender, Chronic disease history, and residing country directly affect the incubation period of the COVID-19 patients.

C. Supervised learning classification algorithms used in COVID-19 domain

One study has identified factors such as patients' age, residing country, if from Wuhan, if they have visited Wuhan and gender directly affect the death/recovery of COVID-19 patients using 100 confirmed laboratory cases in China [10]. This study has used the Naïve Bayes approach to classify the death/ recovery of COVID-19 patients and achieved 93% accuracy. Another study has used the Logistic Regression approach to detect COVID-19 using clinical text data. Authors have labeled 212 clinical records into four categories named COVID, SARS, ARDS, and both (COVID, ARDS). Various text features such as TF/IDF, a bag of words has been extracted from these clinical reports to classify them. This study has reached 94% precision, 96% recall, and 95% f1 score using Logistic regression approach [11]. Support Vector Machine (SVM) has been used in the COVID-19 domain to classify the X-ray images of COVID-19 suspected cases. The study in [12] has used this method to identify the X-ray images of COVID-19 patients by comparing normal X-ray images with X-ray images showing pneumonia. [12] This study has reached an accuracy of 97% by classifying the X-ray images into classes using SVM approach. Another study has used the decision-tree classifier to identify COVID-19 patients by referring to their Chest x-ray (CXR) images [13]. They have used three binary trees to identify the abnormality of the CXR images, identify the symptoms of tuberculosis and to identify COVID-19 symptoms. They have achieved an accuracy of 98% and 80% for the first two decision trees respectively, whereas the average accuracy of the third decision tree has been 95%. One of the studies have used the Random Forest algorithm to identify if a person is infected with the SARS-Cov2 virus and the type of hospitalization (regular ward, semi-ICU, or ICU) needed, based on the hematological parameters such as red blood cells, hemoglobin, neutrophils, lymphocytes, etc. collected from blood tests.. Authors have achieved 92.8% accuracy in identifying the type of hospitalization patients needed based on the hematological parameters from blood tests [14].

III. METHODOLOGY

The key purpose of the study is to identify the factors affecting the incubation period and to design a model that can classify the incubation period of the suspected cases based on patients' characteristics. Machine learning techniques were used to build the classification models. Next, the modelling techniques were compared on validation and model accuracy, to select the best technique. At last, the best classification technique was fine-tuned using a boosting algorithm to achieve higher accuracy.

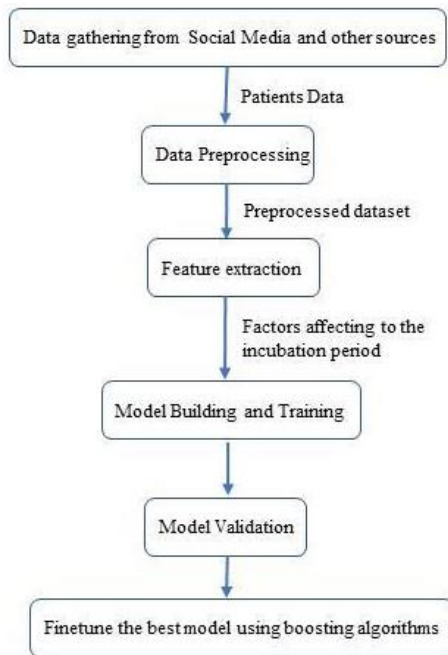


Fig. 1. The methodology of the proposed solution

Publicly available patient data and clinical records were used for this study. The following information about patients was gathered by analyzing the records manually.

- i. Age
- ii. Gender
- iii. Residing Country
- iv. Chronic disease history
- v. Direct/ indirect contact with the affected cases
- vi. Symptom onset date
- vii. Exposure date/Travel dates
- viii. Hospitalized date

Most of the data were collected from social media posts and status related to the COVID-19 patients. Chinese social media WeChat accounts are one of the major data sources which release daily information on the list of COVID-19 cases. Other than social media and WeChat accounts, following sources were used to collect data.

- Kyodo News
- Weibo.com
- Kaggle

In some of the cases, precise information was not recorded to identify the type (direct/indirect) of contact with the affected persons. If the patients travelled together with affected ones or if they got the virus from a family member, those scenarios are considered direct contact with the affected cases. Otherwise, an assumption was made - that they had indirect contact with the affected persons.

The incubation period was calculated using the date difference between symptom onset date and the exposure date. Since the incubation period of the selected population

ranges from 5 to 24 days, it was divided into four classes as below, for classification.

- Class A: 20 - 24 days
- Class B: 15 - 19 days
- Class C: 10 -14 days
- Class D: 5 - 9 days

The incubation class was added to the dataset by creating a new column named 'Incubation Class'. The median age of each incubation class was used to fill the missing values of the age column. Finally, label encoding was performed on the dataset. For analyzing the data, descriptive statistics were used. Bar charts were used to identify the distribution of the incubation period across patients' age, gender, residing country, direct/indirect contact with the affected cases and chronic disease history. Next, Pearson's Correlation Coefficient (PCC) was used to identify the variables which have the strongest relationship with the incubation period.

A number of supervised learning classification models were compared in this study to identify the best model for this particular problem. Models were implemented using Google Collab platform which provides a Jupyter notebook environment that requires no setup and runs entirely in the cloud with the accessibility of powerful computing resources from the browser. Classification algorithms such as multiple regression, support vector machine, random forest, K- nearest neighbor algorithm, naive bayes, and decision tree were compared to find the best model with highest accuracy, to classify the incubation period class based on patients' demographics and other characteristics. In order to validate the classification models, percentage split technique was used. The dataset was divided into two categories randomly, mainly 20% for testing and 80% for training. Furthermore, performance metrics such as Precision, Recall and F1 Score were used to compare model performance.

Boosting algorithms were used in this study to achieve higher accuracy in machine learning algorithms. Boosting algorithms are very useful to create high accuracy models by combining low accuracy models. AdaBoost algorithm was used in this study to improve the accuracy of the best performing classification model.[18] The main concept of AdaBoost is that it assigns weights to classifiers and training the data samples in each iteration such that it ensures the accurate predictions of unusual observations.

IV. RESULTS AND DISCUSSION

This section mainly describes the details related to the results obtained from the implementation process and the discussion of the results.

The gathered dataset for the study consists of 500 patient records with the age ranging from 5-80 years. Out of those records, 285 were male and 215 were female. The dataset includes patients' information from most of the countries around the world with the majority of cases from China Singapore, France, Germany, Taiwan, Japan, Malaysia, United States, and South Korea. Following is the incubation period distribution for the dataset.

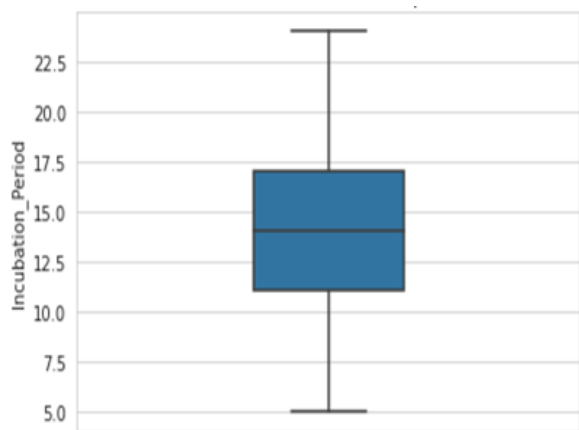


Fig. 2. Incubation period distribution for the dataset

The incubation period of the selected population ranges from 5-24 days with a median value of 13.86 days. The highest number of patients (51) have their incubation period as 14 days. Out of the 500 patient records, 31 of them (7.3% of the overall population) have their incubation period more than or equal to 20 days. 79 patients (15.8% of the overall population) have their incubation period less than or equal to 9 days. Majority of the patients have their incubation period between 10-19 days which is 76.8% of the overall population.

Correlation analysis was used in this study to identify the variables which have the strongest relationship with the incubation period. Based on the results of the correlation analysis, patients' age and the incubation class have a very strong positive relationship which is 0.819. When it comes to the direct contact with the affected cases, it also has a moderate positive relationship with the incubation class which is 0.360. Having a history of chronic diseases such as cardiac, respiratory and metabolic diseases also have a strong positive relationship with the incubation class. Patients' residing country also has a weak relationship with the incubation class which is 0.029.

Results based on descriptive statistics and the correlation analysis suggest that men's COVID-19 cases tend to decrease as the incubation period increases. This implies that men's COVID-19 cases tend to show symptoms quickly than women's cases do. Patients with chronic disease history such as Serious heart conditions, heart failures, coronary artery disease, cardiomyopathies, sickle cell disease, type 2 diabetes mellitus tend to show symptoms quicker than others. The different incubation periods can be the result of different types of inflammation and immune responses. When it comes to the method of exposure to the virus, results specify that patients who got direct exposure to the virus have a shorter incubation period than others. This implies that, if the patients had close contact with someone who has COVID-19 and got exposed to the virus directly, they tend to show symptoms very quickly than others who have got indirect exposure to the virus.

Number of supervised learning classification algorithms were compared in this study to identify the best model to classify the incubation class based on patients age, gender, chronic disease history, direct/indirect exposure to the virus and the residing country. The

following figure explains the accuracy of each model in classifying the incubation class.

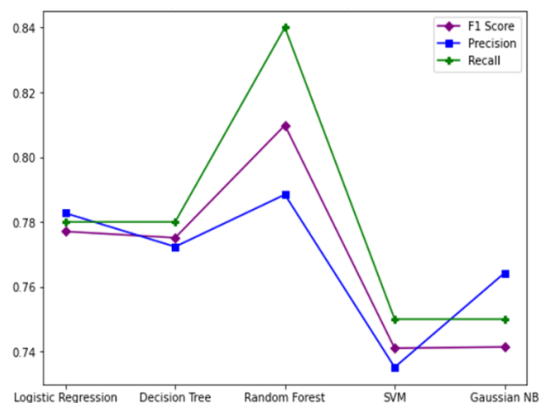


Fig. 3. Comparison of model performance without boosting algorithms

The above figure specifies that the Random forest algorithm performed better in classifying the incubation class by achieving higher precision, recall, and F1 score. Since the F1 score provides the harmonic mean between precision and recall, it was considered the best performance metric to evaluate the models. Following is the model performances in tabular format.

TABLE I. COMPARISON OF MODEL PERFORMANCE IN TABULAR FORMAT

Classifier	Precision	Recall	F1 Score
Naïve Bayes	0.764	0.750	0.741
SVM	0.735	0.750	0.741
Logistic Regression	0.780	0.782	0.777
Random Forest	0.788	0.840	0.809
Decision Tree	0.772	0.780	0.775

AdaBoost algorithm was used in this study to improve the accuracy of the classification algorithms. Since the AdaBoost algorithm needs a base classifier, random forest was used as the base classifier since it outperforms other classification algorithms.

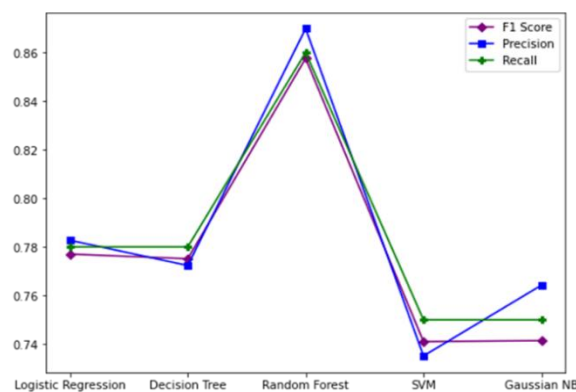


Fig. 4. Comparison of model performance with AdaBoost algorithm

Figure 4 displays the model performance after implementing the AdaBoost algorithm with the random forest algorithm as the base classifier

From figure 4, we can identify that the performance of the random forest algorithm increased after the application of the AdaBoost algorithm. Before applying the AdaBoost algorithm Random Forest algorithm outperformed other algorithms achieving a 0.78 Precision, 0.84 Recall, and a 0.80 F1 score. After applying the AdaBoost algorithm the performance metrics of the Random Forest algorithm increases up to 0.87 Precision score, 0.86 Recall Score, and a 0.86 F1 score.

V. CONCLUSION

This study implies that patients' age, gender, residing country, the method of exposure to the virus (direct/indirect exposure), and the history of chronic diseases such as cancer, chronic kidney disease, COPD, serious heart conditions, type 2 diabetes directly affect the incubation period of the SARS-CoV-2 virus. When it comes to age, older people tend to show symptoms quicker than younger people and they have a shorter incubation period compared to others. Gender wise, male cases tend to show symptoms quicker than others. Patients who have chronic diseases and immunocompromised states have a shorter incubation period than others and show symptoms quicker. The people who got direct exposure to the virus and who had a closer relationship with the affected cases tend to show symptoms quicker than people who got indirect exposure to the virus.

In this study, several supervised learning classification algorithms such as SVM, naïve naves, logistic regression, random forest, and decision tree were compared to find the best model with the highest accuracy to classify the incubation period. Random forest algorithm outperformed in classifying the incubation period achieving higher precision, recall, and F1 score. Finally, boosting algorithms such as the AdaBoost algorithm was integrated with the random forest algorithm to achieve 0.87 Precision, 0.86 Recall, and a 0.86 F1 score in classifying the incubation period.

This study mainly focused on the symptomatic transmission of COVID-19. Symptomatic transmission refers to transmission from a person while they are experiencing symptoms such as fever, cough, tiredness, etc. In a symptomatic case, we are able to track the incubation period by the date difference, between exposure to symptom onset. There are some cases showing asymptomatic transmission of COVID-19. Asymptomatic transmission can be defined as the transmission of virus from person to person, without showing symptoms of being infected. Very few asymptomatic transmission cases have been reported as a result of contact tracing efforts in some countries. Since asymptomatic patients do not show symptoms, it is relatively difficult to identify the incubation period. This study was conducted using only 500 patient records from several countries around the world. If there is larger number of patient records representing all the countries around the world with patients' clinical information, a comprehensive study can be carried out. Further, unsupervised machine learning algorithms such as artificial neural networks can be implemented with a larger dataset in order to achieve higher accuracy.

As future work, chest X-ray images of COVID-19 affected persons can be combined with geographic and

healthcare data processing models which will then be integrated into applications that will support the decision-making process for the authorities and for the growth of the healthcare systems. This will finally lead to the development of semi-autonomous classification systems that can provide the facility to detect the incubation period of COVID-19 patients accurately and prepare us for future outbreaks.

REFERENCES

- [1]. Symptoms of Coronavirus, Retrieved from <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>, September 2020
- [2]. X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, M Leung, E. Lau, J. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia", The New England Journal of Medicine, 2020.
- [3]. K. Grantz, Q. Bi, F. Jones, Q. Zheng, H. Meredith, A. Azman, N. Reich, J. Lessler, "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application", American College of Physicians Public Health Emergency Collection, 2020
- [4]. N. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. Akhmetzhanov, S. Jung, B. Yuan, R. Kinoshita, H. Nishiura, "Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data", Journal of Clinical Medicine, 2020
- [5]. J. Backer, D. Klinkenberg, J. Wallinga, "Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travelers from Wuhan, China, 20–28 January 2020", Europe's journal on infectious disease surveillance, epidemiology, prevention and control, 2020
- [6]. T. Kong, "Longer incubation period of coronavirus disease 2019 (COVID-19) in older adults" Aging Medicine journal, 2020
- [7]. J. Jin, P. Bai, W. He, F. Wu, X. Liu, D. Han, S. Liu, J. Yang, "Gender Differences in Patients With COVID-19: Focus on Severity and Mortality", Frontiers in Public Health Journal, 2020
- [8]. Coronavirus: Why Men May Suffer From Severe Symptoms Of COVID-19 Than Women, According To Studies, Retrieved from <https://timesofindia.indiatimes.com/>, January 2020
- [9]. People with Certain Medical Conditions, Retrieved from <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html>, September 2020
- [10]. I. Sudirman and D. Nugraha, "Naive Bayes classifier for predicting the factors that influence death due to COVID-19 in china.", Journal of Theoretical and Applied Information Technology, 2020
- [11]. A. handay, S. Rabani, Q. Khan, N. Rouf, M. Din, "Machine learning-based approaches for detecting COVID-19 using clinical text data", Nature Public Health Emergency Collection, 2020.
- [12]. D. Novitasari, R. Hendradi, R. Caraka, Y. Rachmawati, "Detection of COVID-19 chest X-ray using support vector machine and convolutional neural network", Communications in Mathematical Biology and Neuroscience, 2020
- [13]. S. Yoo, H. Geng, T. hui, S. Yu, D. Cho, J. Heo, M. Choi, I. Choi, C. Van, N. Nhung, B. Min, H. Lee, "Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging", University of Medicine and Health Sciences, United Arab Emirates, 2020
- [14]. V. Barbosaa, J. Gomesb, M. Santanab, C. Limab, R. Caladoc, "Covid-19 rapid test by combining a random forest-based web system and blood tests", Department of Mechanical Engineering, Federal University of Pernambuco, Recife, Brazil, 2020
- [15]. Transmission of COVID-19 by asymptomatic cases, Retrieved from <http://www.emro.who.int/health-topics/coronavirus/transmission-of-covid-19-by-asymptomatic-cases.html>, January 2020
- [16]. Covid-19 Coronavirus Pandemic, Retrieved from <https://www.worldometers.info/coronavirus/>, July 2020
- [17]. Transmission of SARS-CoV-2: implications for infection prevention precautions, <https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions>, July 2020

- [18]. E. Prabhakar, C. Nalini “Boosted Adaboost to Improve the Classification Accuracy”, Department of Information Technology, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India, 2012
- [19]. Coronavirus disease (COVID-19), Retrieved from <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>, October 2020

Student concentration level monitoring system based on deep convolutional neural network

U. B. P. Shamika*

Department of Statistic and Computer Science,
University of Kelaniya, Sri Lanka
shamikapawani137@gmail.com

W. A. C. Weerakoon

Department of Statistic and Computer Science,
University of Kelaniya, Sri Lanka
chinthanie@kln.ac.lk

P. K. P.G. Panduwawala

Department of Statistic and Computer Science,
University of Kelaniya, Sri Lanka
pkgkpanduwawala@gmail.com

K. A. P. Dilanka

Sri Lanka Telecom PLC
Research & Development Division, Sri Lanka
pasindu@slt.com.lk

Abstract - As synchronous online classrooms have grown more common in recent years, evaluating a student's attention level has become increasingly important in verifying every student's progress in an online classroom setting. This paper describes a study that used machine learning models to monitor student attentiveness to distinct gradients of engagement level. Initially, the experiments were conducted using a deep convolutional neural network of student attention and emotions exploiting Keras library. The model showed a 90% accuracy in predicting attention level of the student. This deep convolutional neural network analysis aids in identifying crucial emotions that are important in determining various levels of involvement. This study discovered that emotions such as calm, happiness, surprise, and fear are important in determining a student's attention level. These findings aided in the earlier discovery of students with poor attention levels, allowing instructors to focus their assistance and advice on the students who require it, resulting in a better online learning environment.

Keywords - Convolutional Neural Network, emotion, Keras, Machine Learning, online learning, student involvement

I. INTRODUCTION

Emotions have an important role in education and in many facets of human existence. Emotions are widely accepted to exist and to be judged. Student involvement is an essential notion in today's education system, and how much information the student receives is equally significant for learning.

The advancement of sophisticated teaching approaches, along with greater computational power, has investigated and resolved numerous research challenges linked to student involvement in the traditional classroom setting, with favorable outcomes. Despite these benefits, current global events have forced students to adjust to the online classroom model. A normal in-person classroom format helps students extend their concentration, develop their critical thinking, and reinforce their meaningful learning experience.

As a result, the research component has expanded to include the issues and obstacles encountered by students during synchronous online classes. Online learning has exploded in popularity in recent years, and it has become a necessary method of continuous learning in the midst of a crisis. Knowing the attention level of students in an online classroom setting is critical for creating an adaptive learning system. Emotions and facial expressions are

important indicators that instructors use to determine a student's attention level, but this is not feasible when learning takes place in a digital environment.

Because of the COVID-19 epidemic, online learning and synchronous online classrooms have become a means of education in recent days. Recognizing students' attention levels with the system they are engaged in working with can change how any teacher interacts with their pupils. Identifying student attention levels will allow you to have a better picture of how they interact with the system and modify your teaching techniques. It also aids in recognizing and categorizing kids depending on their degree of attentiveness. The success of online classrooms is dependent on the outcome of students' knowledge and participation.

Other studies in this field focus on recognizing students' varied emotions (happy, sad, angry, puzzled, disgusted, astonished, calm, neutral) during lectures, laboratories, and class research. The majority of current research in this sector has largely focused on measuring a student's emotional state. Because there is no association model between a student's degree of involvement in class and their emotional state, such research is restricted in its value to teachers.

As a result, in order to make things easier for the teachers, research was conducted to determine if a student is attentive or not throughout class (binary classification on attentiveness). It is always useful to know if students are attentive or inattentive, but most of the time, students are not at either of these extremes. In practice, a student might be half attentive during lectures as well. As a result, a student's attention level may not always be restricted to 0 or 1.

A. Background

We broadened the study to see if there are several categories for classifying student involvement. Therefore, we used a multi-level categorization of student attention level (attentive, partly attentive, and inattentive) in an online classroom setting. The benefit of this approach is that it allows teachers to detect inattentive and partially attentive students early on and offer the necessary assistance, resulting in a better online learning environment.

We suggested a system architecture that makes use of machine learning techniques and a computer vision service. Machine learning techniques are utilized to create a prediction model for each degree of student attention. The computer vision service is utilized to determine the pupils' emotional states. A model is constructed to link emotional states with the level of attentiveness of the pupil.

The first result is the output of one of the most common machine learning models, the Deep Convolutional Neural Network (CNN). Based on their facial expressions, this model was utilized to recognize student involvement. CNN scored the greatest average accuracy of 90.4 percent in the model, suggesting that it is absolutely possible to construct a prediction model for varying degrees of student involvement using information acquired from a recorded video.

The final result highlights the importance of emotion analysis and the prediction model of student attention levels in an online class environment using regression analysis. The rest of the paper is organized as follows. Section II presents the related work. Section III introduces three algorithms and web scraping techniques are used. In section IV, the results and discussions are presented. Final Remarks and References are mentioned in Section V and VI, respectively.

II. RELATED WORK

Monitoring the student learning process and delivering feedback to teachers in the classroom is a recent breakthrough in automated learning analytics. This notion of real-time feedback is made feasible by building the feature set with kinetic data collected from the Kinect One sensor device. In this study, seven different classifiers were evaluated to predict student attentiveness across time and average attention levels [1].

A methodology for detecting student emotions from student interactions with a cognitive tutor for mathematics was described. Cognitive tutors are programmed to respond to the student's actions inside the user interface. The software's log data was gathered, and observations were carried out in the school's computer lab. To evaluate the collected data, classification techniques such as decision tree, step regression, and naive bayes were utilized. The detectors evaluated on re-sampled data obtained 19% more accuracy than the set base rate [2].

A study was undertaken to increase student engagement in E-learning platforms by extracting mood patterns from their facial characteristics. The study aids in the assessment and identification of gaps in sustained attention by a student during an E-learning session. Analyzing moods based on a student's emotional states during an online lecture yielded data that could be easily used to improve the efficacy of the content delivery mechanism inside the E-learning platform. The study looks at whether facial expressions are the most important form of nonverbal communication and identifies the most prevalent facial characteristics that reflect a student's interest in a lecture. To train the models, such as the radial-based Neural Network (NN) model, the Hidden Markov Model, and the Support Vector Machine, a neural network method was employed (SVM). The outcome demonstrates a significant connection with feedback and a success rate of more than 70% in measuring the student's mood [3].

Early on, researchers established a link between visual attention and sadistic eye movement [4], employing the Viola-Jones algorithm to recognize face pictures [5]. To categorize the activities of eye movements, the Support Vector Machine (SVM) was used. These traditional principles served as the foundation for the development of different machine learning approaches.

III. METHODOLOGY

A. Dataset

The "DAISEE: Dataset for Affective States in E-Learning Environments" dataset contains 9068 video snippets captured "in the wild" from 112 users using an HD webcam setup to recognize user affective states, which are raw crowd annotated and associated with a standard annotation built by an expert team of psychologists. According to [6] research, each video was 10 seconds long since this length offered enough information for the labeling action. To mimic an E-learning environment, each participant was shown two separate 20-minute-long films. To capture a focused and comfortable atmosphere, one of the films was instructional and the other was recreational. It enables the capture of natural shifts in user attention levels. The students in this research ranged in age from 18 to 30 years old.

Because it was designed as an E-learning environment, the films were shot in a variety of settings, including dorm rooms, a busy lab area, and a library with varying lighting levels (light, dark and neutral). The video dataset was tagged with several emotional states such as boredom, confusion, engagement, and frustration. Each effect was further categorized into four labels: very low, low, high, and very high.

B. Data pre-processing

The first stage in our study project was to create a dataset from student photos collected in an E-learning environment. The video files were used to extract image frames. Every video is divided into 28 frames with a 20 minute gap. Fig. 1 shows the frames that we got from the videos. The picture dataset was difficult to set up since the films were shot in diverse places with varied lighting conditions. The difficulties included dark picture frames, students who were not within crucial proximity of the webcam, and students who were not within the image frame owing to other distractions. As a result, we concentrated on data collection by centralizing and cropping the face portions at identical pixel size for each frame.

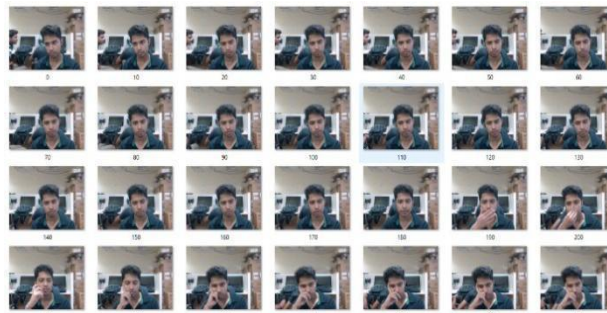


Fig. 1. Cropping the face portions

F. Build a LSTM to take the output

We created a LSTM with an accuracy of 54.4% in order to increase student engagement in the E- learning platform. We used 10 epochs.

IV. EXPERIMENT AND RESULTS

A. Model evaluation outcome for CNN classifier

By adjusting the validation split to 0.2, we were able to employ 10 epochs with the default batch size of 50. For every validation loss inside the model, the density was set to 100, 50, and 10 in the early stopping callback method. It monitors the loss quantity and, when it finds improvements, it is 2% when three densities are used at the same time. Before finishing the 10 epochs, the loss function for our model hit a saturation point of around 0.22, and the total accuracy achieved a high of 90.6 percent. The CNN model was evaluated using the accuracy and loss graphs. When the dataset is balanced across classes, evaluating the model efficiency solely on accuracy and loss value obtained from the validation set may cause difficulties. Figure 8 depicts the construction of the CNN model.

```
Epoch 1/10 [-----] - 283s 118ms/step - loss: 0.6299 - accuracy: 0.7062 - val_loss: 0.4908 - val_accuracy: 0.7728
Epoch 2/10 [-----] - 266s 118ms/step - loss: 0.4673 - accuracy: 0.7830 - val_loss: 0.4862 - val_accuracy: 0.8282
Epoch 3/10 [-----] - 267s 118ms/step - loss: 0.3856 - accuracy: 0.8214 - val_loss: 0.3493 - val_accuracy: 0.8392
Epoch 4/10 [-----] - 270s 112ms/step - loss: 0.3333 - accuracy: 0.8464 - val_loss: 0.3213 - val_accuracy: 0.8335
Epoch 5/10 [-----] - 275s 119ms/step - loss: 0.3011 - accuracy: 0.8610 - val_loss: 0.2918 - val_accuracy: 0.8753
Epoch 6/10 [-----] - 279s 116ms/step - loss: 0.2755 - accuracy: 0.8739 - val_loss: 0.3001 - val_accuracy: 0.8787
Epoch 7/10 [-----] - 276s 119ms/step - loss: 0.2565 - accuracy: 0.8829 - val_loss: 0.2877 - val_accuracy: 0.8797
Epoch 8/10 [-----] - 283s 118ms/step - loss: 0.2455 - accuracy: 0.8903 - val_loss: 0.2533 - val_accuracy: 0.8970
Epoch 9/10 [-----] - 281s 117ms/step - loss: 0.2339 - accuracy: 0.8953 - val_loss: 0.2378 - val_accuracy: 0.9082
Epoch 10/10 [-----] - 283s 118ms/step - loss: 0.2166 - accuracy: 0.9019 - val_loss: 0.2283 - val_accuracy: 0.9060
```

Fig. 8. Build a CNN model

By adjusting the validation split to 0.4, we were able to employ 10 epochs with the default batch size of 10. When it finds improvements, it is 2% when three densities are used at the same time. Before finishing the 10 epochs, the loss function for our model hit a saturation point of around 0.13, and the total accuracy achieved a lower of 54.4 percent. The LSTM model was evaluated using the accuracy and loss graphs. When the dataset is balanced across classes, evaluating the model efficiency solely on accuracy and loss value obtained from the validation set may cause difficulties. Figure 9 depicts the construction of the LSTM model

```
Training Progress
Epoch 1/10 [-----] - 12012s 5s/step - loss: 0.1349 - acc: 0.5404 - val_loss: 0.1335 - val_acc: 0.5447
Epoch 2/10 [-----] - 35806s 15s/step - loss: 0.1334 - acc: 0.5517 - val_loss: 0.1339 - val_acc: 0.5447
Epoch 3/10 [-----] - 35857s 15s/step - loss: 0.1338 - acc: 0.5475 - val_loss: 0.1336 - val_acc: 0.5447
Epoch 4/10 [-----] - 8081s 3s/step - loss: 0.1336 - acc: 0.5476 - val_loss: 0.1337 - val_acc: 0.5447
Epoch 5/10 [-----] - 9093s 4s/step - loss: 0.1336 - acc: 0.5476 - val_loss: 0.1337 - val_acc: 0.5447
Epoch 6/10 [-----] - 13220s 6s/step - loss: 0.1335 - acc: 0.5474 - val_loss: 0.1336 - val_acc: 0.5447
Epoch 7/10 [-----] - 8897s 4s/step - loss: 0.1335 - acc: 0.5476 - val_loss: 0.1335 - val_acc: 0.5447
Epoch 8/10 [-----] - 18161s 8s/step - loss: 0.1335 - acc: 0.5476 - val_loss: 0.1336 - val_acc: 0.5447
Epoch 9/10 [-----] - 7937s 3s/step - loss: 0.1336 - acc: 0.5476 - val_loss: 0.1336 - val_acc: 0.5447
Epoch 10/10 [-----] - 8003s 3s/step - loss: 0.1335 - acc: 0.5476 - val_loss: 0.1335 - val_acc: 0.5447
```

Fig. 9. Build a LSTM model

Figure 10 depicts the connection between the accuracy of the training set and the validation set for each epoch. The graph shows that the accuracy of both the training and validation sets has risen with each epoch. It is not always necessary to take into account the validation learning curve's last data point with the best accuracy of the model. The greatest accuracy of the model reached epoch in our investigation was epoch 10.

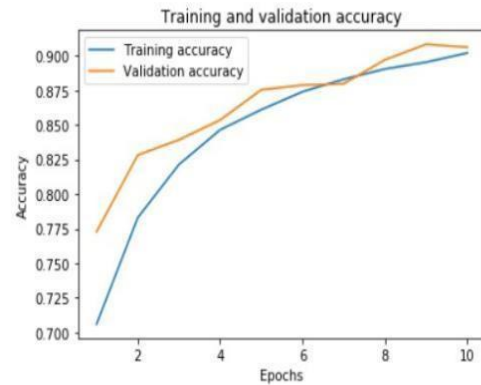


Fig. 10. Accuracy graph: Training vs. validation of CNN

Figure 11 depicts the connection between the accuracy of the training set and the validation set for each epoch. The graph shows that the accuracy of the training set has risen up and after that it has gone down and in the same value, and validation sets have the same value with each epoch. It is not always necessary to take into account the validation learning curve's last data point with the best accuracy of the model. The greatest accuracy of the model was reached each epoch in our investigation.

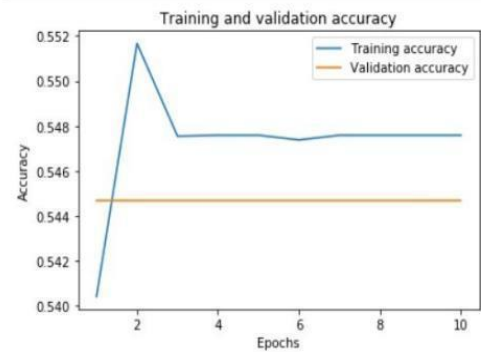


Fig. 11. Accuracy graph: Training vs. validation of LSTM

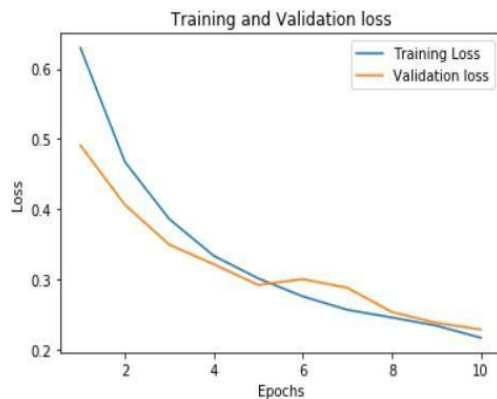


Fig. 12. Loss graph: Training vs. validation of CNN

Fig. 12 depicts the loss function connection between training and validation sets for each epoch. The graph terminates at epoch 10 with the patience parameter set to 6 due to the callbacks adjusted in the CNN model, since the validation loss function detected no progress.

Figure 13 depicts the loss function connection between training and validation sets for each epoch. The graph terminates at epoch 10 with the patience parameter set to 1 due to the callbacks adjusted in the LSTM model, since the validation loss function detected no progress.

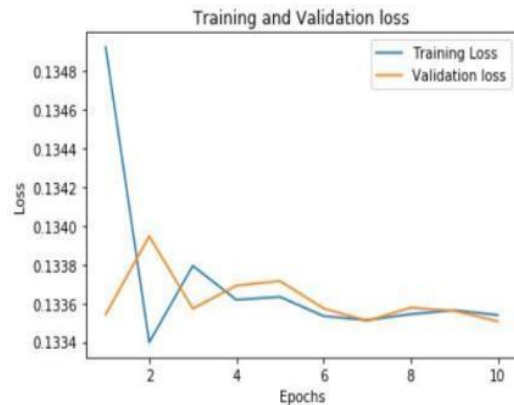


Fig. 13. Loss graph: Training vs. validation of LSTM

B. Result discussion

The efficiency and accuracy of forecasting student involvement levels was investigated in this study using a machine learning model. The machine learning model was evaluated using a balanced dataset. Based on the performance measures, it is possible to infer that the deep learning CNN model is more effective than LSTM. Despite its superior accuracy, the CNN model requires more time to train but the LSTM model wants more time than CNN. When compared to the LSTM models, the CNN model produced the largest proportion of erroneous classifications. In summary, the CNN model outperformed all other models in all measures, with the greatest accuracy of 90.6 percent.

V. CONCLUSION

The outcomes of this study enable teachers to properly detect inattentive and partially attentive pupils, which contributes to a better online learning environment. It enables teachers to help students in need, resulting in a better learning experience. Our research looked at three machine learning models for measuring student involvement based on their emotions. The CNN model was chosen as the appropriate machine learning model to measure a student's attentiveness based on their emotional state by the research methodology employed in this study, with a prediction accuracy of 90.6 percent. The influence of emotion state rage on the connection between emotion states and student engagement levels was also investigated in this study. Understanding the confounding influence of rage on other emotional states has enabled us to identify important emotions displayed by inattentive and partially attentive pupils statistically. Based on the findings of this study, we can infer that the deep CNN model provides a dependable and accurate platform for assessing different

gradients of student involvement based on their facial expressions.

This study effort can be developed in a variety of ways. For future studies, the CNN model may be modified to use computational resource-intensive architectures such as VGG16, VGG19, and ResNet, which would increase the machine learning model's prediction accuracy.

This study could be expanded by incorporating a broader range of engagement levels to gain a more detailed understanding of students' attention levels and facial expressions. Furthermore, the research platform may be enhanced by integrating a web-based application that converts live video files into pictures, providing real-time data to the prediction model. A student survey may be included at the end of each online session to produce user-driven feedback data points to enhance and validate the machine learning models' prediction metrics.

Another goal of the research is to do picture auto-labeling rather than manual labeling. Once the relationship and relevance of emotions and engagement levels has been painstakingly determined, the cloud-based program may function as an AI expert in the labeling process. This approach is useful for dealing with big datasets.

REFERENCES

- [1] K. Janez Zaleteli, "Predicting students' attention in the classroom from Kinect facial and body features," 2017.
- [2] S. j. R. S. P. B. A. C. D. Doborah Rudnick, "The Role of Landscape Connectivity in Planning and Implementing Conservation and Restoration Priorities. Issues in Ecology".2012.
- [3] A. K. B.K. Poornima, "Predicting learner preferences from emotions using Deep Learning Techniques," 2016.
- [4] S. Heiner Deubel Werner, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," 1996.
- [5] L. H. W. W. W. Thomas A. Dingush, "Development of models for on-board detection of driver impairment," 1987.
- [6] Z. S.C. L. F. J. R. M. Jacob Whitehill, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," 2014.
- [7] S. E. H. Erin S. Lane, "A New Tool for Measuring Student Behavioral Engagement in Large University Classes," 2015.

TrackWarn: An AI-driven warning system for railway track workers

M. I. M. Amjath*

Division of Information Technology
Institute of Technology, University of Moratuwa, Sri Lanka
amjathm@itum.mrt.ac.lk

S. Kuhanesan

Department of Physical Science
Faculty of Applied Science, University of Vavuniya, Sri Lanka
kuhan9@yahoo.com

Abstract - This contribution focuses on developing an AI-driven warning device to ensure the safety of railway track workers. Recent studies clearly show that track workers safety has become a major challenge for the railway industry despite many precautionary measures that are implemented. In this regard, many technological solutions have been proposed and developed to warn track workers of the approaching trains. However, the cost and complexity are the drawbacks of these systems. Therefore, we introduce TrackWarn, a low-cost portable smart gadget that detects the sounds of the approaching trains and provides a warning signal to track workers via a phone call. TrackWarn uses a state-of-art Convolutional Neural Network (CNN) that utilizes environmental sounds and spectrograms to classify if the train is approaching or not. This model achieves an average classification accuracy of 92.46%. With the help of Arduino Nano 33 BLE Sense microcontroller, the whole system becomes very handy and potable. This paper addresses the design of the TrackWarn and the results obtained with respect to the various test cases. Further, the performance and communication challenges are also described in detail.

Keywords - Arduino Nano33 BLE sense, CNN, smart, track workers, spectrograms

I. INTRODUCTION

Railway track workers play a crucial role in helping to ensure safe train transport. They usually carry out mechanical work associated with railroad systems without any automated safety systems in place. Due to improper safety measures, train accidents among railway track workers are frequent. These unforeseen accidents ultimately result in loss of life and severe injuries. Although the modern rail industries implement various efforts to mitigate track workers accidents, the accident rate escalates every year unevenly. The rail accident investigations reports reveal that the unawareness of approaching train is one of the primary causes for these unforeseen accidents [1]–[3].

At present, there are two techniques that are widely used to warn people of the approaching trains: automatic track warning systems (ATWS) and lookout-operated warning system (LOWS). Based on deployment ATWS can be classified as train/wayside mounted device and portable zone device. While the train/wayside mounted devices are permanently installable devices, the portable zone devices are temporarily affixed on the railroad corridor. These devices notify the arrival of trains by communicating the specific device carried by the track workers. Although many commercialized automatic track warning systems (ATWS) [4][5] are available in the market, most developing countries still rely heavily on a lookout-operated warning system (LOWS) for ensuring the safety of track workers. In LOWS, a member of the

team is assigned to monitor and alert the arrival of trains. Moreover, the protection of track workers solely depends on the lookout operator. As part of this contribution, we figure out the problems of the existing ATWSs and propose a novel technique to ensure the safety of track workers with the help of AI.

The train detection task is generally considered the most challenging part of any ATWS devices. At present, there are two different techniques that are generally carried out to detect the trains: track circuit and axle-counter [6]. In the track circuits, occupancy of a section of the track has been determined by continuous sensing the short circuit. This continuous sensing technique, can also be used in condition monitoring, for example to detect broken rails. However, power failure, leaves on the track, rusting, contaminants on railheads can cause the faulty result. In addition to this, the track circuit requires continuous maintenance for prolonged use.

On the other hand, the axle-counters count the axles of the trains by measuring the inductance changes [7]. The latest axle-counters have the capability of finding the directions and speeds of the trains as well. However, power supply failures and wheel rocks are the two causes that make this system fail in counting axles. In addition to this, they are more expensive and require long installation times.

In addition, various low-cost technological solutions have been proposed and developed to address the problem of accurately detecting the locations of trains on the railways. These include systems based on global positioning system (GPS) technology [8]–[10], RFID technology [11], wireless sensor networks (WSNs) [12], [13], GSM technology [14], Image processing with vibration sensors [15], [16], and weighing detectors [17], accelerometers sensors [18], [19], coding and transmitting signal measured in track circuits [20]. In particular, the adoption of GPS technology may fail when the trains travel under bridges or within long tunnels [21], [22]. However, all of these methods yield a high error rate for critical decisions. Therefore, we decided to apply the sound classification technique to detect the approaching trains.

With the advent of high-performance computing, deep learning algorithms such as neural networks, recurrent neural networks, convolutional neural networks yield negligible error rates. Especially in automatic voice recognition and computer vision, deep learning has been reached human levels of detection.

The convolutional neural networks are the popular multi-layer architecture that specially applied in computer vision associated projects. However, recent studies prove CNNs are also applicable for automatic voice recognition using spectrogram images. Therefore, we employ a state-

of-art CNN architecture that utilizes sound and spectrograms to classify if the train is approaching or not.

The machine learning techniques enable the Internet of Things (IoT) to achieve its extreme level in a wide variety of applications ranging from tiny insect tracking to planets monitoring. Therefore, we analyzed several AI-enabled microcontrollers to successfully execute our deep learning algorithm. As the result of this study, we chose the Arduino nano 33 BLE sense microcontroller board to deploy our deep learning algorithm. Arduino Nano 33 BLE Sense microcontroller has a variety of built-in sensors such as accelerometer, compass, temperature, microphone, etc. In addition to this, it also supports wireless connections such as radio, Bluetooth [23].

Seamless communication is one of the crucial parts of the ATWS. We use a SIM800L GSM module that supports quad-band GSM/GPRS networks. Low cost and small footprint make this module suitable for any embedded projects that require long-range connectivity. It well operates at 3.7V with an external antenna.

The rest of the paper is organized as follows: In Section 2, we describe the existing automated solution that use acoustic features to detect the trains. In Section 3, we detail the methodology that we used to build TrackWarn. In Section 4, we showcase our results and discuss possible explanation. Finally, we draw our conclusion and future work in Section 5.

II. LITERATURE REVIEW

The trains produce various types of sounds such as the horn, whistle, traction, rolling and aerodynamic effects. Based on this, various acoustic feature-based automated systems have been proposed to detect the trains. Sato et al. proposed a system to detect passing trains using the mobile devices of commuters [14]. This system analyses the environment sounds and predicts the probability of train passing by the use of a logistic regression model and hysteresis thresholding. Before the analysis, a low-pass filter is applied to reduce the environmental noise. Furthermore, the location calculated by the GPS sensor at the train detected point is shared with registered authorities through a central server. However, the authors fail to discuss the detection efficacy with the distance between mobile devices and railroad.

In [22], a mobile phone-based train-localization system is proposed with the help of acceleration and microphone sensors. The microphone captures the high frequency distinct sounds of the train passing the rail joint to estimate the speed of the trains.

Singhal et al. proposed a level crossing warning system to alert road drivers of approaching trains [24]. The system takes composite sound signals (train and surrounding sounds) as input and filters out sound pressure levels between 0 to 65 dB using a band stop and equiripple filters. The filtered signal is then compared with the average sound pressure level (given by $-0.241 * \text{distance}(\text{vehicle}) + 85.78$ dB) to detect the approaching trains at level crossings. Although the authors mention the accuracy of this system is 95%, the various test cases and the ways of affixing circuits on the road vehicles were not discussed properly.

A group of researchers applied Recurrent Neural Network (RNN) based sound recognition system to detect the trains at the level crossings [25]. The system utilizes the

mixing sounds and Mel Frequency Cepstral Coefficient (MFCC) to classify the presence of trains. First, the Authors capture specific sounds such as aircraft, car, train, rain, thunder from online corpus as well as live recordings. Subsequently, with the use of NCH software, the train sound is mixed with other sounds into two categories such as two sound mixture and three sound mixture. Thereafter 12 coefficients per frame from both categories are extracted. Finally, these features are used to train RNN with the backpropagation algorithm. Moreover, the scaled conjugate gradient algorithm (SCG) is designed to reduce the time consumed in line-search. Further, the authors stated that high accuracy (90%) found in both train+rain and train+aircraft+car mixtures.

As per the literature reviews, we believe using deep learning algorithms, the sound sample of trains and the environment (noise) can be analyzed further to produce a robust prediction model with high accuracy.

III. MATERIALS AND METHODS

A. Data acquisition

First, we determined five various locations such as remote areas, busy surroundings (near the market), seaside, near the airport, and tunnels to collect the recordings. Further, we decided to use Samsung galaxy grand prime and Apple iPhone 9 to collect the trains' sound within the 10m range from the railway track. In each location, 10 different trains' sound were recorded, which is 7 min long in total. In addition, to make a more robust classification model, environmental sounds such as thunderstorms, helicopter, aeroplane, road traffic, and background sounds also downloaded from the Kaggle corpus and labelled as noise. The recordings collected for classification are shown in Table I.

TABLE I. TYPES OF SOUNDS AND THEIR DURATION

Train Sound	Length
Train (50)	7 min
Noise	Length
Thunderstorms	1 min
Aeroplane	1 sec
Road traffic	2 min
Helicopter	1 min
Background	5 min

B. Sample preprocessing

Since the sample rate of mobile recordings is 48 kHz, we used Audacity 3.0.2 to resample them to 16 kHz, which is the actual sampling rate of Arduino nano 33 BLE sense. Subsequently, the resampled recordings were exported as a .wav format with 32-bit depth encoding.

C. Model configuration

In the model building process, a window with the size of 1sec with a window increase of 100 milliseconds is used to extract unique features from each raw sample. These windows (Spectrograms) are fed into the CNN model during the training process. Further, the number of epochs, learning rate, and the confidence for our CNN set as 30, 0.005, and 0.7 respectively based on the experiments. The feature extraction process for a raw data is shown in Fig. 1.

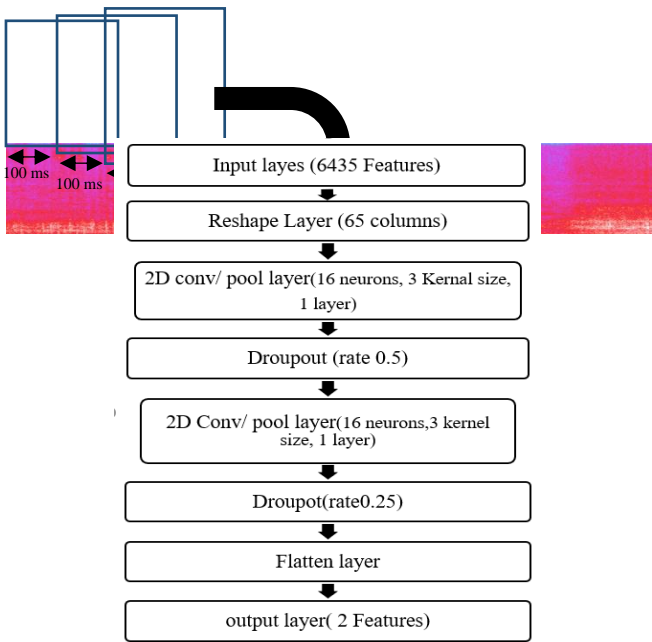


Fig. 1 Feature extraction process

D. Device setup

The SIM800L GSM module and Arduino Nano 33 BLE Sense microcontroller board are powered up using two separate 9V batteries. Two LM2596 DC-DC step-down buck converters modules are used to provide 3.7V and 3.3V to the GSM module and microcontroller respectively. The circuit diagram of TrackWarn is depicted in Fig. 2. Since Dialog Axiata PLC has many subscribers [26], we decided to use Dialog SIM for the GSM module. Two predefined mobile numbers (Dialog) are stored in the EEPROM of the Arduino board to give alert calls when the gadget detects a train. Further, the trained CNN model is deployed to the microcontroller board to detect the trains. Finally, all the components are fixed in a compact box to use

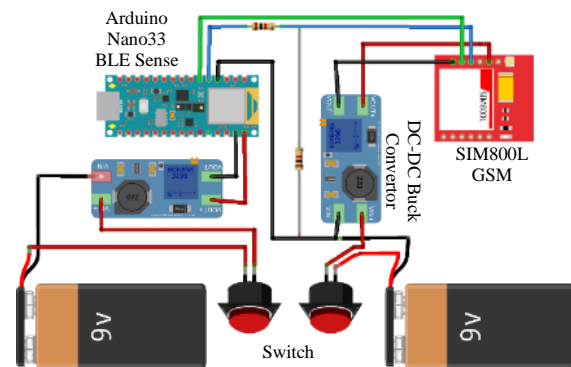


Fig. 2 Circuit diagram

E. Workflow of TrackWarn

The system is set to send an active SMS to the stored numbers every 15 min to ensure the system is kept working without any system failure. In addition, we introduced a

counter variable to ensure the approaching train. In every correct target (train sound) prediction, the counter value increases by 1. When the counter value equals 5, the system confirms the passing of a train. In consequence, the alert calls have been triggered to respective track workers successively. Finally, the system reverts to its initial state. In case the counter value is not increased by 1 within 1.5 sec, the system reset the counter to 0. The clear workflow of TrackWarn is shown in Fig. 3.

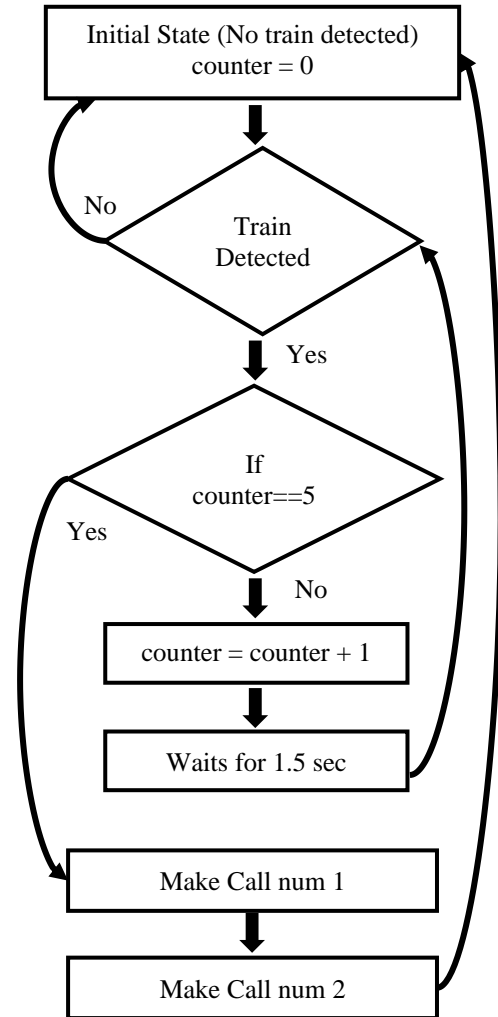


Fig. 3 Workflow of TrackWarn

IV. RESULTS AND DISCUSSIONS

A. Model performance

The gadget is tested in a real environment to calculate the model efficacy. The model achieves 92.46% accuracy for unseen data with the feature extraction and inferencing times 77ms and 508ms respectively in the Arduino nano 33BLE sense. In addition, the peak RAM usage is calculated as 129.7KB. This interprets the model is optimally working for the Arduino nano 33 BLE Sense microcontroller. However, the significant accuracy loss occurred during the thunderstorms. Therefore, various thunderstorms raw data is required to improve the accuracy level.

B. Connectivity test

First, we selected the Western and Central provinces of Sri Lanka to conduct the communication test since, as shown in Fig. 4 the coverage (Dialog) of Western province is comparatively higher than other provinces whereas many tunnels are found in the central province according to [27].

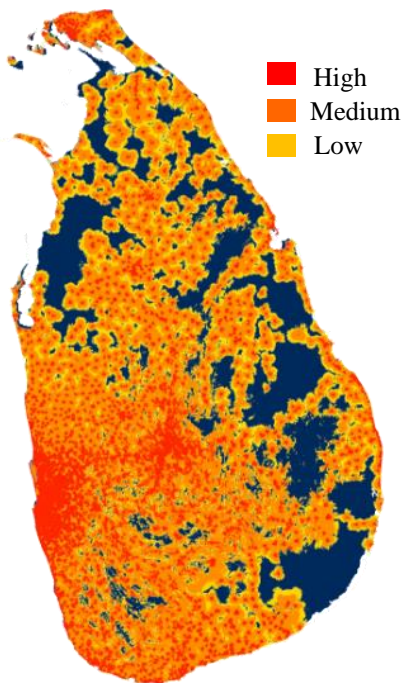


Fig. 4 Dialog coverage map in 2021

Further, we selected 20 different coordinates for various areas from both provinces. The areas and connected calls are depicted in the TABLE II.

TABLE II. CONNECTED CALLS

Area	Call Test
Remote Areas	10
Towns	18
Seaside	16
Tunnels	6

In the remote areas, 10 calls were connected successfully. In the towns and seaside, 18 and 16 calls were connected respectively. However, in the tunnels, the system was able to connect only 6 calls due to poor signal. In remote areas and tunnels the gadget experience poor connectivity. This system can be tested with various SIMs or any other specific radio frequency transmitters to avoid these connectivity issues.

V. CONCLUSION

The safety of track workers is a major concern for the railway industry nowadays. Unawareness of approaching trains causes many fatal accidents among the track workers community. Since the existing automated systems are complex and costly, track workers prefer the look-out method (manual) to alert the track workers. Our TrackWarn uses state-of-art CNN architecture to detect the

trains and alert track workers via a phone call. The testing results of our CNN model shows the trains' sound and noises can be successfully classified with an accuracy of 92.46% within the 10m recording range from the railway track. Further, this outperforms the existing complex systems. Since this gadget is inexpensive and simple, anyone can handle it easily. Since this system contains a fail-safe mechanism, the failures in any components can be easily identified with the constant interval SMSs. In the future, we will use a keypad with an LCD to add dynamic numbers and to change the internal configurations. According to the Table II, some points in various areas have signal problems due to less coverage. With the use of appropriate SIM, multiple SIMs, or specific frequency transmitter, this problem can be solved in future. Further, the parallel call features will also be included for various SIMs to alert at the same time. We ensure the usage of this smart gadget will mitigate track workers' accidents and help to save the country's economy.

REFERENCES

- [1] D. Moy, "Rail Accident Report," Transport, no. November 2005, 2006.
- [2] "Roade rail worker was killed by train while walking along track -BBC News." <https://www.bbc.com/news/uk-england-northamptonshire-57426850> (accessed Jul. 11, 2021).
- [3] "Track worker killed after becoming 'habituated' to train warning horns | New Civil Engineer." <https://www.newcivilengineer.com/latest/track-worker-killed-after-becoming-habituated-to-train-warning-horns-09-06-2021/> (accessed Jul. 11, 2021).
- [4] "Track Warning Systems | Rail Sector | RSS Infrastructure." <https://www.rssinfrastructure.com/track-warning-services/> (accessed Jun. 20, 2021).
- [5] "BitFox Site Safety Division." <http://www.bitfox.it/?id=8&lang=en> (accessed Jun. 20, 2021).
- [6] A. Solution, "Track Circuits vs . Axle Counters," 1872.
- [7] "Axle Counter |." <http://www.railsystem.net/axle-counter/> (accessed Jul. 11, 2021).
- [8] "GPS.gov: Rail Applications." <https://www.gps.gov/applications/rail/> (accessed Apr. 14, 2021).
- [9] R. I. Rajkumar, P. E. Sankaranarayanan, and G. Sundari, "GPS and ethernet based real time train tracking system," Proc. 2013 Int. Conf. Adv. Electron. Syst. ICAES 2013, pp. 282–286, 2013, doi: 10.1109/ICAES.2013.6659409.
- [10] G. Hemanth Kumar and G. P. Ramesh, "Intelligent gateway for real time train tracking and railway crossing including emergency path using D2D communication," 2017 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2017, no. Icices, 2017, doi: 10.1109/ICICES.2017.8070779.
- [11] B. K. Cho, "RFID antenna for position detection of train," Lect. Notes Electr. Eng., vol. 309 LNEE, pp. 903–908, 2014, doi: 10.1007/978-3-642-55038-6_136.
- [12] P. Fraga-Lamas, T. M. Fernández-Caramés, and L. Castedo, "Towards the internet of smart trains: A review on industrial IoT-connected railways," Sensors (Switzerland), vol. 17, no. 6, 2017, doi: 10.3390/s17061457.
- [13] E. Berlin and K. Van Laerhoven, "Sensor networks for railway monitoring: Detecting trains from their distributed vibration footprints," Proc. - IEEE Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2013, pp. 80–87, 2013, doi: 10.1109/DCOSS.2013.38.
- [14] K. Chetty, Q. Chen, and K. Woodbridge, "Train monitoring using GSM-R based passive radar," 2016 IEEE Radar Conf. RadarConf 2016, 2016, doi: 10.1109/RADAR.2016.7485069.
- [15] D. Wang and Y. Ni, "Wireless sensor networks for earthquake early warning systems of railway lines," Lect. Notes Electr. Eng., vol. 148 LNEE, pp. 417–426, 2012, doi: 10.1007/978-3-642-27963-8_38.
- [16] M. I. M. Amjath and T. Kartheeswaran, "An Automated Railway Level Crossing System," 2020.
- [17] "AMTAB Advanced Measurements Technologies AB." https://www.amtab.se/?gclid=EAIaIQobChMIur7J24P97wIVwX8rCh0BogMrEAAyASAAEGLDEfD_BwE (accessed Apr. 14, 2021).
- [18] L. Angrisani, D. Grillo, R. Schiano Lo Moriello, and G. Filo,

- “Automatic detection of train arrival through an accelerometer,” 2010 IEEE Int. Instrum. Meas. Technol. Conf. I2MTC 2010 - Proc., no. i, pp. 898–902, 2010, doi: 10.1109/I2MTC.2010.5488089.
- [19] H. Ardiansyah, M. Rivai, and L. P. E. Nurabdi, “Train arrival warning system at railroad crossing using accelerometer sensor and neural network,” AIP Conf. Proc., vol. 1977, no. June 2018, 2018, doi: 10.1063/1.5042999.
- [20] P. Donato, J. Ureña, J. J. García, M. Mazo, and Á. Hernández, “Use of coded signals to wheel train detection,” IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA, vol. 2, no. January, pp. 685–691, 2003, doi: 10.1109/ETFA.2003.1248765.
- [21] K. Sato, S. Ishida, J. Kajimura, S. Tagashira, and A. Fukuda, Intelligent Transport Systems for Everyone’s Mobility. Springer Singapore, 2019.
- [22] D. Su, S. Sano, T. Nagayama, H. Tanaka, and T. Mizutani, “Train localization by mutual correction of acceleration and interior sound,” Int. Conf. Adv. Exp. Struct. Eng., vol. 2015-Augus, 2015.
- [23] “Arduino Nano 33 BLE Sense | Arduino Official Store.” <https://store.arduino.cc/usa/nano-33-ble-sense> (accessed Apr. 14, 2021).
- [24] V. Singhal, S. S. Jain, and M. Parida, “Train sound level detection system at unmanned railway level crossings,” Eur. Transp. - Trasp. Eur., no. 68, pp. 1–18, 2018.
- [25] S. Ajibola Alim, N. K. B. A. Nahrul Khair, and M. Mozasser Rahman, “Level crossing control: A novel method using sound recognition,” Eng. J., vol. 17, no. 3, pp. 113–118, 2013, doi: 10.4186/ej.2013.17.3.113.
- [26] “Group Overview.” <https://www.dialog.lk/browse/aboutPromo.jsp?id=onlinefld70023> (accessed Jul. 14, 2021).
- [27] “RailwayTunnels in Sri lanka.” <https://www.podimenike.com/2010/11/railwaytunnels-in-sri-lanka.html> (accessed May. 10, 2021).

Application of AlexNet convolutional neural network architecture-based transfer learning for automated recognition of casting surface defects

Shiron Thalagala*

Dept. of Electromechanical Engineering
University of Macau, China
shironceylon@gmail.com

Chamila Walgampaya

Dept. of Engineering Mathematics
University of Peradeniya, Sri Lanka
ckw@pdn.ac.lk

Abstract - Automated inspection of surface defects is beneficial for casting product manufacturers in terms of inspection cost and time, which ultimately affect overall business performance. Intelligent systems that are capable of image classification are widely applied in visual inspection as a major component of modern smart manufacturing. Image classification tasks performed by Convolutional Neural Networks (CNNs) have recently shown significant performance over the conventional machine learning techniques. Particularly, AlexNet CNN architecture, which was proposed at the early stages of the development of CNN architectures, shows outstanding performance. In this paper, we investigate the application of AlexNet CNN architecture-based transfer learning for the classification of casting surface defects. We used a dataset containing casting surface defect images of a pump impeller for testing the performance. We examined four experimental schemes where the degree of the knowledge obtained from the pre-trained model is varied in each experiment. Furthermore, using a simple grid search method we explored the best overall setting for two crucial hyperparameters. Our results show that despite the simple architecture, AlexNet with transfer learning can be successfully applied for the recognition of casting surface defects of the pump impeller.

Keywords - automated inspection, casting defect detection, convolutional neural networks, hyperparameters, transfer learning

I. INTRODUCTION

Cost and time effective quality management [1] in a manufacturing operation is a significant aspect regardless of the domain. Nevertheless, producing higher quality products that yield higher customer satisfaction with the least cost and time has been a challenging task for manufacturing firms. Product visual inspection for defects, being a crucial element in quality management, is increasingly automated in present manufacturing firms due to numerous benefits [2] which ultimately result in higher business performance.

Metal casting is a manufacturing process where molten metals are solidified in a mold to obtain the required shape [3]. Though metal casting processes span across a wide variety of metals and several specific techniques, the most common defect types can be categorized as blowholes, shrinkages, cracks, sand inclusions, defective surfaces, and mismatches [4]. Proper identification of casting defects effectively is vital as unnoticed defective finished products which go to the customers' hand can cause fatal mechanical failures [5]. Automating the process of visual inspection of metal castings with the aid of intelligent systems [6] is beneficial in terms of accuracy, inspection time, and cost. Especially, it prevents the facilitation of human labor in

hazardous environments including costly concerns of the safety of such employees.

The visual identification process of defects in metal castings needs to entertain two main requirements during the process of inspection. One is the identification of surface defects on the casting, and two is the identification of defects located inside the cast product which are not visible to the naked eye. The latter is relatively complicated and expensive, commonly accomplished by non-destructive testing (NDT) methods such as ultrasonic testing, eddy-current testing, magnetic particle testing, and radiographic (X-ray) testing [7].

The main purpose of non-destructive testing is to identify defects located inside the test object by the naked eye without damaging the object. X-ray computer tomography (XCT) is a widely used non-destructive casting inspection method that generates two-dimensional/three-dimensional images of the object interior structure [8]. Inspecting such interior images along with the inspection of casting surfaces of every manufactured product is necessary to maintain lower defect levels. Not only the interior images generated by XCT but also the conventional photographs of the casting surfaces can be fed into intelligent systems that use image processing and machine learning techniques for recognition, categorization, and localization of casting defects [6].

Convolutional neural networks (CNNs), which lie in the domain of machine learning have been well studied for their appropriateness in computer vision applications [9]. The structure of CNNs is analogous to that of the connectivity pattern in the visual cortex of the human brain. CNNs are capable of extracting features by themselves and there is no need to perform manual feature extractions in the input images which, however, is essential in some primitive machine learning techniques. Fig. 1 illustrates the difference in image classification approach between primitive machine learning methods and CNNs. Hence, over the last decade, CNNs have successfully applied for automated inspection of casting defects with varying performances [10]–[12]. Since the onset of the CNNs, numerous architectures have been generated by carrying out structural reformulations, regularizations, parameter optimizations, etc. [13]. AlexNet [14] is a prominent CNN architecture that performs competently in the tasks of image recognition. While CNNs perform better in the realm of images over traditional machine learning techniques still some common hindrances for lack of generalization of models are not fully conquered by research. Specifically, models trained for the same feature space and the same distribution drastically reduce their performance when

tested on a different dataset with different feature distribution.

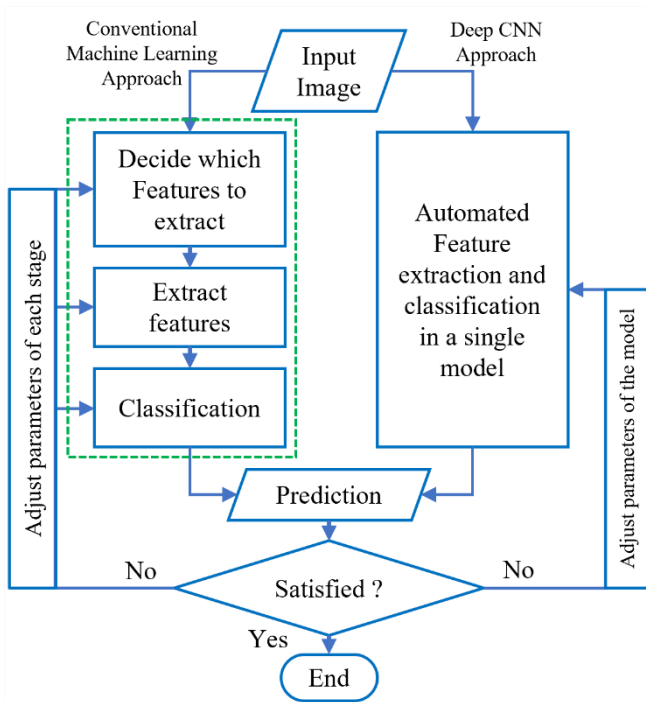


Fig. 1. Difference in image classification approach between conventional machine learning techniques and CNNs

Transfer learning has significantly addressed the issue of using a single CNN model for the recognition tasks in different image fields. Transfer learning in CNNs is the use of knowledge gained by training a model in one domain, on another in a dissimilar domain [15]. It helps not only to mitigate the computational cost in training but also to generalize the CNN models over different domains. Moreover, transfer learning is beneficial in situations when adequate data is lacking for learning from scratch. Despite the successful applications of transfer learning in automated recognition of casting defects, selection of the unique CNN model parameters (hyperparameters) [16] relevant to each casting image dataset is still necessary.

This paper focuses on: (1) investigating the application of an AlexNet CNN model which is pre-trained on an entirely different larger dataset to recognize images of casting surface defects, and (2) optimizing hyperparameters for best performance. The pivot of this study is a classification task to segregate faulty casting products in a manufactured batch through pattern recognition. Further classification of defect types or localization of defects, however, are out of the scope of this study. The dataset [17] used in the study comprised only two classes named ‘defect’ and ‘defect-free’ representing images with one or more defects, and images without any visible defect, respectively.

II. RELATED WORK

Recognition and localization of manufacturing defects using machine learning techniques are explored in numerous studies over the recent years with the focus of achieving high-performing robust models. Several primitive computer vision techniques were used by several authors at the early stages of the pattern recognition field. A

background subtraction method followed by a thresholding algorithm is proposed in [18]. The idea is to generate an image with the same pixel intensities as the original image except defective regions using low-pass filtering [19]. The newly constructed image is then subtracted from the original image resulting in a residual image containing only defective regions. In [20] the Modified Median filter, MODAN-Filter, is proposed to identify contours of the casting defects from non-defective areas with a function to calculate the pixel values of the background image. Furthermore, equations of the MODAN-Filter are generalized in [21] to achieve higher robustness. These filtering-based methods that depend on optimum filter parameters, however, can be unreliable when image noise is present substantially. In [22], the wavelet transform method is described as a potential technique to identify certain casting defect types.

Feature-based detection of casting defects is another trending approach that can be seen applied in [10], [23]. During this process, each pixel is classified as a defect or not based on the features calculated using sets of nearby pixels. Common features include statistical descriptors such as mean, standard deviation, skewness, kurtosis, energy, and entropy [24]. In [25], a hierarchical and a non-hierarchical linear classifier has been implemented based on six geometric and gray value features namely contrast, position, aspect ratio, width-area ratio, length-area ratio, and roundness. A Fuzzy logic-based method for the detection and classification of defects that appear in the radiographic images is proposed in [11].

Many modern studies have tested numerous CNN architectures in terms of the performance and accuracy of casting defect recognition tasks. Among those, Region-Based Convolutional Neural Networks (R-CNNs) are used for the automatic localization of casting defects significantly [12]. R-CNNs are capable of setting bounding boxes around categorical patches in the images where this can be implemented easily to mark the defects in the casting defect images. In [10], a new CNN architecture called Xnet-II is introduced which comprises five convolutional and fully connected layers. Moreover, they have used a dataset generated through simulation using Generative Adversarial Networks (GAN) [27] instead of real casting defect images.

Lack of sufficient data is a common problem in the machine learning domain. Data augmentation where new images are generated by augmenting the existing images of casting defects efficiently and accurately with low background noise is proposed in [28]. This mechanism is based on a traditional image enlargement technique, precisely forcing the CNN to learn more in the regions of the image that need high attention in order to perform better in the classification task. On the other hand, transfer learning is effective not only in the lack of data scenarios but also in respective to the robustness of the model. In [5], the authors use ResNet CNN architecture for the recognition of casting defects. When compared to AlexNet, due to the architectural complexity, ResNet needs a significantly larger number of computations which ultimately consumes higher computational resources.

III. METHODOLOGY

In this section, we explain the approach used to recognize casting surface defects of an industrial product using AlexNet CNN architecture and transfer learning.

Improving the accuracy and the robustness of the AlexNet architecture using transfer learning in the context of casting defect detection is the major objective of this study.

A. Description of the dataset

The dataset, obtained from Kaggle datasets [17], consists of images of a submersible pump impeller which is manufactured as a casting product. All the images depict the top view of the impeller and belong to two classes. The images that exhibit at least one casting defect on the surface of the impeller are labeled as defect while all the other images, conversely, are labeled as defect-free. i.e., Any casting defect on the surface that cannot be identified by the naked eye from the images is labeled as defect-free.

This dataset is collected under stable lighting conditions with a Canon EOS 1300D DSLR camera. The dataset contains a total of 1300 gray-scaled images with the dimensions of each as (512×512) pixels. Among those, 781 images are labeled as defect, and the remaining 519 images are labeled as defect-free. Fig. 2 shows eight sample images (size and the resolution is altered in order to adhere to paper guidelines) and corresponding labels which are randomly picked from the two classes. All the images acquired for this study from the original dataset are only the raw images and the augmentation is done as a part of this study.

B. Image augmentation

In this section, we discuss the image data augmentation techniques applied for the dataset before the experimentation. As in [29], several classical techniques that belong to geometrical and color-based transformations were applied randomly to yield higher variability. As per geometric transformations, rotation, shearing, mirroring, scaling (zoom-in/out) and translation were applied. Nevertheless, color space transformations were limited only to change of apparent brightness as the dataset already contains grayscale images. Moreover, apparent brightness change (performed randomly) in each pixel intensity of an image was restricted to a maximum of 20% (either increase or decrease) of the current intensity. It prevents introducing new defect regions which were not in the original image or disappearing significant regions of the image with low intensities by further decreasing the intensity.

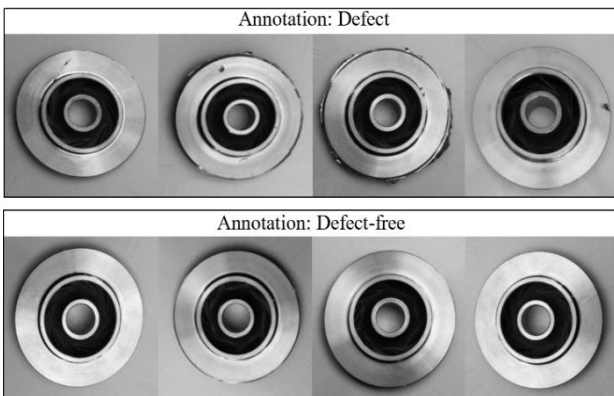


Fig. 2. Randomly picked eight number of sample images from the dataset annotated as defect and defect-free

Fig. 3 shows one sample image (annotated as defect-free) and corresponding images synthesized by augmenting that image using all the techniques used in this study.

Among synthesized images, 5814 are annotated as defect-free and 7668 are annotated as defects. At last, all the images were resized to (224×224) pixels. Throughout all the experimentation, training and validation data split is diversified by changing the amount of training data to 20%, 40%, 50%, 60%, and 80% to understand the capacities of generalization of the used models [30]. Hereinafter, the ratio between the training image set and the validation image set will be referred as train-test split ratio.

C. Non-parametric classification using the k-nearest neighbor algorithm

K-Nearest Neighbor (KNN) algorithm, which is a basic supervised machine learning algorithm, is used to investigate the capability of performing the classification task using raw pixel intensities as the input and without any sophisticated feature extraction techniques.

In the context of computer vision, the KNN algorithm performs classification of the data points (pixel values) based on the distance between them and with the assumption that similar features exist nearby. Common methods of calculating the distance include the Euclidean distance:

$$d(p, q) = \sqrt{\sum_{i=1}^N (q_i - p_i)^2} \quad (1)$$

and the Manhattan/city block distance:

$$d(p, q) = \sum_{i=1}^N |q_i - p_i| \quad (2)$$

where $d(p, q)$ is the distance between two p and q points in the image spatial domain with N pixels.

In this study, the KNN algorithm is performed with the raw pixel intensities of casting images without any feature extraction with the Manhattan distance calculation metric and the k value equals to five. The variation of precision, recall, and f1-score is observed by varying the train-test split ratio.

D. CNN architecture

Despite the emerging CNN architectures, we base our model around AlexNet architecture due to three reasons. (1) To the best of our knowledge, application of AlexNet based transfer learning in recognition of casting defects is not addressed in past literature, (2) AlexNet is applied in a diverse set of deep learning problems witnessing promising results [30], [31], (3) AlexNet, which was proposed in 2012, is regarded as the first deep CNN architecture which showed pioneering results in image recognition and classification tasks [32]. We show that AlexNet is sufficiently deep and reliable for a modest classification of casting surface defects when compared to other deeper sophisticated architectures born after AlexNet, if hyperparameters are properly optimized.

AlexNet consists of five 2D convolutional layers (Conv2D) followed by three fully connected layers (FC). The build of the AlexNet architecture is illustrated in Table I and it is constructed with several common CNN components

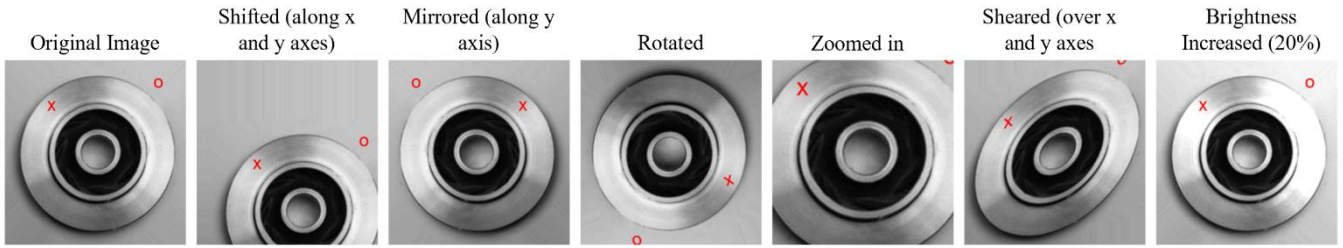


Fig. 3. Six transformations applied to a single original image (the symbols 'x' and 'o' in red color are used to understand the transformation in respect to the original image). Relevant transformation is labeled on top of the image.

a) Convolution layers

Each convolutional layer consists of a set of filters known as convolutional kernels where each neuron plays the role of a kernel. The kernel is a matrix of integers where it will multiply its weights with corresponding values of a subset of pixels of the input image. The selected subset of pixels of the input image has a similar dimension to the kernel. Then, the resulting values are summed up to generate one value that represents the value of a pixel in the output (feature map). The kernel strides across the input image producing the output (feature map of the entire image) of the convolution layer. In each layer, the kernel strides over a varying number of pixels at a time in both dimensions (height and width). The convolution process can be mathematically expressed as [33]:

$$f_l^k(p, q) = \sum_c \sum_{x,y} i_c(x, y) \cdot e_l^k(u, v) \quad (3)$$

where, $i(x, y)$ is an element of the input image tensor with x and y coordinates, which is element-wise multiplied by $e(u, v)$ index of the k^{th} convolutional kernel of the l^{th} layer. u and v are the rows and columns of the kernel matrix. $f(p, q)$ is the corresponding output feature map with p columns and q rows while c is the image channel index.

b) Pooling layers

Pooling operation sums up identical information in the local region of the feature map generated by a convolutional layer and outputs a single value within that region [34]. AlexNet consists of three pooling layers followed by the first, second and last convolution layers.

c) Activation function

Use of Rectified Linear Unit (ReLU) as a non-linear activation function of each layer is a significant characteristic in AlexNet. ReLU activation function is:

$$R(z) = \max(0, z) \quad (4)$$

where z is the function input and $R(z)$ is the function output which equal to the input when the input is positive and equal to zero otherwise.

d) Batch normalization

As a countermeasure for the overfitting, batch normalization is performed after several layers of the AlexNet.

TABLE I. LAYERS OF THE ALEXNET ARCHITECTURE

ID	Layer Type	Layer Parameters (f=no. of feature maps, k=kernel size, s=strides, act=activation function)	Size of Feature Map
0	Input layer	Input image size=(224x224) pixels, Channels=1	224x224x1
1	Conv2D	f=96, k=(11 x 11), s=4, act=ReLU	55x55x96
2	Max Pool	f=96, k=(3 x 3), s=2,	27x27x96
3	Batch normalization	N/A	27x27x96
4	Conv2D	f=256, k=(5 x 5), s=1, act=ReLU	27x27x96
5	Max Pool	f=256, k=(3 x 3), s=2,	13x13x256
6	Batch normalization	N/A	13x13x256
7	Conv2D	f=384, k=(3 x 3), s=1, act=ReLU	13x13x384
8	Batch normalization	N/A	13x13x384
9	Conv2D	f=384, k=(3 x 3), s=1, act=ReLU	13x13x384
10	Batch normalization	N/A	13x13x384
11	Conv2D	f=256, k=(3 x 3), s=1, act=ReLU	13x13x256
12	Max Pool	f=256, k=(3 x 3), s=2,	6x6x256
13	Batch normalization	N/A	6x6x256
14	Dropout	Rate=0.5	6x6x256
15	FC	f, k, s are N/A, act=ReLU	4096
16	Dropout	Rate=0.5	4096
17	FC	f, k, s are N/A, act=ReLU	1024
18	Dropout	Rate=0.5	1024
19	FC	f, k, s are N/A, act=softmax	2

e) Fully connected layer

At the end of the feature extraction stage (accomplished by convolutional layers), three fully connected layers are introduced which perform classification globally [35].

f) Dropout

To achieve generalization, some units or connections with a certain probability within the network are randomly

skipped (dropout) [36]. The AlexNet model executes dropout after several fully connected layers in it.

g) *Output layer*

The final layer of AlexNet architecture which acts as the output layer uses the softmax activation function [37]. The softmax function is given by:

$$S(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^n \exp(y_j)} \quad (5)$$

where y^i is the i^{th} element of the input vector, n is the number of classes which, in our case is two—defect and defect-free.

In our study, four modifications were carried out on the original AlexNet model creating an AlexNet variant. The modifications are: (1) Number of channels in the input convolutional layer is changed from three to one as our dataset consists of only grayscale images, (2) Dropout is imposed after each fully connected layer, (3) Batch normalization is performed after third and fourth convolutional layers, and (4) Number of output features of the second fully connected layer changed from 4096 to 1024.

E. *Application of transfer learning and optimizing model hyperparameters*

ImageNet dataset [38] is used for pre-training of the AlexNet model and the influence of the transfer learning is tested using three experimental configurations (EC):

- EC1: AlexNet is trained with the casting surface defect dataset without any pre-training with weights initialized randomly (training from scratch).
- EC2: the same process in the previous configuration repeated, but the weights initialized with the ones found from the pre-trained model instead of random weights.
- EC3: the exact weights of all the feature extraction layers pre-trained on the ImageNet dataset were used.
- EC4: the entire model parameters (including both parameters of convolutional and fully connected layers) of the pre-trained model on the ImageNet dataset is used on the casting surface defect dataset.

In each configuration, two types of hyperparameters including optimizer [39] and learning rate are optimized using the grid search method to achieve higher accuracy with modest robustness. In the grid search method, all the possible combinations of the selected hyperparameters are tested in multiple trials. The grid search methods suffers from the curse of dimensionality [40] where the number of trials grows exponentially with the increase of the number of hyperparameters. Nevertheless, the other sophisticated optimizations are not used as we obtained sufficient accuracies by varying only the two aforementioned hyperparameters.

F. *Implementation*

Training of the AlexNet model is accomplished using the Google Collaboratory tool—a free online python programming environment specially designed for machine learning tasks. CPU is composed of a single core hyper threaded Intel Xeon Processors at 2.3Ghz speed and 13GB RAM while GPU is a Tesla K80 GPU with a 12 GB GDDR5 VRAM.

For the implementation of the AlexNet model and KNN, TensorFlow [41] and Scikit-learn [42] open-source tools are used. TensorFlow is an open-source framework designed for the implementation and experimentation of machine learning-related tasks while Scikit-learn is a high-level machine learning library for python programming language. Furthermore, pre-trained models including the weights are acquired using PyTorch—an open-source deep learning framework [43].

All the experiments ran for ten epochs, where epochs are the number of training iterations where each neural network accomplishes one learning instance over the dataset. The selection of ten epochs is based on the empirical observation that conveys all the training in each experiment is always converged with ten epochs with optimal hyperparameters.

TABLE II. PRECISION, RECALL AND F1-SCORE OF THE TWO CLASSES OBTAINED AFTER CLASSIFICATION USING KNN ALGORITHM

<i>Test: Train</i>	Defect			Defect-free		
	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
0.2:0.8	0.86	0.87	0.88	0.86	0.81	0.83
0.4:0.6	0.85	0.88	0.86	0.84	0.79	0.81
0.6:0.4	0.85	0.88	0.86	0.84	0.79	0.81
0.8:0.2	0.84	0.82	0.83	0.77	0.79	0.78

IV. RESULTS AND DISCUSSIONS

This section presents the results obtained by following the methods discussed in the previous section and related interpretations.

A. *Classification without learning*

The results of the KNN classification of the casting surface defect dataset are presented in this section. Table II shows precision, recall, and the f1-score corresponding to each class (defect and defect-free) obtained after performing the KNN algorithm with varying the train-test split ratio. With the reduction of the training set percentage, there is no significant gradual change in the accuracy as there is no learning that occurred during the training process by the KNN algorithm unlike the learning models discussed in this paper.

The overall average accuracy of the classification of casting surface image data using the KNN algorithm is relatively lower when compared to the results of CNN models discussed in the future sections. This lower accuracy reveals that the classification using raw pixel intensities and their proximities to neighbor values in the casting surface defect images are not significant. This phenomenon discloses that all the images in each class are unique up to a certain extent in respect of pixel intensities which in return, induces the importance of the feature extraction. On the

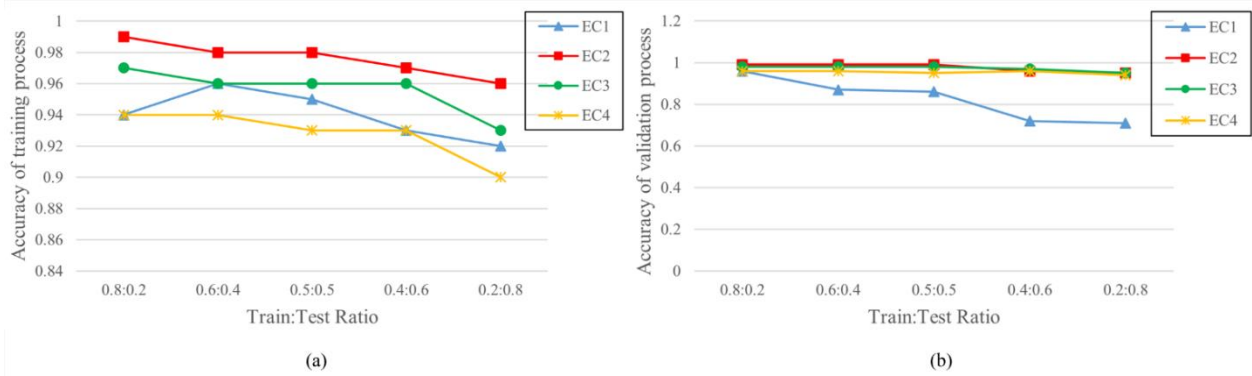


Fig. 4. (a) and (b) are the accuracies of training process and the validation process, respectively over different experimental configurations (ECs) (which are mentioned under methodology of this paper) and varying train-test split ratios.

other hand, when observed with a perspective of the accuracies (i.e., All the accuracies are around 0.8 which is regarded as a significant performance in image classification tasks) it reveals that the image dataset has lower levels of noise.

B. Classification with learning

Classification endorsed by the application of CNNs manifested higher accuracies when compared to the classification performed by the KNN algorithm. Fig. 4 illustrates the variation of accuracy with different train-test split ratios and different experimental configurations.

For each experimental configuration, training accuracy (as shown in Fig. 4-a) is dropped when the training image portion decreases while increasing the number of validation images. In fact, demonstrating the common idea that lesser training in deep learning models causes lesser accuracies. Nevertheless, the size of the drop is negligible as all the accuracies are above 0.9 (or equal to 0.9) in each scenario. The highest overall accuracy is achieved when the training weights are initialized from the pre-trained model (EC2) instead of random initialization (EC1).

Specifically, even with 20% training images, the use of the exact feature extractor of the pre-trained model for training (EC3) induced higher accuracy than training from scratch. In the instance where both feature extractor weights and classifier weights (weights of the fully connected layers) of the pre-trained model are used on training, an overall accuracy of 0.9 is achieved.

On the contrary, validation process accuracy (as shown in Fig. 4-b) does not fluctuate considerably over the variation of train-test split ratio regardless of the experimental configurations except where training is done from scratch. All the transfer learning schemes (EC2, EC3, and EC4) show improved validation accuracies when compared to training from scratch (EC1) on the casting surface image dataset.

Table III indicates the possible combinations of the hyperparameters used for the grid search method and related accuracies for EC3 with 20% of training images. During optimization of hyperparameters, first, we picked a random learning rate (0.0001) and performed a grid search with seven optimizer types. The best performance is gained by setting the optimizer to the RMSprop algorithm [39]. Fixing the optimizer as RMSprop algorithm, then we tested several learning rates which resulted in 0.0001 as the optimum value. Overall best hyperparameters (i.e., optimizer type and learning rate) found by the grid search method with the other hyperparameters found from the literature were standardized as shown in Table IV over the final run of each experiment.

TABLE III. RESULTS OF THE GRID SEARCH METHOD PERFORMED TO FIND BEST OPTIMIZER AND LEARNING RATE

Search 1: Learning Rate is Randomly Selected (=0.0001) and Fixed to Test Several Optimizer Types			
Learning Rate	Optimizer	Training accuracy	Training time (seconds)
0.0001	Adam	0.94	742
	Adadelta	0.57	757
	AdamW	0.90	484
	Adamax	0.89	518
	ASGD	0.57	505
	RMSprop	0.93	635
	SGD	0.58	744
Search 2: Best Optimizer (RMSprop) from Search 1 is Fixed and Tested Several Learning Rates			
Optimizer	Learning rate	Training accuracy	Training time (seconds)
RMSprop	0.1	0.55	630
	0.01	0.57	634
	0.001	0.94	641
	0.0001	0.93	637
	0.00001	0.93	642

TABLE IV. OPTIMIZED HYPERPARAMETER SETTINGS/VALUES STANDARDIZED THROUGHOUT ALL EXPERIMENTS

Hyperparameter	Setting/Value	Obtained with Grid Search (GS) Method/Using Literature (LT)
Optimizer	RMSprop	GS
Learning Rate	0.0001	GS
Learning rate policy	Step (decay over epoch)	LT
Momentum	0.9	LT
Batch Size	16	LT

V. CONCLUSIONS AND FUTURE WORK

Maintaining quality standards is vital in the casting product manufacturing industry for better business

performance and for the safety of the end-users who consume products with critical mechanical components fabricated by casting. Automated inspection of casting defects leads to lesser inspection times and circumvents safety problems of employees working in hazardous environments.

In this paper, we discussed the application of AlexNet CNN architecture-based transfer learning for automated inspection of surface defects of a submersible pump impeller manufactured by casting. Over the last decade, for the task of casting defect recognition, numerous sophisticated architectures were proposed with higher architectural complexity and better performance compared to the AlexNet architecture. Using the results of our study, we show (limited to the dataset used) that a simpler architecture like AlexNet can perform better when it is implemented with transfer learning and optimized model parameters. As future work, methods discussed in this study can be tested over other datasets containing images of casting surface defects of different products.

Over the several experimental configurations tested, the use of the exact feature extractor of the pre-trained model for training demonstrated the best performance in terms of training accuracy and the training time (Although training with weights initialized from the pre-trained model resulted in the overall highest accuracy the training time is higher in contrast to using the entire feature extractor).

Several recommendations for a casting surface defect detection system can be made based on the results of this study. Nevertheless, as future work, the practical usability of such a system needs to be tested prior to implementation as several dataset-specific parameters still need to be adjusted depending on the circumstance. The process of capturing the surface images of the casting products is vital including, but not limited to: (1) adhering to proper lighting conditions, and (2) maintaining unique and plain background when capturing. As shown in the results, transfer learning can be implemented to reduce the training time and enhance the robustness of the model. Moreover, transfer learning is beneficial when the number of training images is lower. Specifically, the use of a feature extractor from the pre-trained model and limiting the training only with the classification layers (fully connected layers) with casting defect data is advantageous instead of using all the parameters of the pre-trained model. Furthermore, fine-tuning the model hyperparameters is crucial for obtaining better results.

REFERENCES

- [1] W. Barkman, In-process quality control for manufacturing. CRC Press, 1989.
- [2] R. T. Chin and C. A. Harlow, "Automated visual inspection: A survey," IEEE transactions on pattern analysis and machine intelligence, no. 6, pp. 557–573, 1982.
- [3] M. Sahoo, Principles of metal casting. McGraw-Hill Education, 2014.
- [4] T. V. Sai, T. Vinod, and G. Sowmya, "A critical review on casting types and defects," Engineering and Technology, vol. 3, no. 2, pp. 463–468, 2017.
- [5] M. K. Ferguson, A. Ronay, Y.-T. T. Lee, and K. H. Law, "Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning," Smart and sustainable manufacturing systems, vol. 2, 2018.
- [6] D. Mery, T. Jaeger, and D. Filbert, "A review of methods for automated recognition of casting defects," INSIGHT-WIGSTON THEN NORTHAMPTON-, vol. 44, no. 7, pp. 428–436, 2002.
- [7] S. Gholizadeh, "A review of non-destructive testing methods of composite materials," Procedia Structural Integrity, vol. 1, pp. 50–57, 2016.
- [8] Q. Wan, H. Zhao, and C. Zou, "Effect of micro-porosities on fatigue behavior in aluminum die castings by 3D X-ray tomography inspection," ISIJ international, vol. 54, no. 3, pp. 511–515, 2014.
- [9] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv:1511.08458, 2015.
- [10] H. Strecker, "A local feature method for the detection of flaws in automated X-ray inspection of castings," Signal Processing, vol. 5, no. 5, pp. 423–431, 1983, doi: [https://doi.org/10.1016/0165-1684\(83\)90005-1](https://doi.org/10.1016/0165-1684(83)90005-1).
- [11] Z. Górný, S. Kluska-Nawarecka, D. Wilk-Kołodziejczyk, and K. Regulski, "Diagnosis of casting defects using uncertain and incomplete knowledge," Archives of Metallurgy and Materials, vol. 55, no. 3, pp. 827–836, 2010.
- [12] M. Ferguson, R. Ak, Y.-T. T. Lee, and K. H. Law, "Automatic localization of casting defects with convolutional neural networks," in 2017 IEEE international conference on big data (big data), 2017, pp. 1726–1735.
- [13] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," Journal of big Data, vol. 8, no. 1, pp. 1–74, 2021.
- [14] Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.
- [15] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.
- [16] Koutsoukas, K. J. Monaghan, X. Li, and J. Huan, "Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data," Journal of cheminformatics, vol. 9, no. 1, pp. 1–13, 2017.
- [17] R. Dabhi, "Casting product image data for quality inspection," Kaggle.com. <https://kaggle.com/ravirajsinh45/real-life-industrial-dataset-of-casting-product> (accessed Jun. 14, 2021).
- [18] Gayer, A. Saya, and A. Shiloh, "Automatic recognition of welding defects in real-time radiography," Ndt International, vol. 23, no. 3, pp. 131–136, 1990.
- [19] Eckelt, N. Meyendorf, W. Morgner, and U. Richter, "Use of automatic image processing for monitoring of welding processes and weld inspection," in Non-destructive testing, Elsevier, 1989, pp. 37–41.
- [20] Filbert, R. Klatte, W. Heinrich, and M. Purschke, "Computer aided inspection of castings," in IEEE-IAS Annual Meeting, 1987, pp. 1087–1095.
- [21] Mery, "New approaches for defect recognition with X-ray testing," Insight, vol. 44, no. 10, pp. 614–15, 2002.
- [22] X. Li, S. K. Tso, X.-P. Guan, and Q. Huang, "Improving automatic detection of defects in castings by applying wavelet technique," IEEE Transactions on Industrial Electronics, vol. 53, no. 6, pp. 1927–1934, 2006.
- [23] Kehoe and G. A. Parker, "An intelligent knowledge based approach for the automated radiographic inspection of castings," NDT & E International, vol. 25, no. 1, pp. 23–36, 1992.
- [24] D. Wang, B. Wang, H. Yao, H. Liu, and F. Tombari, "Local image descriptors with statistical losses," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 1208–1212. doi: 10.1109/ICIP.2018.8451855.
- [25] R. R. Da Silva, M. H. S. Siqueira, L. P. Calôba, and J. M. Rebello, "Radiographics pattern recognition of welding defects using linear classifiers," Insight, vol. 43, no. 10, pp. 669–74, 2001.
- [26] Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53–65, 2018.
- [27] L. Jiang, Y. Wang, Z. Tang, Y. Miao, and S. Chen, "Casting defect detection in X-ray images using convolutional neural networks and attention-guided data augmentation," Measurement, vol. 170, p. 108736, 2021.
- [28] Shorten and T. M. Khoshgofaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.
- [29] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Frontiers in plant science, vol. 7, p. 1419, 2016.

- [30] Abd Almisreb, N. Jamil, and N. M. Din, "Utilizing AlexNet deep transfer learning for ear recognition," in 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), 2018, pp. 1–5.
- [31] M. Z. Alom et al., "The history began from alexnet: A comprehensive survey on deep learning approaches," arXiv preprint arXiv:1803.01164, 2018.
- [32] Khan, A. Sohail, U. Zahoor, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.
- [33] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Artificial intelligence and statistics*, 2016, pp. 464–472.
- [34] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.
- [35] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.
- [36] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks.," in *ICML*, 2016, vol. 2, no. 3, p. 7.
- [37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.
- [38] S. R. Labhsetwar, S. Haridas, R. Panmand, R. Deshpande, P. A. Kolte, and S. Pati, "Performance Analysis of Optimizers for Plant Disease Classification with Convolutional Neural Networks," arXiv preprint arXiv:2011.04056, 2020.
- [39] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.
- [40] M. Abadi et al., "Tensorflow: A system for large-scale machine learning," in 12th symposium on operating systems design and implementation, 2016, pp. 265–283.
- [41] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [42] Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

SYSTEMS ENGINEERING

An exploratory evaluation of replacing ESB with microservices in service-oriented architecture

L. D. S. B. Weerasinghe*
Department of Computer Science & Engineering
University of Moratuwa, Sri Lanka
weerasingheldsb.20@uom.lk

Indika Perera
Department of Computer Science & Engineering
University of Moratuwa, Sri Lanka
indika@cse.mrt.ac.lk

Abstract - With the continuous progress in technology during the past few decades, cloud computing has become a fast-growing technology in the world, making computerized systems widespread. The emergence of Cloud Computing has evolved towards microservice concepts, which are highly demanded by corporates for enterprise application level. Most enterprise applications have moved away from traditional unified models of software programs like monolithic architecture and traditional SOA architecture to microservice architecture to ensure better scalability, lesser investment in hardware, and high performance. The monolithic architecture is designed in a manner that all the components and the modules are packed together and deployed on a single binary. However, in the microservice architecture, components are developed as small services so that horizontally and vertically scaling is made easier in comparison to monolith or SOA architecture. SOA and monolithic architecture are at a disadvantage compared to Microservice architecture, as they require colossal hardware specifications to scale the software. In general terms, the system performance of these architectures can be measured considering different aspects such as system capacity, throughput, and latency. This research focuses on how scalability and performance software quality attributes behave when converting the SOA system to microservice architecture. Experimental results have shown that microservice architecture can bring more scalability with a minimum cost generation. Nevertheless, specific gaps in performance are identified in the perspective of the final user experiences due to the interservice communication in the microservice architecture in a distributed environment.

Keywords - *microservice, performance, scalability, SOA*

I. INTRODUCTION

Since the world is more inclined towards new technology, it has ultimately resulted in an information system-driven society. People are concerned about attending to their routine tasks in the most efficient, easy, and fastest method possible. Because of this driving need to achieve efficiency and effectiveness, the necessity to successfully build systems to win over these real-world problems was considered vital by software engineers. Researching and proposing new software architectural concepts by the software industry were initiated to develop the most reliable software in the world [1]. These architectures give a better view of the software to provide the services and evolve the quality of its life cycle. Architecture is responsible for providing the bridge for the software functionalities and the system quality attributes necessary for the business needs. As a first step, the engineers develop object-oriented architecture patterns that cater to the small-scale software run on the host machines.

Historically, the software industry developed monolithic software for enterprise-level solutions. The traditional monolithic application encapsulates all the

components, functions into one single package and deploys as a single application. Most of the service-oriented monolithic applications are developed using the C, C++, Java, and Python languages. Those languages by default support creating the single executable artifact. Some of the monolithic systems are deployed in the distributed environment using the RMI, Network Object, and CORBA concepts. However, it's tough to maintain the monolithic in the distributed environment [2].

On the contrary, there are many advantages of using the monolithic systems such as easy deployment because all the modules are in the same code base, supportive nature of the entire IDEs, ease of testing the entire system as there's no requirement to set up various components, and the ease of scaling since monolithic application comes up with the option of a single distribution. However, the monolithic application has significant drawbacks, which are mostly related to business growth and technology adaptations. For instance, all the components are packed together in monolith architecture with a vast codebase; hence, it's complicated to make modifications. Also, the application patching process and understanding the monolithic applications are quite challenging. On the other hand, one single failure of the application can cause the collapse of the entire system. Therefore, it can be derived that those monolithic applications are not suitable for deployment in the containerization environment. Monolith applications are cumbersome, and it takes a considerable amount of time to startup. Continuous integration and continuous delivery pipeline are complicated to maintain with monolithic systems because of the heaviness of the systems. One single change needs to test the overall system functionalities as of the tightly coupled components. Hence overall time to test and the cost generated for deployment will be considerably high.

With the concept of the "separation of concerns," component-based software engineering comes into the world, which leads to better implementation, design, and evolution of software systems. Then the Service-Oriented Computing (SOC) paradigm comes into context. People moved to distributed software development and deployed that software in the distributed environment [3]. In SOC, each component's functionalities are shared using the message passing through those distributed components. The SOC architectural concept brings several advantages to the software industry, such as "dynamism" which can introduce the same component based on the system load, modularity which can be reused across the components, and distributed development.

In the mid-'90s, Gartner Group researchers introduced a reference architecture for the industry called service-oriented architecture (SOA) [4]. In SOA architecture, both the service consumers and service providers get together

and provide the business needs. Services are the distributed components, and they have published the interfaces to do the communication via middleware. Those interfaces abstract all business logic. One of the main components of service-oriented architecture is the enterprise service bus (ESB) which serves as middleware. ESB's main task is to enable communication between those services and govern them. Most of the SOA systems use the Simple Object Access Protocol (SOAP) for communication. SOA architecture data sources are shared with the components deployed in the same environment. That means the same database is open for both Data Definition Language (DDL) and Data Manipulation Language (DML) and all the components residing inside the SOA architecture.

The difference between SOA and monolithic architecture is that SOA architecture consists of the component as a service, but the monolithic builds all the logic in one package. In the monolithic architecture, all the logic is based on sharing one single hardware resource. Nevertheless, in SOA architectures, each component uses its hardware resources to provide the service. Compared to the monolith applications, SOA brings more advantages to the software industry, such as enabling the system's growth to the enterprise level, bringing component-wise scalability to the whole environment, and reducing operational costs.

The term "Microservice" was initially introduced in 2011 at an architectural workshop conference [2]. Microservice architecture comes into the world as a new architectural paradigm that can be illustrated as tiny services running independently and communicating with each other and satisfying the business requirement. The microservice architecture was widely used by people in the past few years, which can be considered as a positive behavior to the software industry. With time, most software firms arrived at the notion that using the microservice architecture developments brings high productivity to the company and produces a successful end product for the clients [5]. Microservice architecture also takes advantage of cloud services such as on-demand provisioning, serverless functions, and elasticity as well as a lot of quality attributes such as scalability, maintainability, performance and many more.

People who intend to move away from the monolithic to SOA architecture should particularly comprehend the quality attributes generated by it. In this paper, our acute concentration is on evaluating and coming up with the architectural conclusion on the extremely critical quality attributes which diverge from the most common SOA architecture with ESB and the Microservice architecture.

II. BACKGROUND AND RELATED WORK

Microservice architecture is derived from the concept of the SOA. Microservices are now considered the new software architecture for highly scalable and highly maintainable distributed systems. Nevertheless, when the system functionalities grow day by day, microservice architecture tends to get complex because of the large set of independent services it has as functions. Developing and deploying the microservices independently to each other brings high cohesion and loosely coupled modules [6].

The reason behind the popularity of the microservices architecture is the quality attributes associated with the microservices. We identified the most concerning quality attributes on the microservices architecture, such as

scalability, performance, availability, maintainability, and security [7, 8].

A. Quality attributes in microservice architecture

Several definitions can define "Quality" in a microservice architecture. Some people denote it by the software's capability to meet the required requirements, and some of the people define it as the "reality of the objectives" [9]. In the context of software engineering, quality refers to the relationship between the business and the product. This software quality contains two types;

Software functional quality – Describes the functional requirements with the current system design. Functional quality attributes show how the system matches the business requirement. Using this quality, people can decide whether the developed software is acceptable or not.

Software structural quality – Describes the software non-functional requirements that support in providing the functional requirement on the system. Those non-functional requirements bring more value addition to the software ecosystem.

The software stakeholders are primarily concerned about the system requirements. Based on the stakeholder requirements, we can divide software quality into two main groups; the development phase and the operations phase. In the development phases, we need quality requirements that are very important for software developers, such as maintainability, modularity, and understandability. Quality requirements for the operations related to the system end-users and system supporting teams include usability, traceability, availability, and performance.

Those quality requirements have differed from the software domain, priorities of the developers, and the end-users. We can see the quality attributes when the system has been implemented.

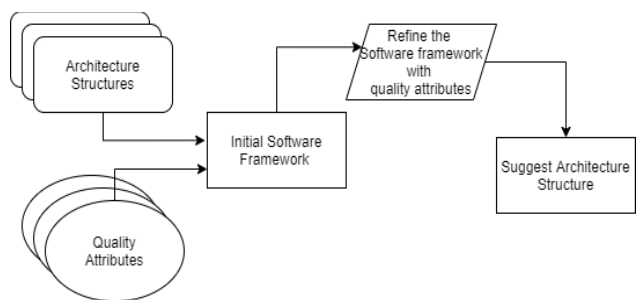


Fig. 1. How quality attributes influence to software architecture

According to Figure 1, all the quality attributes are depending on the software architecture [10]. It is mandatory to review the software architecture before the software development or use the reference architecture to develop the software. The qualities cannot be added to the system architecture ad-hoc. Therefore, developers need to build those qualities from scratch on the software.

B. Scalability

The scalability quality attribute is one of the primary critical features in the microservice architecture. The scalability attribute was initially introduced to enhance software performance and control high traffic. Scalability

quality also ensures the system fault tolerance. There are two main parts of scaling.

Horizontally Scaling- This method ensures that the application's performance is increased by adding another application instance over it. For example, we have one web server before scaling, and after scaling, we have multiple web servers that serve traffic. Load balancers help to distribute the traffic load among those web servers [11].

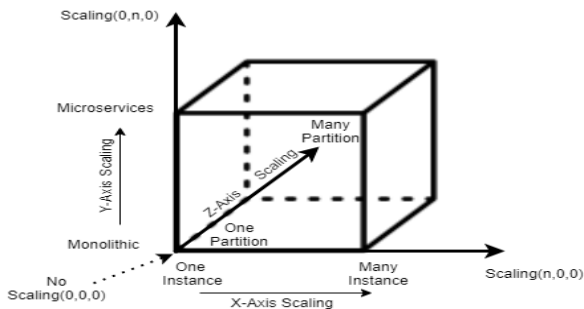


Fig. 2. Scaling cube

Vertically Scaling- This means increasing the hardware resource to improve the application performance, such as increasing the RAM, increasing the CPU, and using the SSD instead of HDD storage [12]. Vertical scaling is a very traditional method, and most people use computers to do this kind of scaling. For instance, vertical scaling is majorly used when the personal computer is slow and the need to increase the computer hardware occurs. Nevertheless, this scaling is bound to a limited area, and there's no possible way to increase the hardware resource as we want. Because particular hardware only supports the specific ranges only. As an example, some motherboards' maximum supported RAM is 64GB.

Scaling cube shows scaling model for the software applications [13]. We also refer to this concept when scaling the application in our research. Figure 2 X-Axis scaling is referred to as horizontally scaling, work evenly distributed scaling, and horizontally duplication. The simple meaning is that running the software application behind the load balancer. The Load balancer is responsible for the equal distribution of the load among the number of applications connected to the load balancer rules. X-Axis scaling is mostly used by monolithic applications with shared databases and caches.

Y-Axis scaling applications are decomposed to the small binaries by considering the functions/services called microservices. (0,0) indicates the monolithic application, which contains all the services as one single binary. Y-Axis scaling gives more value to the software architecture because services behave independently. Therefore, people can only scale the relevant services using this concept.

The microservice architecture is a combination of both X and Y-Axis scaling. This helps bring more scalable software architecture to the deployments.

Z-Axis scaling is somewhat similar to the X-Axis scaling, but it differs from the data used by the application. For instance, assume that we have a significant number of students, and according to the admission number, they are segregated into groups. In each group, the same application is running and doing the same service but using different data. This is primarily applicable to B2C applications. The

load balancer should need to be intelligent to recognize the correct data partition server to route the traffic. Otherwise, we need to put the router before those servers.

When it comes to a cloud-native architecture, most cloud providers such as Amazon Web Services (AWS), Google Cloud Service (GCP), and Azure develop various vertical and horizontal scaling solutions. Most prominent players, such as Netflix, Uber, WhatsApp, and Instagram, also deploy their applications in cloud-native environments [11]. Using the virtualization technology, the cloud providers introduce vertical and horizontal scaling on the cloud resources such as servers, storage, and databases. They have introduced AI technologies like machine learning to perform predictive analysis on the scaling part and automatic scaling. Day by day, those reactive scalings become seamless with the help of those AI technologies. Most of the cloud-native applications developed as containerized applications and deployed on container orchestration engines like Kubernetes. Cloud providers also give services to cloud consumers by enabling the container orchestration engine. For example, the AWS cloud provider gives Amazon Elastic Container Service (Amazon ECS) and Google cloud to provide the Google Kubernetes Engine (GKE). Those services will take care of managing the whole container orchestration part. The developer needs only to develop the application which is suitable for cloud-native environments. In this cloud-native environment, containers are warped as small pods that allow the scaling up and down in a simple way.

C. Performance

Performance is one of the most critical quality attributes. Both software consumers and the developer care about application performance during the run time. Performance is measured by the measurable factor of the system when performing the given functionalities within given constraints such as accuracy, latency, and resource consumption. A simple way to define the performance in the software is how software behaves on time, which is called responsiveness [12]. Most people move away from manual work to digitalized platforms with the belief that such work can be done in lesser time and minimum effort. The outcome of the software system should always be; consumption of less amount of time with more accuracy. The main objective of the real-time system is to give a response in real-time. For that, system architectures and software design also need to be well established. In the past decade, most of the performance issues were identified in the production environments since unpredictable behaviors of the users who are using the software and the unpredictable behaviors in the environment are found to be the root causes for performance issues. To reduce the above issue, the performance factor is considered when the system is in the design phase.

There are several criteria to check the performance of the software system.

a) Latency / response time

This refers to how much time is taken to complete the task and respond. If the time difference between start time and end time is low, that means the system performance is good. API-based synchronized system's API response time measure using the microseconds and milliseconds.

b) *Throughput*

Throughput refers to the number of tasks that have been completed within the given time interval. In other words, it is the software process rate or the time frame as seconds. It's also called transactions per second (TPS). Measurement of the throughput is different from application to application. High throughput means software performance is in a good state.

c) *Capacity*

This means how much work software can perform. The maximum throughput is considered as system capacity. In other words, the maximum number of events the software can perform within a unit of time and total resource consumption. For example, software A can support a maximum of 250 TPS with 1s latency backend AWS m4.large VM (8GB RAM, 2vCPU) and network perspective bandwidth means the capacity. When the capacity is getting immense value, then we can consider that the software performance is high.

III. RESEARCH METHODOLOGY

This research will talk about the most concerning quality attribute variation when converting software architecture from SOA to microservice architecture. By critically reviewing the software architecture, we identified that scalability and performance are the most critical quality attributes in the software industry [8][9]. After the monolithic architecture, software architects introduced the SOA. However, we can identify some limitations on the scaling and the performance quality attributes by reviewing the SOA. There were several problems identified when scaling the SOA-based system. All the services are decoupled in the SOA-based system and exchange the required data via the enterprise service bus (ESB). ESB is responsible for the service orchestration, and it acts as a backbone of the SOA system. When scaling the SOA-based system, at one point, people need to scale the ESB also. So scaling ESB requires high-end specification servers that will generate a considerable amount of cost. ESB servers contain many features and modules, and in some cases, the software ecosystem did not use all of the features carried on the ESB servers in SOA. Because of that, performance-wise, it has some impact on the SOA systems during run time. With those factors, people are moving from Software Oriented Architecture to microservice-based architecture. This research evaluates how scalability and the performance quality attributes vary when transforming SOA to the microservice-based architecture.

We have developed the SOA system that can talk with the legacy backend, and at the same time, we have developed business functionalities using microservice-based architecture, which can also communicate with the legacy backends.

Fig. 3., shows how the SOA system integrates with the databases, backend, and clients. ESB is responsible for catering the message routing and publishing all the communication to the data source.

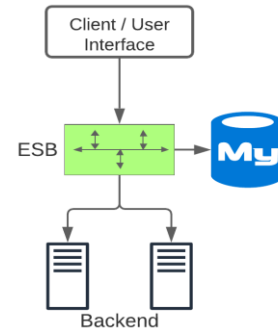


Fig. 3. SOA architecture

Here we use the WSO2 Enterprise Service Bus, an open-source product, and most of the well-known software companies use this product for their software systems as well [14]. We choose WSO2 ESB as it generates many features like better performance and user-friendly nature compared to other ESBs. Also, in WSO2 ESB, the lightweight mechanism is introduced, and also it is an open-source product [14] [15]. With the WSO2 ESB, we wrote the business logic using the Apache Synapse language [17] and deployed it as Carbon applications in the ESB servers [18]. All the products of WSO2 are based on the Carbon platform. This is a form of middleware platform that stores business IT projects on the cloud, and on-premises servers [19]. With the help of the WSO2 developer studio, WSO2 ESB has created the opportunity for the software developers to swiftly orchestrate applications, business processes, and the services such as data service, proxy-based service, message routing service, etc. With this kind of development, software companies can deliver the services promptly to the clients. Moreover, the technical and the business services can be integrated with the legacy systems and any kind of SAAS services in SOA architecture. Backend is a legacy that one can communicate using the REST protocol. Clients/User interface communicates to the ESB using the REST protocol by exposed APIs.

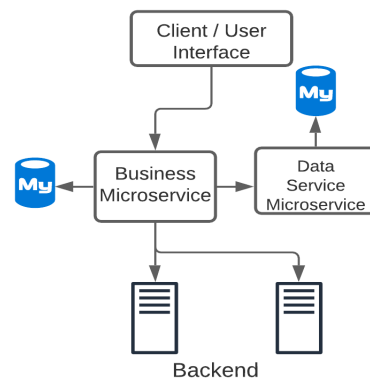


Fig. 4. Microservice architecture

Figure 4 shows how the microservices replace the SOA System. We have identified the ESB server's required services and made those services into individual components and deployed them as microservices. Business microservice consists of all business logic, and data service

is responsible for publishing data. Here we used the same legacy backend, which can communicate with the REST protocol with the microservices. This microservices architecture is developed using JAVA language with the help of the Spring boot framework. REST client libraries are used for inter-service communication with the microservice to microservice and other services. Business logic microservice has exposed the APIs using the request controllers to communicate with the clients/user interfaces.

IV. RESULTS AND EVALUATION

The developed two systems were evaluated in the real environment with two main quality attributes: performance and scalability. In scalability, we are more concerned about the hardware footprint and the cost. There are several aspects of performance. In this, we evaluated the latency, throughput, and capacity with the allocated hardware. Throughout the experimental time, we collected statistics about the load average of the server, memory usage on the server, overall response time of the application, and throughput of the application using the JMeter [20]. Applications' ramp-up time frame and the steady-state time frame are included in those statistics. Firstly, we hosted the application in the different servers which are having different footprints. Then we collected the above statistics in those different environments by sending the 1KB size POST JSON payload to the applications. Upon collecting the statistics and sending the payload, backend servers returned the 1KB size JSON response. We use the Amazon Web Services (AWS) environment for all the environments. As a client, we used Apache open source JMeter [20] to generate the load toward the deployed servers. For all stress tests, we used 350 concurrent threads. In the AWS environment, T2 type resources were used in our experiment because of the following several reasons: It has Intel Xeon processors with high frequency that can be burstable, its coherent baseline performance is suitable for the general-purpose application deployments [21], and it is capable of balancing the overall server resources (CPU/memory/network).

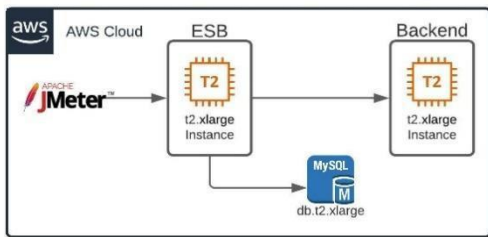


Fig. 5. 1st Test suite architecture

As the first test suite shows in Figure 5, we used the AWS t2.xlarge EC2 instance with four virtual CPUs and 16GB RAM. Also, the Solid-State Drive (SSD) was used to store the application. Then we deployed the WSO2 ESB application with customized development using the synapse language to cater to business logic. The ESB server connects with the AWS RDS MYSQL database service, which is deployed in the same VPC to reduce network latency. We used db.t2.xlarge, which has four virtual CPUs and 16GB RAM. Simultaneously, we provisioned the 100GB storage size for this RDS.

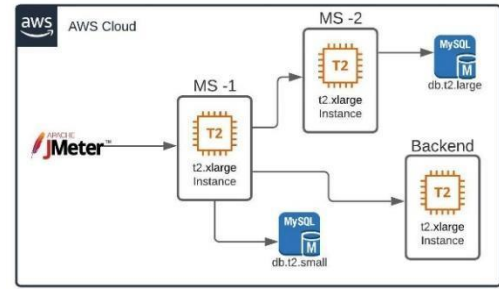


Fig. 6. 2nd Test suite architecture

The microservices for the second test suite, as shown in Fig. 6, that can perform the same ESB business logic relevant to this deployment, was developed. It had two microservices, and those microservices are deployed in the AWS t2.xlarge EC2 instances with Solid-State Drive (SSD) storage. Following the microservice concept, two different databases which are deployed in the same internal network. db.t2.large type RDS with 100GB storage was used for the data service microservice, and db.t2.small type RDS with 20GB storage was used for business microservice.

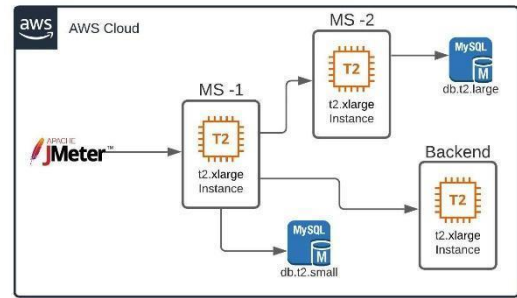


Fig. 7. 3rd Test suite architecture

For the third test scenario shown in Fig. 7, we reduced the server footprint after analyzing the statistics we collected on the 2nd test suite. For both the microservice deployments, we used the t2.medium AWS EC2 instances, which have 2 virtual CPUs and 4GB RAM. We used the Solid-State Drive (SSD) in both servers to store the application. The same database type was used in the 2nd test suite without any modifications. All the servers and the database were placed in the same internal network.

The backend servers and the client server (JMeter) were not changed for any of the testing scenarios. For storage, AWS t2.xlarge EC2 instances with Solid-State Drive (SSD) were used for both backend servers and the client servers. These two servers were also placed in the same internal network as the other servers.

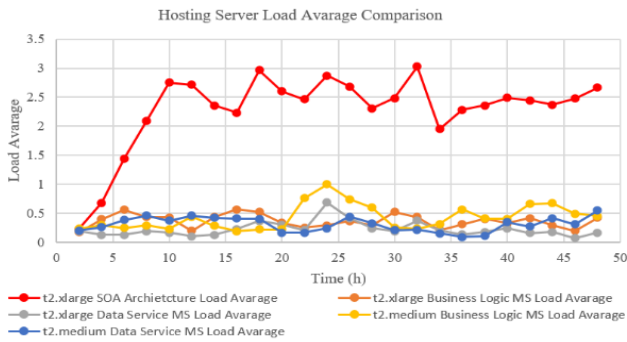


Fig. 8. Load average comparison

Figure 8 shows how the average server load varies on the SOA architecture and microservice architecture systems on the different hardware footprints. In the SOA architecture, the ESB node consumes many load averages to process the client requirement. However, none of the microservices deployed in the two different server types went for more than one load average.

If we group and add up the t2.xlarge two microservices load averages, those added up values will not be higher than the SOA architecture load average values. This is the same for the t2.medium microservices load average as well. It was found that Microservice architecture deployment was able to work with less resource consumption once we were vertically scaled-down the servers. On the contrary, ESB servers could not vertically scale down because they have fully utilized the current server resources.

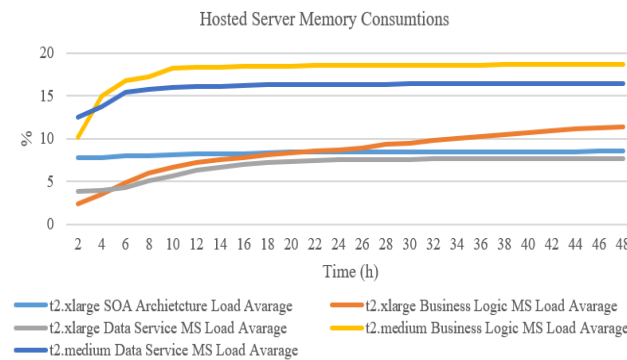


Fig. 9. Memory usage comparison

Figure 9 shows the memory consumption on the SOA architecture system and the microservice architecture systems. None of the servers consume the 20% server RAM. When vertically scaling down the microservices, it can be seen that there is a slight improvement in the throughput in figure 11 when vertically scaling the hardware footprint in the microservice architecture was observed that it increases the memory by nearly 5% on both the data service microservice and the business logic microservice.

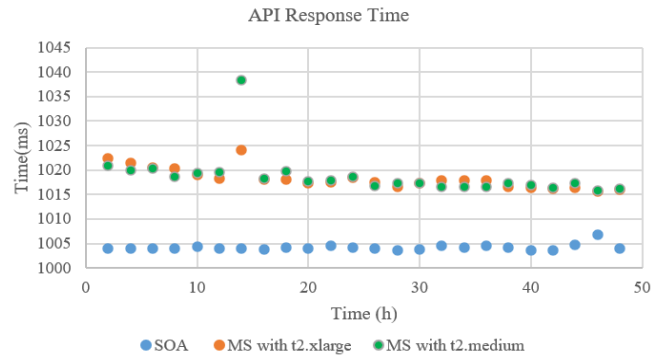


Fig. 10. Response time variation

Figure 10 shows the overall response time on each system with the deployed environment. SOA system performs with less response time in comparison to the microservice architecture system. It does not deviate much from the environment, and its software architecture. In the SOA system, all the modules we packed in the ESB server and no network calls for satisfy the full business function. All the logic is handled inside the single JVM. Because of that, response time is lower than the microservice architecture. The reason behind having a higher response time in the microservice architecture is because of the network call to the separate services. It introduces the additional time for the overall response time.

SOA system shows high performance by producing within a less response time. However, system throughput is less than the microservice. At a single time, the slice system only handles the smaller number of concurrent requests rather than the microservices. Because the SOA system consists of all the modules in the same JVM, and it takes all the resources on the JVM. So, the server does not accept the high number of requests to the single run time environment.

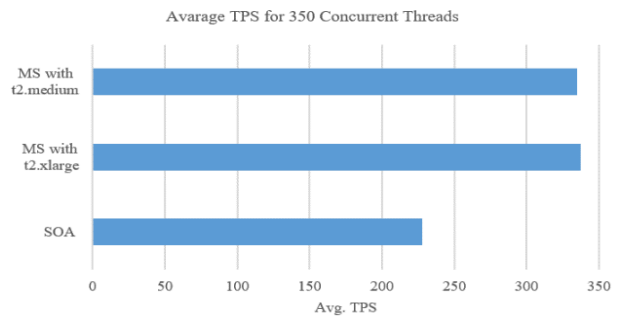


Fig. 11. Throughput comparison

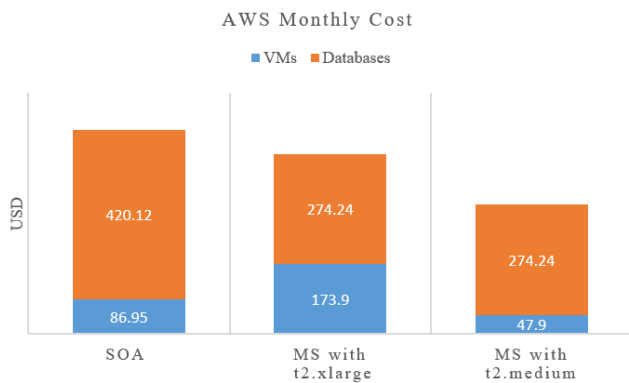


Fig. 12. Cost comparison

Fig. 12 graph only considers the dynamic values we have used in different test suites. Comparing the cost of both SOA and the microservice architecture shows that SOA generates a higher cost for the entire end-to-end deployments [22]. Experienced system architects can determine the exact footprint for the developed microservice by considering the user requirements. Using the optimal hardware footprint, we can save much money on software deployment projects. Those microservice can deploy the Kubernetes environments without putting more effort. From that, we can do the auto-scaling as per the traffic load. With this also we can save the overall cost.

V. CONCLUSION AND FURTHER WORK

This topic unfolds the factors to evaluate the research problem, which is the most concerning quality attributes of scalability and the performance variate between the Service Oriented Architecture and microservice architecture. Most organizations expect microservice architecture to move their current monolithic architecture or SOA. The main concern with the current monolithic and SOA architecture is the cost of scalability. Their current deployment footprint is also high, and it already involves a considerable cost. When we were going to scale that current environment, it made the cost nearly double. Nowadays, all the systems are deployed as contained in a cloud-native environment.

Nevertheless, monolithic and SOA-based architecture systems are not suitable for cloud-native environments. Because those applications are enormous and take a considerable amount of time to startup and serve the traffic, if we put those kinds of applications in the Kubernetes environments as pods, we cannot get the advantages provided by the container orchestration engines. Nevertheless, when converting to cloud-native microservices, some of the performance factors get affected. Before converting the monolithic / SOA system, we need to think about what performance factor requires enhancement. In terms of capacity and cost-effectiveness, microservice could be considered a better approach. When we move to the microservice architecture, we have flexible scalability. Through Microservice architecture, people have the option of only scaling the necessary services rather than the entire application. The previous chapter shows the fundamental analysis, and this could assist researchers in concluding microservice architecture.

In summary, we could state that microservice architecture is a better approach in terms of scalability and performance in comparison to SOA and monolithic

architecture. The research study results clearly showed that microservice architecture gives more performance in terms of the throughput and the application's capacity. Moreover, it is a cost-effective solution when scaling the applications. With this study, architects can redesign existing microservice architecture applications and adhere to cloud-native environments. Future work needs to find a solution for reducing the performance impact on latency in the microservice architecture.

REFERENCES

- [1] R. Flygare and A. Holmqvist, "Performance characteristics between monolithic and microservice-based systems," *Blekinge Inst. Technol.*, 2017.
- [2] N. Dragoni et al., "Microservices: Yesterday, Today, and Tomorrow," in *Present and Ulterior Software Engineering*, M. Mazzara and B. Meyer, Eds. Cham: Springer International Publishing, 2017, pp. 195–216. doi: 10.1007/978-3-319-67425-4_12.
- [3] MacKenzie, K. Laskey, F. McCabe, P. Brown, and R. Metz, "Reference model for service oriented architecture 1.0," *Public Rev Draft*, vol. 2, pp. 1–31, Aug. 2006.
- [4] R. Mohan, T. Ramanathan, G. Rajendran, and D. N. MohanRaj, "Gartner Research Reviews on Middleware," *Int. J. Sci. Res. Publ.*, vol. 5, no. 9, p. 2, 2015.
- [5] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: a proposal and a discussion," *Requir. Eng.*, vol. 11, no. 1, pp. 102–107, Mar. 2006, doi: 10.1007/s00766-005-0021-6.
- [6] N. Alshuqayran, N. Ali, and R. Evans, "A Systematic Mapping Study in Microservice Architecture," in *2016 IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA)*, Macau, China, Nov. 2016, pp. 44–51. doi: 10.1109/SOCA.2016.15.
- [7] S. Li, "Understanding Quality Attributes in Microservice Architecture," in *2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW)*, Nanjing, Dec. 2017, pp. 9–10. doi: 10.1109/APSECW.2017.33.
- [8] S. Li et al., "Understanding and addressing quality attributes of microservices architecture: A Systematic literature review," *Inf. Softw. Technol.*, vol. 131, p. 106449, Mar. 2021, doi: 10.1016/j.infsof.2020.106449.
- [9] A. Chandrasekar, M. SudhaRajesh, and M. P. Rajesh, "A Research Study on Software Quality Attributes," *Int. J. Sci. Res. Publ.*, vol. 4, no. 1, p. 4, 2014.
- [10] M. Svahnberg, C. Wohlin, L. Lundberg, and M. Mattsson, "A Method for Understanding Quality Attributes in Software Architecture Structures," in *Proceedings of the 14th international conference on Software engineering and knowledge engineering*, Jan. 2002, p. 8. doi: 10.1145/568760.568900.
- [11] N. Kratzke, "A Brief History of Cloud Application Architectures," *Appl. Sci.*, vol. 8, no. 8, p. 1368, Aug. 2018, doi: 10.3390/app8081368.
- [12] U. Smith and L. G. Williams, "Software performance engineering: a case study including performance comparison with design alternatives," *IEEE Trans. Softw. Eng.*, vol. 19, no. 7, pp. 720–741, Jul. 1993, doi: 10.1109/32.238572.
- [13] Marquez, M. M. Villegas, and H. Astudillo, "An Empirical Study of Scalability Frameworks in Open Source Microservices-based Systems," in *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*, Santiago, Chile, Nov. 2018, pp. 1–8. doi: 10.1109/SCCC.2018.8705256.
- [14] S. Sasono, F. R. Rumambi, R. Priskila, and D. B. Setyohadi, "Integration of pharmacy and drug manufacturers in RSUD Dr Samratulangi Tondano by ESB WSO2 to improve service quality: (A case study of RSUD Dr Samratulangi Tondano, Minahasa Regency, North Sulawesi)," in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Oct. 2017, pp. 249–254. doi: 10.1109/ICITACEE.2017.8257712.
- [15] softawarewiki, "7 Excellent Open Source ESB (Enterprise Service Bus) Alternatives." <https://www.fromdev.com/2012/03/7-excellent-open-source-enterprise.html>
- [16] Chanaka Fernando, "Five Reasons Why WSO2 is Better Than Mule." WSO2, Sep. 23, 2020. [Online]. Available: <https://wso2.com/blogs/thetsource/five-reasons-why-wso2-is-better-than-mule/>

- [17] J. Ma, H. Yu, and J. Guo, "Research and Implement on Application Integration Based on the Apache Synapse ESB platform," *AASRI Procedia*, vol. 1, pp. 82–86, 2012, doi: 10.1016/j.aasri.2012.06.015.
- [18] WSO2, "Quick Start Guide - Enterprise Service Bus 5.0.0 - WSO2 Documentation." WSO2 Inc. [Online]. Available: <https://docs.wso2.com/display/ESB500/Quick+Start+Guide>
- [19] J. Krein, "Web-based application integration: advanced business process monitoring in WSO2 carbon," 2011, doi: 10.18419/opus-2719.
- [20] R. B. Khan, "Comparative Study of Performance Testing Tools: Apache JMeter and HP LoadRunner," p. 57.
- [21] "Amazon EC2 Instance Types - Amazon Web Services," Amazon Web Services, Inc. <https://aws.amazon.com/ec2/instance-types/>
- [22] "AWS Pricing Calculator." <https://calculator.aws>

Comparison of supervised learning-based indoor localization techniques for smart building applications

M. W. P. Maduraga*
IIC University of Technology,
The Kingdom of Cambodia
m.w.pasan@iic.edu.kh

Ruvan Abeysekera
IIC University of Technology,
The Kingdom of Cambodia
ruvan@iic.edu.kh

Abstract - Smart buildings involve modern applications of the Internet of Things (IoT). Intelligent buildings could include applications based on indoor localization, such as tracking the real-time location of humans inside the building using sensors. Mobile sensor nodes can emit electromagnetic signals in an ambient sensor network, and fixed sensors in the same network can detect the Received Signal Strength (RSS) from its mobile sensor nodes. However, many works exist for RSS-based indoor localization that use deterministic algorithms. It's complicated to suggest a generated mechanism for any indoor localization application due to the fluctuation of RSSI values. This paper has investigated supervised machine learning algorithms to obtain the accurate location of an object with the aid of Received Signal Strengths Indicator (RSSI) values measured through sensors. An available RSSI data set was trained using multiple supervised learning algorithms to predict the location and their average algorithm errors were compared.

Keywords - indoor positioning, Internet of Things (IoT), Supervised Learning

I. INTRODUCTION

Integrating technological advances into a building can be combined with many applications to improve humans' living standards. For example, tracking a person's location in a shopping complex, tracking the daily activity of an elderly person living alone in a house, tracking autonomous robots in an indoor environment, etc. In the recent development of the Internet of Things (IoT), wearable smart devices are built on wireless technologies such as Wi-Fi, Bluetooth Low Energy (BLE), Zigbee, LoRaWAN, etc. These devices can communicate data with the IoT network. Such data transmitted through the web could be information on building health, weather conditions, or other sensing information. When a connection is established between a sensor and the base station, the signal strengths of each wireless link can be measured. In indoor localization, it uses the signal strength as an input to compute the geographical location of that mobile sensor.

An indoor positioning system is used to locate stationary or moving objects and devices in an environment where the Global Positioning System (GPS) cannot be applied. GPS is appropriate when it is used in outdoor positioning-related applications. However, it consumes much energy, and implementation is costly for each node in an extensive network. Moreover, GPS is highly dependent on line-of-sight (LOS), and GPS cannot be used indoors. In addition, GPS allows only a maximum of 5 meters. Therefore, this may be suitable for the outdoors. Many applications initiate indoor positioning systems in areas such as hospitals that can perform indoor positioning

to track patients, where the doctor will accurately know a patient's location within the building.

Another example is real-time tracking of elderly people inside the home. The guardians could monitor the real-time location of elderly people using their mobile phones through IoT servers. In the farming industry [1], indoor positioning can be used for animal tracking, military applications, etc. [2][3]. Implementation costs of this technique is very low compared to the other monitoring mechanisms such as image processing-based systems. In image processing-based systems the camera has to be always focused on objects, and the object and camera should always be in the line of sight.

Most IoT devices are small in size. Thus, hardware requirements are usually minimal. They have limited capacity for storage, low processing power, and fundamental communication capabilities. Therefore, the localization algorithm needs to adapt to these features of the apparatus. To make an indoor positioning system successful, it requires to track multiple targets at once.

Various wireless technologies have been proposed and tested to perform indoor positioning in literature. The most commonly used technologies are Wi-Fi, Bluetooth, Radio Frequency Identification (RFID), Bluetooth Low Energy (BLE), Zigbee, and LoRaWAN. But, each of them has strengths and weaknesses. Due to the high availability of access points in the building, Wi-Fi has become the most straightforward option in such solutions. However, the purpose of deploying Wi-Fi access points is usually to provide maximum coverage to Internet users. In this case, signal coverage is not sufficient for a localization application.

Furthermore, Wi-Fi also consumes a lot of power. Compared to Wi-Fi, Zigbee and LoRaWAN have a perfect sensing range. But when these devices are used, implementation costs are high

This article compares indoor positioning accuracy using multiple supervised algorithms for IoT systems developed using Zigbee, BLE, and LoRaWAN. Zigbee is considered a long-range and low-power technology and is typically used in IoT applications. LoRaWAN is a new technology and is not as popular as the previous technology, transmitting at 915MHz with high data Speed. LoRaWAN nodes can reach a distance of 15000 meters, limiting the number of nodes required for the sequence.

The remaining content of the paper is organized as follows. Section II presents recent related work in the literature on signal strength-based indoor localization, and Section III discusses the different wireless technologies experimented with, in this work. The experimental setup

used to collect data is explained in section IV. Section V presents the supervised learning algorithms trained to estimate the locations of the results analyzed in section VI. Finally, the discussion and concluding remarks are presented in Section VII.

II. RELATED WORKS

Based on related literature, indoor localization primarily uses time-based, angle-based, RSS-based, or a combination of these technologies to obtain their signal measurements. The relationship between RSSI and distance is the key to wireless ranging and localization systems, where length is measured based on the signal strength received from each transmitting node. According to RSSI-based indoor positioning applications, mobile node position estimation is primarily achieved by triangulation and trilateration techniques. The Time of Arrival (TOA) and Time Difference of Arrival (TDOA) are time-based measurements related to transmission time. The Angle of Arrival (AOA) -based position estimation system requires a very complex directional antenna as a beacon node for angle measurement [1]. In literature, RSS-based multilateration positioning technology is the most popular algorithm used due to its simplicity.

Moreover, Kalman filters and extended Kalman filters have been used to filter RSSI data, and several Bayesian algorithms are investigated for estimating the locations. Machine learning is very suitable for predicting the expected target output using sample data, and algorithms such as neural networks, to identify WSNs. Furthermore, Payal et al. used FFNN to develop WSN-based ANN localization techniques, a cost-effective localization framework [4].

An experiment on localization uses RSSI based on Wi-Fi. RSSI values have been obtained from 32 different locations in an indoor environment and a supervised learning algorithm has been used to obtain accurate locations. Their results show that Decision Tree Regressor, Support Vector Regressor, and Random Forest Regression show fewer errors in location estimations [5].

Sebastian and Petros contributed to indoor positioning based on Zigbee, LoRaWAN, Wi-Fi, and BLE. They have designed individual systems in indoor environments and obtained RSSI values. They have used a deterministic algorithm in the localization phase, trilateration to get the accurate location, and presented error comparisons [6] [7].

The RSSI measurements are volatile in terms of time and position, so it is difficult to generally propose a stable and accurate positioning algorithm for all kinds of indoor localization applications. Further, related works presented in the literature for deterministic algorithms based on localization have low accuracy. The proposed study explores open issues in the literature by simplifying the hardware architecture while minimizing the complexity of the deterministic algorithms used to find mobile nodes in an indoor environment.

The proposed solutions for indoor localization based on deterministic and probabilistic algorithms are impractical to be implemented on real hardware devices. This is due to the complexity of proposed algorithms and hardware incompatibility. However, recently developed hardware devices such as programmable sensor nodes and single-board computers for IoT, support machine learning computations.

III. WIRELESS TECHNOLOGIES

This work has considered three types of wireless technologies used in IoT systems to collect RSSI data.

A. BLUETOOTH LOW ENERGY – BLE

Bluetooth Low Energy (BLE) is considered a low-power wireless communication technology used in short distance communication applications. Specific smart wireless devices that work every day (smartphones, smartwatches, fitness trackers, wireless headphones, computers, etc.) use BLE to create a seamless connection between devices.

For the experiment testbed in [7], the ten beacon nodes are designed using Gimbal Beacon. The Gimbal Beacon is from the Apple iBeacon protocol. iBeacon data packet structure defines three fields: a universal unique identifier (UUID), a 16-byte lot used to identify a group of beacons. The second and third fields are the "primary" and "secondary" values.

B. ZIGBEE - IEEE 802.15.4

Zigbee is low-cost, energy-saving, and can create mesh networks. It is a communication protocol based on the IEEE 802.15.4 standard for creating personal area networks with small antennas. The XBee is a type of sensor node based on Zigbee technology where XBee has low latency requirements and is easy to use, a device that allows you to create a multipoint Zigbee network quickly. In the experimental testbed in [6], it has used 2mW wired antenna XBees. Due to the limited processing power of XBees, Microcontrollers are essential for controlling the flow of information. Therefore, the microcontroller selected is Arduino Uno, due to its easy integration with XBee and low power consumption [6][7].

C. LoRAWAN

At lower transmission speeds, this technology was initially developed as LongRange by the LoRa Alliance Local Area Network (LoRaWAN) Protocol. The frequency is 915MHz [8]. Benefits of using frequency lower than 2.4GHz, is because longer wavelengths are possible. Then this makes the signal reach far distances. The frequency of 915MHz is LoRaWAN is relatively free and does not interfere.

Therefore, the node communicates with other transmission equipment. When used, it is less susceptible to noise. LoRaWAN is safer than other wireless technologies in IoT because encrypted data can be sent to various places frequently. A wide transmission range makes it very suitable for applications such as smart cities. The disadvantage of using such low frequencies is reduced data rates between nodes.

In terms of cost, it's pretty high for LoRaWAN based devices. Moreover, a large antenna and additional hardware are needed to access the media. Very effective for remote outdoor positioning, but short-range indoor positioning may present some challenges. In terms of range, each wireless technology has its sensing ranges, as shown in Table I.

TABLE I. TRANSMISSION RANGE OF THE WIRELESS COMMUNICATION TECHNOLOGIES

Wireless Technology	Range(m)
LoRaWAN	10,000
BLE	60
Zigbee	100

IV. EXPERIMENTAL SETUP

This work has used the data set in Sebastian and Petros [9]. The original experiment has been conducted in two different environments, and two datasets are available. However, this experiment uses the dataset related to environment 1 [9]. The experiment setup has been implemented in a laboratory room, as shown in figure 2. The environment is non-line-of-sight (NLOS). An experiment was conducted to eliminate interferences from other wireless devices such as Wi-Fi hotspots and mobile phones in the evening. Beacon nodes are placed at positions A, B, and C, as shown in figure 1, and mobile nodes are placed at positions D1, D2, and D3, respectively, to collect RSSI data. A series of tests were conducted to test positioning accuracy when positioning short and long distances between receivers and transmitters in all indoor systems., All experiments are done at night to minimize interference caused by other devices using the same media for transmission. Because RSSI values are vulnerable to interference, a controlled environment can generate more consistent readings for all tests performed.

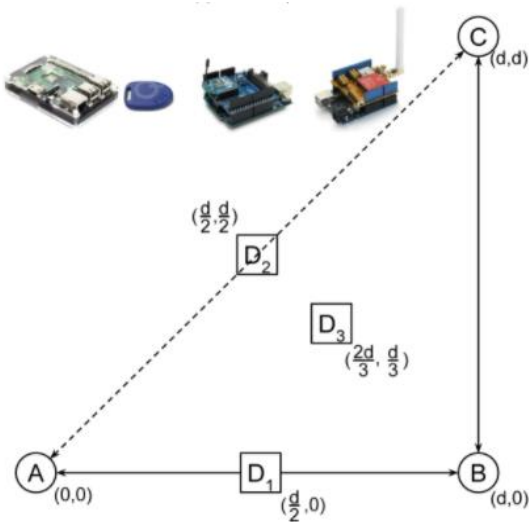


Fig 1. Arrangement of sensor nodes and positions [9]



Fig 2. Experiment environment [9]

V. INDOOR LOCALIZATION USING SUPERVISED LEARNING

A. RSSI based indoor localization

RSSI is recommended as one of the best approaches for indoor localization [7]. The main reason for its popularity is that RSSI does not require any additional hardware for signal measurement. The RSSI levels are measured by the received from the transmitter end of the device. In localization scenario, reference node detecting the RSSI levels receiving from the mobile sensor node, that we need to estimate the location. It is often used to determine the distance between a transmitter and a receiver because the signal strength decreases as the signal moves outward from the transmitter. Because the propagated signal is susceptible to environmental noise, RSSIs usually lead to inaccurate values and errors in positioning systems—the relationship between the distance and RSSI is expressed in equation 1 [6].

$$RSSI = -(10n) \log_{10}(d) + A, \quad (1)$$

where n is the signal propagation constant, d is the distance in meters, and A is the offset RSSI reading at one meter from the transmitter.

B. Support Vector Regressor

Support Vector Regression (SVR) uses the same classification principles as Support Vector Machine (SVM), with some differences. First, because the output is accurate, the information at hand is difficult to predict and has endless possibilities. SVR is a robust supervised learning algorithm that allows selecting an error tolerance by accepting the margin of error and adjusting the margin of error that exceeds the margin of error. For regression, the margin of error (ϵ) is set to approximate the SVM requested by the problem [5] [10].

C. Decision Tree Regressor

In Decision Tree Regressor, decision trees form a learning tree structure for solving classification or regression problems. The model divides the training data into several labels according to the creation rules. After creating the tree structure, it predicts the new data label by traversing the input data in the training tree. The information flow in the decision tree is so transparent that users can easily correlate assumptions without any background analysis [5][10].

D. Random Forest Regression

Random Forest Regression (RFR) is a supervised machine learning algorithm that uses ensemble learning methods for classification and regression. It works by creating many decision trees during training and testing each tree's class (classification) or average prediction (regression) model. This is one of the most accurate learning algorithms available. Many datasets produce very accurate classifiers when this algorithm is used. It could be run efficiently on large databases. It can handle thousands of input variables without removing the variables [10] [11].

VI. MODEL TRAINING AND RESULTS

The RSSI values received from the mobile sensor node at positions D1, D2, and D3 are used as the feature to train

models. These RSSI values are collected by reference nodes placed at fixed points, as shown in figure 9. In this work, RSSI data were trained using supervised algorithms DTR, RFR, and SVR, and a comparison of errors of each location D1, D2, and D3 shows in Table I, Table II, and Table III, respectively. The errors of positioning are calculated based on equation 1. The Jupyter Notebook (Python 3) was used to train the algorithms [12]. The experimental results present valuable insights in terms of accuracy. BLE was the most accurate wireless technology compared to the other two. However, BLE has a minimal distance of operation. Therefore, BLE is suitable for short-range indoor localization applications.

Further, BLE consumes very little power [7]. Thus, it prolongs the sensor uptime. While Zigbee showed average errors, LoRaWAN had the highest estimation errors.

$$Error = \sqrt{(x_{predict} - x_{real})^2 - (y_{predict} - y_{real})^2} \quad (2)$$

TABLE II. ERROR COMPARISON FOR BLE

Test Point	Actual Coordinates		Errors (m)		
	x	y	DTR	RFR	SVR
D1	0.500	0.000	0.116	0.089	0.189
D2	0.500	0.500	0.013	0.011	0.602
D3	0.667	0.333	0.167	0.124	0.478
Average			0.432	0.323	0.423

TABLE III. ERROR COMPARISON FOR ZIGBEE

Test Point	Actual Coordinates		Errors (m)		
	x	y	DTR	RFR	SVR
D1	0.500	0.000	0.193	0.223	0.394
D2	0.500	0.500	0.113	0.299	0.403
D3	0.667	0.333	0.303	0.982	0.384
Average			0.536	0.501	0.393

TABLE IV. ERROR COMPARISON FOR LORAWAN

Test Point	Actual Coordinates		Errors (m)		
	x	y	DTR	RFR	SVR
D1	0.500	0.000	0.993	0.523	1.932
D2	0.500	0.500	1.093	0.521	0.928
D3	0.667	0.333	0.890	0.732	1.993
average			0.992	0.592	1.617

VII. CONCLUSION

This paper compared RSSI-based indoor localization based on the wireless technologies BLE, LoRaWAN, and Zigbee for use in indoor localization systems. The experiments used RSSI data received from three reference nodes built on the above wireless technologies. Supervised learning techniques were used to estimate the geographical location of a mobile node. When comparing the localization accuracy, all algorithms tested in this experiment give fairly good error values less than one meter. When comparing the technologies BLE outperformed the other two technologies based on the results, achieving the lowest error from all the supervised algorithms experimented with. It is observed that one algorithm cannot be proposed as the best because different algorithms perform differently with each technology. Moreover, BLE is considered the minimal power-consuming technology. This experiment only considers 2D environments. Study on localization for 3D environments would be an interesting future research direction.

REFERENCES

- [1] M.W.P Maduranga and Ruwan Abeysekera "Machine Learning Applications in IoT Based Agriculture and Smart Farming: A Review," In Proc. International Journal of Engineering Applied Sciences and Technology, 2020, Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 24-27
- [2] Obeidat, H., Shuaieb, W., Obeidat, O. et al. A Review of Indoor Localization Techniques and Wireless Technologies. *Wireless Pers Commun* (2021). <https://doi.org/10.1007/s11277-021-08209-5>
- [3] H. Ahn and S. Rhee, "Simulation of a RSSI-Based Indoor Localization System Using Wireless Sensor Network," 2010 Proceedings of the 5th International Conference on Ubiquitous Information Technologies and Applications, 2010, pp. 1-4, doi: 10.1109/ICUT.2010.5678179.
- [4] Payal C. S. Rai and B. V. R. Reddy, "Artificial Neural Networks for developing localization framework in Wireless Sensor Networks," In *Proc.2014 ICDMIC*, New Delhi, 2014, pp. 1-6.
- [5] M W P Maduranga and Ruwan Abeysekera. "Supervised Machine Learning for RSSI based Indoor Localization in IoT Applications" *International Journal of Computer Applications* 183(3):26-32, May 2021
- [6] S. Sadowski and P. Spachos, "Comparison of RSSI-Based Indoor Localization for Smart Buildings with Internet of Things," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 24-29, doi: 10.1109/IEMCON.2018.8614863.
- [7] S. Sadowski and P. Spachos, "RSSI-Based Indoor Localization With the Internet of Things," in *proc. of IEEE Access*, vol. 6, pp. 30149-30161, 2018, doi: 10.1109/ACCESS.2018.2843325.
- [8] M. Rizzi, P. Ferrari, A. Flammini, E. Sisinni and M. Gidlund, "Using LoRa for industrial wireless networks," in *Proc. of 2017 IEEE 13th International Workshop on Factory Communication Systems (WFCS)*, 2017, pp. 1-4, doi: 10.1109/WFCS.2017.7991972.
- [9] Sebastian Sadowski, Petros Spachos, May 30, 2018, "RSSI-Based Indoor Localization with the Internet of Things", *IEEE Dataport*, doi: <https://dx.doi.org/10.21227/H21Q18>.
- [10] A. Nessa, B. Adhikari, F. Hussain and X. N. Fernando, "A Survey of Machine Learning for Indoor Positioning," in *IEEE Access*, vol. 8, pp. 214945-214965, 2020, doi: 10.1109/ACCESS.2020.3039271
- [11] S. Bozkurt, G. Elibol, S. Gunal and U. Yayan, "A comparative study on machine learning algorithms for indoor positioning," 2015 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2015, pp. 1-8, doi: 10.1109/INISTA.2015.7276725.
- [12] "Python Machine Learning" www.w3schools.com accessed on 20.01.2021.

Solution approach to incompatibility of products in a multi-product and heterogeneous vehicle routing problem: An application in the 3PL industry

H. D. W. Weerakkody*

Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
dilshan606@gmail.com

D. H. H. Niwunhella

Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
hirunin@kln.ac.lk

A. N. Wijayanayake

Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
anni@kln.ac.lk

Abstract - Vehicle Routing Problem (VRP) is an extensively discussed area under supply chain literature, though it has variety of applications. Multi-product related VRP considers about optimizing the routes of vehicles distributing multiple commodities. Domestic distribution of goods of multiple clients from a third-party logistics distribution centre (DC) is one example of such an application. Compatibility of products is a major factor taken into consideration when consolidating and distributing multiple products in the same vehicle. From the literature, it was identified that, though compatibility is a major consideration, it has not been considered in the literature when developing vehicle routing models. Therefore, this study has been carried out with the objective of minimizing the cost of distribution in the multi-product VRP while considering the compatibility of the products distributed, using heterogeneous vehicle types. The extended mathematical model proposed has been validated using data obtained from a leading 3PL firm in Sri Lanka which has been simulated using the Supply Chain Guru software. The numerical results showcase that cost has been reduced when consolidating shipments in a 3PL DC. The study will contribute to literature with the finding that the compatibility factor of products can be considered when developing vehicle routing models for the multi-product related VRP.

Keywords - compatibility of products, consolidation, simulation, third-party logistics, vehicle routing problem

I. INTRODUCTION

The Vehicle Routing Problem (VRP) is a well-known problem in the field of Operations Research, in which a set of geographically dispersed customers are served using a fleet of vehicles based in one or several warehouses [1]. This is an extension of the traveling salesman problem [2]. The objective of VRP is to find the optimal set of routes to deliver a set of customers with known demands at an optimized cost, where the vehicle routes are originated and terminated at a destination. Reference [3] states that VRP is an important problem in the fields of transportation, distribution, and logistics. Furthermore, it states that the context in VRP is to plan the routes to deliver goods from a central depot to customers who have placed orders for goods. VRP is an NP (non-deterministic polynomial-time) hard problem that has got a lot of attention in research work, and several techniques on exact methods and heuristics have been proposed and developed in solving the VRP [4].

There are many variants of VRP found in the literature such as Capacitated Vehicle Routing Problem (CVRP), Vehicle Routing Problem with Time Windows (VRPTW), studies which were conducted on the areas considered in this study.

Multiple Products, and Compartment related Vehicle Routing Problem and so on [5][6]. In CVRP, the capacity of the vehicle is imposed as a constraint while in VRPTW, the customer must be served within a specific time interval. In Multi-Product related VRP, multiple commodities are distributed to several customer locations.

One practical example where multi-product VRP can be applied is the domestic distribution of products of multiple clients from a 3PL DC. In this context, 3PL firms could consolidate the goods of multiple clients; thus, the problem can be treated as a multi-product related VRP. Consolidation is the coupling of shipments of different clients into the same vehicle. When considering the 3PL firms in Sri Lanka, most of them are currently not consolidating the shipments of different clients in the domestic distribution, whereas they separately distribute the shipments of those clients. Though consolidation can be identified as cost-effective, it has been challenging for them due to several reasons such as compatibility of different products, the unwillingness of clients to share the same vehicle with another client, and so on. Here the compatibility of the products is a major factor which should be considered when consolidating shipments. As an example, though a detergent product may be compatible with another chemical product, it is not compatible to transport in the same vehicle with a food product, because food items and detergent items are not compatible. Though, this compatibility factor has to be considered when developing the models in the multi-product related VRP, a gap in the existing literature was identified where the compatibility of products has not been considered when developing vehicle routing models. Therefore, this study has been carried out with the objective of considering the compatibility of products in a multi-product and heterogeneous vehicle routing problem where it has been applied to a real-world scenario in the 3PL industry.

II. LITERATURE REVIEW

In order to understand how the problem has been addressed in the previous studies, a thorough literature review was conducted in the areas of VRP where the detailed focus was given to multi-product related VRP and consolidation. It was noted that the compatibility of the products has not been taken into consideration when developing the vehicle routing models in the multi-product related VRP. This section will provide insights on few

Reference [7] has considered the VRP with multi-compartments in which the authors have considered the

problem, where customers can order several products, and the vehicles contain several compartments, but one compartment is dedicated to one product. The authors of that study have proposed a memetic algorithm and a tabu search algorithm. A study has been conducted by [8] on split VRP with capacity constraints for multi-product cross-docks. In VRP with split deliveries, customers can receive goods in multiple shipments, so the customer can be served by more than one vehicle. A mathematical model has been proposed to optimize the total operational and transportation cost. GAMS software has been used to obtain solutions for this problem in small-sized instances.

Reference [9] has conducted a study on multi-size compartment VRP with a split pattern where the distribution of multiple types of fluid products to customers has been considered. The authors have mainly focused on splitting the order quantities and loading each split demand to the compartments with different capacities and then determining the optimal routes. The paper has proposed three mathematical models and solution procedures of an optimization approach using CPLEX, 2-opt algorithm, and clustering technique. The study conducted by [10] on VRP in the frozen food distribution has proposed a model to optimize the total cost including transportation, refrigeration, penalty, and cargo damage cost. A heuristic-based Genetic Algorithm (GA) has been proposed to solve the model. The paper concludes that the proposed GA method can provide sound solutions in a reasonable time.

A study has been conducted by [3] on VRP for multiple product types, compartments, and trips with soft time windows. In soft time window, a penalty is being charged when the time windows are violated. The mathematical model proposed in the study has been developed in 3 cases: as in the first case, VRP for multiple product types, compartments, and trips is done without considering time windows. In the second case, time window is considered while in the final case, a soft time window is considered. The model proposed in this study contains a lot of constraints since the study deals with several aspects of VRP. A set of data obtained from literature was used to validate the model while AIMMS software has been used to obtain the solution.

The study conducted by [11] on a multi-compartment VRP with a heterogeneous fleet of vehicle has proposed a model to minimize total driving distance using a minimum number of vehicles. A heuristic algorithm has been proposed in the paper which had shown effective results in solving the model. A study conducted on Fuel Replenishment Problem by [12] has considered the multi-compartment VRP with multiple trips to determine the routing of vehicles and the allocation of multiple products to vehicle compartments. The proposed MILP model in the study has been solved using CPLEX and an Adaptive Large Neighborhood Search (ALNS) heuristic algorithm which had given optimal solutions much faster than the exact MILP model using CPLEX.

In conclusion, the authors were unable to locate any model which has considered the compatibility aspects of the product. Thus the study will focus on the compatibility of the product categories when consolidating multiple products.

III. PROBLEM DEFINITION AND MODE DEVELOPMENT

The problem addressed considers a domestic distribution, which consists of a central 3PL DC, distributing goods of multiple clients to different customer locations in the same region. These customer locations can be regional distributors or supermarkets that have ordered goods of different clients. It is considered that a fleet of heterogeneous vehicles is allocated to distribute the goods. Here the orders are assumed to be given in Cubic Meter (CBM) units and the truck capacities are also given in the same units. This study proposes a model where the goods of multiple clients are consolidated into vehicles considering the compatibility which depends on the nature of the products. However, the method of arranging the allocated orders in the vehicles is not considered in this study. Since the compatibility of products is considered, it is assumed that the products can be arranged in vehicles where there will be no requirement for separate compartments for the products.

Fig. 1, illustrates the problem with a situation where a central 3PL DC is distributing products of 5 different clients to 9 customer locations in a particular region. The products of these 5 clients may belong to 6 different product categories as shown in Fig. 1. The compatibility among the product categories may be different as shown in Table I. If the products are compatible, then shown in 1 if not 0. Currently, the 3PL providers do not consolidate shipments of different clients though they are compatible in nature. Therefore, it is required to build up a model which consolidates these goods considering the compatibility as given in Table I.

Model Assumptions

- Customers are divided into clusters/regions and there will be no movements between clusters.
- The location of the distribution center and customers are constant.
- Distribution centers can adequately satisfy the demands of the customers.

Notations

n	Number of customers
m	Number of product categories
l	Number of brands/clients
v	Number of delivery vehicles
Q_k	Load capacity of k^{th} truck (CBM)
q_{igb}	Quantity demanded by i^{th} customer, for g^{th} product category of b^{th} brand (client)
OC_k	Fixed operating cost of k^{th} vehicle
TC_k	Transportation cost per kilometer (delivery cost per distance unit of k^{th} vehicle)
L_{ij}	Distance between client i and client j

$$\min Z = \sum_{k=1}^v OC_k + \sum_{i=0}^n \sum_{j=0}^n \sum_{k=1}^v (L_{ij} * TC_k * x_{ijk}) \quad (1)$$

$$\sum_{i=0}^n \sum_{g=1}^m \sum_{b=1}^l \sum_{t=1}^m (q_{igb} * y_{ik} * z_{gk} * z_{tk}) \leq Q_k \quad \forall k \quad (2)$$

$$\sum_{i=0}^n x_{ijk} = y_{jk} \quad \forall j,k \quad (3)$$

$$y_{ik} = \begin{cases} 1 & i^{th} \text{ customer is served by } k^{th} \text{ truck} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$x_{ijk} = \begin{cases} 1 & k^{th} \text{ truck drives from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$z_{gk} = \begin{cases} 1 & g^{th} \text{ product type is transported in truck } k \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Note:

$$z_{gk} * z_{tk} \begin{cases} 1 & \text{if 2 product types are compatible } t = 1 \dots m \\ 0 & \text{if 2 product types are incompatible } t = 1 \dots m \end{cases}$$

Here the expression (1) is the objective of the model, which is to minimize the overall cost of transportation including the fixed operating cost of vehicles and delivery cost per distance unit of vehicle type. Expression (2) ensures that the vehicles are not overloaded in terms of capacity. Expression (3) ensures that the route for each vehicle is considered. Expressions (4), (5) and (6) reflect the integer constraints related to assigned truck k or product type is transported in kth truck. Note: ensures the compatibility constraint where only compatible product categories are transported in a vehicle.

TABLE I. COMPATIBILITY MATRIX

	Apparel	Chemical	Detergent	Food	Pharma	Stationary
Apparel	1	0	1	1	1	1
Chemical	0	1	1	0	0	1
Detergent	1	1	1	0	0	1
Food	1	0	0	1	1	1
Pharma	1	0	0	1	1	1
Stationary	1	1	1	1	1	1

Compatibility Matrix

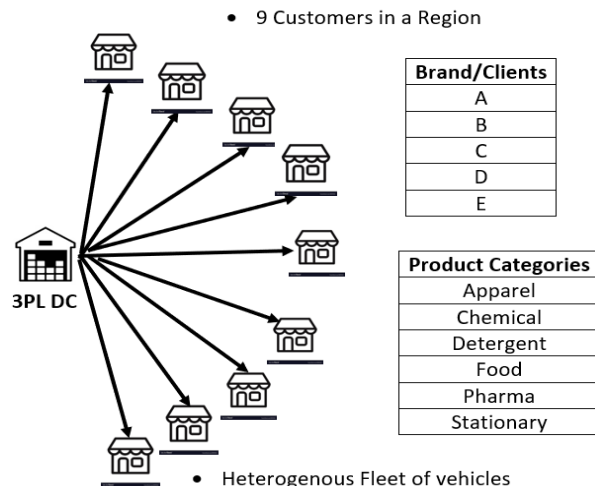


Fig. 1. Problem identification

Expressions (4), (5) and (6) reflect the integer constraints related to assigned truck k or product type is transported in kth truck. Note: Ensures that the route for each vehicle is considered.

IV. DATA ANALYSIS AND RESULTS OBTAINED

The extended mathematical model was validated using the data obtained from a leading 3PL provider in Sri Lanka. Customer locations were first divided into regions according to the distance. Then a particular region was selected, and the model was applied considering that region. Supply Chain Guru modelling and simulation software was used to simulate a real-world scenario taken from the 3PL provider. As mentioned earlier, since they are currently not consolidating the shipments of multiple clients, this current scenario was modelled as the baseline case and several other scenarios such as the consolidation scenario were created.

The real-world example considered here consists of a 3PL DC situated in Colombo, Sri Lanka, distributing six different categories of products to 9 different customers in the Southern region (same region). Fig. 1, shows the locations of the customers and the 3PL DC. It was assumed that a fleet of trucks with 10 vehicles of different capacities is available for the delivery process and their relevant delivery cost per km and the fixed costs were fed to the model. Geocode in Supply Chain Guru software was used to obtain the locations of the customers and the site. Data tables which were used include customers, sites, products, transport assets, asset availability, relationship constraints, rate, etc.

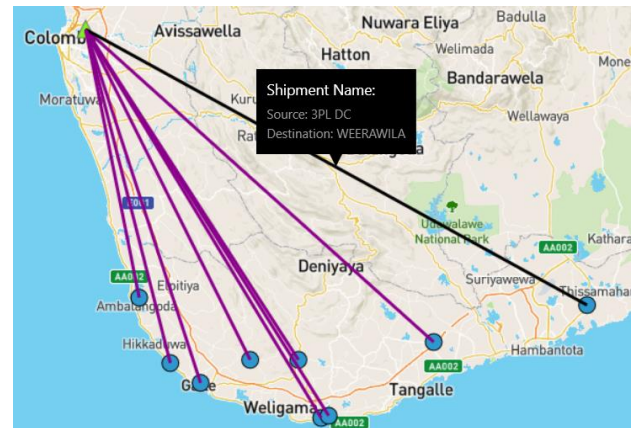


Fig. 2. Customer & 3PL DC Locations

The baseline case was compared with other scenarios based on the cost of travelling, travelling distance, use of vehicles, etc. Since the latitudes and longitudes of the locations are given as inputs here, the distance between locations is considered as the direct distance between locations. The results obtained from the simulated model using Supply Chain Guru software for the above scenario are discussed here.

The baseline model represents the situation where shipments of different clients are distributed separately, even to the same region though there are compatible shipments which can be distributed together. Fig. 3 depicts the aforementioned baseline scenario.

The consolidation scenario was created where compatible shipments are allowed to be shipped in the same vehicles. It was observed that the total transportation cost could be reduced. The route map of this scenario is shown in Fig. 4.

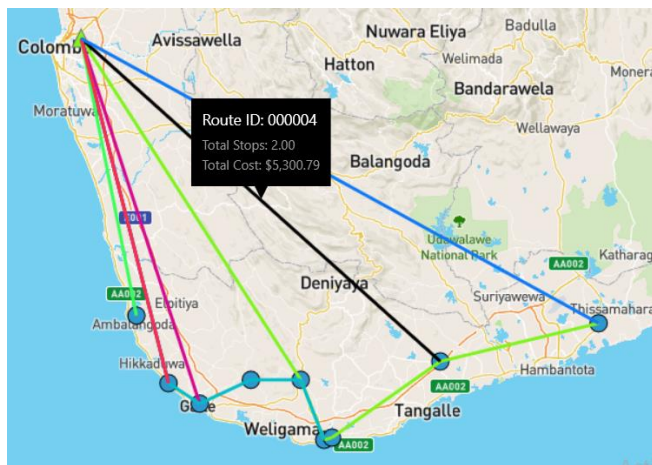


Fig. 3. Current scenario



Fig 4. Consolidated scenario

Table II presents the comparison of the baseline model, where the shipment of different clients were not being consolidated with the consolidation scenario where compatible shipments are consolidated and distributed. It was evident from this model that, for the above example, a percentage cost reduction of 11% could be achieved when consolidating shipments of different clients. Further, the distance travelled could be reduced significantly. During the simulation of the proposed model using Supply Chain Guru software, it was experienced that the percentage of cost reduction varied between 10% and 18%.

TABLE II. COMPARISON OF CURRENT & CONSOLIDATION SCENARIOS

	Baseline		Consolidation	
Total cost	79,360.08		70,488.38	
# Trucks used	Truck 1	1	2	
	Truck 2	0	1	
	Truck 3	3	1	
	Truck 4	2	2	
	Truck 5	1	1	
Total distance travelled	788.88		745.59	

$$\begin{aligned} \% \text{ Cost Reduction} &= (79,360.08 - 70,488.38) / 79,360.08 \\ &= 11.18\% \end{aligned}$$

$$\begin{aligned} \% \text{ Distance Reduction} &= (788.88 - 745.59) / 788.88 \\ &= 5.48\% \end{aligned}$$

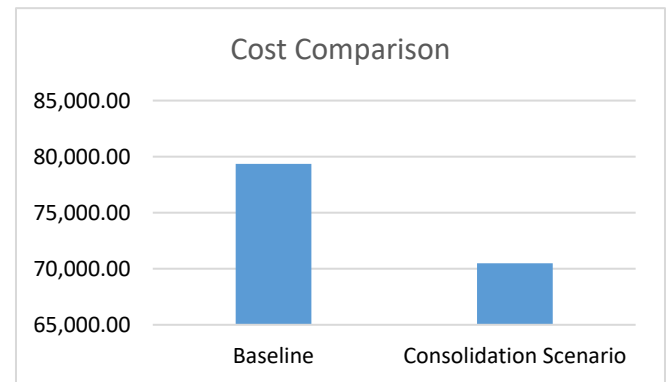


Fig. 5. Cost comparison

Fig. 5 depicts the comparison of the total costs in the baseline model and the consolidation model.

V. CONCLUSION

Multi-product related VRP is a variant of VRP which is a well-discussed problem in the literature. Since multi-product related VRP considers multiple commodities, a practical example for such a scenario could be identified as the domestic distribution of products of multiple clients from a 3PL DC. In Sri Lanka, most of the 3PL firms do not consolidate the shipments of multiple clients, though it could be found as cost-effective. One major factor which avoids 3PL firms from consolidation is the compatibility factor of products which should be definitely taken into consideration. From the referred literature, it was identified that a gap is existing where the compatibility of products has not been taken into account when developing models. Therefore, this study was conducted in order to address the above-mentioned gap. The extended mathematical model proposed in this study has been developed considering the compatibility of products. A real-world scenario taken from a leading 3PL provider in Sri Lanka has been used to validate the model which has been simulated using Supply Chain Guru modelling and simulation software. The numerical results have shown that the cost could be reduced nearly by 11% when consolidating the shipments, considering the compatibility. The study can be further expanded by adding more complexity to the model, considering different constraints such as order cutoff times, etc.

REFERENCES

- [1] S. Birim, "Vehicle Routing Problem with Cross Docking: A Simulated Annealing Approach", *Procedia - Soc. Behav. Sci.*, vol 235, no October, pp 149–158, 2016, doi: 10.1016/j.sbspro.2016.11.010.
- [2] C. Sabo, P. C. Pop, and A. Horvat-Marc, "On the selective vehicle routing problem", *Mathematics*, vol 8, no 5, 2020, doi: 10.3390/MATH8050771.
- [3] P. Kabcome and T. Mouktonglang, "Vehicle routing problem for multiple product types, compartments, and trips with soft time windows", *Int. J. Math. Math. Sci.*, vol 2015, 2015, doi: 10.1155/2015/126754.
- [4] A. A. Ibrahim, N. Lo, R. . Abdulaziz, and Ishaya J.A, "Capacitated Vehicle Routing Problem", *Int. J. Res. -*

- GRANTHAALAYAH, vol 7, no 3, bll 310–327, 2019, doi: 10.29121/granthaalayah.v7.i3.2019.976.
- [5] M. Hoàng Hà, N. Bostel, A. Langevin, and L. M. Rousseau, “An exact algorithm and a metaheuristic for the generalized vehicle routing problem with flexible fleet size”, *Comput. Oper. Res.*, vol 43, no 1, bll 9–19, 2014, doi: 10.1016/j.cor.2013.08.017.
- [6] H. Nazif and L. S. Lee, “Optimized crossover genetic algorithm for vehicle routing problem with time windows”, *Am. J. Appl. Sci.*, vol 7, no 1, bll 95–101, 2010, doi: 10.3844/ajassp.2010.95.101.
- [7] A. El fallahi, C. Prins, and R. Wolfler Calvo, “A memetic algorithm and a tabu search for the multi-compartment vehicle routing problem”, *Comput. Oper. Res.*, vol 35, no 5, bll 1725–1741, 2008, doi: 10.1016/j.cor.2006.10.006.
- [8] A. Hasani-Goodarzi and R. Tavakkoli-Moghaddam, “Capacitated Vehicle Routing Problem for Multi-Product Cross-Docking with Split Deliveries and Pickups”, *Procedia - Soc. Behav. Sci.*, vol 62, no 2010, bll 1360–1365, 2012, doi: 10.1016/j.sbspro.2012.09.232.
- [9] K. Asawarungsangkul, T. Rattanamanee, and T. Wuttipornpun, “A multi-size compartment vehicle routing problem for multi-product distribution: Models and solution procedures”, *Int. J. Artif. Intell.*, vol 11, no 13 A, bll 237–256, 2013.
- [10] Y. Zhang and X. D. Chen, “An optimization model for the vehicle routing problem in multiproduct frozen food delivery”, *J. Appl. Res. Technol.*, vol 12, no 2, bll 239–250, 2014, doi: 10.1016/S1665-6423(14)72340-5.
- [11] W. Chowmali and S. Sukto, “A novel two-phase approach for solving the multi-compartment vehicle routing problem with a heterogeneous fleet of vehicles: A case study on fuel delivery”, *Decis. Sci. Lett.*, vol 9, no 1, bll 77–90, 2020, doi: 10.5267/j.dsl.2019.7.003.
- [12] L. Wang, J. Kinable, and T. van Woensel, “The fuel replenishment problem: A split-delivery multi-compartment vehicle routing problem with multiple trips”, *Comput. Oper. Res.*, vol 118, 2020, doi: 10.1016/j.cor.2020.104904.
- [13] H. D. W. Weerakkody, D. H. H. Niwunhella, and A. Wijayanayake, “Cost minimization model through consolidation: application to a third party logistics distribution center”, in *International Conference on Applied and Pure Sciences*, 2020, bl 115.
- [14] S. Ahmed, M. Salman, and M. ASIM, “Factors Affecting the Selection of Third-Party Logistics Service Providers in the Edible Oil Industry of Karachi”, *CenRaPS J. Soc. Sci.*, vol 2, no 1, bll 88–102, 2020.
- [15] E. U. I. S. Byeon, “Simulation Study of Consolidated Transportation”, *Proc. 6th WSEAS Int. Conf. Simulation, Model. Optim. Lisbon, Port.*, bll 582–585, 2006.
- [16] M. L. F. Cheong, R. Bhatnagar, and S. C. Graves, “Logistics network design with supplier consolidation hubs and multiple shipment options”, *J. Ind. Manag. Optim.*, vol 3, no 1, bll 51–69, 2007, doi: 10.3934/jimo.2007.3.51.
- [17] N. Ghaffari-Nasab, M. Ghazanfari, and E. Teimoury, “Hub-and-spoke logistics network design for third party logistics service providers”, *Int. J. Manag. Sci. Eng. Manag.*, vol 11, no 1, bll 49–61, 2015, doi: 10.1080/17509653.2014.992994.
- [18] A. S. Hanbazazah, L. Abril, M. Erkoc, and N. Shaikh, “Freight consolidation with divisible shipments, delivery time windows, and piecewise transportation costs”, *Eur. J. Oper. Res.*, vol 276, no 1, bll 187–201, 2019, doi: 10.1016/j.ejor.2018.12.043.
- [19] A. S. Hanbazazah, L. E. Castro, M. Erkoc, and N. I. Shaikh, “In-transit freight consolidation of indivisible shipments”, *J. Oper. Res. Soc.*, vol 71, no 1, bll 37–54, 2019, doi: 10.1080/01605682.2018.1527191.
- [20] Z. C. Hua, H. Xin, and Z. Wei, “Logistics distribution routing optimization algorithm”, *Appl. Mech. Mater.*, vol 513–517, no 1, bll 1740–1743, 2014, doi: 10.4028/www.scientific.net/AMM.513-517.1740.
- [21] M. A. Khan, “Transportation Cost Optimization Using Linear Programming”, *Int. Conf. Mech. Ind. Energy Eng.*, no February, bll 1–5, 2014.
- [22] Z. Li, Z. Ma, W. Shi, and X. Qian, “Research on Medicine Distribution Route Optimization for Community Health Service Institutions”, *Math. Probl. Eng.*, vol 2016, 2016, doi: 10.1155/2016/6153898.
- [23] N. L. Ma, K. W. Tan, E. Lik, M. Chong, and K. W. Tan, “Improving Carbon Efficiency through Container Size Optimization and Shipment Consolidation”, *Proc. Int. Conf. Logist. Transp. 8th ICLT 2016*, 2016.
- [24] H. D. W. Weerakkody, A. Wijayanayake, and D. H. H. Niwunhella, “Vehicle Routing and Shipment Consolidation in a 3PL DC: A Systematic Vehicle Routing and Shipment Consolidation in a 3PL DC: A Systematic Literature Review of the Solution Approaches”, in *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, 2021, no March, bll 932–943.
- [25] R. Mesa-arango and S. V. Ukkusuri, “Benefits of in-vehicle consolidation in less than Truckload freight transportation operations”, *Procedia - Soc. Behav. Sci.*, vol 80, bll 576–590, 2013, doi: 10.1016/j.sbspro.2013.05.031.
- [26] M. D. Simoni, P. Bujanovic, S. D. Boyles, and E. Kutanoglu, “Urban consolidation solutions for parcel delivery considering location, fleet and route choice”, *Case Stud. Transp. Policy*, vol 6, no 1, bll 112–124, 2017, doi: 10.1016/j.cstp.2017.11.002.
- [27] A. Uzorh and N. Innocent, “Supply Chain Management Optimization Problem”, *Int. J. Eng. Sci.*, vol 3, no 6, bll 01–09, 2014.
- [28] S. U. Wijayadasa and D. G. N. D. Jayarathna, “Cost Optimization of Distribution network in Coca Cola Beverages Sri Lanka Limited”. 2017

Model to optimize the quantities of delivery products prioritizing the sustainability performance

A. P. K. J. Prabodhika*

Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
jpjinadari2@gmail.com

D. H. H. Niwunhella

Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
hirunin@kln.ac.lk

A. N. Wijayanayake

Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
anni@kln.ac.lk

Abstract - Many manufacturers and retailers often outsource their logistics functions to Logistics Service Providers (LSPs) to focus more on their core business process. Due to the competitiveness and the popularity of the sustainability concept, those organizations evaluate their prospective LSPs not only based on economic aspects like cost, service quality but also on social and environmental aspects as well when selecting LSPs. This paper proposes a methodology that can be used by organizations when evaluating and selecting LSPs based on their sustainability performance. Analytic Network Process (ANP) is used in evaluating the LSPs' sustainable performance since multiple dimensions and indicators need to be incorporated when measuring the sustainability performance. A Linear Programming Problem (LPP) model was proposed which allows the organizations to decide both desired number of LSPs and the volume to be allocated for those selected LSPs. The proposed methodology is flexible as it depends on the sustainability requirements of the organization when selecting LSPs. Both the indicators and their relative importance are up to the organization to decide.

Keywords - analytic network process, linear programming problem, logistics service providers, sustainability, sustainability indicators

I. INTRODUCTION

Logistics Service Providers (LSPs) which are also called 'Contract Logistics', 'Third-Part Logistics', 'Logistics Alliances', and 'Logistics Outsourcing' are firms that provide logistics services that are often integrated or bundled together for use by customers [1]. The role of LSPs has changed over time from providing transportation services to a wide range of services including warehousing, inventory management, freight forwarding, cross-docking, technology management, etc. At present many manufacturers and retailers often outsource their logistics functions to LSPs as they want to focus more on their core business processes.

Today business organizations are more towards sustainability and sustainable development and focus on making themselves and their supply chain partners economically, socially, and environmentally sustainable. Due to the competitiveness and the popularity of the sustainability concept, those organizations evaluate their prospective LSPs not only based on economic performance like cost, service quality but also on social and environmental performance as well. Although there are studies on one or two dimensions of sustainability performance (Economic and Environmental to be precise), the studies which incorporate social dimension are still lagging [2]. Relatively few studies done on the

environmental sustainability of the LSPs [3] and often sustainability dimensions are addressed in isolation [4].

The objective of this paper is to propose a methodology that can be used by organizations when evaluating their LSPs based on their sustainability performance and select the most suitable LSPs as the logistics partners. The proposed methodology is flexible as it depends on the sustainability requirements of a particular organization when selecting LSPs. Both the indicators and their relative importance are up to the organization or the decision-maker to decide.

A. Justification of the research

Many manufacturers and retailers often outsource their logistics functions to LSPs. The Sri Lankan logistics services sector has developed throughout the past few decades providing their customers a satisfactory service. The competitiveness has increased which resulted in LSPs becoming more integrated with their customers. And the research has found that the usage of logistics services will increase to a large extent in the near future. The competitiveness between the Sri Lankan LSPs has increased which has resulted in them being more integrated with customers [5].

LSPs are mainly dependent on both transport vehicles and employees, managing them from the viewpoint of social sustainability as well as from environmental sustainability has become a crucial issue [6],[7]. Selecting the best LSP for an organization is a crucial step. According to Pareto Analysis, [8] commonly used criteria when selecting a LSP are cost, relationship, services, quality, information systems, flexibility, and delivery. But with the popularity of the topic of sustainable development, organizations are now focusing on environmental and social criteria as well.

In general, the research focused on the evaluation of all three dimensions of sustainability are rare to find. Although many studies have been done on the areas of logistics outsourcing and logistics strategies, but relatively few studies on environmental sustainability. The majority of the studies measure the sustainability performance of the upstream supply chain and studies on the sustainability performance of LSPs are minimal [2].

Both quantitative and qualitative approaches have been used in evaluating and measuring sustainability performance. Mathematical models are used under the quantitative approach [9]. Widely used qualitative approaches are AHP, ANP, Fuzzy Set Approach, Balance Score Card, and DEA [2].

There is a need to develop research aimed at identifying standard metrics to measure LSP's environmental performance [3], [7].

B. Objectives of the research

- RO1: To identify the sustainability performance measures/indicators/criteria in LSPs
- RO2: To develop a methodology to evaluate the sustainability performance of LSPs
- RO3: To develop an LPP model to select the most suitable LSPs based on sustainability performance and other constraints.

II. LITERATURE REVIEW

A. Sustainability and LSPs

The economic dimension of sustainability is the aspect that is often evaluated in an organization. Studies that focus on measuring the performance of supply chains or LSPs traditionally have focused on economic aspects of it with cost minimization (Profit maximization) and service level maximization [10]. The study of [11], in their framework, covers the economic performance evaluation in five fields: Reliability, Responsiveness, Flexibility, Finance, and Quality. These five fields are further categorized into subfields with an extensive review of the literature. Further, this study highlights that the 'Finance' field was the field that was analyzed often.

From the business and management perspective, the environmental dimension of the sustainability concept involves all activities and decisions needed to minimize environmental pollution caused by an organization. In the logistics sector, the environmental concern has become a buzz topic due to many factors. Logistics and transport activities are the 2nd biggest contributor to GHGs (Greenhouse Gases) after electricity production. Demand for moving and delivering goods has grown exponentially in recent years and is expected to grow in the coming years which in turn will increase the demand for logistics services. Recent economic crisis and global warming have urged for more environmentally sustainable logistics services [12].

There are relatively few studies done on environmental sustainability in the logistics service industry. [12] in its descriptive analysis of literature has identified that there is a need to develop research aimed at identifying standard metrics to be used to measure green 3PL's environmental performance. And it suggests that future research should be aimed at developing frameworks and applications that may quantify 3PL's environmental commitment and its impact on finance and operational performance. Further the analysis suggests that future research should better evaluate the efficiency of green measures by using alternative performance indicators as well.

Using an extensive review of the literature [13] identified that Triple Bottom Line (TBL) and Global Reporting Initiatives (GRI) applications are the two main frameworks in measuring logistics environmental sustainability. [13] propose a set of environmental indicators for city logistics using the GRI framework as the evaluation basis. The proposed set of indicators falls under five categories: Energy, transport and infrastructure, noise, congestion, and emissions, effluents, and waste.

Social sustainability of LSPs means to operate its services considering their impact on internal and external stakeholders (i.e., society and employees) in terms of welfare, safety, and wellness. [11] includes the social dimension of sustainability to its analytical assessment model with five (5) social fields/categories: Work conditions, human rights, social commitment, customer issues, best practices. Further, the research categorizes the five fields into subfields/categories as well. The proposed composite index by [14] also includes the social dimension. The taken social performance measures are corruption risk and sourcing from local suppliers.

By using an extensive literature review [6] selected frequently adopted sustainability criteria with the help of industry experts. The study proposes price, service, and social sustainability as main criteria. Social sustainability criteria are sub-categorized into philanthropy and average salary which are quantitative measures and management policy which is a qualitative measure. Management policy is further categorized into organizational learning/training process or programs, human rights and participation, occupational health and safety, and vehicle safety.

Although the definition of sustainability consists of three dimensions and the need for such research papers is high, sustainability dimensions are addressed in isolation and quantified indicators for a social dimension are underdeveloped. [4] mentions the challenges when conducting sustainability logistics services including a wide range of sustainability indicators, measuring and quantifying the indicators – Especially social dimension indicators, integrating sustainability dimensions, trade-offs between the dimensions, influence from the stakeholders, time perspective, and contextual considerations.

B. Sustainability performance management and evaluation

To be more competitive, organizations need to measure and manage their supply chain sustainability effectively and efficiently. Through measuring and evaluating sustainability performance organizations can identify the gaps and areas to be improved for further development. Many research studies have proposed metrics and frameworks to measure sustainable supply chain performance.

Sustainability performance management approaches include environmental management standards like ISO 14001, international Reporting Standards (Global Reporting Incentive - GRI), SCOR framework, Life Cycle Assessment, Multi-Criteria Decision Making (MCDM) tools (AHP, ANP, DEA, etc.), Rough Set Theory, Fuzzy Set approach, Composite indicators, and conceptual frameworks. Industry-specific studies are sparsely present in the literature. The majority of the studies are focused on developing general frameworks to assess supply chain sustainability. Even Though there are studies with all three dimensions of sustainability, still the social dimension is lagging. Math-focused methods and tools used to measure sustainability are exponentially increasing. The majority of the studies focused on measuring the sustainability performance between suppliers and manufacturers [2].

Through an extensive analysis of literature [15] has found out that traditional research has focused on measuring supply chain performance in terms of cost, quality, speed, flexibility, and reliability refers to the

economic dimensions of sustainability. Further, the analysis has found out that in the last decade a considerable amount of research was based on green supply chains or green logistics referring to environmental sustainability. But little research has shown the social dimension performance of supply chains. It also highlights the importance of developing research models and frameworks that are country and industry-specific as the sustainability dimension impacts are context-dependent and technology-related.

[16] proposes a framework for environmental sustainability assessment by analyzing the literature which consists of seven macro-areas and these seven macro areas are divided into two as inter-organizational and intra-organizational environmental practices. Distribution strategies and transport execution, warehousing and green building, reverse logistics, packaging management, and internal management belong to the intra-organizational practices in the context of the logistics industry while collaborating with customers and external collaborations belong to inter-organizational environmental practices. A study found that LSPs have adapted many sustainability initiatives related to distribution and transportation activities while initiatives related to internal management are less. Internal management initiatives include environmental compliance and auditing programs, environmental performance measuring and monitoring, use of green IT, promotion of environmental awareness among managers, incentives, and benefits for green behaviors, and development of formal environmental sustainability standards of the company. It also highlights the lack of standard methodology for measuring the environmental impact and the need of developing effective performance measurement systems. With the case study conducted, [16] found that the main driver for the environmental sustainability initiatives for LSPs is customers. The case study also revealed that government rules and regulations are also an important driver, but it is often considered as a barrier by the LSPs.

There are many tools to assess Supply Chain Management practices like Odette ENALOG, Efficient Consumer Response (ECR), Oliver Wight Class A Checklist for Business Excellence, and SCOR model. Among them, the most sustainability-oriented model is the SCOR model. The SCOR model has become more mature with GREENSCOR, but still, it lacks the integration of all three dimensions of sustainability.

[17] proposes the ASSC framework (Assessment of Sustainability in Supply Chains Framework) that allows qualitative and quantitative indicators to be employed in assessing environmental and social dimensions. It also allows the aggregation of relevant indicators into KPIs (Key Performance Indicators) with respect to specific aspects of sustainability. The proposed ASSC framework and the aggregation method are stable, but the content or the sustainability indicators used are adaptable which will be able to reflect the dynamics of sustainable development.

[6] has been using Analytical Hierarchy Process (AHP) for its sustainability performance evaluation framework due to its ease of use and applicability in real-world scenarios. For further preciseness fuzzy theory has been incorporated into the AHP to overcome the high degree of fuzziness and uncertainty of the answer.

TABLE I: SUMMARY OF THE SUSTAINABILITY DIMENSIONS AND THE RESPECTIVE MODELS USED IN THE LITERATURE REVIEW

Authors	Sustainability Dimension			Output
	Economic	Environment	Social	
[16]	—	√	—	Conceptual Model
[11]	√	√	√	Analytical Assessment Model
[10]	√	√	√	Multidimensional Model
[18]	√	√	√	Mathematical Model
[19]	√	√	√	Conceptual Model
[14]	√	√	√	Composite Index
[17]	√	√	√	Conceptual Model (ASSC Model)
[9]	√	√	√	Composite Index
[6]	√	—	√	Multi-Criteria Evaluation Model using Fuzzy AHP
[20]	√	√	—	Network Data Envelopment Analysis (NDEA) Model
[2]	√	√	√	Conceptual Model
[21]	√	√	√	3 rd Party Logistics Green Logistics Model (3PL GIF) Index

The proposed framework was used to evaluate three 3PL providers of an e-commerce company. Also, the study has proved that by changing the relative position of the criteria/sub-criteria in the proposed framework, decision-makers can determine the effect of such a change. Although the results show that the proposed framework is a good and a viable alternative to evaluate the social sustainability of 3PL providers, the exclusion of the environmental dimension in the framework is a major drawback.

[21] proposes the Green Innovative framework, 3PL GIF (Third Party Logistics Green Innovative Framework) based on social, economic, and environmental indicators. 3PL GIF checks the implementation of the business policies in all three dimensions of sustainability and helps the LSPs by altering them to use quality standards, measure them and continuously improve them. 3PL GIF provides an easy comparison between organizations and helps to identify lacking fields. 3PL GIF compares the progress in sustainable development between organizations and can be applied to a logistics company of any size.

According to Table I past authors have used different combinations of sustainability dimensions in their studies and their outputs were of different models and methodologies.

III. METHODOLOGY

Sustainability performance indicators for LSPs under each dimension are identified with the literature review, Global Reporting Initiatives (GRI), and expert opinions. Analytic Network Process (ANP) has been used to create a model and give weights or priorities for each dimension/indicator and then the sub-dimensions or sub-

indicators under each dimension and to rank the pool of LSPs available.

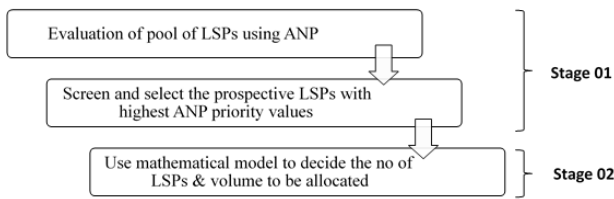


Fig. 1. Flow diagram of the methodology process

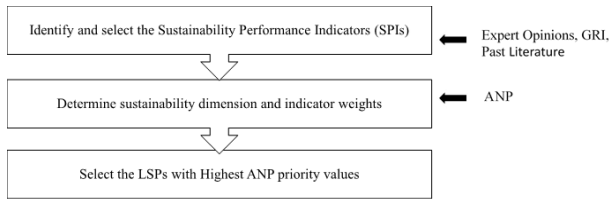


Fig. 2. Flow diagram of stage 1 of the methodology process

After getting the ranks of LSPs using ANP, the desired number of LSPs will be selected using a mathematical optimization model which was formulated as a Linear Programming Problem (LPP) with an objective of maximization of the volume allocated to LSPs with the highest rank while satisfying the constraints. Using the proposed LPP model, both the desired number of LSPs and the capacity to be allocated for those selected LSPs can be determined.

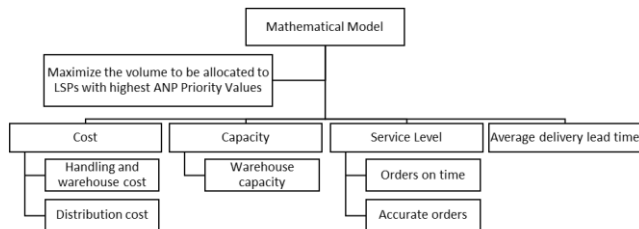


Fig. 3. Optimization Model Structure

C. Evaluation of LSPs using ANP.

As the initial step the sustainability performance indicators (sub-criteria) for LSP were identified with the help of literature review, Global Reporting Incentives (GRI), and expert opinions. Opinions on sustainability performance indicators were extracted from the logistics service industry experts through interviews (Table II). All three dimensions of sustainability were considered when selecting the indicators. Ten (10) industry experts from five leading 3PL service providers and a leading apparel manufacturing firm in Sri Lanka. The number and the types of indicators selected depend on the requirement of the organization which provides the flexibility for the proposed model.

The proposed methodology was applied to an apparel organization that uses multiple LSPs. Questionnaires were given to the logistics experts in the organization to determine the relative importance of four selected

dimensions and sustainability performance indicators. By the results of the questionnaire, weights of the dimensions of sustainability and sustainability performance indicators were determined using ANP based pairwise comparison using “Super Decision” software.

Then using the data acquired, the execution of the mathematical model was done.

TABLE II. SELECTED SUSTAINABILITY PERFORMANCE INDICATORS WITH THEIR SOURCES

No .	Economic dimension	Environmental dimension	Social dimension
1.	E1 - Direct economic Values generated and distributed (GRI 201-1)	EN1 - Adhering to Environmental laws and regulations (GRI 307-1)	S1 - Number of incidents of corruption reported and investigated (GRI 205-1)
2.	E2 - Market Share (Oršič et al., 2019)	EN2 - Directing waste for reuse/recycle or other recovery operations (GRI 306-4)	S2 - Incorporation of minorities in the workforce (GRI 405-1)
3.	E3 - R&D Expenditure (Salvado et al., 2015)	EN3 - Controlling GHG emissions (GRI 305-5)	S3 - Incorporation of women in the workforce (GRI 405-1)
4.		EN4 - Directing wastewater for recycling/reuse or other recovery operations (GRI 303-3)	S4 - Investments in local community development programs (GRI 413)
5.		EN5 - Controlling energy consumption through conservation and efficiency initiatives (GRI 302-4)	S5 - No of accidents and work-related ill health reported (GRI 403-1)

Table II shows the indicators selected under each sustainability dimension along with the sources of selection.

D. Development and Implementation of Mathematical Optimization Model

Assumptions:

- LSPs have unlimited distribution capacity.
- Supply from the manufacturer and the demand by the DCs are equal for the calculating period.

TABLE III. NOTATIONS AND DEFINITIONS OF THE MATHEMATICAL MODEL

Notations	Definitions
<i>Indices</i>	
n	<i>number of product types</i>
m	<i>number of distribution centers</i>
l	<i>number of LSPs</i>

Input Variables	
C_k	ANP priority value of LSPs
R_{ik}	if LSP k can distribute the product i, then R_{ik} is 1, otherwise 0
P_i	volume of the one unit of product i
DC_{jk}	total cost of delivery from LSP k to distribution centre j
HC_k	total cost of of handling at the LSP k
FD_{jk}	fixed cost of delivery from LSP k to distribution centre j
FH_k	fixed cost of handling LSP k
VD_{ijk}	variable cost of delivery of 1 CBM of product i from LSP k to distribution centre j
VH_{ik}	variable cost of handling 1 CBM of product i at LSP k
L_k	service level of LSP k
Q_k	average lead time of LSP k
V_k	LSP'k maximum capacity or volume
D_{ij}	demand or order qty for product i by the distribution centre j
W_i	supply quantity of product i
B	available budget for logistics outsourcing
N	desired no. of LSPs to have by the manufacturer
Decision Variables	
X_{ik}	delivery qty of product i from manufacturer to the LSP k
Y_{ijk}	delivery qty of product i from LSP k to distribution centre j
Z_k	if LSP k is considered, then Z_k is 1, otherwise 0

Objective Function

$$\text{Maximize } \sum_{k=1}^l \sum_{i=1}^n C_k * Z_k * L_k * Q_k * X_{ik}$$

Constraints

$$\sum_{k=1}^l X_{ik} \leq W_i \quad \text{for } i = 1 \dots n \quad (1)$$

$$\sum_{i=1}^n \sum_{j=1}^m (Y_{ijk} * R_{ik}) = D_{ij} \quad \text{for } j = 1 \dots m, i = 1 \dots n \quad (2)$$

$$\sum_{i=1}^n (P_i * X_{ik}) * Z_k \leq V_k \quad \text{for } k = 1 \dots l \quad (3)$$

$$DC_{jk} = FD_{jk} + \sum_{i=1}^n VD_{ijk} * Y_{ijk} * P_i \quad \text{for } k = 1 \dots l, j = 1 \dots m \quad (4)$$

$$HC_k = FH_k + \sum_{i=1}^n VH_{ik} * X_{ik} * P_i \quad \text{for } k = 1 \dots l \quad (5)$$

$$\sum_{k=1}^l \sum_{i=1}^n DC_{jk} + \sum_{k=1}^l HC_k \leq B \quad (6)$$

$$X_{ik} \geq \sum_{j=1}^m Y_{ijk} \quad \text{for } k = 1 \dots l, \text{ for } i = 1 \dots n \quad (7)$$

$$\sum_{k=1}^N Z_k \leq N \quad (8)$$

$$Z_k \in \{1,0\} \quad \text{for } k = 1 \dots l \quad (9)$$

Defining the constraints:

1. All the units of product i allocated to LSPs should be less than or equal to the manufacturers production capacity of that product i.
2. All the products distributed/ delivered to the distribution centers by the LSPs should be more than or equal to the demand from each distribution center and if LSP k can distribute the product I, then R_{ik} is 1, otherwise 0.
3. The volume that is allocated to the LSP k should be less than or equal to its capacity.
4. Total cost of delivery from LSP k to distribution center j
5. Total handling cost at LSP k.
6. Total cost (Delivery and handling) should be less than or equal the available budget for logistics outsourcing.
7. Quantity of products i allocated to each LSP should be equal or more than the amount of that product distributed by that LSP.
8. Sum of the allocated number of LSPs does not exceed the desired number of LSPs to have by the organization.
9. Binary variable If LSP k is considered, then Z_k is 1, otherwise 0.

IV. DATA ANALYSIS

A. Calculation of weights and priorities using ANP.

The final weight of each indicator was calculated by multiplying the indicator (sub-criteria) weight by the relevant dimension (criteria) weight as shown in Table IV.

TABLE IV. FINAL WEIGHTS OF SUSTAINABILITY PERFORMANCE INDICATORS

Dimension	Indicator	Indicator Weight	Dimension Weight	Final Weight	Rank
Economic	E1	0.5810	0.5630	0.3271	1
	E2	0.1954	0.5630	0.1100	4
	E3	0.2236	0.5630	0.1259	3
Environment	EN1	0.3061	0.1763	0.0539	5
	EN2	0.0741	0.1763	0.0131	13
	EN3	0.1734	0.1763	0.0306	10
	EN4	0.1834	0.1763	0.0323	9
	EN5	0.2631	0.1763	0.0464	6
Social	S1	0.5410	0.2608	0.1411	2
	S2	0.0791	0.2608	0.0206	12
	S3	0.1071	0.2608	0.0279	11
	S4	0.1411	0.2608	0.0368	7
	S5	0.1318	0.2608	0.0344	8
				1.0000	

Here three (3) prospective LSPs of the apparel manufacturing firm were considered and using ANP ranks were given to them based on their sustainability performance.

TABLE V. PRIORITY AND RANK CALCULATIONS OF 3LSPS

LSP	Priority	Rank
LSP1	0.32789	3
LSP2	0.33188	2
LSP3	0.34023	1

According to the results, the highest weighted and least weighted sustainability dimensions and the sustainability performance indicators of the organization can be identified. The prospective LSP with the highest priority/rank can be selected as the best alternative.

The following are the results of the calculations done for the data collected from the apparel manufacturing organization. According to the results, the highest importance is given to the economic dimension (0.5630) by the decision-makers, then to social (0.2608), then environmental (0.1763). Priorities of the LSP based on the weights are 0.32789, 0.33188, 0.34023 for LSP 1, LSP 2, LSP 3, respectively. Among them, the highest values were obtained by LSP 3 which is 0.34023 and it is the best selection among the three alternatives. The reason LSP 3 got the highest rank is, it has performed best in the highly weighted sustainability performance indicators by the organization.

B. Results of the mathematical optimization model

Execution of the mathematical model using the data acquired was done using IBM ILOG CPLEX Optimization Studio version 12.9.

Optimization was done with the implementation of the model in the Optimization Programming Language (OPL). The optimization results summary is shown in Table VI(A) and (B). The data used, and the detailed results tables are shown in the Appendix.

Only two prospective LSPs were considered for the execution of the model in CPLEX. According to the results, both LSPs were selected.

TABLE VI(A). OPTIMIZATION RESULT SUMMARY

LSPs	Product (Units)					Total
	1	2	3	4	5	
1	1000	1500	2200	0	800	5500
2	0	0	0	0	0	0
3	0	0	300	500	0	800
Total	1000	1500	2500	500	800	6300

TABLE VI(B). OPTIMIZATION RESULT SUMMARY

LSP	DC	Product	Qty (Units) delivered
1	1	3	1000
1	3	3	750
3	2	5	500
1	4	2	500
1	3	2	500
1	2	3	500
1	2	1	500
1	1	2	300
1	4	3	250
3	4	5	200

1	4	4	200
1	4	1	200
1	2	2	200
1	1	1	200
3	1	5	100
1	3	1	100
1	3	4	50
1	2	4	50
1	1	4	200

V. CONCLUSION

This paper uses Analytic Network Process (ANP) to evaluate the LSPs based on their sustainability performance. Analytic Network Process (ANP) provides the opportunity to the organization to evaluate its prospective logistics partners based on their requirements and priorities and the different sustainability dimensions and indicators. The criteria (Sustainability dimensions) and sub-criteria (Sustainability Indicators) used to select the LSP can be different from company to company and this methodology enables such options and provides the flexibility to select criteria and sub-criteria accordingly. The relative importance of the dimensions and sustainability indicators was determined through pairwise comparison. The LSP with the highest priority value is selected as the best sustainability performer.

As the next step, the desired number of LSPs will be selected using a mathematical optimization model which was formulated as a LPP with an objective of maximization of the volume allocated to LSPs according to the rank obtained during the Analytic Hierarchy Process values while satisfying the constraints. Using the proposed LPP model, both the desired number of LSPs and the capacity to be allocated for those selected LSPs can be determined.

Due to the difficulty in the collection of actual figures or quantitative values for the performance levels of sustainability, performance indicators were measured using a 9-point Likert Scale for getting data to do pairwise comparison which made the results subjective to the person who is giving the scores for the relevant performance. This requires future studies to collect the real quantitative indicator values of the prospective LSPs when using the model to get a more accurate outcome.

The proposed model enables not only to identify the best LSPs who meet the sustainability performance criteria at their best levels but also enables them to distribute the goods to different warehouses or distribution centers after considering all relevant constraints. Though the validity of the model was tested to an apparel industry this could be applied to many other industries.

In the LPP model, two assumptions were incorporated for ease of calculations and to reduce the optimization model complexity. One was considering LSPs have an unlimited distribution capacity which was not true in real life. And researcher has assumed that the supply from the manufacturer and the demand by the Distribution Centers (DCs) are equal for the calculating period. If future research can overlook these limitations and incorporate more constraints into the LPP model to get more accurate results.

REFERENCES

- [1] A. Ali, K. Chauhan, M. Barakat, and A. Eid, "The Role of Sustainability for Enhancing Third-Party Logistics Management Performance," *J. Manag. Sustain.*, vol. 9, no. 1, p. 14, Jan. 2019, doi: 10.5539/jms.v9n1p14.
- [2] A. Qorri, Z. Mujkić, and A. Kraslawski, "A conceptual framework for measuring sustainability performance of supply chains," *Journal of Cleaner Production*, vol. 189. Elsevier Ltd, pp. 570–584, Jul. 10, 2018, doi: 10.1016/j.jclepro.2018.04.073.
- [3] P. Evangelista, L. Santoro, and A. Thomas, "Environmental sustainability in third-party logistics service providers: A systematic literature review from 2000-2016," *Sustainability (Switzerland)*, vol. 10, no. 5. MDPI AG, May 11, 2018, doi: 10.3390/su10051627.
- [4] M. Björklund and H. Forslund, "Challenges addressed by swedish third-party logistics providers conducting sustainable logistics business cases," *Sustain.*, vol. 11, no. 9, May 2019, doi: 10.3390/su11092654.
- [5] W. Premarathne, "Issues and trends of third party logistics (3PL) market in Sri Lanka," *International Conference on Asian Studies*, 2012. [Online]. Available: <https://www.researchgate.net/publication/315656660>.
- [6] H. Jung, "Evaluation of third party logistics providers considering social sustainability," *Sustain.*, vol. 9, no. 5, 2017, doi: 10.3390/su9050777.
- [7] E. Sweeney, P. Evangelista, M. Hüge-Brodin, and K. Isaksson, "The Role of Third Party Logistics Providers (3PLs) in the Adoption The Role of Third Party Logistics Providers (3PLs) in the Adoption of Green Supply Chain Initiatives of Green Supply Chain Initiatives." [Online]. Available: <https://arrow.tudublin.ie/nitloth/84>.
- [8] A. Aguezzoul, "Third-party logistics selection problem: A literature review on criteria and methods," *Omega (United Kingdom)*, vol. 49. Elsevier Ltd, pp. 69–78, 2014, doi: 10.1016/j.omega.2014.05.009.
- [9] A. Haddach, M. Ammari, L. B. Allal, K. AZAAR, and A. Laglaoui, "How to measure sustainability of uppl chain.," *Int. J. Adv. Res.*, vol. 5, no. 5, pp. 401–418, May 2017, doi: 10.21474/IJAR01/4125.
- [10] M. Varsei, C. Soosay, B. Fahimnia, and J. Sarkis, "Framing sustainability performance of supply chains with multidimensional indicators," *Supply Chain Manag.*, vol. 19, no. 3, pp. 242–257, May 2014, doi: 10.1108/SCM-12-2013-0436.
- [11] E. Chardine-Baumann and V. Botta-Genoulaz, "A framework for sustainable performance assessment of supply chain management practices," *Comput. Ind. Eng.*, vol. 76, pp. 138–147, 2014, doi: 10.1016/j.cie.2014.07.029i.
- [12] P. Evangelista, C. Colicchia, and A. Creazza, "Is environmental sustainability a strategic priority for logistics service providers?" *Journal of Environmental Management*, vol. 198(Pt 1):353-362, August 2019, doi: 10.1016/j.jenvman.2017.04.096
- [13] E. Zhang, X., Valantis Kanellos, N., Plant, "Environmental Sustainability of Logistics Service Providers: a Systematic Literature Review on Indicators for City Logistics, 24th International Symposium on Logistics: Supply Chain Networks vs Platforms: Innovations, Challenges and Opportunities, 2019, pp. 405–413, [Online]. Available: <https://arrow.tudublin.ie/beschspcon>.
- [14] S. G. Azevedo, H. Carvalho, L. M. Ferreira, and J. C. O. Matias, "A proposed framework to assess upstream supply chain sustainability," *Environ. Dev. Sustain.*, vol. 19, no. 6, pp. 2253–2273, Dec. 2017, doi: 10.1007/s10668-016-9853-0.
- [15] P. Taticchi, F. Tonelli, and R. Pasqualino, "Performance measurement of sustainable supply chains: A literature review and a research agenda," *Int. J. Product. Perform. Manag.*, vol. 62, no. 8, pp. 782–804, Oct. 2013, doi: 10.1108/IJPPM-03-2013-0037.
- [16] C. Colicchia, G. Marchet, M. Melacini, and S. Perotti, "Building environmental sustainability: Empirical evidence from Logistics Service Providers," *J. Clean. Prod.*, vol. 59, pp. 197–209, Nov. 2013, doi: 10.1016/j.jclepro.2013.06.057.
- [17] J. P. Schögl, M. M. C. Fritz, and R. J. Baumgartner, "Toward supply chain-wide sustainability assessment: A conceptual framework and an igation method to assess supply chain performance," *J. Clean. Prod.*, vol. 131, pp. 822–835, Sep. 2016, doi: 10.1016/j.jclepro.2016.04.035.
- [18] P. Ahi and C. Searcy, "Assessing sustainability in the supply chain: A triple bottom line approach," *Appl. Math. Model.*, vol. 39, no. 10–11, pp. 2882–2896, 2015, doi: 10.1016/j.apm.2014.10.055.
- [19] S. Santiteerakul, A. Sekhari, A. Bouras, and A. Sopadang, "Sustainability performance measurement framework for supply chain management," *Int. J. Prod. Dev.*, vol. 20, no. 3, pp. 221–238, 2015, doi: 10.1504/IJPD.2015.069325.
- [20] T. Badiezadeh, R. F. Saen, and T. Samavati, "Assessing sustainability of supply chains by double frontier network DEA: A big data approach," *Comput. Oper. Res.*, vol. 98, pp. 284–290, Oct. 2018, doi: 10.1016/j.cor.2017.06.003.
- [21] J. Oršič, B. Rosi, and B. Jereb, "Measuring sustainable performance among logistic service providers in supply chains," *Teh. Vjesn.*, vol. 26, no. 5, pp. 1478–1485, Oct. 2019, doi: 10.17559/TV-20180607112607.
- [22] A. P. K. J. Prabodhika, D. H. H. Niwunhella, and A. Wijayanayake, "Framework for Measuring Sustainability Performance of Logistics Service Providers- A Systematic Review of Literature," in *11th International Conference on Business & Information ICBI, University of Kelaniya, Sri Lanka*, 2020, no. November, pp. 615–626, doi: 10.2139/ssrn.3862225.
- [23] A. P. K. J. Prabodhika, A. Wijayanayake, and D. H. H. Niwunhella, "Measuring Sustainability Performance of Logistics Service Providers using AHP", *International Conference on Industrial Engineering and Operations Management Singapore, March 9-11, 2021*, pp 599-610.

A MILP model to optimize the proportion of production quantities considering the ANP composite performance index

N. T. H. Thalagahage*
Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
Nayomi1013@gmail.com

A. N. Wijayanayake
Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
anni@kln.ac.lk

D. H. H. Niwunhella
Dept. of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
hirunin@kln.ac.lk

Abstract - The apparel industry is considered as one of the most labor-intensive industries where Production Planning and Control (PPC) is considered as an important function, because of its involvement from scheduling each task in the process to the delivery of customer demand. Line planning is a sub-process within PPC, through which the production orders are allocated to production lines according to their setting and due dates of production completion. The decisions that address line planning functions still heavily rely on the expertise of the production planner. When production planners are required to select production lines for the production of a particular type of product, little emphasis has been placed on ways to apportion certain production orders to the most appropriate production system. In this research, a framework is developed using Analytical Network Process (ANP) which is a Multi-Criteria Decision Making (MCDM) method, enabling the incorporation of all the planning criteria in the selection of a production line. The weighted scores obtained by the best alternative production lines are used in a Linear Programming model to optimize the resource allocation in an apparel firm.

Keywords - Analytical Network Method (ANP), apparel production planning, linear programming, multi-criteria decision making (MCDM), production line planning

I. INTRODUCTION

Clothing is the quintessential worldwide industry wherein the world's biggest retailers, marked advertisers, and producers without processing plants are the dominant players. The clothing and material industry area is consistently under steady tension and where rivalry is fierce, there is an opportunity for opponent firms standing by to challenge them. Even though the apparel and textile business may be buyer-focused to fulfill retail procedures and shopper needs, the clothing manufacturing system is at the core of any cut-and-sew activity. The production system, as the center of an assembling undertaking, shapes a huge capital venture for any organization. As clothing organizations face the requests of things to come, capital speculations turn into a genuine budgetary issue.

[1] Discusses capital intensity, energy intensity, and competitive market as the three main factors which make production planning an essential activity in the quest for improvements in operational efficiency. In the apparel industry, production line planning is the process of scheduling and allocating production orders to production lines according to product setting (product is being made in the line) and due dates of production completion. A line plan defines when a style is going to be loaded to the line, how many pieces are to be expected (target) from the line

and when an order is to be completed. Production planning usually assumes a perfect environment in terms of resource availability and process quality. Resource unavailability during the production process will increase production costs and affect inventory levels needed to satisfy customer demand. Production planning is done as part of a hierarchical planning process, where the production plan is cascaded down to a more detailed production schedule.

[2] A production line has the capability to produce a number of different product types. There exists a large number of process constraints from one production system to the other due to the varied capabilities and processing requirements of a given production order. Some of the production orders can be produced on more than one production line and some of the sub-processes require sharing of special tools and machinery. Some products have constraints with regards to the precedence of operations that should be performed for the production while others have similar production conditions that should be scheduled for consecutive production. Switching from one production line to another for the same product style or switching in between different styles within the same production line leads to a reduction in efficiency and it wastes lots of machine and labor production hours of the manufacturing firm. Current practices on scheduling daily production in the production lines are based on the experience of the management. At present, scheduling daily production in the manufacturing process is subjectively based on the manager's experience. With an increasing emphasis on the multiple objectives of on-time shipment, low inventory, and production quality; the management of the plant needs a scheduling tool to improve the production scheduling for better system performance.

To improve the process of line planning, decision-makers need to understand the impacts of the characteristics of apparel production systems and parameters in the manufacturing environment on production system performance which can thus provide insights into the selection decision. However, it is difficult to anticipate the impact of the parameters in the manufacturing environment on production system performance through observation or experimentation because it is costly and time-consuming. In such circumstances, Multi-Criteria Decision Making frameworks can be used because of its ability to explicitly model multiple and possibly conflicting factors.

In this research, a Multi-Criteria Decision Making (MCDM) framework is constructed with the objective of

finding the best suitable production line to minimize the total costs, including the production costs, inventory holding costs, idle time costs and lateness costs. Therefore, this research will focus on finding the solution for on, how to select the best production line for a particular production order through a collaborative decision-making framework and increase the production planning efficiency in the apparel sector in Sri Lanka. The main objectives of the research are to

RO1: Identify the production line selection criteria of an apparel manufacturing firm

RO2: Identify the most suitable MCDM method for the research

RO3: Develop a framework to select the most suitable production line in the apparel sector

II. LITERATURE REVIEW

In this section, the existing achievements of the industry and work by academic scholars in the intersecting fields of the scope of the research are being reviewed. The literature review was done under the topics of capacity planning and line selection approaches, MCDM frameworks used for different research problems in different industries with their pros and cons comprehensive review on ANP method and applications of Linear Programming in production planning.

A. Production planning approaches

This section reviewed the literature which mainly focused on capacity planning and scheduling function in different industries. Those results were used to identify the main criteria and sub-criteria in the ANP framework. [3] Discuss 3 main parameters that are considered to have the most significant effect on the selection of production systems in real-life which are, product complexity, production order size, and operator competence level. [4] Also discusses how the operator's performance is affected in a production line and recommends that it should be taken into consideration in the line planning process. It also investigated flexible flow line problems with sequence-dependent setup times and different Project Management policies to minimize the make span in parallel machines. [5] Mentions that, when scheduling orders in the paper and pulp industry managers have to use a base sequence of grades. Customers place orders for reels of different widths and grades therefore, the lot sequencing approach can be used to verify the earliest available slot for a lot size and hence commit to the due date. Also, he discusses the fact that there's a priority level for different orders based on the logistic model. The maximum priority is given to those that travel by ship since the company has to schedule containers in advance and commit to a given due date. Also, the important costs related to production stability must be taken into account when defining production plans.

[4] Discusses the production family concept. Family set-up time reflects the need to change a tool for each class of styles and even sizes within the style. This set-up time is large compared to the average processing time of the production order. In general, therefore, large batches have the advantage of high machine utilization because the number of setups is small. On the other hand, processing a large batch may delay the processing of an important job

belonging to a different family, resulting in customer dissatisfaction due to tardy deliveries. [2] Discusses a solution to product family setup time, under Group Technology (GT) concept. In the GT approach, some parts of different products which involve similar manufacturing processes are combined in the production process. This method reduces inventories and Work in Progress (WIP). Since workers are producing similar products all the time, throughput time and setup time can be largely reduced. [6] Also, addresses the need for diminishing switch over methodology between production systems, which is a current administration concern. It is referenced that production run length (the number of days a handling line is booked to deliver a similar item type) should be long enough to deliver completed items with predictable quality. Regular item switchovers in the preparing line can bring about quality issues. Be that as it may, a run-length bigger than should be expected can expand the stock level. [6] showed calculations to produce everyday creation plans considering two different goals which are, to limit shipment delays and to limit normal stock levels. The administration fabricating frameworks faces the issue of meeting client conveyance dates while working the system productively. This includes clashing targets. The contention emerges because improvement in one target can be made to the disservice of at least one of different goals. In addition, different creation and quality requirements should be fulfilled.

The literature showed that making a decision based on multi criteria is a considerably complex task. In general, scheduling problems imply that a set of rules should be evaluated and ranked according to different criteria which are conflicting to each other. These facts emphasize the need of a Multi Criteria Decision Making framework to be used in the production planning process. Given a client request, two practical heuristic or successive streamlining calculations are created to produce every day creation plans for two essential destinations: limit shipment delays (pull-in reverse strategy) and limit normal stock levels (push-forward technique). A third heuristic calculation (decrease switch-over method) which depends on the current administration practice is additionally evolved to fill in as a benchmark. Analysis of literature showed that there's a large set of criteria that needs to be considered. Therefore, when selecting the best production line for a given production order, proper balance between each criterion should be considered.

B. Multi criteria decision making methods

Many methodologies were discussed in the literature under the selection process, which involve building alternatives, identifying selection criteria, and evaluating alternatives against the criteria. This approach in selection is developed as MCDM, which has the ability to reveal the complexity of the problem with decisive attributes, to make appropriate trade-offs among conflicting factors, and to recommend well-balanced solutions to different stakeholders [7]. When considering MCDM methods, the criteria interactivity should be concerned since there're several forms of interactions among criteria that might occur in real world problems. According to the classification done by [8] there're distinct philosophies under criteria interactivity. Alternative selection methods fall under the structural dependency which implies the

dominance and dependency relations in the structure of the criteria. The structural dependency is prevalent in AHP, ANP, and hierarchical TOPSIS methods.

C. AHP and ANP methods

AHP technique is one of the multi-criteria decision making methodologies. The AHP is a typical methodology of numerous models dynamic in operations management [9]. A primary preferred position of this technique is to beaten impromptu choices of supervisors which are regularly founded on encounters or emotions. Numerous dynamic issues can't be organized in a hierarchal manner as a result of the connections and conditions between standards. In such cases the structure of the issue ought to be inherent the type of an organization. ANP is the general type of the AHP, and can help in managing conditions and collaborations in complex dynamic issues. Throughout the most recent decade there have been many studies considers that were led utilizing ANP in various ventures for various purposes. This follows a review of such work to recognize the diverse emphasizing points of interest and weaknesses of AHP and ANP strategies. Since ANP is developed from AHP, the two techniques were examined and contrasted with the utilization of ANP strategy for the advancement of production line planning system.

[10] Did an exploration study to give decision support to supervisors concerning the determination issue. The cutting-machine determination rules were controlled by thinking about the related literature, and by counseling the industrial experts. After that, selected criteria weights were determined by fuzzy AHP and ranking cutting machine alternatives by fuzzy MOORA method. The investigation recommended for the most appropriate cutting machine for the firm. [6] Utilized AHP for the arrangement of assembling techniques to client necessities. [11] Analyzes AHP and ANP through a use of key dynamic in an assembling organization. It specifies that numerous choice issues can't be organized progressively when they involve the interaction and dependence of higher level elements in a hierarchy on lower level elements [12]

While the AHP represents a framework with a uni-directional hierarchical AHP relationship, the ANP allows for complex interrelationships among decision levels and attributes. [13] Used ANP method to develop an Evaluation Indices System for product line selection process for ERP. This framework was built to facilitate ERP system of the organization to make decisions on how to organize production rationally to achieve the highest profit and the lowest cost given limited resources. [14] Presented the ANP to explore the relationship among lead time, cost, quality, and service level in a supply chain to select one strategy among a lean, agile or Leagile (i.e., combining lean and, agile) supply chain. [8] also developed an MCDM support for a sustainable supply chain. They used sustainable dimensions of a supply chain and selected the best alternative practice using AHP method. The research then developed the framework to an ANP method and compared the results of each method. The change in final result of each method implies that the earlier AHP model had been an over-simplification of the problem and that the interdependencies of the elements had not been properly and adequately captured by the model. The addition of the network influence of alternatives on criteria

in the model has made the model more comprehensive and realistic, reflecting the relationships among the elements.

[15] Utilized ANP technique to choose the best methodology for reducing risks in a supply chain. Supply risk, process risk, demand risk, and disturbance risks were considered as risk factors in this paper. The ANP is applied to represent the significance of the supply chain risk factor and to assess the appropriate arrangement out of the other options, Total quality administration, Lean, Alignment, Adaptability and Agility. A hybrid MCDM approach is developed by [16] to assess aircraft administration quality in Iran. Fuzzy DEMATEL was applied to decide the level of impact one criteria has on one another and that helped in ranking criteria based on the relationship. ANP network map was developed dependent on the connection map created from Fuzzy DEMATEL examination. Fuzzy ANP approach helped with organizing criteria based on the requirement for development and enabled in a more exact estimation in decision making. In the research [17] ANP strategy was utilized to decide the relationship among the measures for investigating the green building rating framework in Taiwan. DEMATEL and best worst method (BWM) was utilized to build up the system.

D. Linear programming for production planning

When making decisions using MCDM methods, the Decision Maker has to face problems relating to the use of limited resources considering how to decide on which resources would be allocated to obtain the best result, which may relate to profit or cost or both. MCDM methods are characterized by subjectivity, where the framework can be different from person to person and apparel firm to firm. Therefore, a Linear Programming model can be formulated and solutions can be derived to determine the best course of action within the constraints that exist.

Linear Programming is a method of allocating resources optimally. It is one of the most widely used operations research tools to determine optimal resource utilization. Therefore, this research develops a model which consists of the objective function and certain constraints. In past researches, Linear Programming is heavily used in microeconomics and company management such as planning, production, transportation, technology, and other issues.

[18] presents a model to be applied in the consumer goods industry consisting of multiple manufacturers, multiple production lines, and multiple distribution centers which integrates the production and distribution plans. Number of products, number of product groups, number of production lines, number of plants, number of production lines at plant, number of products that can be processed on a line, number of distribution centers, number of periods have been used as the constraints for the model while the capacity of the production line, external demand of the product, time consumed to produce a product, minor setup time of product group, the major setup time of product group, processing cost of the product, minor setup cost of the product, major setup cost of product group, inventory holding cost of the product are used to decide the optimum production methods and distribution means.

[[19] used equipment technology options, timing constraints, lot assignment constraints, the capacity of the equipment in the batch tasks, maximum campaign length in continuous tasks, changeover procedures, setup time,

corresponding due dates, storage and shelf-life constraints, maintenance operations as the constraints in the planning optimization model for the biopharmaceutical industry.

[20] used the data collected from the industries like monthly held or available resources but a company procures resources like fabrics and threads as the production requirement. Monthly available time also can be variable because the number of workers may be increased or decreased as per the production plan. In this paper, the cost minimization along with increasing profit using the same resources used at present is proposed. Using a linear programming method, the optimal, or most efficient, way of using limited resources to achieve the objective of the situation was found out.

III. METHODOLOGY

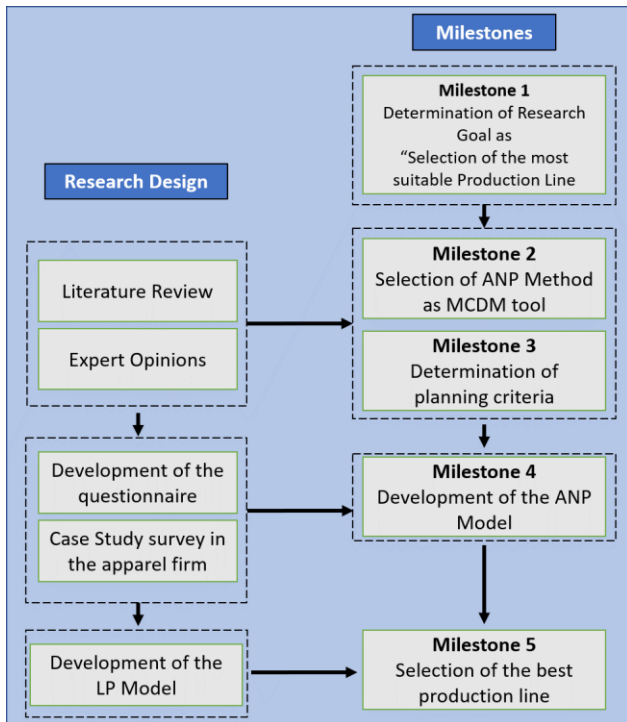


Fig. 1: Flow diagram of the methodology process

The research was started with a literature review to find out the current situation in production planning function in apparel firms, the proposed approaches for the production line selection problem, and to identify the gaps and limitations in the past research. Identification of the research gap led to form the research objectives and then the research questions were formulated.

This study was carried out as a mixed approach study, which is a quantitative study together with qualitative features. This research study provides a quantitative solution for the critical issue in production line selection. It mainly focuses on the optimization of machinery and human resource allocation for production orders. Here, the ANP MCDM technique was used to build the production line selection framework and to select the best production line among the potential alternatives. Qualitative data (production line selection criteria and sub-criteria) were gathered through literature review and interviews with the professionals in the planning function of the study apparel firm. Expert opinions were conducted with professionals

from 5 different apparel firms to identify the line selection criteria and it was used in the ANP conceptual framework.

A. Criteria identification

Table I shows the 5 criteria, 19 sub-criteria which were identified for the production line selection decision. The table also shows the references used for the criteria identification. (LR- Literature Review/ EO- Expert Opinion)

TABLE I. CRITERIA AND SUB CRITERIA OF PRODUCTION LINE SELECTION

Goal: To select the most suitable production line under different criteria	
Criteria	Sub Criteria
(C1) Characteristics of the product	(C1.1.) Standard Minute Value (LR/EO)
	(C1.2.) Labor time (EO)
	(C1.3.) Style efficiency (EO)
	(C1.4.) Supervisory control (EO)
	(C1.5.) Throughput time (LR/EO)
	(C1.6.) Number of operations (LR)
(C2) Characteristics of the Production Order	(C2.1.) Delivery Date (LR/EO)
	(C2.2.) Order Quantity (LR)
	(C2.3.) Size Quantities (EO)
	(C2.4.) PCU Date (EO)
(C3) Characteristics of the Production Line	(C3.1.) Technical Infrastructure (LR)
	(C3.2.) Ability to adopt changeovers (LR)
	(C3.3.) Efficiency of the Production Line (EO)
	(C3.4.) Skills inventory of the Production Line (LR)
	(C3.5.) Availability of the Production Line (EO)
(C4) Technical support	(C4.1.) Infrastructure support by the technical team (EO)
	(C4.2.) Machine service requirements (EO)
(C5) Quality and IE concerns	(C5.1.) Expected quality parameters (EO)
	(C5.2.) Cadre (EO)
Alternatives: Potential production lines – 6 were selected	

B. Development of Linear Programming model

Next objective of this study is to optimize the resources required for the production of the order by considering the relevant constraints. Here, Linear Programming will be used to build the optimization model.

Constraints for the LP model were selected through the interviews conducted with the respondents of the case study organization. The business unit produces products with simple to mid-level complexity. The variety between product types are high, therefore, one of main objectives of the planning process is to minimize the number of changes done to a production line from one product to the other. Unless the objective is achieved, the machine idle time, operator idle time go high due to frequent switching between one product to the other.

- Decision variables

The planner's task is to set a production amount for each production line ranked through the ANP method. Therefore, the decision variables will be whether the

production line is selected, and if selected, what is the amount of production order assigned for each production line.

- Objective function

To maximize the ANP weighted composite performance index through selection of production lines. In here calculated competency score values will be used as coefficient values of the objective function.

Assumptions:

In the formulation it is assumed that,

- Production output is stored in one main warehouse, therefore there are no inventories per individual production line and the previous production output has been transferred to the warehouse at the beginning of the Production Order.
- Raw materials are supplied to all production lines without any delay and limit, as per the production plan and raw materials are consistently provided without any disruption.

Constraints

The following constraints were identified and will be used for the formulation of Linear Programming model.

- Machine hours
- Trimming labor hours
- Sewing labor hours
- Machine set up time
- Machine set up cost
- If x_i is zero then no line is assigned for production
If $x_i > 0$, then y_i is assigned for production
- Total number of production lines should be below the desired number of lines

Development of the model

Indices

- i - i th Production Line
- n - Number of Production lines
- $x_i = q_i / Q_i$ - Proportion of Production allocated to the i th production line (this number is a fraction of the total production quantity of the order)
- y_i - Production Line i , if selected 1, otherwise 0, a binary variable

Parameters

- m_i - ANP Composite performance index of the i^{th} production line
- M - Maximum number of machine hours allowed for order completion on Production End Date (PED)
- m_i - Number of machine hours required for production completion in each i^{th} production line
- T - Maximum number of trimming labor hours allowed for production completion on PED

- t_i - Number of trimming labor hours required for production completion in each i^{th} production line
- S - Maximum number of sewing labor hours allowed for production completion on PED
- S_i - No of sewing labor hours required for product completion in each i^{th} production line
- Q_i - Total production Order quantity
- q_i - Production quantity allocated for i^{th} production line
- N - Desired number of production lines
- $POS_i - POC_i$ - Maximum set up time allowed from each i^{th} production line for the Production Order to start production on Production Start Date (PSD)
- POS_i - Next order Production Start Date
- POC_i - Current order Completion Date
- st_i - Setup time required for each i^{th} prod. line
- SC - Maximum set up cost allowed for the Production Order to start production within the profitable range
- SC_i - Set up cost required for each i^{th} production line

Objective Function:

$$\text{Max } Z = \sum_{i=1}^n c_i x_i$$

Subjected to

$$\sum_{i=1}^n m_i x_i \leq M \quad (1)$$

$$\sum_{i=1}^n t_i x_i \leq T \quad (2)$$

$$\sum_{i=1}^n s_i x_i \leq S \quad (3)$$

$$\sum_{i=1}^n sc_i . y_i \leq SC \quad (4)$$

$$st_i y_i \leq POS_i - POC_i \quad i=1, \dots, n \quad (5)$$

$$x_i \leq y_i \quad (6)$$

$$\sum_{i=1}^n y_i < N \quad \sum_{i=1}^n y_k \leq N \quad (7)$$

$$x_i \geq 0 \text{ and } y_i = (0,1)$$

IV. DATA COLLECTION AND ANALYSIS

Data collection was done through a questionnaire which was designed to feed pairwise comparisons to the matrix form. The survey was conducted as a case study, therefore professionals from production planning department of an apparel manufacturing firm in Sri Lanka were involved.

An instance was created with one of the frequently manufactured styles in the organization. Questionnaires were given to 11 experts in the Planning Department of the organization to determine the relative importance of each sub criteria. The experts were selected based on the years of experience they have in the Planning Department and

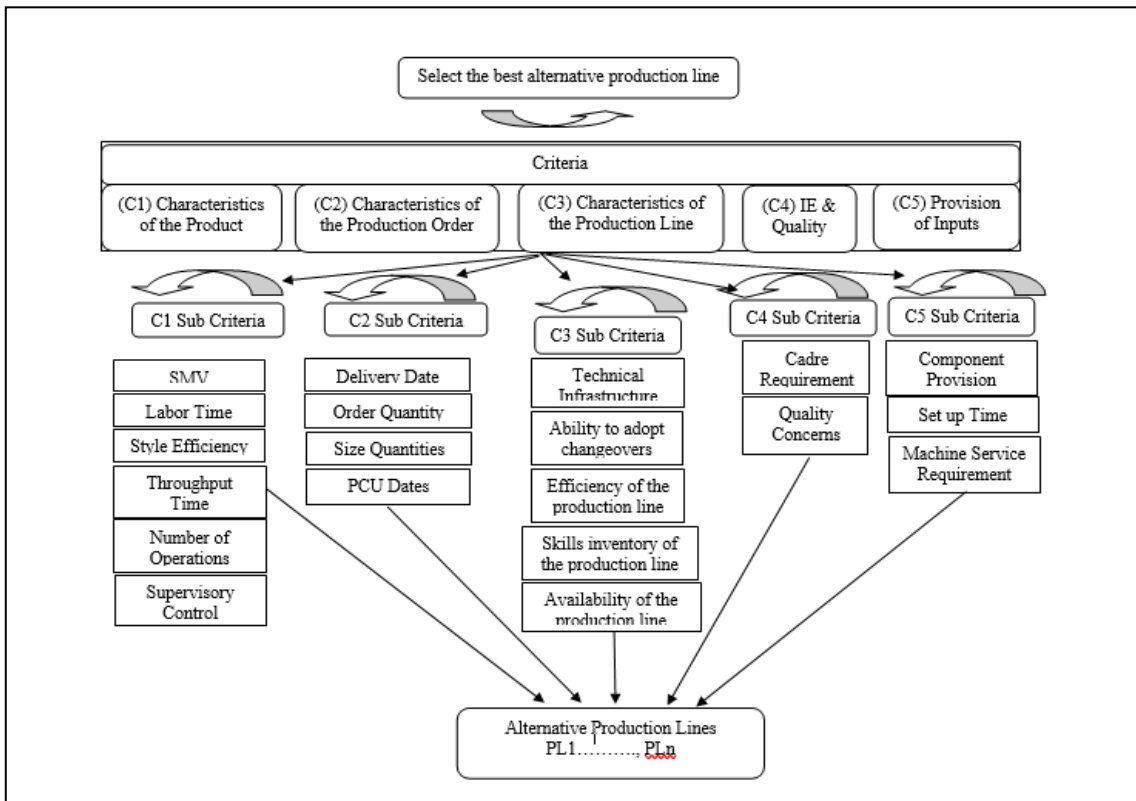


Fig 2: ANP framework

also based on the product type/style specialization. 1 Senior Manager, 2 Assistant Managers, 4 Senior Executives and 4 Executives were selected from the organization. They were instructed to relate the answers to the instance described prior to the questions.

V. DATA ANALYSIS AND DISCUSSION

After the case study was conducted, the ANP framework was developed using SuperDecisions Software. SuperDecisions is a free educational software that implements AHP and ANP methods and was developed by Thomas Saaty’s team who created the method.

Compared to other software tools SuperDecisions is known as a simple, easy to use software package for constructing decision models with dependence and feedback. Moreover, it is an opensource software which was designed to run in many different environments from Windows to Macintosh to Unix systems as Linux.

Multiplication of criteria weight and sub criteria weight were recorded as the final weight for each sub criteria. When calculating the weighted supermatrix in ANP, the pairwise comparisons at the node level must be done based on Saaty’s scale. Comparisons should be done between,

- i. The criteria and the goal
- ii. Criteria with respect to other criteria
- iii. Alternatives with respect to each criterion
- iv. Each cluster

The ANP framework calculates the priorities for each production line and the following result (Table II) was received for the case study.

TABLE II: ANP RANKING

Alternative	Normalized Priority Score	ANP Rank
Production Line 1	0.03587	6
Production Line 2	0.04856	5
Production Line 3	0.22997	2
Production Line 4	0.43481	1
Production Line 5	0.15413	3
Production Line 6	0.09666	4

Weights received from the ANP model were taken as the coefficients for the Mixed integer Linear Programming (MILP) model and it was simulated using MS Excel. Then the model was solved using solver. Result has been recorded in Table III.

TABLE III: RESULTS OF THE MILP MODEL

Alternative	Quantity Assigned	Set up time (Mins)	Set up cost (Rs.)
PL 1	0	0	0
PL 2	0	0	0
PL 3	21877.5	540	120
PL 4	53122.5	720	200
PL 5	0	0	0
PL 6	0	0	0

VI. CONCLUSION

Ranks received can be used to prioritize the production lines for a particular production order. When implemented once, the same framework can be used for similar production orders. Most of the times, the case study organization receives orders from same client with same style and quantity specifications for repeating months. When such occasions arise, only the alternative comparison

has to be performed because pairwise comparison of criteria and sub criteria are same. Then the ANP weights can be applied to the LP model and optimize the alternative production lines.

When totally new styles are received, the whole process should be carried out again along with the pairwise comparisons. However, all the criteria will be covered and visibility to every detail will be made sure, therefore the probability of re-planning is very low.

Addition of Mixed Integer Linear Programming model will help the planner to decide on the quantity that should be manufactured in each alternative production line. This way, the production line which consists of materials, machines and equipment, manpower can be optimized.

VII. FUTURE WORK

The framework was primarily developed for apparel sector industrial setting, but with the extensive study done under production planning in general, this framework covers most of the considerations of other industries as well. Therefore, with slight modifications, this methodology can be applied for any industry in production line selection function.

In order to have a more reliable result, it is suggested that in the future Fuzzy ANP be applied to guide decision making toward a more constructive and consolidated line allocation.

With the usage of Linear Programming, continuous planning model can be developed with product routing and line balancing implications. Furthermore, an algorithm can be developed to solve the ILP model using CPLEX which is much more efficient and accurate.

The model was implemented in only one organization as a case study. Therefore, more scenarios can be used under different contexts and the framework can be further validated.

REFERENCES

- [1] Figueira, G., Oliveira Santos, M., & Almada-Lobo, B. (2013). A hybrid VNS approach for the short-term production planning and scheduling: A case study in the pulp and paper industry. *Computers and Operations Research*, 40(7), 1804–1818. <https://doi.org/10.1016/j.cor.2013.01.015>
- [2] Shin, H. & Leon, V. Jorge. (2004). Scheduling with product family set-up times: An application in TFT LCD manufacturing. *int. j. prod. res.* 15. 4235-4248. [10.1080/00207540410001708461](https://doi.org/10.1080/00207540410001708461).
- [3] Mok, P. Y., Cheung, T. Y., Wong, W. K., Leung, S. Y. S., & Fan, J. T. (2013). Intelligent production planning for complex garment manufacturing. *Journal of Intelligent Manufacturing*, 24(1), 133–145. <https://doi.org/10.1007/s10845-011-0548-y>
- [4] Song, B. & Wong, W. & Fan, J. & Chan, S.. (2006). A recursive operator allocation approach for assembly line-balancing optimization problem with the consideration of operator efficiency. *Computers & Industrial Engineering*. 51. 585-608. [10.1016/j.cie.2006.05.002](https://doi.org/10.1016/j.cie.2006.05.002).
- [5] Figueira, G., Amorim, P., Guimarães, L., Amorim-Lopes, M., Neves-Moreira, F., & Almada-Lobo, B. (2015). A decision support system for the operational production planning and scheduling of an integrated pulp and paper mill. *Computers and Chemical Engineering*, 77, 85–104. <https://doi.org/10.1016/j.compchemeng.2015.03.017>
- [6] Ben Hmida, J., Lee, J., Wang, X., & Boukadi, F. (2014). Production scheduling for continuous manufacturing systems with quality constraints. *Production and Manufacturing Research*, 2(1), 95–111. <https://doi.org/10.1080/21693277.2014.892846>
- [7] Poh, K. L., & Liang, Y. (2017). Multiple-Criteria Decision Support for a Sustainable Supply Chain: Applications to the Fashion Industry. *Informatics*, 4(4), 36. <https://doi.org/10.3390/informatics4040036>
- [8] Gogolcuka & Baykasoglu (2016). An Analysis of DEMATEL Approaches for Criteria Interaction Handling with ANP Expert Systems Application
- [9] Hofmann, E., & Knébel, S. (2013). Alignment of manufacturing strategies to customer requirements using analytical hierarchy process. *Production and Manufacturing Research*, 1(1), 19–43. <https://doi.org/10.1080/21693277.2013.846835>
- [10] Vatanserver, K., and Kazançoğlu, Y., “Integrated usage of fuzzy multi criteria decision making techniques for machine selection problems and an application”, *International Journal of Business and Social Science*, vol. 5, no. 9, pp. 12–24, 2014.
- [11] Görener, A., (2012) Comparing AHP and ANP: An Application of Strategic Decisions Making in a Manufacturing Company *International Journal of Business and Social Science* <https://www.researchgate.net/publication/267857709>
- [12] (Saaty and Özdemir, 2005), The Analytic Hierarchy and Analytic & Analytic Network Processes for the Measurement of Intangible Criteria and for Decision Making
- [13] International Series in Operations Research & Management Science book series (ISOR, volume 78)
- [14] Wei, Jin-Yu & Bi, Ran. (2008). Knowledge management performance evaluation based on ANP. *Int Conf Mach Learn Cybernet.* 1. 257 - 261. [10.1109/ICMLC.2008.4620414](https://doi.org/10.1109/ICMLC.2008.4620414).
- [15] Agarwal, Ashish & Tiwari, Manoj. (2006). Modeling the metrics of lean, agile and leagile supply chain: An ANP-based approach. *European Journal of Operational Research*. 173. 211-225. [10.1016/j.ejor.2004.12.005](https://doi.org/10.1016/j.ejor.2004.12.005).
- [16] Ivanov, Dmitry & Hosseini, Seyedmohsen & Dolgui, Alexandre. (2019). Review of quantitative methods for supply chain resilience analysis. *Transportation Research Part E Logistics and Transportation Review*. 125. 285-307. [10.1016/j.tre.2019.03.001](https://doi.org/10.1016/j.tre.2019.03.001).
- [17] Navid, H. (2017). Evaluating Airline Quality Using Fuzzy DEMATEL and ANP. *Strategic Public Management Journal*. <https://doi.org/10.25069/spmj.351296>
- [18] Liu, P. C. Y., Lo, H. W., & Liou, J. J. H. (2020). A combination of DEMATEL and BWM-based ANP methods for exploring the green building rating system in Taiwan. *Sustainability (Switzerland)*, 12(8), 3216. <https://doi.org/10.3390/SU12083216>
- [19] Bilgen, B. (2010). Application of fuzzy mathematical programming approach to the production allocation and distribution supply chain network problem. *Expert Systems with Applications*, 37(6), 4488–4495. <https://doi.org/10.1016/j.eswa.2009.12.062>
- [20] Vieira, M., Pinto-Varela, T., & Barbosa-Póvoa, A. P. (2019). A model-based decision support framework for the optimisation of production planning in the biopharmaceutical industry. *Computers and Industrial Engineering*, 129(January), 354–367. <https://doi.org/10.1016/j.cie.2019.01.045>
- [21] Woubante, Gera. (2017). The Optimization Problem of Product Mix and Linear Programming Applications: Case Study in the Apparel Industry. *Open Science Journal*. 2. 10.23954/osj.v2i2.853.
- [22] Thalagahage N.T.H., A. Wijayanayake, and D. H. H. Niwunhella, Developing a Multi Criteria Decision Making Framework to Select the Most Suitable Production Line in Apparel Firms :Use of ANP Method *International Conference on Industrial Engineering and Operations Management Singapore, March 9-11, 2021*

Reduce food crop wastage with hyperledger fabric-based food supply chain

Dewmini Premarathna*

Department of Software Engineering
University of Kelaniya, Sri Lanka
dewminic@kln.ac.lk

Abstract - Food is the utmost important thing for every living being. The quality and safety of food has become a crucial factor in the food industry. Most of the customers tend to pay more attention to food safety and seek to get food from verifiable resources. To improve this trustworthiness Distributed Ledger Technology (DLT) - based Food Supply Chain (FSC) plays a vital role because of its traceability. There are multiple actors involved throughout the journey of FSC and with the high visibility of data in DLT, everyone can ensure trust. The transparency of data itself is a reason for some to opt-out because some of their private data can be exposed to others. Hyperledger Fabric (HF) based FSC can address that matter as it supports permissioned network solutions. Though there are a lot of solutions available in a similar kind of approach, whether the crops take their journey throughout the FSC without any wastage, is still questionable. This study focuses on reducing wastage of food crops as they take a long journey in their raw state and possible hazards are high. It discusses farmers' behavior based on the Sri Lankan context and how it accompanies food crop wastage. Further, this paper ruminates the other possible crop wastage that can take place in FSC and how to eliminate it with the proper involvement of knowledgeable and authorized parties. Then, the study explores how all the parties can collaboratively join the FSC based on HF so that everyone can benefit. Finally, it concludes on how such design is effectively contributing to reducing food crop wastage in Sri Lanka (SL).

Keywords - crop, farmer, food, hyperledger fabric, lock chain

I. INTRODUCTION

The food industry is a globally widespread industry where agriculture plays a main role. Most of the humans' food needs are met by crops like vegetables, fruits, potatoes, and grains [1]. From production to consumption, crops go through many different stages in the FSC. FSC has become an extremely complex and long process as it involves many actors like farmers, transporters, wholesalers, retailers, end consumers, and many more on different scales [2]. As the participation of different parties increases, many issues such as lack of communication, transparency, accuracy, mutual trust, and traceability arise. There is a growing interest in the technology world to design systems based on DLT for FSC to address such issues [3] – [7].

DLT is involved in a distributed style with no central governance for the data [8]. All the technologies engaged with DLT work in a similar manner which ensures tamper-proof data. Blockchain is one of the DLT types and has great potential in various industries with the availability of vast technology platforms. As it supports immutability and traceability, those who join the blockchain network can expect high trust. It is an ideal solution for any e case where

trust is required as a key feature [8]. So, it plays a vital role in FSC in which contributors can have mutual trust. Food consumers can ensure their food safety and nutritional value, and anyone can know the path food has taken from its origin to destination.

Blockchain's high data transparency has benefited some industries, but some are reluctant to get involved. Because some people are afraid that their information will be passed on to the competitors [5]. In that case, it would be better if they could interact with the FSC while keeping their data confidential and HF could do the same. HF provides an authorized way for each actor to join the network, and the literature explores it the most.

Although we can improve the privacy and trustworthiness of FSC with HF-based supply chains, there are many stages throughout the chain where food wastage can take place. Due to the high complexity associated with FSC, measuring food loss through the chain is a difficult process. Farmers can be known as the heart of the FSC, and the initial point of food wastage starts there. The issues faced by farmers and their behaviors have a great impact on their harvest which may indirectly cause food loss. The post-harvest period is another main stage where food wastage happens. With the proper involvement of actors in the DLT based FSC, food wastage and loss can be minimized. Further, this study is based on the SL context in discussing the problems associated with food cultivation, transportation, and marketing. There can be information that can suit the global context. But, in developing countries, most of the issues are very different due to factors like poverty, improper education, and cultural challenges [9] [10].

The SL government has taken various measures and legislation to prevent food loss and wastage [9]. But, the problem lies in the challenges they face in implementing them. If everyone meets at FSC, to get together and go on this journey, they can solve a lot of problems in a way that is profitable for everyone involved. To facilitate that, HF-based FSC is a great solution because there we can make the participants work more credibly in a way that is transparent.

Therefore, this paper provides a solution on how to use the HF-based FSC to minimize food crops wastage in the process of supplying food in Sri Lanka from the beginning of growing crops to the consumer's home. To incorporate that, the rest of the paper is arranged as follows in a sectioned order; literature review, solution overview based on the literature, results and discussion, and conclusion.

II. LITERATURE REVIEW

DLT systems have a distributed database shared among each node in the network with no central authority

[8]. Among the notable popular DLT platforms, blockchain is the frequently associated one, having gained its popularity with the cryptocurrency bitcoin introduced by Satoshi Nakamoto [11]. Because bitcoin works so reliably in transactions where fraud is almost impossible, many people are curious about how to use the underlying technology when developing applications where trust plays a crucial role [12].

A. Blockchain

Blockchain is a decentralized, distributed ledger that facilitates recording and tracking of transactions. Like any other database, blockchain also stores data. The key difference of blockchain with a typical database is that it stores transactions in a data structure called blocks instead of a predefined table structure or file format. If it is described from its simplest form, the block consists of transactions data, nonce, the hash of the previous block as well as the hash of itself [6] [8] [11] [13]. So, each block in the chain is cryptographically linked together. This type of chain is called a ledger and there are multiple copies of the same ledger stored in a distributed peer-to-peer network. When a new transaction occurs all the peers work upon an inbuilt consensus protocol, and it is approved upon 51% agreement of the peers. When the network grows, it becomes more robust and it is almost impossible to tamper with other data although someone spends more time and applies computational power more than 51% [8] [14]. So, the data immutability offers a high tamperproof nature and can rely more trust on the data.

Although blockchain is often identified to have two or three main types, it could use the four types below: public, private, consortium, and hybrid [13] [15] [16]. In a public blockchain, no restrictions are applied, anyone can engage with transactions, running nodes, and mining. A private blockchain is a closed network and is operated by certain members only, but everyone has visibility over the data within the network. Consortium blockchain differs from other blockchain types. It is not only a closed network but also members have accessibility over a permission manner. Hybrid blockchain is a combination of both public and private blockchains. With the evolution of blockchain, many platforms have emerged. Among them, Ethereum and Hyperledger frameworks are popular at enterprise level [16].

Ethereum's main network is a public blockchain and it can be deployed as a private network also [17]. But it cannot control its data visibility in a permissioned way across the participants i.e. someone to be visible and someone to not. In such cases, HF is the ideal solution provided by the Hyperledger platform which is an open source community that provides frameworks, libraries, and tools for enterprise blockchain solutions [18]. HF is the most active and mature project in Hyperledger projects backed by Linux Foundation with a strong development community. HF is more suitable for multi-stakeholder businesses due to its unique features associated with the identity of the participants, data privacy, confidentiality, and performance than other platforms [19].

B. Hyperledger Fabric (HF)

HF is a DLT platform that has a pluggable modular architecture [20]. Therefore, it can be easily adapted to satisfy most of the business's needs. Also, its permissioned

nature allows businesses to operate in a more confidential manner which is a major concern enterprises pay attention to, related to their data privacy [21]. With its latest version 2.x, HF provides a new architecture for the transactions, called execute-order-validate. Over the earlier approach, the order-execute new approach has a huge impact on the performance [20] [22]. It first executes the transaction using chaincode. According to the endorsement policy when enough peers agree upon the correctness of the transaction, transactions are ordered with consensus protocol which is also pluggable. Ordered transactions are validated by peers against the specified endorsement policy. So, it eliminates non-determinism rather than being limited to domain-specific languages, it allows writing smart contracts in standard programming languages such as Java, Go, and Node.js [20].

When creating a HF network understanding the functions of its components and how they work collaboratively to form a secure network is very important. Although there are many components involved with HF, ten identified key points are discussed here which describe HF architecture in detail.

- 1) *Ordering Service*: Every HF network consists of at least one ordering service. When clients send endorsed transactions to the ordering nodes, they come to a consensus on the order of the transaction by executing a consensus algorithm. The consensus algorithm is pluggable and Raft is the recommended one. After the transaction order is confirmed, they form them into blocks and send those to the endorsing peers which are pre-defined in the endorsement policy. The earlier versions of HF used the Kafka and Solo consensus algorithm to order the transaction, and it is deprecated with the HF version 2.x whereas Kafka makes additional overhead to the system administration and Solo is for test only and consists only of a single ordering node [23].
- 2) *Peers*: Peers are the fundamental element in the HF network. They are owned and maintained by a relevant organization. They host the ledger and smart contracts specific to them. Peers can hold multiple smart contracts (when packaged it is called chaincode) and multiple ledgers. Peers validate and commit the transaction blocks into the ledger [24]. So peers basically read, write operations to the ledger by running chaincode [25].
- 3) *Applications*: Applications can execute chaincode hosted in peers by connecting them. When they send the proposal to the peers to read or write data, peers check its correctness by endorsing it, and a response is sent to the application. Then the application sends a request to ordering nodes to order the transaction. Ordered transactions blocks are sent to the peers and peers update their ledger and the application receives the ledger update event [24].
- 4) *Organization*: An organization is a logical entity in a HF network and is also known as a member. The organization is defined by the root certificate specific for the organization and is stored in Certificate Authority (CA). The organization represents a physical separation of their Certificate Authority (CA), Membership Service Provider (MSP), and peers. Each

organization added to the channel at the channel creation time is a part of a consortium which is again a collection of organizations. The HF network can consist of one or many organizations[25] [26].

- 5) *Certificate authority (CA)*: CA is responsible for giving certificates to components of its organization. CA issues key-value pairs (public and private key) and can be used to prove the identity components like peers [25], [27].
- 6) *Membership Service Provider (MSP)*: MSP is a directory that includes certificates and private keys for each identity that is generated by the CA. So MSP contains a list of files and directories representing those permissioned identities to the fabric network. It allows organizations to manage their members under MSP. When organizations perform different business modules in multiple channels they can have multiple MSPs by properly naming them [27].
- 7) *Channel*: Channel is like a sub-network within the HF network that allows organizations to communicate privately[25]. The organizations are invited to join their peers to the channel for validating the transaction on the channel. Organizations can only access the data of the channels they have joined, the channels they have not joined are restricted [28]. Within a channel also there can be one or more private data collection (PDC). This allows the organization to expose certain data to all channel members while keeping some part confidential within another subset of members in the channel [29]. It minimizes the number of channel creations with extended privacy.
- 8) *Smart contracts and chaincode*: Smart contract contains the business logic and executes upon ledger to read and write data[30]. The related smart contracts are packaged before they are deployed to the blockchain network. Packaged smart contracts are known as chaincode. Chaincode is installed on peers and invoked by the client application through HF Software Development Kit (SDK). When a smart contract generates a transaction, the endorsement policy associated with the chaincode defines which members should approve the transaction against its validity. When the transaction is signed by a required number of members, the transaction is indicated as valid or invalid. Then that information is added to the distributed ledger. But only valid transactions are updated to the world state which represents the current state of the latest transactions. To be able to execute efficient queries word state supports state databases, level DB, and CouchDB [31].
- 9) *Ledger*: In HF network ledger can be identified into two pieces i.e. the blockchain immutable ledger with all history of transactions distributed in the peers and world state with the current value [31].
- 10) *Policies*: Policies make HF distinguished from other networks. Unlike the other blockchain platforms, HF cannot use any node to validate the transaction. “Who is going to do what” can be clearly defined as a set of rules [32]. Policies containing those rules are stored in a configuration file. So access to the resources within

the network is restricted and only permissioned ones can access them. Policies can be defined before the network is launched or at the time the network is functioning. So those are implemented in different levels of the HF network. Policies in the system channel configuration govern the consensus used by the ordering service and which members are allowed to create new channels. Policies in the application channel configuration govern which members are allowed to join the channel and which members can approve the chaincode to be committed to the channel. Policies defined in Access Control Lists (ACLs) refer to policies defined in an application channel configuration and extended to control additional resources. Smart contract endorsement policies define how many peers need to execute and validate a transaction against a given smart contract [33]. So the default policies in the HF at its network first stage can be overridden at any time according to the business requirement and provide governance over the privacy.

As a summation to all these, since the HF network is highly configurable it allows any component to act in a pluggable manner. Also, with proper endorsement policies, data can be shared within the network on a need-to-know basis [19]. As of this modular architecture, anyone can design their network in high-performance, scalable, and confidential ways [34]. More importantly, in the HF network trust is not dependent only on its immutable ledger., Since the well-identified participants are engaging all the time, more trust can be ensured and any fraud can be easily identified which prevents them from tampering the data.

III. ISSUES IN SRI LANKAN FOOD SUPPLY CHAINS

FSC in SL is mainly built on farmers, wholesalers, transporters, retailers, and end consumers. Normally wholesalers buy crops from the farmers. Then wholesalers use transporters or their own transportation to receive goods. Retailers buy crops from wholesalers or directly from farmers and then go to the end consumer. Most of the time this supply chain takes place based on everyone’s knowledge and experience. The educated people are not involved in this supply chain process and hence a lot of misbehavior can occur in various stages of FSC [35].

Fig.1 displays the exact problem of the current supply chain. Red lines indicate how intermediate parties directly involve farmers and it will indirectly affect the synchronized supply chain process. Green lines display the ideal flow and still isolated educated resources, and regulatory bodies are not involved.

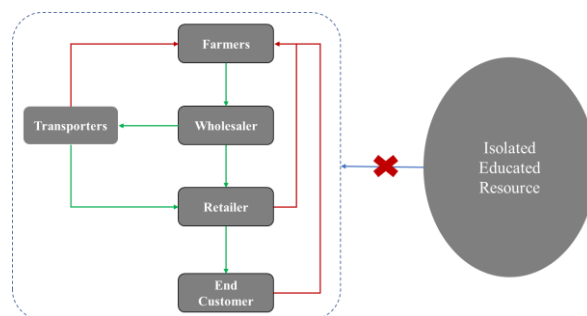


Fig. 1. Sri Lanka supply chain in high-level

In SL most of the farmers cultivate any crop according to the current trend and they do not foresee the future demand. Many of them count on facts like how the fellow farmers made profit during the past and tend to cultivate the same crop. During the harvest season, it will result in the same product being so abundant in the market and demand gets lower. According to the equilibrium theory in economics, when the price is high, the supply increases and it lowers the demand and ends up with a low price. The same theory is applicable here and farmers get low profit and their motivation to sell their harvest is also lowered [36]. With this disappointment, they sometimes destroy their harvest. Sometimes farmers tend to commit suicide due to debt [37]. Not only does it cause huge food wastage in the country, but that causes economic loss too. Further in parallel to growing crops, supplying fertilizers, insecticides, pesticides and herbicides are also important. If it is not received on time farmers suffer low harvest. Although they receive those, there can be a lack of knowledge on how to use them properly. Such activities often engage with the help of oral knowledge. Although there are many regulatory bodies established in SL to help farmers, because of the lack of communication, this knowledge transfer does not properly happen. All of these factors contribute to a lower yield than what they are able to obtain. So, the wholesalers will not be able to fulfill the required demand. In this situation, wholesalers will search for alternative solutions and will end up finding low-quality crops. Farmers also suffer from less ROI (Return on Investment).

As per the study, most of the food supply chains in SL have no proper methodology, and instead, it is a kind of ad-hoc process [35]. Many problems in the process take place post-harvest [36] [38]–[43]. Especially when loading and unloading harvest there is no defined process to check how the quality of such work is carried out. Overloading the sacks of crops and sometimes throwing the sacks into the vehicle without properly stocking, usually occur. This entire food transportation is not properly monitored and regulated. So much food loss and wastage happens during transportation [36] [38] [40] [43]. This causes not only food wastage but also food safety is at risk. Raw food such as vegetables and fruits are perishable, and the shelf life is severely reduced. Customers have to buy poor-quality food with lower nutritional value. Sometimes customers are even tempted to throw them away once they bring them home. When this happens in many homes, there is a huge food wastage in the country. It is a pity that when a significant number of people in the country are starving, they are not able to utilize the product for other reasons. So, there is a need for a supply chain eco system to minimize the food (mainly crops) wastage by improving the quality of it. The solution for this issue should come as a global solution and should involve each minor party who is directly or indirectly involved with the FSC. Also, there should be strong technological solutions where transparency, trustworthiness, immutability, and privacy are major concerns [3] [5].

IV. SOLUTION OVERVIEW

This solution is guided by the Design Science Research (DSR) methodology to take the various decisions over the designed artifact, HF based FSC which is used by the context of farmers, wholesalers, transporters, retailers,

and regulators. When considering the relevance of the solution some of the technical barriers were identified in between context and the design and those were overcome based on the output obtained from the literature review. So the following solution demonstrates in detail how farmers, wholesalers, transporters, retailers, and regulators are successfully joined to the HF upon an invitation from the network initiator. Later it presents how more parties like knowledgeable persons, fertilizer, or chemical suppliers are also included in this FSC.

The diagram in Fig.2 explains the basic structure of the HF-based blockchain network for the food crops supply chain. Six organizations are identified as main contributors, and channels are identified based on the data privacy requirement on the organizations. Five main applications are identified to support end-users to interact with the network. Four smart contracts are deployed to support storing private data separately and one contract is used to handle common queries required for all the nodes. Seven separate ledgers are used to maintain private data and it is bound with peers connected with the channels. When the number of components increases, complexity will be added to the design. But once the network is consistent there is no development complexity as HF provides pluggable modules in a configurable manner.

A. Organizations

Followings are the main organizations in this design.

- Farmer – Grows the crops.
- Wholesaler 1 – Buys crops from the farmer.
- Wholesaler 2 – Buys crops from the farmer.
- Transporter – Transport crops between locations
- Retailer – Buys crops from wholesalers.
- Regulator – Controls the quality of other organizations and provides quality certificates.

B. Chaincode

The followings explain details of the chain codes.

- Price and private data negotiation between farmer and wholesaler.
- Price and private data negotiation between wholesaler and transporter
- Price and private data negotiation between wholesaler and Retailer.
- Crop transferring

1) *First smart contract (S1)*: Following common functions will be available for seven organizations on smart contract 1.

a) *Farmer*: Access to the following functions.

- Record Crop
- Query Demand
- Update Demand
- Update Price

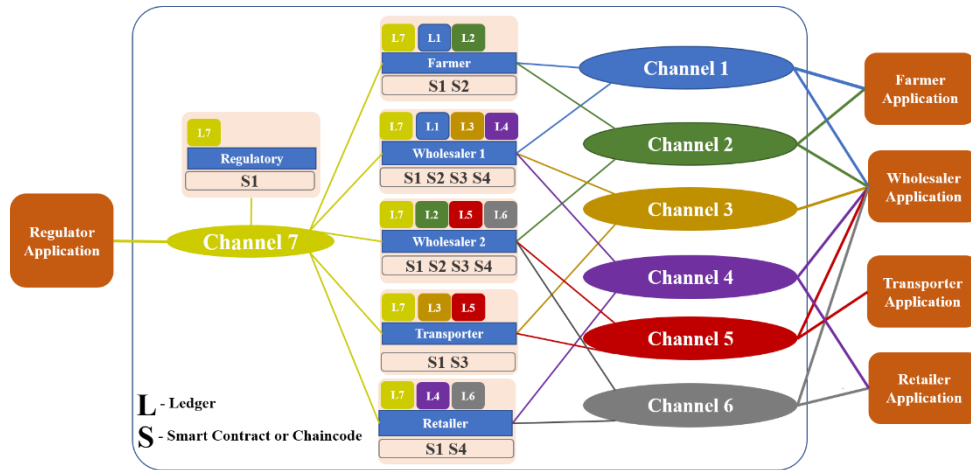


Fig. 2. HF based blockchain network for crops supply chain

- b) *Wholesaler*: Access to the following functions.
 - Record Demand
 - Buy Crop
 - Update Demand
 - Query Crop
 - Pay Transport
- c) *Transporter*: Access to the following functions
 - Query Transport
 - Record Transport
 - Update Transport
 - Pickup Demand
- d) *Retailer*: Access to the following functions.
 - Buy Crops
 - Mark Purchase
 - Query Crops
- e) *Regulator*: Access to the following functions.
 - Query Farmers
 - Query Wholesalers
 - Query Transporters
 - Query All Crops

- 2) *Second smart contract (S2)*: Mark price and private data between farmer and wholesaler.
- 3) *Third smart contract (S3)*: Mark price and private data between wholesaler and transporter.
- 4) *Fourth smart contract (S4)*: Mark price and private data between wholesaler and retailer.

C. Channels

- Channel 1 – Price and private data negotiation between farmer and wholesaler 1.
- Channel 2 – Price and private data negotiation between farmer and wholesaler 2.

- Channel 3 – Price and private data negotiation between wholesaler 1 and transporter.
- Channel 4 – Price and private data negotiation between wholesaler 1 and retailer.
- Channel 5 – Price and private data negotiation between wholesaler 2 and transporter.
- Channel 6 – Price and private data negotiation between wholesaler 2 and retailer.
- Channel 7 – Crop transfer.

D. Applications

- Farmer application – Farmer will use this to execute the function defined above.
- Wholesaler application – Wholesaler will use to execute functions defined above.
- Transporter application – Transporter will use to execute functions defined above.
- Regulator application – Regulator will use to execute functions defined in above
- Retailer application – Retailer will use to execute function defined in above.

E. Ledgers

There are six ledgers defined in the solution and peers in each organization will use ledgers as follows.

- 1) *L1*: This ledger maintains data private to the farmer and wholesaler 1
- 2) *L2*: This ledger maintains data private to farmer and wholesaler 2
- 3) *L3*: This ledger maintains data private to wholesaler 1 and transporter
- 4) *L4*: This ledger maintains data private to wholesaler 1 and retailer
- 5) *L5*: This ledger maintains data private to wholesaler 2 and transporter
- 6) *L6*: This ledger maintains data private to wholesaler 2 and retailer.

F. Example Message Sequence

- 1) *Step:* Farmer records crop (the available harvested stock) using farmer app.
- 2) *Step:* Farmer updates different prices for wholesaler 1 and wholesaler 2.
- 3) *Step:* Wholesaler 1 buy the crop from the farmer
- 4) *Step:* System update crop as bought, price and make available for transporters.
- 5) *Step:* Transporter picks the demand and delivers the crop into location.
- 6) *Step:* Transporter updates the demand and marks the price in the ledger.
- 7) *Step:* Regulator is doing continuous monitoring and removes Transporter or wholesaler from the network if any misbehavior has taken place.

G. Implementation

This design involves other key components in HF such as MSP, Order Service, Policies, CA, etc. For implementation of this solution, a network configuration file (NCF) is created after identifying the network initiator. In this design, it is the regulator. NCF contains channel configurations, policies, chaincode details, peer details, etc. So once successfully implemented the solution needs to be tested properly to identify performance and functional errors. Based on the performance test result implementation can be considered to fine-tune the number of channels and maintain private data collections which minimize the overhead of channel administration and provide commit and query private data without having to create separate channels.

This solution is to involve more organizations who are indirectly involved with this supply chain. Such as fertilizer suppliers, agriculture instructors, field officers (Fig.3 below). Further, this solution can be enhanced by implementing a loyalty platform where organizations can give feedback to each other, and with the transparency and immutability of HF each can get quality of works and goods provided.

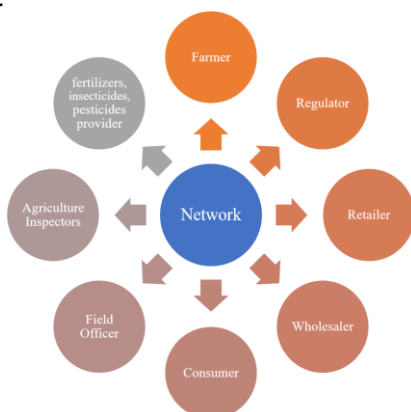


Fig. 3. Network contributors (Organizations) to enhance the solution

Another important factor is to enable an alerting system so that everyone will get alerts on various stages of the supply chain and it will help organizations give prompt responses rather than waiting till the last minute.

V. RESULT AND DISCUSSION

This section discusses how the above solution benefited to reduce food wastage and improve food quality in SL. This is a very powerful solution to align all the ad-hoc processes, entities in a very disciplined manner and build consumer trustworthiness while it supports saving food and reducing hunger.

A. Responsibility of the regulator

DLT platforms are primarily based on the feature of no central governance. HF also adopts that feature while allowing privacy over the data among the group of parties. All data visibility can be retained, only within certain groups, if desired. So, it would be good to have some common party who can track the activities of others to some extent. The regulator is the one who can perform such monitoring over the entire network. If the regulator is a representative of the government, it can be ensured whether the rules defined by the government are followed in this FSC. They can identify if something goes wrong within a channel or a PDC and take action against it. For example, if a transporter uploads a nice photo of transporting food even if it was improperly packed it can be notified by the farmer or the person accepting the transportation. Then they can add their comment or complaint to the system. Then the regulator can view those and warn the transporter. If it continuously happens from the same party, the regulator can remove them from the network. Regulators can also issue certifications to the involved parties throughout the chain which will be visible to others. It provides an extra layer of trust other than the built-in trust we can get with HF. When actors of the FSC are getting certified, it will cause the system to be more robust and food quality also improves, also, reduces food wastage.

B. Farmer to wholesaler transaction

Farmers are the most valuable entity in this chain, the starting point would always be farmers. Once they join this network they have two options to start farming. The first option is, they can choose their own crop to grow and update the network with the same information. Another option is they can check the demand in the network and start growing crops by accepting the demand.

In option one, once a farmer marks that he is starting cultivation it will be visible to all parties in the network. So that agricultural instructors and fertilizers, insecticides, pesticides providers are notified by the network and they can start to provide required knowledge and supply required items on time till the farmer finishes growing crops. So, these organizations also need to update the network with information including images and details of provided fertilizers, etc. So this information is visible to everyone and no one can alter them due to immutability in blockchain technology. Regulators can do continuous monitoring to maintain the quality of the cultivating process. By working with the HF network in this way it can reduce a lot of issues farmers are facing in traditional harvesting which results in minimizing food crops wastage and improving the quality of the same.

In the second option, everything is similar, other than the farmer starts cultivation once the farmer accepts the already created demand by the wholesaler. Since all the farmers and wholesalers are connected with the network

and because of the permissioned feature in HF-based blockchains, wholesalers will be able to provide demand to farmers in private channels at an agreed price. In this scenario, there is no visibility for other farmers and wholesalers about this transaction, but regulators, agricultural instructors, and other raw material providers will have visibility about the transaction but not the prices and private data. That is the capability of a well-designed HF-based FSC network. Anyhow in both options wholesalers will have well-managed high-quality crops to provide transporters and then retailers.

The diagram in Fig.4 explains a summary of what was discussed in option one above.

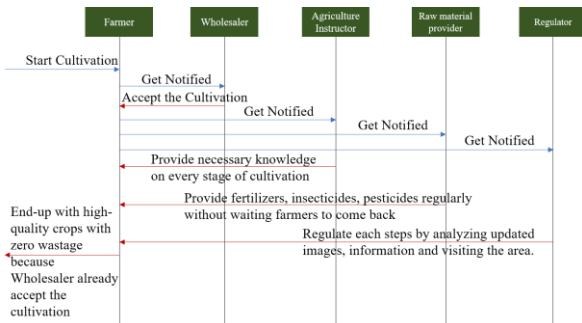


Fig. 4. Farmer to wholesaler transaction flow works in HF network

As per the diagram in Fig.4 farmers will have quality crops to sell to wholesalers who have already agreed on the price. But there can be two problematic situations where farmers will not be able to provide mentioned crops and wholesalers will not be able to buy the agreed crops. So, in both scenarios, the network can help to resolve this issue. Wholesalers can open the crop for another wholesaler who already joined the network. Farmers also can search for other farmers who are having similar crops and seeking to sell. So, the network itself connects each other to fulfill everyone's requirements. On the other hand, regulators can either remove such organizations from the network or give warnings if any repeated problematic situations occur.

C. Wholesaler to transporter transaction

This section discusses how the wholesaler and transporter are involved with the network to maintain the same quality maintained by farmers and wholesalers to reduce food crops wastage. Once the wholesaler is ready with the crops, the system is updated with the same information, and transporters are alerted. In this case, the wholesaler will have a choice to update a particular transporter in a private channel or visible the transaction to the entire transporter network. But in any case, the regulator is notified with transaction information except prices and private data. Then the most important part is how the transporter packs, loads and unloads the crops. Transportation plays a crucial role in maintaining crop quality and freshness as much as possible. So, regulators need to play a vital role here because transportation needs to be closely monitored. So, the transporter's responsibility is to update the network with how they pack the crops and load the crops into vehicles. In this case images and videos, evidence is mandatory to update the system with geolocation tags. The regulator's responsibility is to remove transporters who are not following standards or not

providing evidence to the system. Because of the immutability of HF-based networks, this information cannot be altered and that will build trust among the network members. On the other hand, transporters try to do their best to maintain the quality of transportation, otherwise, the organization's reputation gets damaged since it will be visible to the other parties on the network (transparency). Once transportation quality is maintained food crops' quality will not be damaged till it is provided to the wholesaler's storage location or retailer and food wastage will be minimum when considering traditional food transportations where sacks are not properly packed and loaded while in transition. The diagram in Fig.5 demonstrates the summary of this.

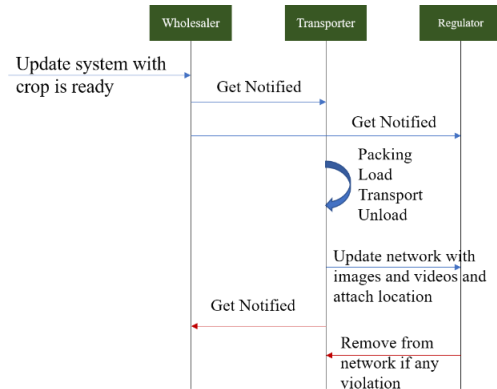


Fig. 5. Wholesaler to transporter transaction flow works in HF network.

D. Transporter to retailer and consumer transaction

Though the network controls food wastage up to transportation, there can be various reasons that food gets wasted due to various reasons such as poor storage and poor maintenance from the retailer end. If the Retailer did not receive the crops in good condition, they can update the network with status which will notify other members in the network. Because of that transparency, transporters will be careful on handling crops. Retailers need to update the network with how they keep crops in the market and these updates need to be monitored by regulators to identify unhealthy processes to reduce food wastage and increase consumer satisfaction. They also can remove retailers from the network if they are not doing a good job. The diagram in Fig.6 demonstrates the summary discussed above.

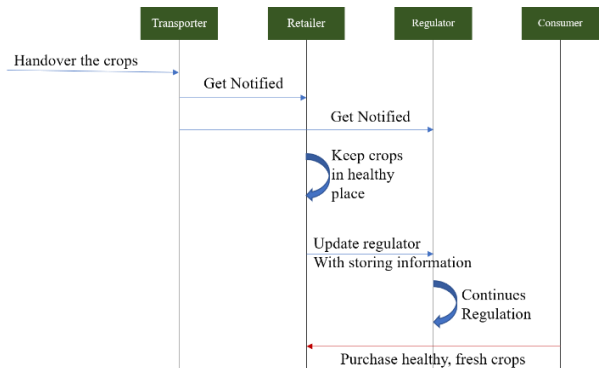


Fig. 6. Transporter to retailer and consumer transaction flow works in HF network.

E. Practical challenges

As crops are grown seasonally, there can be significant time gaps between supply and demand, which can lead to changes in market demand, making it difficult to enter into long-term contracts, and so on. As the number of farmers and wholesalers connected to the system increases, those factors may change further. Farmers and wholesalers can then identify their behavior patterns as they mature from the system and adjust their trade. Such factors can be further evaluated once the system is used in the particular context.

Another challenge that might be faced when a technical solution to a problem is introduced to the non-technical people is, how far they will accept that. The protagonist here is the farmer who may deviate from that technical acceptance. Farmers rely more on traditional methods especially in developing countries [44]. So, the farmers are provided a simple mobile application with a user-friendly interface while hiding the technical complexity. Although, in the initial stage there can be little denials, once the farmers or other participants identify how far they can get benefitted, the same will attract them towards this solution. Especially when it comes to farmers, very few people in countries like SL ever think of becoming farmers because of the uncertainties associated with it. Most people want to do a professional job. But nowadays those of the younger generation are the ones who use smartphones. Therefore, when technology is involved in traditional farming, they may also be interested in it. There may be some similarities, but the behavior of farmers can vary according to their country. Therefore, a country-based survey can be conducted for the technical recognition of non-technical individuals and the results can be used to improve the solution.

The practical implementation of this research can be further evaluated using qualitative and quantitative methods and enhance the HF based FSC towards the maximum reduction of food wastage in SL. Eventually, consumers can be satisfied with quality crops which came through a process where transparency, trustworthiness, and immutability played a major role.

VI. CONCLUSION

There is a lot of research and implementation based on how to use DLT in FSC. Though this has great value, still organizations are having a low tendency to join with such chain networks. This is mainly due to transparency where all the network members will have visibility on each one's data. But HF is playing a vital role to break that concept where organizations can make private channels to hide data when they need privacy. At first glance, the HF architecture looks complex as a lot of components are associated with it. Once the components are properly identified, we can easily handle an HF network the way we want. According to this study, various parties can join the HF-based FSC and it presents how to actively contribute to minimizing food crop wastage while maintaining the privacy they want. Though there are a lot of discussions on DTL-based FSC none of them have focused on reducing food crops wastage while keeping data privacy in each party. Throughout this research, we focused on how everyone can contribute to reducing food crops wastage on FSC after analyzing the current ad-hoc process in SL, how the crops come from farmers to end consumers. Also, literature on HF

technology is well supported to resolve practical problems that arise while implementing such FSC. Not only in SL if any country is having such an ad-hoc supply chain, from farmers to consumers, they can use this analysis to support the development of HF-based FSC to reduce food wastage and finally reduce world hunger.

REFERENCES

- [1] C. Bennett, "PLANTS AS FOOD."
- [2] FAO, "The future of food and agriculture – Trends and challenges. Rome.," 2017.
- [3] Iftikhar, X. Cui, M. Hassan, and W. Afzal, "Application of Blockchain and Internet of Things to Ensure Tamper-Proof Data Availability for Food Safety," *Journal of Food Quality*, vol. 2020, 2020, doi: 10.1155/2020/5385207.
- [4] P. Gonczol, P. Katsikouli, L. Herskind, and N. Dragoni, "Blockchain Implementations and Use Cases for Supply Chains-A Survey," *IEEE Access*, vol. 8, pp. 11856–11871, 2020, doi: 10.1109/ACCESS.2020.2964880.
- [5] P. Gonczol, P. Katsikouli, L. Herskind, and N. Dragoni, "Blockchain Implementations and Use Cases for Supply Chains-A Survey," *IEEE Access*, vol. 8, pp. 11856–11871, 2020, doi: 10.1109/ACCESS.2020.2964880.
- [6] M. Kumarathunga, "Improving Farmers' Participation in Agri Supply Chains with Blockchain and Smart Contracts," in *2020 7th International Conference on Software Defined Systems, SDS 2020*, Apr. 2020, pp. 139–144. doi: 10.1109/SDS49854.2020.9143913.
- [7] "Global Trade & Supply Chains | IOTA." <https://www.iota.org/solutions/global-trade-and-supply-chains> (accessed Jul. 01, 2021).
- [8] M. Rauchs et al., "DISTRIBUTED LEDGER TECHNOLOGY SYSTEMS A Conceptual Framework," 2018. [Online]. Available: <https://ssrn.com/abstract=3230013>
- [9] M. Aheeyar et al., "Food waste in Sri Lanka: an analysis of the applicable urban regulatory framework Seond draft-Pending FAO feedback 2," 2020.
- [10] R. Anjum, "Design of mobile phone services to support farmers in developing countries," 2015.
- [11] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System." [Online]. Available: www.bitcoin.org
- [12] "Technology Innovation Management Review," 2021.
- [13] H.-N. Dai, M. Imran, and N. Haider, "Blockchain-Enabled Internet of Medical Things to Combat COVID-19," *IEEE Internet of Things Magazine*, vol. 3, no. 3, pp. 52–57, Oct. 2020, doi: 10.1109/iotm.0001.2000087.
- [14] S. Shalaby, A. A. Abdellatif, A. Al-Ali, A. Mohamed, A. Erbad, and M. Guizani, "Performance Evaluation of Hyperledger Fabric," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies, ICIoT 2020*, Feb. 2020, pp. 608–613. doi: 10.1109/ICIoT48696.2020.9089614.
- [15] Zhong, H. Wu, L. Ding, H. Luo, Y. Luo, and X. Pan, "Hyperledger fabric-based consortium blockchain for construction quality information management," *Frontiers of Engineering Management*, vol. 7, no. 4, pp. 512–527, Dec. 2020, doi: 10.1007/s42524-020-0128-y.
- [16] L. Wu, W. Lu, and F. Xue, "Construction inspection information management with consortium blockchain 'BIM Square': Blockchain and i-Core-enabled Multi-stakeholder Building Information Modelling Platform for Construction Logistics and Supply Chain Management in Hong Kong View project From point cloud to building and city information model (BIM/CIM): A study of architectonic grammar optimization View project SEE PROFILE Construction Inspection Information Management with Consortium Blockchain," Springer. [Online]. Available: <https://www.researchgate.net/publication/346463411>
- [17] "Enterprise on Ethereum mainnet | ethereum.org." <https://ethereum.org/en/enterprise/> (accessed Jul. 12, 2021).
- [18] "Hyperledger – Open Source Blockchain Technologies." <https://www.hyperledger.org/> (accessed Jul. 12, 2021).
- [19] "Welcome Hyperledger Fabric 2.0: Enterprise DLT for Production – Hyperledger." <https://www.hyperledger.org/blog/2020/01/30/welcome-hyperledger-fabric-2-0-enterprise-dlt-for-production> (accessed Jul. 12, 2021).
- [20] "Introduction — hyperledger-fabricdocs main documentation." <https://hyperledger-fabric.readthedocs.io/en/latest/whatis.html> (accessed Jul. 12, 2021).

- [21] Ma, X. Kong, Q. Lan, and Z. Zhou, "The privacy protection mechanism of Hyperledger Fabric and its application in supply chain finance," *Cybersecurity*, vol. 2, no. 1, Dec. 2019, doi: 10.1186/s42400-019-0022-2.
- [22] Androulaki et al., "Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains," *Proceedings of the 13th EuroSys Conference, EuroSys 2018*, vol. 2018-January, Apr. 2018, doi: 10.1145/3190508.3190538.
- [23] "The Ordering Service — hyperledger-fabricdocs master documentation." https://hyperledger-fabric.readthedocs.io/en/release-2.2/orderer/ordering_service.html (accessed Jul. 12, 2021).
- [24] "Peers — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/peers/peers.html> (accessed Jul. 12, 2021).
- [25] "Glossary — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/glossary.html> (accessed Jul. 12, 2021).
- [26] "Adding an Org to a Channel — hyperledger-fabricdocs master documentation." https://hyperledger-fabric.readthedocs.io/en/release-2.2/channel_update_tutorial.html (accessed Jul. 12, 2021).
- [27] "Membership Service Provider (MSP) — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/membership/membership.html> (accessed Jul. 12, 2021).
- [28] "Channels — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/channels.html> (accessed Jul. 12, 2021).
- [29] "Private data — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/private-data/private-data.html> (accessed Jul. 12, 2021).
- [30] "Smart Contracts and Chaincode — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/smartcontract/smartcontract.html> (accessed Jul. 12, 2021).
- [31] "Ledger — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/ledger/ledger.html> (accessed Jul. 12, 2021).
- [32] "Policies — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/policies/policies.html> (accessed Jul. 12, 2021).
- [33] "Endorsement policies — hyperledger-fabricdocs master documentation." <https://hyperledger-fabric.readthedocs.io/en/release-2.2/endorsement-policies.html> (accessed Jul. 12, 2021).
- [34] "What's new in Hyperledger Fabric v2.x — hyperledger-fabricdocs main documentation." <https://hyperledger-fabric.readthedocs.io/en/latest/whatsnew.html> (accessed Jul. 12, 2021).
- [35] M. Perera, S. S. Kodithuwakku, and J. Weerahewa, "Analysis of Vegetable Supply Chains of Supermarkets in Sri Lanka," *Sri Lankan Journal of Agricultural Economics*, vol. 6, no. 1, p. 67, Aug. 2011, doi: 10.4038/sjae.v6i1.3471.
- [36] M. Perera, S. S. Kodithuwakku, and J. Weerahewa, "Analysis of Vegetable Supply Chains of Supermarkets in Sri Lanka," *Sri Lankan Journal of Agricultural Economics*, vol. 6, no. 1, p. 67, Aug. 2011, doi: 10.4038/sjae.v6i1.3471.
- [37] N. Booth, M. Briscoe, and R. Powell, "Suicide in the farming community: Methods used and contact with health services," *Occupational and Environmental Medicine*, vol. 57, no. 9, pp. 642–644, 2000, doi: 10.1136/oem.57.9.642.
- [38] M. Aheeyar et al., "Food waste in Sri Lanka: an analysis of the applicable urban regulatory framework Seond draft-Pending FAO feedback 2," 2020.
- [39] F. Omar and M. Z. MatJafri, "Principles, methodologies and technologies of fresh fruit quality assurance," *Quality Assurance and Safety of Crops and Foods*, vol. 5, no. 3, pp. 257–271, Sep. 2013, doi: 10.3920/QAS2012.0175.
- [40] M. Reitemeier, M. Aheeyar, and P. Drechsel, "Perceptions of food waste reduction in sri lanka's commercial capital, Colombo," *Sustainability (Switzerland)*, vol. 13, no. 2, pp. 1–16, Jan. 2021, doi: 10.3390/su13020838.
- [41] O. P. Chauhan, S. Lakshmi, A. K. Pandey, N. Ravi, N. Gopalan, and R. K. Sharma, "Non-destructive Quality Monitoring of Fresh Fruits and Vegetables," *Defence Life Science Journal*, vol. 2, no. 2, p. 103, May 2017, doi: 10.14429/dlsj.2.11379.
- [42] J. Munasinghe, A. de Silva, G. Weerasinghe, A. Gunaratne, and H. Corke, "Food safety in Sri Lanka: Problems and solutions," *Quality Assurance and Safety of Crops and Foods*, vol. 7, no. 1, Wageningen Academic Publishers, pp. 37–44, 2014. doi: 10.3920/QAS2014.x007.
- [43] J. L. Mangal and A. Dhyani, "Post Harvest Technology of Fruits and Vegetables"
- [44] S. Kariuki, "Factors determining adoption of new agricultural technology by smallholder farmers in developing countries." [Online]. Available: <https://www.researchgate.net/publication/303073456>.

Application of Game Theory on financial benefits and employee satisfaction: Case study of a state bank of Sri Lanka

D. D. G. T. Jayasekara*
Department of Mathematics
Faculty of Engineering
University of Moratuwa, Sri Lanka
trevince.jayasekara@gmail.com

A. N. Wijayanayake
Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
anni@kln.ac.lk

A. R. Dissanayake
Department of Mathematics
Faculty of Engineering
University of Moratuwa, Sri Lanka
mayake@gmail.com

Abstract - The principal agent problem revolves around the competing interest between shareholders and the employees. The organization focus is on maximizing shareholder wealth, while employees try to obtain the maximum benefits for themselves. As per the motivational theories, people have different types of needs. Therefore, management should focus on a wide range of factors to motivate the employees to work to their full potential in the interest of the organization. The study focuses on both employee and the management of a state bank. The organization is always eager to minimize the cost and maximize the profit. Game Theory was used to provide a mathematical framework for understanding the optimal outcome and what the tradeoffs are to achieve that outcome. The objective is to find the right balance between financial gains and employee satisfaction. To fulfill that objective, one needs to evaluate the benefits given to employees, the effectiveness of those benefits on employees and finally recommend an effective benefits allocation mix to the organization, which will address both employee and the top management of the bank.

Keywords - *employee satisfaction, Game Theory, optimization*

I. INTRODUCTION

In any business organization, there are two parties. The main party is the stake holders or the management, the second party is the employees or the workers. Management always looks at the business by their perspective, which is to maximize their profit and be the market leader. The employees desire also same which is to be the leader of the market and elevate their workplace brand at the top. But their main target is to upgrade their financial stability.

Therefore, the employee's perspective their ambition is to elevate the earnings, if the management fulfill their targets, then the employees are eventually motivated to work effectively and efficiently. If the earnings are increased, it will be a cost to the organization and it will affect to the profit of the organization as well. Therefore, the organizations are always focusing on the variable earnings than the fixed income such as salaries of the employees, to elevate if the targets are achieved. Then the organization can survive in the market easily.

Many business entities as well as state banks, the allowances provided are not effective and attractive to employees. From an employee perspective, it is not properly allocated. Therefore, most of the employees are working to get their salary and attend to other additional jobs to fulfill their financial requirements. If

this scenario continues, it is difficult to cater to the customers to fulfill their satisfaction because employees are not focusing on customer expectations but intend to achieve their personal targets in financial benefits. But most of the private institutions and banks have recognized and resolved their employees' non-salary benefits and allowances by allocating the funds effectively. Therefore, the employees of private banks are willing to give the maximum output to the organization and get the maximum benefits from the employer. Eventually, compare to state banks the growth rate and the services are higher in these banks or institutions [1]. Because of the government security and the deposits of the government institutions are hold by the state banks. Therefore, the brand value and the profitability are high in these institutions [2]. To achieve the targets, state banks need to motivate the employees to fulfill the required expectations. In Maslow's hierarchy of needs, a theory of motivation, states that five categories of human needs dictate an individual's behavior. Those needs are physiological needs, safety needs, love and belonging needs, esteem needs, and self-actualization needs. The theory explains what is important to fulfill the needs in each level of a human being. According to the theory, each banker has already achieved the first need out of five. That is physiological needs. Therefore, the bankers are always focusing on the second stage which is safety needs. At this stage, Maslow clearly mentioned that emotional security, financial security (e.g. employment, social welfare), law and order, freedom from fear, social stability, property, health and wellbeing are to be satisfied [3]. Therefore, many employees in state banks are considering the safety needs or financial security in this stage.

Therefore, the financial benefits should be rescheduled according to the set goals and requirements of the bank while fulfilling the present requirements of the employees for their hard work as a reward system. Therefore, it is needed to observe that the allowances of the state banks, provided to elevate the effectiveness and the quality of the work. Hence, it is a suitable time for state banks to revise their allowances for employees and to introduce new allowances to satisfy the employees and get the maximum output of them to increase the profit of the bank while having a good balance between minimum cost and maximum benefit [4]. The requirements of the employer and the employee are contradictory to one another. Because if the employer allocates more money for the employee, the profitability

will be decreased and if the employees are not satisfied with the payments for their hard work, then they are not motivated to work more and that will have an adverse effect on bank performance.

Therefore, the authors assume that this could be a game between two players who try to maximize their profit or financial benefits while the opponent try to minimize the loss or cost to the bank. This problem can be defined as a game between the employer and the employee. If an employee tries to maximize its profits by limiting the allowances and financial benefits to its employees, the limiting amount would be the maximum gain to the employer as well as minimum loss to the employee [5].

Therefore, it can be defined as game between employer and employee. Application of Game Theory would be the appropriate technique to solve this issue and this game can be defined as Zero-sum game between those two. Here the authors carefully assume that the loss of one party is similar to the gain of the other party. The authors have recognized the critical allowances to upgrade the profitability and employee satisfaction in the organization and introduced the best model by using Game Theory to find the effectiveness of these benefits.

II. LITERATURE REVIEW

The article “Burnout and customer satisfaction” discussed about the service provider’s dissatisfaction should be taken in to consideration for the success of the organization. This is because it connects to the most important outcome for the organization which is customer satisfaction. Considering the results of this research, shows there is a positive correlation between service provider’s service and customer satisfaction [6]. There should also be empowerment in the employees who serve the customers.

The empowerment can be done in many steps. Providing a high salary is one method but there are many other ways than increasing salary. The organization can provide training programs. The empowerment of the employees can mitigate the problems coming in day-to-day activities. Also, by satisfying the employees, lead them to serve their customers pleasantly. Before implementing the empowerment programmes, the organization should look at how the employees are satisfied. This article explains that the over empowerment of employees also affects badly in treating customers because if more power comes to the employees that they may reject the customers and they treat some selected customers only [7].

The game theory was used to find the corruption survey. The theory is performed an ineffective manner to find the best solution and make the decision on which the corrupt people react. This is based on the bribery commission and the company. They tried to find the corrupted people by giving some questions and collecting the responses. It is a simple model which is presented that bribery might be the dominant strategy. This is the same approach as a prisoner’s dilemma type of situation. In game theory, it is difficult to predict the winning party, but this has taken various parameters like legal remedies. This paper then reviews the principal general equilibrium effects and concludes that they are negatively effect on economic development [8].

The article described how the private and the public transportation systems mitigated their risk factor by using the game theory. This game also has the options for the private sector to make the decision. The public sector has introduced some strategies. By considering those strategies and the payoff values, it can be seen that the authors had made an assumption that the priority will be given to player 2 and as a result of that, the payoff values have been taken according to the consideration of the private sector. Therefore, in this game, it can be seen that the negotiation can be made, and the real values must be presented to another party [9].

The author has introduced a mathematical model to look at the teams and select players. The author used the game theory for those findings. He has taken every player’s salary or allowance for each league as payoff values and found the most suitable player for each club. Also, it is mentioned that the authors have made some realistic assumptions to address the limitations in the practical world to incorporate them to the objective in the model [10].

III. MODEL DEVELOPMENT

The model development of game theory in the proposed system is on the satisfaction of the employees for the allowances providing the employer get the maximum gain for the financial benefits given to employees for a minimum cost. To fulfill the objective, it is being prepared a set of questions for the employees, and evaluated their preference in Likert scale. The questions have been made referring to the non-salary benefit circulars and that would directly affect the reliability of the research. As player 1, the banks will introduce many allowances to satisfy the employees. But only the main benefits have been taken as player 1 strategies, because other benefits are claimed by some groups. Satisfaction is a mental process, but in this case, the employee should scientifically argue with themselves to find the best allowance mix that they should utilize.

It has been taken only the allowances which have been allocated to the officer grades and above employees. The minor staff has been omitted as they get only the OT allowances. The allowances allocated to officer grades are directly affected to the bank’s profitability. They are medical allowance, difficult station, key holding, disturbance and cash loading are some of them considered in this research. Those allowances are considered as different strategies (Str) proposed by the player or management of the bank.

TABLE I: PAYOFF TABLE FOR PLAYER VS OPPONENT

		Opponent		
		Str 1	Str 2	Str 3
Player	Str 01	a_{11}	a_{12}	a_{13}
	Str 02	a_{21}	a_{22}	a_{23}
	Str 03	a_{31}	a_{32}	a_{33}
	Str 04	a_{41}	a_{42}	a_{43}

The summary of these can be incorporated into a cross-tabulation table as shown above. The main assumption in this methodology is that the game between the employees and employers is considered as a zero-sum game. The benefit that player 1 gets is equal to the loss of the opponent. In another way that the benefit that the management of the bank or the employer earn is equivalent to the loss of the employee. The value is measured in a Likert scale and weights are given according to that.

Assumed that there are no saddle points in this game. Therefore, in the long-run decision was mixed and used the mixed strategy to find the ultimate optimal value of the game.

Assume that the probability of the strategies are $P_1, P_2, P_3, \dots, P_n$

Where $\sum P_i = 1$

Assume that the optimal value of the game is V

Therefore, the above problem could be developed as a linear programming problem as follows,

Obj Max V

$$a_{11} P_1 + a_{21} P_2 + a_{31} P_3 + a_{41} P_4 + a_{51} P_5 \geq V$$

$$a_{12} P_1 + a_{22} P_2 + a_{32} P_3 + a_{42} P_4 + a_{52} P_5 \geq V$$

$$a_{13} P_1 + a_{23} P_2 + a_{33} P_3 + a_{43} P_4 + a_{53} P_5 \geq V$$

$$a_{14} P_1 + a_{24} P_2 + a_{34} P_3 + a_{44} P_4 + a_{54} P_5 \geq V$$

$$a_{15} P_1 + a_{25} P_2 + a_{35} P_3 + a_{45} P_4 + a_{55} P_5 \geq V$$

$$P_1 + P_2 + P_3 + P_4 + P_5 = 1,$$

$$P_i \geq 0$$

V value can be found in Linear Programming. Therefore, values for P_1, P_2, P_3, P_4 & P_5 can be found values.

TABLE II: PAYOFF VALUE FOR DIFFERENT STRATEGIES

	Strongly disagree	Disagree	Neither agree nor	Agree	Strongly agree	Min
Medical Allowance (P1)	123	166	306	112	210	112
Difficult Station (P2)	102	214	156	368	125	102
Key Holding (P3)	24	60	309	632	315	24
Disturbance (P4)	91	176	189	296	310	91
Cash loading (P5)	37	118	186	568	390	37
Max	123	214	309	632	390	

IV. DATA ANALYSIS AND PRESENTATION

The proposed model was validated from the data collected from the employees and employers of a state bank. According to the output values obtained after

running the model in MS Excel Solver using actual data as well as simulated data, it can be seen that, there is no pure solution for this game as expected. Therefore, it has been taken the “Minimax” and “Maximin” principles to fulfill the objectives.

$$\text{Raw Min, Value} = \{112, 102, 24, 91, 37\}$$

$$\text{Max} = 112$$

$$\text{Column Max, Value} = \{123, 214, 309, 632, 90\}$$

$$\text{Out of that the Min} = 123$$

Therefore,

$$\text{Maximin Value} \neq \text{Minimax Value}$$

It indicates that there is no saddle point, and the value of the game lies between 112 and 123. To find the effective way to allocate funds among these benefits, it is needed to assign the probabilities for that. Therefore, Linear Programming technique has been used to solve the problem.

Expected Payoff for Player 01, When Player 02 or opponent choose, Strongly Disagree (SD), Disagree (D), neither agree nor disagree (AD), Agree (A), Strongly Agree (SA) under different strategies as shown in Table II.

Objective Functions: Max V

$$123 P_1 + 102 P_2 + 24 P_3 + 91 P_4 + 37 P_5 \geq V \quad (1)$$

$$166 P_1 + 214 P_2 + 60 P_3 + 176 P_4 + 118 P_5 \geq V \quad (2)$$

$$306 P_1 + 156 P_2 + 309 P_3 + 189 P_4 + 186 P_5 \geq V \quad (3)$$

$$112 P_1 + 368 P_2 + 632 P_3 + 296 P_4 + 568 P_5 \geq V \quad (4)$$

$$210 P_1 + 125 P_2 + 315 P_3 + 310 P_4 + 390 P_5 \geq V \quad (5)$$

Assume that the minimum value of game = V

And, $V > 0$

$$\text{Subject to, } P_1 + P_2 + P_3 + P_4 + P_5 = 1$$

Expected pay-off equations in the model,

1. Expected Payoff for Player 01, When Player 02 Choose, Strongly Disagree (SD)
2. Expected Payoff for Player 01, When Player 02 Choose, Disagree (D)
3. Expected Payoff for Player 01, When Player 02 Choose, neither agree nor disagree (AD)
4. Expected Payoff for Player 01, When Player 02 Choose, Agree (A),
5. Expected Payoff for Player 01, When Player 02 Choose, Strongly Agree (SA),

The above linear programming problem was solved using MS Excel Solver and obtained the following output table.

By considering the above MS Excel Solver output spread sheet, the value of the game is 122.166. Further, the probabilities of the decision 1 or the strategy 1 is 96.03% and decision 2 is 3.97% in the long run to achieve the maximum benefit of the game to player 1 or to the management of the bank.

Variable Cells						
Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$B\$2	variables p1	0.960288809	0	0	42.1754386	21
\$C\$2	variables p2	0.039711191	0	0	21	20.03703704
\$D\$2	variables p3	0	-52.07220217	0	52.07220217	1E+30
\$E\$2	variables p4	0	-15.62454874	0	15.62454874	1E+30
\$F\$2	variables p5	0	-44.90974729	0	44.90974729	1E+30
\$G\$2	variables V	122.166065	0	1	1E+30	1

Constraints						
Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
\$G\$12	V	1	122.166065	1	1E+30	1
\$G\$7	s.to V	122.166065	-0.924187726	0	11	60.91346154
\$G\$8	V	167.9061372	0	0	45.7400722	1E+30
\$G\$9	V	300.0433213	0	0	177.8772563	1E+30
\$G\$10	V	122.166065	-0.075812274	0	266	11
\$G\$11	V	206.6245487	0	0	84.45848375	1E+30

Fig. 1. MS excel solver output spread sheet

Ninety six percent of the total allowance should be spent on employee medical scheme and 3.97% to difficult station or for servicing in remote less privileged area. This would satisfy the employees while getting the best benefit to the organization. $P_1=96.03\%$ and P_2 is 3.97 % and all other values of P_1 are equal to zero.

V. CONCLUSION AND RECOMMENDATIONS

According to the P_1 value obtained, it is evident that the medical allowance is the most important allowance among the others. The value of P_1 is 96.03%. This is because most of the employees are utilizing this allowance. The effectiveness is very high. Medical allowances are given to the employees as well as to their families. Most of the employees are satisfied and happy to receive the medical allowances as it covers the medical expenses of the whole family. This will be a big benefit to the employees from their savings. Therefore, according to the management of the bank and employee's perspectives, this may fulfill both player's and the opponent expectations.

According to the output table, the value of P_2 is 3.97%. Therefore, it is a prudent decision to allocate 3.97% of the allowances to the difficult station or working for rural areas or outstations allowances. The obtained output result for P_2 reflects the actual scenario. Since this research was carried during the epidemic lockdown period, most of the employees are not willing to go to outstation areas and work, because of the uncertainty of the lockdowns of the country. Therefore, if the bank increases this allowance that would be useful for both employees and the banks, especially during the lockdown period. This will fulfil the requirements of the bank to give a similar service to the customers in out station while rewarding the few employees still wish to render their services in outstations.

It is evident that P_3 to P_5 values, are equal to 0% out of the total allowance which is not popular among the bank employees. Therefore, by referring to the value it can be justified as this type of allowance may disappoint some employees. For an example strategy 3, allowance could be utilized only one employee at once. Then the others cannot utilize this allowance as only one person is allocated. However, this is an additional allowance

which every employee prefers to get. However only one or maximum two employees will be appointed, and the rest of the members cannot claim that allowance. Therefore, majorities of employees were not satisfied with the P_3 or 3rd strategy that the bank offers. Therefore, by introducing this strategy the satisfaction of the majorities of the employees will be very less

The strategy 4 allowance, P_4 value is also 0%, as there is not much benefit to the majorities of employees. This disturbance allowances rewards the employees who report to work by 6: 30am. Most of the male employees in the bank are between 30-50 years of age who are having school-age children. This segment of employees prefers to report to work after dropping their kids to school. However, a few male employees are willing to come to work early morning to get the benefit of disturbance allowance. Because they like to come in the early morning, fulfill their daily target easily and return home early in the evening to engage in some extra earnings through some other external sources. Further, around 65% of the bank employees are females. Due to the issues related to domestic and family affairs, most of the female employees preferred to wok from 8am and they will not be benefited by having the 4th allowance. Further most of the ladies uses office transport services which arrives to their banks at 8am. This is another reason for the unpopularity of this allowance. Therefore, the bank should seriously reconsider this allowance and review it to make this allowance a very effective and worthy one.

Further the strategy 5 or the value of P_5 is also 0% out of total allowance. Which indicates that the popularity of this allowance is also not significant. This reflects the reality in the practical world. If needed, most the officers can load a small amount of cash and they can frequently go out for loading purpose and claim this allowance many times a day. Then it is a meaningless and additional cost and a burden to the bank. The Game theory application will provide us the best value to allocate many funds on a fair basis, which brings benefits to the player as well as to the opponent. This allowance is a very common allowance in the banking industry and all the employees including minor staff can claim this type of allowance. In addition, the officers, clerical grade employees, and the minor staff members are eligible for this allowance due to the less risk. The riskiest part and the responsibility of this cash loading activity is borne by the security department and the transport department. In addition, this allowance can be claimed by both male and female employees, as this transaction is done during office hours, and anyone can attend for this task without doing overtime work. However, the researchers noted that this allowance is not that popular among staff grade employees as only assigned people can claim these allowances and only a limited number of employees privileged to get the benefit but not all staff grade employees.

However, considering the bank's perspective as well as the majorities of the employees, the state bank should revise their employee benefits to enhance the satisfaction on their staff grade employees by introducing appropriate financial benefits and allowances through various attractive schemes which brings benefit to both employer and employees.

Therefore, the suggested model can be employed to bring right balance between financial benefits and employee satisfaction not only to the above-mentioned state bank but also to other private banks and other organizations.

REFERENCES

- [1] Brand Finance . (2020). Sri Lanka 100 2020 ranking. Retrieved 09 03, 2020, from <https://brandirectory.com/rankings/sri-lanka/table>
- [2] Hari Creations (Pvt) Ltd, 2020. "Top 100 most valuable Sri Lankan brands 2020". [online] available at: <http://www.newswire.lk/2020/05/09/top-100-most-valuable-sri-lankanbrands-2020/>
- [3] McLeod, D. S. (2020, 12 29). "Simply Psychology" . Retrieved 04 23, 2021, from Maslow's Hierarchy of Needs: <https://www.simplypsychology.org/maslow.html>
- [4] Sharp Graphic House (Pvt) Ltd Romualdas Ginevičius, A. K., 2008. "Application of game theory for duopoly market analysis". *Journal of Business Economics and Management*, III (9), pp. 214-216.
- [5] Holler, M. J., 2001. "Classical Game Theory and the Autonomously Rational Player. Classical Game Theory and the Autonomously Rational Player, 1(1), pp. 2 - 8.
- [6] Hallowell, R., 1996. "Southwest Airlines. A case study linking employee needs, satisfaction and organizational capabilities to competitive advantage", *Human Resource Management*, 35(4), pp. 525-529.
- [7] Yagil, D., 2006. "Burnout and customer satisfaction. The relationship of service provider power motivation, empowerment and burnout to customer satisfaction", *International Journal of Service Industry Management*, III(17), pp. 260 - 266.
- [8] Macare, J., 1982. "Underdevelopment and the Economics of Corruption, A game theory approach", 10(8), pp. 677-687.
- [9] Medda, F., 2007. A game theory approach for the allocation of risks in transport. *International Journal of Project Management*, 25(2), pp. 215-218.
- [10] Solberg, K. K. H. & H. A., 2010. "Financial Profit in Football. The Financial Crisis in European Football - a Game Theoretic Approach", *European Sport Management Quarterly*. , 10(5), pp. 553 – 565.

A novel approach for weather prediction for agriculture in Sri Lanka using Machine Learning techniques

J. S. A. N. W. Premachandra*

Department of Computer Science

Gen. Sir John Kotelawala Defence University, Sri Lanka
nishadiwasana833@gmail.com

P. P. N. V. Kumara

Department of Computer Science

Gen. Sir John Kotelawala Defence University, Sri Lanka
nandana@kdu.ac.lk

Abstract - Climate variability in recent years has critically affected the usual aspects of human lives, where the agriculture sector can be considered as one of the most vulnerable. Sri Lanka is also facing these climate changes over the past few decades. It has resulted in rainfall pattern changes where the expected rain may not occur during the expected time and amount. The mismatch between the rainfall pattern and traditional seasonal cultivation schedule has critically affected the agricultural sustainability. Even with the current technological advancements, weather prediction is one of the most technically and scientifically challenging tasks. This paper presents a novel machine learning-based approach for predicting rainfall for precision agriculture in Sri Lanka and it can be recognized as the first attempt to validate machine learning models to predict the weather in Sri Lankan context for precision agriculture. By analyzing the nature of the weather in Sri Lanka, the relationship of weather attributes with agriculture, availability, and accessibility, seven attributes are selected including rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context. For the prediction model, cross-validated data are trained and tested with four machine learning algorithms: Multiple Linear Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest. Currently, Support Vector Machine, K-Nearest Neighbors models have achieved accuracies of 88.57%, 88.66%. Random Forest has been recognized as the best-fitted model with 89.16% accuracy. The results depict a significant accuracy in this novel approach for Sri Lankan weather prediction.

Keywords - data mining, machine learning, precision agriculture, weather prediction

I. INTRODUCTION

As a developing country in the Asian region, Sri Lanka has an economy based on agriculture while emphasizing that the agricultural sector is playing a significant role in the country's current development in both economic and social aspects[1]. Climate variability in recent years has critically affected the usual aspects of human lives, where the agricultural sector can be considered as one of the most vulnerable. According to the report "Sustainable Sri Lanka 2030 Vision and Strategic Path", as a developing country, Sri Lanka is facing potential agricultural risks due to unpredictable climatic changes[2]. The discrepancy between the rainfall pattern and traditional seasonal cultivation due to climatic variabilities is the main problem which is addressed in this research. According to the Intergovernmental Panel of Climate Change, among the sub-regions of Asia, South Asia is facing the most vulnerable climate changes. Sri Lanka has also been facing these changes during the past few decades, which has been

resulted in rainfall pattern changes where the expected rain may not occur during the expected time as well as with the expected amount and intensity. As a result, a mismatch between the rainfall pattern and traditional seasonal cultivation schedule will happen. This problem indicates the current necessity of an advanced weather prediction model that can be used to guide farmers on their cultivation schedules based on weather and make them ready to handle the issues that occurred with the uncertain climate changes.

The climate of Sri Lanka consists of a variety of different conditions which depend on the geographical existence of different locations on the island. Generally, Sri Lanka has been divided into three main climatic zones: wet, dry, and intermediate. This research aims to propose a weather prediction model to predict daily rainfall in Kandy district, Sri Lanka, which belongs to both wet and intermediate climate zones.

As a result of the modern technological advancements in data analysis, variations in weather-related atmospheric conditions such as precipitation/rainfall, humidity, wind speed, wind direction, temperature, etc. are now accessible for any person. Weather can be demonstrated as an atmospheric state based on the above-mentioned parameters at a particular time and location. As Wiston [3] have mentioned in their research article, the scientific estimation of weather conditions for a specific future time can be performed with the following three steps,

1. Observing and collecting the required data related to weather
2. Processing and analyzing collected data
3. Extrapolating for future state prediction of the atmosphere

Combining the above observations analyzed data with designed models integrated with computer systems will produce a prediction model. All these three steps are significant for improving the accuracy of weather prediction. Most of the existing approaches are based on the weather data related to a particular geographical region on which the research has focused. Therefore, when developing a weather prediction model for Sri Lanka, it is vital to identify the most appropriate weather conditions with higher reliability. Also, processing and analyzing collected data is highly affected for obtaining the most accurate results. Required data preprocessing techniques are different based on the nature of the collected data. Therefore, to obtain high-quality predicted weather results through this research, data preprocessing techniques are identified and applied while ensuring that the originality of

raw data is not changed. When selecting the machine learning algorithms, the nature of the input data and the expected output has to be considered[4]. For the weather prediction model developed through this research, the machine learning algorithms have been selected by considering the size, nature of the distribution of the input dataset, speed, and accuracy of the output.

In this study, a historical weather data set has been received from the Central Environment Authority, including hourly data of different weather conditions such as rain gauge, average temperature, etc. By analyzing the nature of the weather in Sri Lanka, the relationship of weather attributes with agriculture, availability, and accessibility, seven attributes are selected including rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context. Daily data has been generated from the collected hourly data by averaging. A sequence of data preprocessing techniques has been used to assure the quality of the predicted output. A Cross-Validation has been done for the preprocessed data by partition the data set as 70% for model training and 30% for testing purposes [5]. Four machine learning algorithms are used for the weather prediction model: Multiple Linear Regression, Support Vector Machine, K Nearest Neighbors, and Random Forest. Based on the performance and accuracy, the best-fitted model for weather prediction is recognized.

The organization of the paper is as follows. Literature Review, Methodology and Results have been demonstrated in sections II, III, and IV, respectively. Finally, section V discusses the conclusion of this study.

II. LITERATURE REVIEW

A comparative study conducted by Medar [6] have stated different weather-predicting techniques as below,

- Synoptic Weather Prediction- weather parameters are observed within a specific time
- Numerical Weather Prediction includes advanced computer programs based on physical and mathematical equations or algorithms related to weather. Variations occur within the weather over time for deriving meteorological predictions
- Statistical Weather Prediction- identified as a part of objective weather prediction and it is generally focused on least square regression procedures

There are numerous existing weather prediction approaches proposed by researchers through their studies about statistical models and data analytic techniques for predicting future weather in terms of different weather-related variables. Some of these attempts are on identifying the most accurate and efficient techniques for data analytics to predict weather are based on statistical models, while some are based on regressions, decision trees, clustering, neural networks, and many other data mining techniques[3].

Data preprocessing can be identified as an integral step in machine learning-based weather prediction. Research on rainfall prediction conducted by Mohapatra [7] recognizes the importance of data preprocessing because of the difficulty of dealing with the existing outliers and inconsistencies of raw data.

Rainfall Prediction based on data mining approaches can be identified as data models that are more data-intensive than compute-intensive. Bayesian prediction model supports in reducing compute overhead while efficiently working with large data sets. In addition, the Bayesian classifier demonstrates a supervised learning methodology and a statistical methodology for the classification process[8].

An application developed for atmospheric temperature prediction based on Support Vector Regression has been able to recognize the better performances of Support Vector Machines in weather Prediction. It is a compulsory practice to select the most suitable parameters for the application since parameter selection significantly affects the overall system performance[9].

Sequential Patterns-based classification for time series and numeric data from multiple sources has become a significant method in the field of data mining. Yasmin [10] has been able to recognize the importance of processing numeric data and classifying the identified sequential patterns in data to mine data with high accuracy. The system has the ability to maintain a good accuracy in terms of not eliminating the original meaning of raw data but the use of limited parameters to reduce the system complexity has a possibility to affect the accuracy of the system.

Air Pollution data has also been used in weather forecasting approaches. One such system has been proposed by Chakraborty [11] to forecast weather with an Incremental K-means clustering algorithm. However, though the accuracy is considerably high, the insights provided by the output results of this system are minimal.

A similar approach based on clustering analysis has been proposed by Kalyankar [12] for analyzing meteorological data. Clustering can be considered as one of the most useful data mining techniques that can be used to identify hidden patterns in large data sets.

Uncertainty can be a significant aspect of weather prediction because it is really difficult to forecast the future without having certainty in data. As Shahi [13] have indicated in their research, Fuzzy C-Mean clustering can improve the accuracy of weather predicting systems based on data mining techniques such as regression models and decision trees.

A rainfall prediction model developed by Joseph [14] based on Artificial Neural Networks is an empirical method-based prediction approach. In these types of approaches, since the amount of time required for model training excessively increased with the number of neurons, it is necessary to carefully determine the number of hidden layer neurons required for the model.

Shah [15] has provided a rainfall prediction model which enhances the accuracy by using a combination of machine learning and data mining techniques. According to their study, the best accuracy was given by Neural Networks and ARIMA models. In contrast, the Random Forest model has given the best accuracy in classification out of several machine learning algorithms used.

All these researches are conducted in different countries based on the relevant geographical context. However, none of them are based on Sri Lankan Agriculture domain and never validated regarding the Sri Lankan context for precision agriculture purposes.

TABLE I: SUMMARY OF LITERATURE REVIEW

No:	Application	Technologies Used	Attributes	Data Set	Remarks
01	Rainfall Prediction By: Mohopatra [7]	<ul style="list-style-type: none"> Linear Regression K-fold Cross Validation 	<ul style="list-style-type: none"> Precipitation Wet day frequency 	<ul style="list-style-type: none"> Monthly 100 years 	Accuracy: 70% Pros: <ul style="list-style-type: none"> Discrepancies in raw data have been removed successfully during data preprocessing. Cons: <ul style="list-style-type: none"> Accuracy will be decreased due to the use of limited attributes.
02	Rainfall Prediction By: Nikam [8]	<ul style="list-style-type: none"> Bayes Method 	<ul style="list-style-type: none"> Pressure Relative Humidity Wind Speed Rainfall 	<ul style="list-style-type: none"> Daily 16000 instances 	Accuracy: 81% - 96% Pros: <ul style="list-style-type: none"> Simplicity Efficient Performance Cons: <ul style="list-style-type: none"> Accuracy depends on the size of the training data set Missing values in an attribute category
03	Temperature Prediction By: Radhika [9]	<ul style="list-style-type: none"> Support Vector Regression 	<ul style="list-style-type: none"> Maximum Temperature 	<ul style="list-style-type: none"> Daily 5 years 	Accuracy: Not mentioned Pros: <ul style="list-style-type: none"> A better performance by SVM Cons: <ul style="list-style-type: none"> System performance depends on the parameter selection
04	Extreme Weather Prediction By: Yasmin [10]	<ul style="list-style-type: none"> Sequential Pattern Mining Progressive Sequence Tree(PS Tree) 	<ul style="list-style-type: none"> Precipitation Wind direction Wind Speed 	<ul style="list-style-type: none"> 10 min. intervals 	Accuracy: Not Mentioned Pros: <ul style="list-style-type: none"> Reduces the data complexity through data categorization. Fast performance with high scalability. Cons: <ul style="list-style-type: none"> Accuracy will be decreased due to the use of limited attributes
05	Weather Category Forecasting By: Chakraborty [11]	<ul style="list-style-type: none"> Incremental K-means Clustering 	Air Pollution elements <ul style="list-style-type: none"> NOx CO2 SO2 RPM 	<ul style="list-style-type: none"> Daily 10 months 	Accuracy: 83.3 % Pros: <ul style="list-style-type: none"> Good Accuracy with a small data set. Cons: <ul style="list-style-type: none"> Not compared with other existing incremental algorithms for clustering. Predicted output is insufficient to make insights on the weather.
06	Analyzing Meteorological Data By: Kalyankar [12]	<ul style="list-style-type: none"> K-means Clustering 	<ul style="list-style-type: none"> Rainfall Pressure Temperature 	<ul style="list-style-type: none"> Daily 4 yrs 	Accuracy: Not Mentioned Pros: <ul style="list-style-type: none"> Can be used to build dynamic and adaptive prediction models. Cons: <ul style="list-style-type: none"> Not compared with other existing incremental algorithms for clustering. Predicted output is insufficient to make insights on the weather.
07	Temperature Prediction By: Shahi [13]	<ul style="list-style-type: none"> Type-1 Fuzzy Logic System Fuzzy C Mean Clustering 	<ul style="list-style-type: none"> Temperature Humidity 	<ul style="list-style-type: none"> 15 min. intervals 4600 instances 	Accuracy: 1.6590 RMSE Pros: <ul style="list-style-type: none"> Higher accuracy by detecting outliers in data Cons: <ul style="list-style-type: none"> Accuracy depends on the size of the data set
08	Rainfall Prediction By: Joseph [14]	<ul style="list-style-type: none"> Artificial Neural Networks 	<ul style="list-style-type: none"> Humidity Temperature Pressure Precipitable water Wind speed 	<ul style="list-style-type: none"> Daily 370 instances 	Accuracy: 87% Pros: <ul style="list-style-type: none"> ANN can be used with both linear and non-linear data. Cons: <ul style="list-style-type: none"> Model training time increase with the number of hidden layer neurons

09	Rainfall Prediction By: Shah [15]	<ul style="list-style-type: none"> • ARIMA model • Holt Winter method • Simple Moving Average model • Seasonal Naive method • Neural Networks 	<ul style="list-style-type: none"> • Max. and Min. temperature • Relative Humidity • Wind Speed 	<ul style="list-style-type: none"> • Daily (Jun. to Dec.) • 35 yrs 	<ul style="list-style-type: none"> • Accuracy: 70.5% • Pros: Good accuracy through few parameters. • Cons: Dataset includes only half of every year (Jun to Dec). • Predict the rainfall only for months with a possibility to rain.
----	--	--	--	--	--

According to the literature review, several limitations exist in the currently available weather prediction approaches. Among them, the problems that occurred during the data collection process can be identified as major issues. In addition, inaccuracy in data where the collected data are not related to the problem domain, a high amount of missing data has affected the accuracy of the existing systems.

Inefficient data preprocessing has also affected accuracy reductions in current weather predicting systems. As a result of not carefully handling the incomplete and inconsistent data, most systems have been unable to obtain a high-quality output.

Weather prediction systems based on a single machine learning algorithm have faced the problem of selecting the best algorithm. However, the systems that have used multiple machine learning algorithms have not focused on selecting the most appropriate algorithms according to the research domain.

The main problem identified through the literature review is that even different features have been used by different researchers to ensure the accuracy and performance of their systems. Therefore, those proposed approaches have not consolidated those advanced features and techniques into a single system. For example, even a system has considered using a large data set for its model train, it does consider systematic data-preprocessing techniques. As a result, even the data set is adequately large, due to insufficient data preprocessing techniques, the expected accuracy and performance of the system will not be achievable. Also, the systems that give a considerable accurate level cannot provide valuable insights through the predicted results. Therefore it is important to carefully recognize the nature of the intended output given through the model while thinking about whether that output can fulfill the purpose of requirements.

III. METHODOLOGY

The proposed architecture of the weather prediction approach comprises a set of interrelated steps such as data collection, data preprocessing, exploratory data analysis, application of machine learning algorithms, evaluation and identification of the best ML algorithm, and analysis of results. This research mainly focuses on identifying the most appropriate technology-based solution for weather prediction for precision agriculture in Sri Lanka. Even numerous advanced technologies are emerging continuously, it is important to select the most appropriate by identifying the nature of the context in which we are trying to apply them.

In Sri Lanka, rainfall is one of the most significant weather conditions required in agriculture-based decision-making. However, due to the high expensiveness of the available weather data in Sri Lanka, we have to perform the

predictions based on a small dataset that does not comprise more than a few thousand records. By considering the nature of the requirement of predicted weather for Sri Lankan Agriculture basis and the available data on different weather parameters, we have proposed a machine learning-based weather prediction approach.

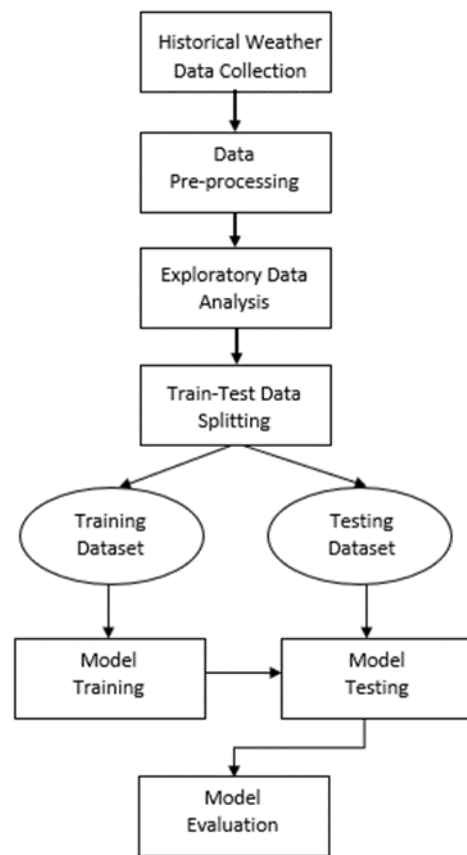


Fig 1. Proposed architecture

As we emphasized in the literature review, it is important to apply different techniques for each proposed architecture step to obtain a better accuracy level. Therefore, the aim of this research is to follow the identified effective practices in the reviewed literature while overcoming the issues that exist within the current approaches in order to build up a better solution for weather prediction using machine learning.

A. Data Collection and pre-processing

In the first part of the proposed weather prediction model, a data set of historical weather data in Sri Lanka is retrieved from the meteorological department and the Central Environmental Authority. Hourly data from 01.01.2019 to 28.02.2021 is collected. . By analyzing the

nature of the weather in Sri Lanka, the relationship of weather attributes with agriculture, availability, and accessibility, seven attributes are selected, including rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context. Daily data is generated from the collected hourly data by averaging.

In order to assure the quality of the predicted results, the collected and structured data are pre-processed through a sequence of data preprocessing techniques as follows,

- **Data Consolidation** – Required data were collected from different sources and therefore, it is required to integrate them into a single table.
- **Data Reduction**- To maintain the prediction model's efficiency, redundant and unnecessary data were removed from the data set.
- **Data Cleansing**- Since the dataset consists of null values and noises, it is very important to handle them carefully. As concluded in the literature review, they are filled with average values instead of replacing missing values with zero.
- **Data Discretization**- To utilize data within machine learning algorithms, rainfall data values are segregated into two intervals: Rain (1) where Rain Gauge is greater than 0mm, and No Rain(0) where Rain Gauge is 0mm.

In order to preprocess data efficiently and accurately, we use python with its libraries including NumPy, Matplotlib, Pandas.

B. Exploratory data analysis

In order to identify the nature of weather condition distributions and correlations, distribution graphs and correlation matrices are used[16]. Correlation matrices can be used to recognize the weather conditions that are most affected by rainfall. In addition to data summarization, data visualization is useful in discovering insights in data effectively and efficiently. In this study, R ggplot2 is used for the exploratory data analysis because it provides better visualization features through its default plots with magnificent graphics.

C. Train-test data splitting

Weather data are usually time series but to prevent unnecessary bias to the machine learning model, we used the Train_Test_Split module. The Train_Test_Split approach, a common cross-validation technique, is done for the pre-processed data by partitioning the data set as 70% for model training and 30% for the testing purpose.

D. Training and testing model

After analyzing the nature of the input dataset and the expected requirements of the output results four supervised machine learning algorithms are used. The purpose of using multiple algorithms instead of a single algorithm is to predict rainfall at a highly accurate level through an evaluation comparison of the results. Multiple Linear Regression has been used as a regression model while Support Vector Machine, K-Nearest Neighbors and Random Forest Models have been used as classification models[17].

a) Multiple Linear Regression (MLR)

Multiple Linear Regression is a machine learning regression approach, which attempts for the relationship modeling between two or more independent variables and response through fitting a linear equation for the observed data. Homogeneity invariance, independence of observations, multi-variate normality, and linearity are the assumptions of the regression model[18].

b) Support Vector Machine (SVM)

In this algorithm, it tries to identify a hyperplane within an x-dimensional space that has the ability to classify the data points in a distinct manner where x means the number of features. Out of all the possibilities, the hyperplane with the maximum margin is selected where the distance between the classes is maximum[19].

c) K-Nearest Neighbors (KNN)

This supervised machine learning algorithm also can be used for both classification and regression problems. K denotes the number of neighbors whose nearest to an unknown new variable is required to predict[20].

d) Random Forest (RF)

Random Forest is a famous and straightforward machine learning algorithm and it is based on ensemble learning that creates an effective model by combining multiple classifiers. This algorithm provides a combination of multiple decision trees and therefore, accuracy is high as well; it reduces overfitting up to a large content[21].

For each algorithm, default parameters are used without performing any modifications. After the model training process, it is used for predicting daily rainfall, based on the data available within the testing dataset. In this study, the weather prediction approach is based on supervised machine learning, including both regression and classification. For the implementation of the proposed solution, sci-kit-learn which is a Python-based module in machine learning also supported by pandas which is a Python library of statistical tools and data structures are used.

E. Model evaluation

For the evaluation of the above machine learning models, a confusion matrix and classification reports are used[22]. Since regression models give a continuous output, before computing the confusion matrix, the predicted output is classified into two categories as below;

- Rain Gauge > 0: Output= 1
- Rain Gauge < 0: Output=0

Accuracy, precision, and recall are the three metrics considered for the model evaluating process. Through the evaluation, the acceptable algorithms for weather prediction are recognized and then the most accurate approach is selected.

IV. RESULTS AND DISCUSSION

In this study, the gathered dataset includes 14000 records and 7 weather attributes were selected from the collected dataset. They are Rain Gauge, Relative Humidity (RH), Average Temperature (AT), Wind Speed (WS Raw),

Wind Direction (WD Raw), Solar Radiation(Solar Rad),
 Ozone Concentration (O3 Conc).

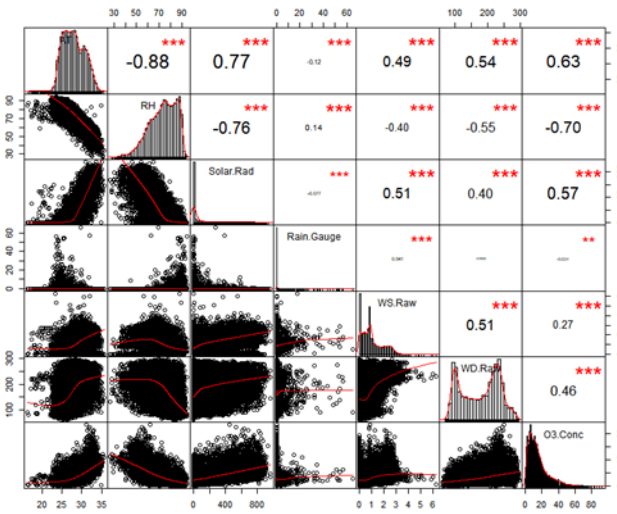


Fig 2. Correlation matrix chart

According to the correlation matrix chart represented in Fig. 2, computed through R, Rain Gauge and Solar Rad are not normally distributed.

	AT	RH	Solar.Rad	Rain.Gauge	WS.Raw	WD.Raw	O3.Conc
AT	1.00	-0.88	0.77	-0.12	0.49	0.54	0.63
RH	-0.88	1.00	-0.76	0.14	-0.40	-0.55	-0.70
Solar.Rad	0.77	-0.76	1.00	-0.08	0.51	0.40	0.57
Rain.Gauge	-0.12	0.14	-0.08	1.00	0.04	0.00	-0.02
WS.Raw	0.49	-0.40	0.51	0.04	1.00	0.51	0.27
WD.Raw	0.54	-0.55	0.40	0.00	0.51	1.00	0.46
O3.Conc	0.63	-0.70	0.57	-0.02	0.27	0.46	1.00

Fig. 3. Correlation matrix

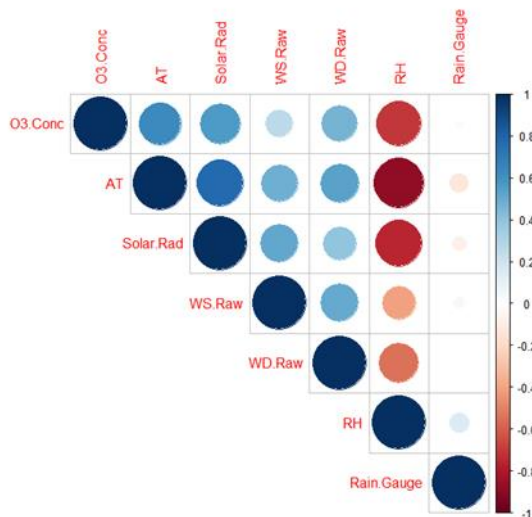


Fig. 4. Correlogram

As represented in Fig. 3 and Fig. 4, correlations among Rain Gauge and other weather parameters are slightly weak, it has computed correlations between Rain Gauge and multiple weather parameters as shown in Figure 5. The correlation between Rain Gauge and the combination of AT, RH, Solar Rad, WS Raw, WD Raw, O3 Conc. is 0.4949 which is a considerable value.

	Rain Gauge
RH + WSRaw	0.1494476
RH + WSRaw + O3	0.1538214
RH + WSRaw + O3 + AT	0.1794619
RH + WSRaw + O3 + AT + Solar Rad	0.1808611
RH + WSRaw + O3 + AT + Solar Rad + WDRaw	0.4949962

Fig. 5. Multiple Correlation

Also when the dataset is large, it is statistically significant even with a weak correlation[23].

A. Evaluation of MLR Model

According to the confusion matrix in Fig. 6 and the classification report in Fig. 7, the accuracy of the predicted output is 44% which is a considerable low accuracy. The accuracy of linear regression is often affected by the normal distribution nature of the data. Since the weather parameters are slightly weak, it is difficult to increase the accuracy of this model.

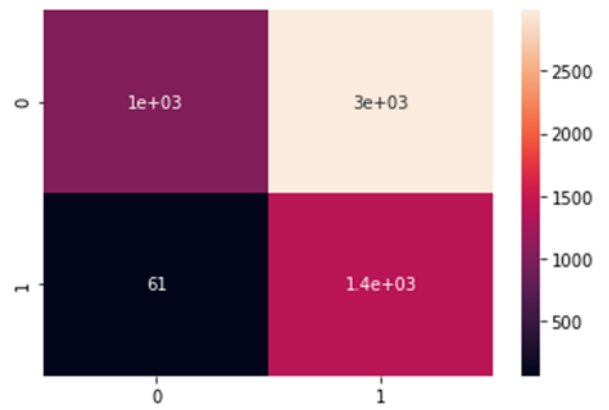


Fig. 6. Confusion matrix- MLR model

	precision	recall	f1-score	support
0	0.0	0.94	0.26	4018
1	1.0	0.32	0.96	1469
accuracy			0.44	5487
macro avg	0.63	0.61	0.44	5487
weighted avg	0.78	0.44	0.42	5487

Fig. 7. Classification Chart- MLR Model

As we concluded in the literature review, the accuracy of regression models depends on the number of variables used. The linear regression model proposed by Mohopatra [7] has acquired 70% accuracy with 2 attributes. In this research, we attempted to predict rainfall using 7 attributes but due to the weaknesses in the normal distribution of the data set which we used, we could reach 44% of accuracy.

B. Evaluation of SVM Model

According to the conclusions made through the literature review, most of the machine learning models including SVM required proper selection of weather parameters. Therefore, in this research, we highly focused on selecting the most suitable weather parameters by studying the domain and performing effective data analysis techniques [9-10].

As depicted in the classification report in Fig. 8, the SVM model has achieved 89% accuracy. This accuracy has been taken by rounding off the value 88.57%. This model has offered a high accuracy compared to the linear regression model. The main reason for achieving good accuracy is the ability of SVM to handle input spaces with non-linear features. Both precision and recall also have achieved greater than 80% where precision is 83% and recall is 89%.

	precision	recall	f1-score	support
accuracy			0.89	5487
macro avg	0.02	0.02	0.02	5487
weighted avg	0.83	0.89	0.85	5487

Fig. 8. Classification chart- SVM Model

C. Evaluation of KNN Model

KNN is a supervised machine learning model which can learn from already labeled data. As we previously mentioned, Rain Gauge values are appropriately labeled as either 1 or 0, and the dataset is properly preprocessed. Since our dataset is considerably small, we selected KNN which will be more applicable in these scenarios. For example, Chakraborty [11] has used a small data set with K-means clustering to forecast weather category and their model has secured 83% accuracy.

In this research, as depicted in the classification report in Fig. 9, the KNN model has achieved 89% accuracy when k=7. This accuracy has been taken by rounding off the value 88.66%. Thus, the precision is 83% and recall is 89%, similar to the KNN model.

	precision	recall	f1-score	support
accuracy			0.89	5487
macro avg	0.02	0.02	0.02	5487
weighted avg	0.83	0.89	0.86	5487

Fig. 9. Classification chart: KNN Model

D. Evaluation of RF Model

Random Forest was selected in this approach since it can be used in both regression and classification problems as well it has a simplified methodology to measure the relative importance of every feature on prediction. As depicted in the classification report in Fig. 10, the RF model has achieved 89% accuracy. This value has been taken by rounding off the value 89.16%. It has an 89% of recall which is similar to both SVM and KNN. However, the precision is 84% which is slightly higher than SVM and KNN models. Since the Random Forest model gives the best overall accuracy compared to the other models, Random Forest can be recognized as the best-fitted model.

	precision	recall	f1-score	support
accuracy			0.89	5487
macro avg	0.04	0.03	0.03	5487
weighted avg	0.84	0.89	0.86	5487

Fig. 10. Classification chart: RF Model

E. Comparison of Machine Learning models

According to the evaluation, the summary showed in TABLE. II, the MLR model has achieved the lowest accuracy at 44%. However, it has been able to achieve a 78% precision. The highest accuracy, 89.16% has achieved by the RF model with the highest precision of 84%. Both SVM and KNN also have been able to achieve high accuracies as 88.57% and 88.66% as respectively.

TABLE II. EVALUATION RESULTS OF FOUR ML MODELS

Evaluation Criteria	Accuracy	Precision	Recall
MLR	44%	78%	44%
SVM	88.57%	83%	89%
KNN	88.66%	83%	89%
RF	89.16%	84%	89%

V. CONCLUSION AND FUTURE WORK

In this research, we comprehensively addressed that, weather plays a significant role in the field of agriculture. However, climate variability is always beyond human control. Sri Lanka is also struggling with the mismatch between weather pattern variations and traditional cultivational schedules. Accurate weather forecasts enable farmers to schedule their cultivation tasks while minimizing weather-based agricultural damages. The proposed architecture attempts to introduce a novel machine learning-based approach for predicting rainfall for precision agriculture in Sri Lanka. Since the weather conditions in Sri Lanka are not perfectly matched with other countries, it is very important to identify the most related weather conditions to predict the weather.

First, we concluded that seven weather attributes could be used to predict rainfall in Sri Lanka for precision agriculture. The selected attributes are rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context.

Secondly, through the exploratory data analysis, we concluded that the multiple correlation of the weather attributes is 0.4949 which is a good value compared to the correlations observed within existing.

Thirdly, we concluded that several data preprocessing techniques are required to enhance the quality of the prediction. Therefore, data consolidation, reduction, cleansing, and discretization were performed on the data carefully.

Fourthly, by studying and analyzing the problem background and the nature of obtained dataset to improve the accuracy, four supervised machine learning algorithms were selected. For the prediction, model cross-validated data were trained and tested with Multiple Linear Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest.

Finally, with the model evaluations, Random Forest was recognized as the best-fitted model that achieved 89.16% accuracy. This can be considered as a better level of accuracy compared to the prevailing weather prediction approaches.

As for future work is expected to increase the size of the dataset and apply more data preprocessing techniques such as feature engineering to enhance the quality of the

dataset. Since SVM and KNN models have also given better accuracy levels, it is important to build and evaluate a hybrid ensemble learning model which combines these machine learning models for this weather prediction approach. Deep learning is a member of the broader community of machine learning and it is based on artificial neural networks with representation learning. It is expected to apply deep learning for predicting the weather with a large dataset and evaluate the accuracy improvement.

REFERENCES

- [1] T. B. Adhikarinayake, "Methodical design process to improve income of paddy farmers in Sri Lanka," [publisher not identified], Wageningen, 2005.
- [2] "Sri Lanka tackles challenges to rice production to end reliance on imports," oxford business group.
- [3] M. Wiston and M. Km, "Weather Forecasting: From the Early Weather Wizards to Modern-day Weather Predictions," *J Climatol Weather Forecasting*, vol. 06, no. 02, 2018, doi: 10.4172/2332-2594.1000229.
- [4] L. H. S. De Silva, N. Pathirage, and T. M. K. K. Jinasena, "Diabetic Prediction System Using Data Mining," presented at the Proceedings in Computing, 9th International Research Conference-KDU, Sri Lanka, 2016.
- [5] Berrar, "Cross-Validation," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.
- [6] R. Medar, A. B. Angadi, P. Y. Niranjan, and P. Tamase, "Comparative study of different weather forecasting models," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, Aug. 2017, pp. 1604–1609. doi: 10.1109/ICECDS.2017.8389719.
- [7] Mohopatra Sandeep, "Rainfall Prediction using 100 years of Meteorological Data." 2017.
- [8] B. Nikam and B. B. Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach," in 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, Seoul, Korea (South), Sep. 2013, pp. 132–136. doi: 10.1109/CIMSim.2013.29.
- [9] R. Yalavarthi and M. Shashi, "Atmospheric Temperature Prediction using Support Vector Machines," 2009. doi: 10.7763/IJCTE, 2009.V1.9.
- [10] R. Y. Yasmin, A. E. Sakya, and U. Merdijanto, "A classification of sequential patterns for numerical and time series multiple source data — A preliminary application on extreme weather prediction," in 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, Nov. 2017, pp. 1–5. doi: 10.1109/ICoDSE.2017.8285845.
- [11] S. Chakraborty, N. K. Nagwani, and L. Dey, "Weather Forecasting using Incremental K-means Clustering," p. 6.
- [12] Meghali A. Kalyankar, "Data Mining Technique to Analyse the Meteorological Data," *IJARCSSE*.
- [13] Shahi, R. B. Atan, and N. Sulaiman, "Detecting Effectiveness of Outliers and Noisy Data on Fuzzy System Using FCM," p. 13.
- [14] J. Joseph and Ratheesh T K, "Rainfall Prediction using Data Mining Techniques," 2013.
- [15] U. Shah, S. Garg, N. Sisodiya, N. Dube, and S. Sharma, "Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques," in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, Dec. 2018, pp. 776–782. doi: 10.1109/PDGC.2018.8745763.
- [16] T. Pham-Gia and V. Choulakian, "Distribution of the Sample Correlation Matrix and Applications," *OJS*, vol. 04, no. 05, pp. 330–344, 2014, doi: 10.4236/ojs.2014.45033.
- [17] Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.
- [18] R. Bevans, "An introduction to multiple linear regression." <https://www.scribbr.com/statistics/multiple-linear-regression/>
- [19] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," *Medium*, Jul. 05, 2018. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> (accessed Jun. 02, 2021).
- [20] "KNN - The Distance Based Machine Learning Algorithm," *Analytics Vidhya*, May 15, 2021. <https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/> (accessed Jun. 02, 2021).
- [21] S. Awasthi, "Random Forests in Machine Learning: A Detailed Explanation," *datamahadev.com*, Dec. 05, 2020. <https://datamahadev.com/random-forests-in-machine-learning-a-detailed-explanation/> (accessed Jun. 02, 2021).
- [22] M. Goonathilake and P. Kumara, "SherLock 1.0: An Extended Version of 'SherLock' Mobile Platform for Fake News Identification on Social Media," *Sri Lanka*, p. 7, 2020.
- [23] "The Correlation Coefficient (r)." <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html> (accessed May 28, 2021).

Design and development of pump based chocolate 3D printer

R. R. A. K. N. Rajapaksha*
Department of Engineering Technology
Faculty of Technology, University of Ruhuna, Sri Lanka
kalaninethmarajapaksha@gmail.com

Yashodha G. Kondarage
Department of Engineering Technology
Faculty of Technology, University of Ruhuna, Sri Lanka
yashodha@etec.ruh.ac.lk

B. L. S. Thilakarathne
Department of Engineering Technology
Faculty of Technology, University of Ruhuna, Sri Lanka
sanjaya@etec.ruh.ac.lk

Rajitha De Silva
RCS2 Technologies (Pvt) Ltd, Sri Lanka
rajitha@rcstotech.com

Abstract - The use of 3-Dimensional (3D) printing, known as Digital fabrication (DF) or additive manufacturing (AM), technology in the food sector has countless potential to fabricate 3D constructs with complex geometries, customization, and on-demand production. For this reason, 3D technology is driving major innovations in the food industry. This paper presents the construction of a chocolate 3D printer by applying the pressure pump technique using chocolate as a printing material. Here the conventional 3D printer's design was developed as a chocolate 3D printer. As an improvement, a new extruder mechanism was introduced. The extruder was developed to print the chocolate materials. In the working mechanism, the 3D printer reads the design instruction and chocolate material is extruding accordingly, through the nozzle of the pump to the bed of the 3D printer followed by the design (layer by layer). The special part of this chocolate 3D printer is the pressure pump in the extruder part. That pressure pump provides pressure on melted chocolate from the chocolate container to the nozzle point. The usability and efficiency of the 3D printer were tested with sample designs. The obtained results were presented and discussed. Together with these advances this 3D printer can be used to produce complex food models and design unique patterns in chocolate-based sweets by satisfying customers.

Keywords - 3D printing, additive manufacturing, food printing, hot melt extruder, pressure pump

I. INTRODUCTION

3D Printing, also known as the additive manufacturing technique, refers to processes used to synthesize a three-dimensional object in which successive layers of material are formed under computer control to create an object. Referring to the present used to synthesize a 3D object layer by layer materials are formed under a complete control system, to create an object. Currently, this technique is applied to make proofs of concept, prototypes, or end-products. Companies are implementing 3D printing at different stages of their manufacturing processes. The modern world has lots of applications of 3D printing technology. Now 3D printers have become more affordable for ordinary consumers.

Food printing manufacturers have realized the potential of 3D food printers in promoting culinary creativity, nutrition, and ingredient optimization, and food

sustainability [1]. 3D printing has a process to manufacture 3D objects, this process can divide into steps such as (i) Create a model using software and convert it to STL (Standard Triangle Language) format. (ii) Fill the storage tank choose to model. (iii) Input STL format to the system. (iv) Operate the 3D printer to extrude the material. (v) Final object using with XYZ movement [1].

This project mainly focuses on chocolate 3D printing. Most chocolate 3D printers can process CAD files, just like normal 3D printers. Currently, chocolate 3D printers use a syringe instead of a filament, load it, and then hold it at a temperature at the time of printing [2]. The extruder head moves and deposits the melted chocolate in the desired shape. The chocolate eventually cools and solidifies. The syringe loading system is safe, clean, efficient, and keeps the chocolate fresh. If the operating temperature is followed, the chocolate will not dry at all in the syringe [3]. In a conventional 3D printing machine, the mechanical parts of the 3D printer have four stepper motors used to drive the XYZ axis. The movement of the Y-axis is independent and is performed mainly by a pair of ball screws with sliding supports that move the platform back and forth. The movements of the X and Z axes are interconnected and are performed mainly by a spherical screw supporting the optical axis. The X-axis is responsible for horizontal movement and the Z-axis is responsible for vertical movement [4]. Fig. 1 shows the conceptual design of the 3D printer XYZ axis.

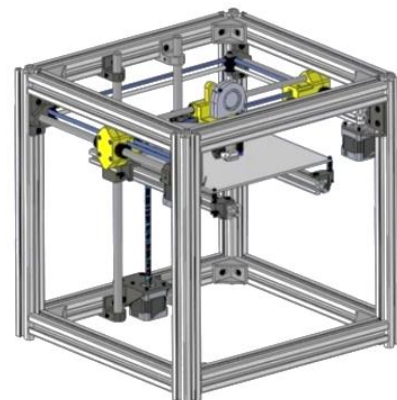


Fig. 1. Conceptual design of 3D printer XYZ axis

II. MATERIALS AND METHOD

The following techniques were applied when developing the chocolate 3D printer. It is the development of a chocolate 3D printer that uses chocolate as ink with a novel pump mechanism.

A. Mechanical platform and controls

Food structure can be deposited/sintered effectively point by point and layer by layer according to a computerized design modeling and route planning. This system uses layer-by-layer extruding.

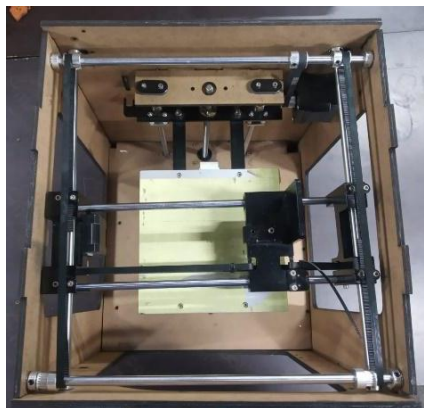


Fig 2. Mechanical platform of the chocolate 3D printer

Fig. 2 shows the mechanical mechanism part without the extruder of the chocolate 3D printer. As above mentioned, the pulleys, rods, and belts were used for the mechanical movement with the stepper motors. The rotational movement provided by the motor is activated by the pulleys, rods, and belts.

3D printer makes products by various kinds of layer-by-layer deposition on the plane surface. 3D printing has different types of layer deposition methods. The extrusion head usually pushes food through the nozzle through compressed air. Typically, the smaller the nozzle, the longer it takes to print the food [6]. In this research, the normal 3D printer was designed like a chocolate 3D printer by introducing a pump base chocolate pressure system. Especially, it is considered extruder mechanism. This proposed system will be focused on chocolate base food products using paste-type ingredients.

Following modifications were done when developing the proposed system. In the extrusion-based printing technique, the pump usually pushes the materials to the nozzle through compressed air [2]. The compressed air means the air inside of the food syringe, but the proposed system has extra pressure on extruding chocolate. Therefore, the storage tank can be larger than the syringe. In the storage tank of the chocolate 3D printer, the food container should provide the chocolate continuously and not need to pause the 3D printing process to refills. Fig 3 shows the complete chocolate 3D printer.

The movements of the 3D printer are using the belt mechanism. The rotational movement provided by the NEMA 17 stepper motor was activated by the pulleys and belts. Four NEMA 17 stepper motors were used for the mechanical part. XYZ axis was powered by three NEMA

17 stepper motors. Another one is used for power to a pump of the extruder system.

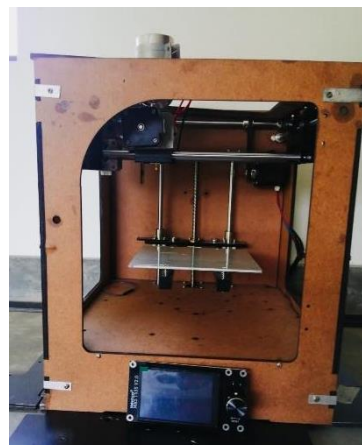


Fig. 3. Chocolate 3D printer

B. Extruder design

The design of the extruder part has three main components, such as chocolate containers, chocolate pumps, and nozzle sets. The chocolate container has a cylindrical shape, and the pieces of chocolate are put into this container. Then melted chocolate goes through the chocolate pump to the nozzle set with pressure. That chocolate pump has two gear wheels, and it has the same mechanism as an oil gear pump. All parts of the extruder set were prepared using stainless steel because all components are in contact with the chocolate. Further extruder set has a temperature system.

A too-small nozzle will lead to too slow extrusion speed, and a too-large nozzle will lead to a rough food surface [7]. Three nozzle diameters such as 1 mm, 1.25 mm, and 1.5 mm were investigated by varying the extrusion rates and nozzle moving speed the best nozzle diameter was found as 1.25 mm in terms of the property of the deposited product.

Stainless steel is a commonly used material in the food industry and is generally resistant to corrosion. [5]. Food Grade Stainless Steel 316 turned into used for the layout of an extruder, the grade 316 SS, can experience severe pitting corrosion when exposed to chocolate, which is often present in food product machines. 316 SS makes for great food-grade stainless steel parts for nearly any food application.

Extreme care must be taken when making the extruder part of the 3D printers which are used for food. Stainless steel has proven itself, time and time again to be a food-safe material. It does not corrode, rust, or provide livable conditions for harmful pathogens. In terms of hygiene and durability, stainless steel was used in the design.

C. Feature-based software

The chocolate 3D printer has used firmware to control all activities called "Marlin". Marlin is an open-source firmware that controls all real-time activities of the machine such as adjust heaters, steppers, sensors, lights, LCD screens, buttons, and everything related to the 3D printing process.

D. Chocolate (as the material intended to be printed)

Normal 3D printers are used plastic materials for printing purposes however chocolate 3D printers use paste-type melted chocolate [2]. This proposed system is based on cooking chocolate. Cooking chocolate is a type of chocolate and uses for decorating foods. Cooking chocolates contain sugar, vegetable fat, and cocoa powder [4]. The main distinction between cooking chocolate and 'normal' eating chocolate is how sweetened it is. Baking chocolate has a higher percentage of cocoa solids and usually contains less or no sugar than regular eating chocolate. Cooking chocolate for tempering or couverture may have more cocoa butter to ensure that it melts evenly and easily. The melting temperature of cooking chocolate is between 38°C and 42°C. The solidifying temperature of cooking chocolate is between 26 °C and 28 °C [4]. The melting temperature is important for the temperature unit of the chocolate 3D printer to melt the chocolate pieces. The temperature is important for solidifying the printing object. When the temperature is at the low 26 °C, chocolate begins to melt. When chocolate is crystallized or tempered, it is liquid and usable between 30 °C and 32 °C (lower for white and milk chocolate, higher for dark), and solidifies quickly at room temperature. Your chocolate melted but didn't get tempered/re-crystallized when it cooled, therefore it stayed liquid due to its lack of crystalline structure.

III. RESULTS AND DISCUSSION

A. Printing technology

A chocolate 3D printer is a machine that can be used to produce prototype products rapidly. This 3D printer used chocolate as a process material. Further, this 3D printer applied a pump-based 3D printing technique using the pressure pump. In normal, 3D food printers use several 3D printing technologies such as Selective Sintering, Hot melt extrusion, Binder jetting, and Inkjet printing [3].

This chocolate 3D printer used hot-melt extrusion technology to extrude the chocolate. In this hot-melt extrusion, the chocolate material was heated up to its melting point. After this melted chocolate is deposited on the bed to build the required object layer-by-layer from the bottom to the top by heating and extruding the filament. This food printer was designed based on the efficient size of the hot-melt extrusion with low maintenance cost. The disadvantages such as the time take to connect the layers, long production time, and delamination caused by temperature variation, need to be further investigated.

In this chocolate 3D printer was used Hot melt extrusion technology with a pressure pump. It is a special feature of this 3D printer. That pressure pump is made up of an external gear pump with two gear wheels. An external gear pump contains two equals, interconnecting gears supported by separate shafts. Generally, one gear is driven by a stepper motor, and this drives the other gear. Fig 4 shows the cross-sectional area of the pressure pump. This chocolate 3D printer compares with syringe-type 3D food printers. This syringe-type 3D printer used a pressure syringe to extrude the printing materials and this chocolate 3D printer uses a pressure pump to extrude the printing materials. That pressure pump provided extra pressure on extruding materials (chocolate) than normal syringe-type food 3D printers. The pressure of the pump can be change

depend on the amount of the chocolate container. Fig. 5 shows the top side is of the extruder with melted chocolate.



Fig. 4. Cross sectional area of pressure pump



Fig. 5. Melted chocolate in chocolate container

Another advantage of this design is the continuation of the printing process. In syringe-type 3D printers, the printing process will stop when finished the materials' contents from the syringe. The syringe plunger wants to remove to refill the syringe. However, that problem can be overcome using this extruder. The extruder is powered by a pressure pump by a stepper motor. The upside of the chocolate container is free; therefore, it can be refilled as soon as it is printing. Fig 6 shows a chocolate extruder with its features.

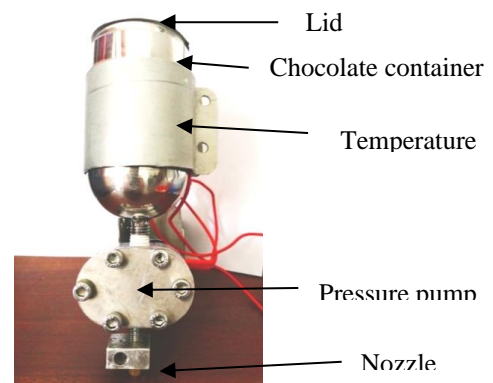


Fig. 6. Chocolate extruder

B. The efficiency of the chocolate 3D printer

The printing rate was calculated by dividing the weight of the printed object overprinting time [7].

$$\text{Printing rate (g/min)} = \frac{\text{Total weight of the printed object (g)}}{\text{Printing time (min)}} \quad (1)$$

The melting and crystallization behaviors of fat present in chocolate will be important to understand from the point of view of deposition temperature and change occurring in deposited chocolate. The physical properties and mouth feel of the 3D printed chocolate product will be dependent on the time and temperature history after deposition. Fig 7 shows the slicing software detail of a printing object.

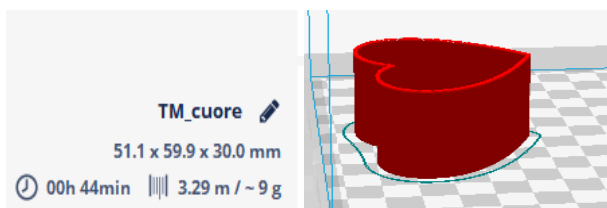


Fig. 7. Slicing software details of print object

TABLE I. CALCULATION DETAIL OF PRINTING RATE OF PRINTER

Printing time	Total weight of the object	Printing rate of printer
44 min	9 g	0.20 g/min

Differential Scanning Calorimetry analysis (DCS) is a thermo-analytical technique in which the difference in the amount of heat is required to increase the temperature of the sample. This analysis is used to study the behavior of the material of the function of temperature or time. Ex: Melting point, Crystallization behavior, Chemical reaction. DCS analysis of the deposited chocolate product, indicating that the viscosities of the chocolates can be relatively constant when the temperature is between 32°C and 40 Celsius and the pressure between 3.5Pa and 7Pa [2].

Some research papers about DCS analysis were studied and get an idea about the temperature behavior of the chocolates. The quality of the print object is depending on printing materials and nozzle type. In this chocolate 3D printer, the quality of the print was compared with nozzle type only. Normal chocolate 3D printers have used the adapter of syringe (endpoint of the normal syringe) but this chocolate 3D printer was used the normal nozzles of normal 3D printers. Therefore, the quality of the print objects can increase.

Further, this chocolate 3D printer can use different size nozzles such as 0.8 mm, 1.0 mm, 1.25 mm, and 1.50 mm. The quality of the print was changed depending on nozzle size. The 1.25 mm nozzle is the best nozzle size for this chocolate 3D printer, and it was selected based on several experiments. Fig. 8 shows printed chocolate objects.



Fig. 8. Printed objects

C. Three-axis mechanism

The structure of the chocolate 3D printer was prepared by using MDF (Medium-density fiberboard). And pulley, rods, and belts were used for constructing the movement of the three-axis. It provides smooth rotation and reduces vibrations. The bed of the 3D printer moves to the Y-axis, the set of the extruder move to the Z-axis, and the extruder move to the x-axis. This system has four stepper motors, three for the XYZ axis and another one is to rotate the pressure pump.

IV. CONCLUSIONS

The developed, chocolate 3D printer has a maximum printing size is 150 mm*150 mm*150 mm. The chocolate 3D printer was developed using a pressure pump. The pressure pump provides the extrusion of the materials with the support of the stepper motor and driving commands. Comparing to other food printing techniques in the market, most of the printing methods have syringe-type extruders. Using this syringe-type extruders, have several limitations. Mainly that continuation of the printing process. The printing process will stop the contains of materials is finish at the syringe. However, that problem can be overcome using this extruder. The advantage of this 3D printer is the ability to get beautiful and creative designs, and it can be used very effectively in the hotel industry. Also, another limitation of this chocolate 3D printer is the lack of an advanced cooling system which is required to printing the bed. The chocolate printer contains a normal cooling system with the cooling fan was attached to the frame, to help the solidification of the printing object. Additionally, in future work, the bed cooling system will be introduced.

ACKNOWLEDGMENT

The authors would like to acknowledge Mr. T. A. Sandakalum at the mechanical workshop, and the staff at Department of Engineering Technology, University of Ruhuna for the support given to the development of the mechanical design.

REFERENCES

- [1] C. Liu, C. Ho and J. Wang, "The development of 3D food printer for printing fibrous meat materials", IOP Conference Series: Materials Science and Engineering, vol. 284, p. 012019, 2018. Available: 10.1088/1757-899x/284/1/012019 [Accessed 10 March 2021].
- [2] F. Yang, M. Zhang, and B. Bhandari, "Recent development in 3D food printing", Critical Reviews in Food Science and Nutrition, vol. 57, no. 14, pp. 3145-3153, 2015. Available: 10.1080/10408398.2015.1094732 [Accessed 10 March 2021]
- [3] J. Sun, Z. Peng, W. Zhou, J. Fuh, G. Hong and A. Chiu, "A Review on 3D Printing for Customized Food

- Fabrication", *Procedia Manufacturing*, vol. 1, pp. 308-319, 2015. Available: 10.1016/j.promfg.2015.09.057 [Accessed 20 March 2021].
- [4] M. Lanaro., "3D printing complex chocolate objects: Platform design, optimization and evaluation", 2021
- [5] M. Jellesen, A. Rasmussen and L. Hilbert, "A review of metal release in the food industry", *Materials and Corrosion*, vol. 57, no. 5, pp. 387-393, 2006. Available: 10.1002/maco.200503953 [Accessed 10 March 2021].
- [6] J. Sun, Z. Peng, W. Zhou, J. Fuh, G. Hong and A. Chiu, "A Review on 3D Printing for Customized Food Fabrication", *Procedia Manufacturing*, vol. 1, pp. 308-319, 2015. Available: 10.1016/j.promfg.2015.09.057 [Accessed 10 March 2021].
- [7] S. Mantihal, S. Prakash, F. Godoi and B. Bhandari, "Optimization of chocolate 3D printing by correlating thermal and flow properties with 3D structure modeling", 2021

Theoretical framework to address the challenges in Microservice Architecture

Dewmini Premarathna*
Department of Software Engineering
University of Kelaniya, Sri Lanka
dewminic@kln.ac.lk

Asanka Pathirana
Department of Software Technology
University of Vocational Technology, Sri Lanka
asanka.pathirana@gmail.com

Abstract - Microservice Architecture (MSA) is a recommended way to introduce the application software in a modularized manner instead of the traditional Monolithic Architecture (MA) approach due to the inherent advantages. The MSA is very much effective considering the true benefits of scalability, flexibility, cost-effectiveness, etc. However, there are significant challenges in the use of MSA as well in the viewpoint of the seniors in the field of Software Engineering (SE). So, the objective of this research is to introduce a theoretical framework to be followed by the SE industries to address the challenges they face in providing MSA-based software solutions. In this research, the literature of MSA is evaluated in detail to understand the influencing factors to cater to the requirements of the software developments. In methodology, two research questions are derived based on the hypothesis of not getting adequate benefit in the process of adopting MSA for software application development; 1. What are the challenges to implementing applications incorporating MSA? 2. How to achieve the exact needs of the clients via MSA? For this study, based on purposive sampling the five SE professionals are selected for interviews to understand the true impact on identified factors through literature for development challenges and client satisfaction. Further, thematic analysis is conducted for evaluating those extracts of the interview qualitatively. Nevertheless, the online questionnaire is distributed among a wide range of SE professionals in the domain of MSA implementation for overall understanding about significant factors filtered out through the literature and the interviews, and those were analyzed descriptively. Based on the findings, a theoretical framework is introduced for successful implementation of MSA assuring the clients' requirements. Eventually, this study confirms how MSA adaptation with the theoretical framework is effective for both organizations and clients.

Keywords - *development, framework, microservices, modularize*

I. INTRODUCTION

At present, the software industry is a bit more complex due to the evolution of the technology, progressive demand of the clients, affordability of the customer, complex business requirements, etc. ultimately, the nature of the solutions is also complex catering to different requirements of different audiences[1], [2]. As a result, the software industry is possibly subdivided into different main categories such as product-based, service-based, solutions-based, and research-based. However, the most important consideration of any software solution is its architecture influencing the quality of the final outcome. There are many ways to introduce solid architecture to incorporate specific requirements of the software solution giving priority to exact requirement(s). But any architecture software industry can decide whether it is a single component or a combination of several modules.

Among the many available software architectures, the MSA is a priority consideration for introducing solution architecture either partially or completely because the MSA allows introducing the solution as a collection of smaller services[3]. On the other hand, the MA provides the entire software solution as a single service but it comprises drawbacks in implementation and maintenance perspectives [4], [5]. However, MSA has been introduced to address those issues effectively.

Moreover, the solutions-based software industries mainly interact with clients to cater to their emerging requirements, and the software solutions are developed by providing the priority for the client requirements [6]. However, the technical decisions over architecture are made by SE professionals. In some situations, the technical background is also communicated with the client, but with the facts in long run, the final outcome of the particular phase of the development is more focused on [7]. As a result, the client may be suffered in the long run due to extended maintenance and extra efforts is different.

It is compulsory for the client to have an entire understanding of the lifecycle of the use of particular software for making a strategic decision towards selecting the right application software [6], [7]. In other words, the effectiveness of the business process should be improved with the involvement of software solutions by increasing productivity to achieve business objectives. The MSA is a priority consideration for such initiatives so the important factors of MSA are identified in detail in different aspects such as maintainability, scalability, reusability, etc. [3], [5], [8]–[10].

There is a trend in the industry to use MSA due to its benefits, but there is uncertainty whether the organization and client are acquired the true benefits of MSA. MSA also has its own drawbacks associated with distributed services, partitioned databases, infrastructure resources allocation which add extra complexity to the software analysis, design, development, and deployment [10]. The unacceptable or improper usage of MSA also prevents getting its advantages towards the organizations. These reasons may cause the software industry to suffer from various shortcomings throughout the Software Development Life Cycle (SDLC) process. There is a possibility this is indirectly transferred as a cost to the client. As a result, the client ends up with high costs in long run [7]. This paper focuses on those situations and proposes how to satisfy both organizations and clients with the use of MSA on their projects.

Section II discusses the literature about MSA incorporating the reviewed research papers. In section III, the methodology is described and the research design is also extracted from the methodology in the same section. The results and discussions are laid down in section IV, whereas the recommendations are illustrated via theoretical

framework next in section V. Then the conclusion is made finally in section VI towards delivering the true benefits of MSA for everyone.

II. LITERATURE REVIEW

The literature is to understand the real value of the MSA and analyze whether those values are properly utilized by the software industry towards delivering appropriate benefits for the clients according to the specific requirements. The main focus here is to have a strong understanding of MSA, its benefits, and its challenges. Literature is mainly categorized into design & implementation, security, deployment, and reporting to understand the benefits and the challenges associate with MSA.

A. Design & Implementation

Incorporating MSA for the software solution is comprised of a mix of both the benefits and the drawbacks as per the requirement of the situation of a client, so it is always challenging to make appropriate use of the required microservices by software engineering professionals [11]–[13]. Some features are required to implement essentially and some others are inherently available with MSA. The important high-level features of MSA are briefly described as follows.

1) *Scalability*: Scaling is a very important aspect of MSA and it is highly supported for utilizing resources as per the dynamic requirements [14]. To achieve scaling, the solution is introduced as a collection of small services assisting to easily allocate resources upon the requirement of the specific service. Resources such as memory, CPU, disk usage, can be shared within services and more resources will be allocated to those who need it, thus reducing cost [11].

2) *Flexibility*: MSA has great flexibility in selecting programming language and introduce new human resources into the project effortlessly [8], [10]. If the solution requires more services to develop, the industry has the flexibility to selecting resources at any given time irrespective of its programming skills and which language is used to developed other services [11]. So this is a great advantage that you couldn't achieve from MA.

3) *Unit Testing and Integrating Testing*: The effect of the unit testing is not much different in MSA and MA domains, but MSA comprises some repeated works (Rahman, and Gao, 2015). Further, the integration testing is relatively more complex in the use of MSA because the involvement of dependent services is significant with respect to MA [4]. As a result, MSA requires more time and effort to complete such testing.

4) *Service Discovery*: The main function of service discovery is to incorporate new services into the solution [4], [8]. It seems service discovery is an essential element to implement with the solution for large-scale microservices-based solutions because it should automatically detect the services added into the echo system and give zero downtime to the entire system.

5) *Circuit Breaker*: Circuit breaker is also an essential feature for a solution and it is the approach to isolate the faults automatically to prevent system failures due to an

issue with one service. So the main functionality of the circuit breaker is to check the availability of independent services and to start sending requests again upon the availability of the dependent services up [14]. So, this is an additional overhead that developers need to do.

B. Security

In one viewpoint, there is a benefit over security when it comes to the MSA, if one service is open for vulnerability, it is a matter of disabling that and allow the system to run as usual with minimum impact [15]. In another viewpoint, MSA influences security negatively due to network security risk because each microservice communicates over the network via messages. As a result, the internal attacker is in a position to easily find out the message format and try to sabotage the system. Following aspects discuss more details about security aspects.

1) *Web Application – Front End*: With the MSA there is a need of considering web applications development as small features called micro-front-ends (MFE). So, with the MFE architecture, if one function breaches security or opens for vulnerability, it is easy to disable such functions and the application is available to users with less effect of user experience. On the other hand, the security of each function needs to be validated separately and all the developers who work in parallel on services must have strong knowledge of web application security such as disabling auto-filling on the text fields, masking sensitive inputs typed on the text fields, handling cookies securely in the browser level, keep token like sensitive information in an encrypted format, etc [15].

2) *Application Level – Back End*: For the micro-front end architecture there are a set of microservices are available to support backend services as well (Rahman, and Gao, 2015). As mentioned earlier it is an advantage to isolate service open for vulnerability and allow the system to work smoothly. But achieving security standards for each microservice is a more time taking task. Developers, designers, and architects need to think about factors like enabling HTTPS (transport layer security) for intercommunications, secure database connectivity, loading secrets and keys from secure stores such as vaults solutions, etc [15]. Further, some application needs to comply with client's security requirement such as banking guidelines for banking solution. Hence applying these things to all the microservices is required extended time and effort.

3) *Source code*: Microservices source code is kind of repeating the same security approaches in multiple places. So, the requirements like keeping passwords in encrypted format in property configurations are going to be a big overhead to the network because it is needed to load from a centralized secure store; like vault solutions [15]. Hence, when it is compared with MA source codes security with MSA, has significant complexity.

4) *Database*: The best practice of MSA is to keep separate databases for each service because when it is required to scale up, the database also can be scaled up separately [8], [15]. If the common database is used for all the microservices, scaling only services is not enough and a bottleneck can occur from the database side. From the

security perspective database, administrators have to apply/configure security for databases separately and which requires more time and effort.

5) *Vulnerability Assessment and Penetration Testing*: For a production-ready application, a final check is to assess vulnerabilities and do a security test which is called penetration testing which covers CSS attacks, SQL injection, CSRF, basically all the security standards are defined by OWSAP application security verification standards [15]. So, the preparation of testing and carry out testing on each developed service and the deployed environment is required extended effort than it is deployed with MA.

C. Deployment

Deployment of MA is very easy because it is required to deploy one or two applications in an application server and high availability can be achieved through horizontally scaling two or three nodes and required a minimum of two databases for failover/replication [5]. But in MSA things are different, it is required more tools like Docker and Kubernetes and the industry needs to build a required skill set to do a successful deployment. The entire deployment process is in five main topics.

1) *Docker*: Docker is a containerized technology that acts as a small machine and its configuration can be defined by the DevOps engineer or architects to match with particular service requirements. So, each microservice developed for a solution can be configured as containers and can run as small servers [16].

2) *Kubernetes*: On average, software solution is comprised of a considerable amount of microservices and if those run as Docker containers the same number of small machines are running on top of the infrastructure and managing them might be an arduous task. Hence Kubernetes technology has introduced the capability of managing docker-containers efficiently [8], [16]. So when compared with MA this requires more works to achieve sustainable MSA deployment.

3) *Continuous integration and deployment (CI/CD)*: CI/CD is a most important concern on any development means it helps to automate the building of application and deploy in test, staging, and then production environment [16]. So, the CI/CD process incorporates automation of the build process from development to production environment. When it comes to the MSA building process it requires more configuration.

4) *Observability*: Observability requirement is consisted of log analytics, distributed tracing, and metrics monitoring. There is a special toolset and most industries use ELK stack for log analytics, elastic APM for metrics, and Zipkin for distributed tracing which is an essential tool for MSA [10]. So, to troubleshoot the issues this setup is required for every deployment, and this is involved more works.

5) *Service mesh*: Service mesh is a dedicated communication layer that ensures reliable and safe communication between services with high observability[14]. It can handle high-volume communication and uses existing persistent connections to

improve performance [17]. Implementing service mesh is not mandatory with MSA but it can add benefits in service discovery, load balancing, encryption, observability, traceability, authentication, and authorization [14]. As service mesh supports circuit breaker, it is no need to develop that feature separately at the sourcecode level.

D. Reporting

Reporting in MSA is a bit complex. Because required data for reporting is in individual microservices [3]. Followings are three approaches that can be used in report generation, and each has its own drawbacks.

1) *API-based Reporting*: In this approach, reporting service will extract data through API calls from each service and it increases network traffic [8]. Further, the system tends to unresponsive service calls due to hanging if users extract data for long period.

2) *Database-based Reporting*: In this approach single report service connect to each database owned by other services [8], [12], [14], [15]. Drawbacks that are arisen with the API approach can be overcome with this, but then it breaks the basic principle of MSA because one service is tightly coupled with all other services. If the developer changes any logic or implementation which affects the data structure on a particular service, the report service also needs to be adjusted to address the changes.

3) *Message Queue(s) based Reporting*: This can be considered as the best approach where each service sends an event to a message queue and report service saves the message into its own database [8]. Then data is available for the reports without affecting any service. Although it is the best approach, extra complexity is added to the environment since additional message brokers need to be managed.

III. METHODOLOGY AND RESEARCH DESIGN

The methodology is introduced for having an overall understanding of the use of microservices to fulfill the requirements of the clients. Then the experiment design is introduced based on derived methodology.

A. Methodology

The background analysis is the initiation for this research with the use of experiences and available literature until enough background understanding is obtained. Then the overall understanding of the influencing factors is achieved for continuing with the interview with SE Professionals. The purposive sampling is used to filter out the 5 key experts who work with MSA due to their comprehensive understanding of MSA, and such data is evaluated based on a thematic analysis approach. Then the questionnaire is introduced incorporating background information and interview findings, and it is shared among the different stakeholders to obtain their opinion in a broader sense. Then responses for the questionnaires are collected for descriptive analysis due to their quantitative nature. As a result, this research approach is a mixed method. Further, findings are organized to recommend a theoretical framework for SE Professionals to use for the betterment of themselves as well as their clients.

B. Experimental design

As per the above methodology, the flowchart in Fig.1 is introduced as an experimental design, and the outcome of this research is a theoretical framework for SE professionals to use as guidance.

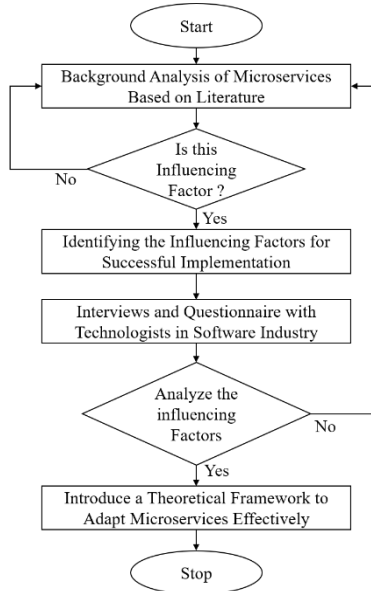


Fig. 1. Experimental design

According to the flow chart in Fig.1, background analysis was carried out to understand key components of the MSA. Then checked whether that is sufficient to influence the solution that going to propose in this study. This cycle was carried out till the background understanding is enough for the solution. Once it is sufficient, further that evidence was confirmed by using interviews and questionnaires. Zoom was used to conduct the interview with SE industry professionals and the survey was delivered as a Google form. This process was repeated until the gathered information is being satisfied to introduce the theoretical framework to adapt to MSA effectively.

IV. RESULT AND DISCUSSION

The interviews with SE professionals are extracted with important information on the use of MSA focusing on the benefits towards the client, and those qualitative data are evaluated based on thematic analysis. Further, the questionnaire is shared among the stakeholders of the software industry to have an overall understanding of their view towards the same goal as in Fig.2 and those quantitative data is analyzed descriptively.

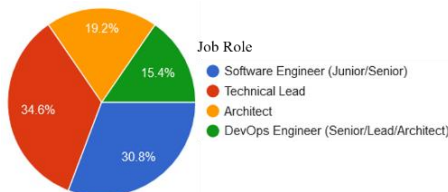


Fig. 2. Contributors for the Questionnaire.

A. Design & implementation requirements

As per the interviews conducted, the following extracts are emphasized to convince the importance of the initial design incorporating the relevant services.

“Representative of the client is a key stakeholder in the software design process” - (SE Professional 1).

The above statement is true once the client is from a non-technical background. However, the level of technical knowledge is reflected on such initiatives as clients can represent themselves throughout the software development lifecycle analysis phase once there is adequate understand of the technology. Such initiatives are positively influenced in addressing the challenges of the MSA implementation.

“Bad designing would cause buying more time for developers”- (SE Professional 2).

Design is important for having a shared understanding between the development team and client from a technical perspective and it streamlines the software engineering development process with clear requirements avoiding reworks. As per the above quotation, it is clear that improper design wastes time due to a poor understanding of the requirements, and it slows down the development process.

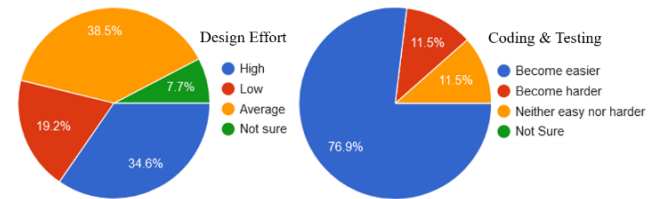


Fig. 3. Design and development phase.

Based on the above understanding, the findings of the survey have also convinced the situation as per Fig.3. 73.1% of responses on design effort are in the average or above level so their primary focus is also on the design. Although there is an extra effort in the designing phase if the industry can manage reusable service repository to reuse the predefined services, it is positively influenced to save more time from coding and testing to deliver true benefit to the client.

B. Security requirements

As per all the interviewees, the required level of security should be achieved via MSA initiatives. The following extract is about the security requirements of the client applications.

“As the services are isolated, securing those are relatively easy but each service should be addressed separately to enrich the level of security” (SE Professional 2).

According to the above statement, the security of each service is assured individually in MSA with the extra effort for the implementation. Eventually, the vulnerability of individual service is not influenced by the others so it is possible to achieve an improved level of security at the end as per the above statement.

Based on the survey findings, Figure 4 also illustrates that better security is achieved in MSA having 80.7% responses on/above the medium level of security. However, the nature of the communication of services by using

messages introduces issues as described in section II, and it reflects here having 11.5% responses for low security.

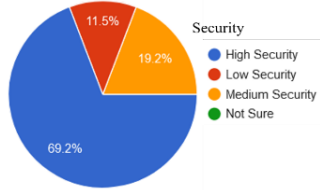


Fig. 4. Security feedback.

C. Deployment container requirements

The requirement for the deployment container is emphasized in the interviews as following.

“Better monitoring strategies should be considered during the deployment with properly planned infrastructure, otherwise, maintenance will be hard.” (SE Professional 3).

As per the statement, it is clear the SE professionals struggle with monitoring, deployment, and infrastructure utilization support provided by MSA influencing the cost factor of the client negatively due to the maintenance. But it also mentions these concepts need to be properly planned, which means there is a way that we can control the above aspect to improve and give a cost-benefit for the client.

Further, the survey extracts the following information as in Fig.5 with respect to the deployment infrastructure, and 73.1% of responses represent on/above average complexity so it is an important finding on the true complexity of the deployment. As a result, deployment complexity should be addressed with proper tools then clients receive the benefit. Further, infrastructure resource utilization is average/above considering 84.7% of responses in that aspect, so client solutions should be finalized with that understandings.

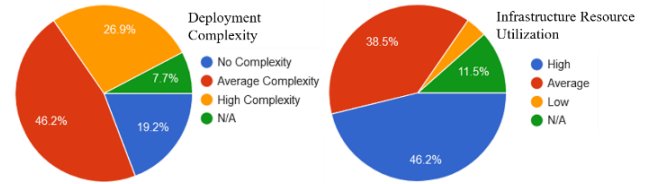


Fig. 5. Deployment Complexity and Infrastructure.

D. Client requirements

It is difficult to judge the client and it extracts per the findings of the interviews as follow

“Client is always worried about the price and quality but not the technology. It depends”

- SE Professional 2

Understanding the above statement is also clearly illustrated in Fig.6 based on survey findings on how industry experts answered their thoughts about the client expectations. Most of the senior leadership accept clients’ most expectation for cost reduction and they do not rely on the underlying technology while some also think infrastructure resource utilization and product quality is equally important.

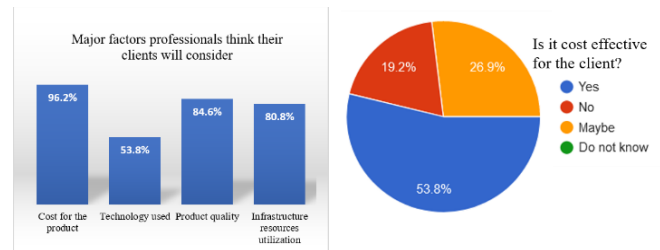


Fig. 6. The perspective of SE professionals about their clients

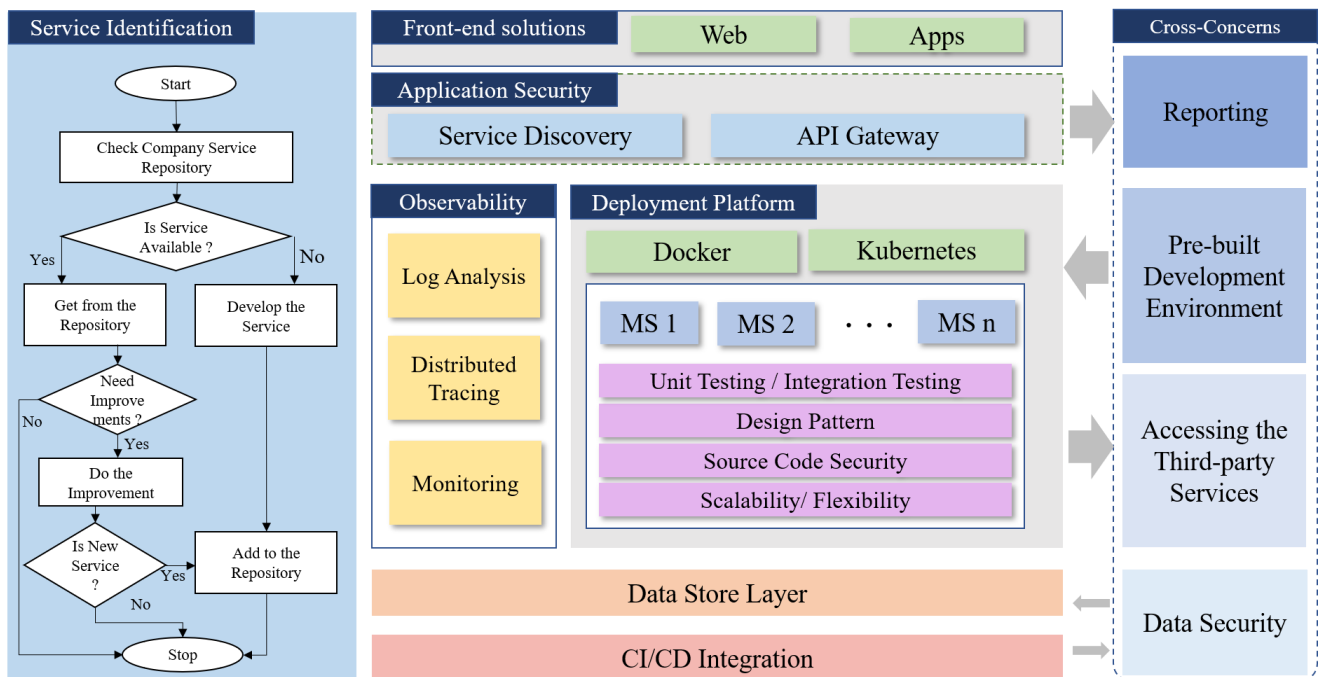


Fig. 7. Theoretical framework for Microservices(MS) developments

As per overall understand, though MSA has challenges, the industry continues on MSA solutions. But those challenges indirectly support to increase in the cost of the projects.

V. RECOMMENDATION

By considering the overall aspects of the use of MSA in different aspects, it is recommended for the organization to follow the basics before moving with MSA addressing the specific requirements of the client in their design of the solution. Incorporating an overall understanding of the findings, a theoretical framework is introduced as a guideline for SE professionals to use for evaluating different possible options available for discussion among all the stakeholders

In Fig.7, the high-level theoretical framework is introduced based on the above literature, survey result, and answers to the interview questions. Further our individual experience is also used when retrieving some components on the introduced framework. SE Professionals can use this framework as a guideline for discussion catering to the exact need of the client appropriately can be adapted. The framework breaks into the following 8 major components; Service Identification, Front-end Solution Layer, Application Security Layer, Observability, Deployment Platform, Data Stored Layer, and CI/CD integration layer. Moreover, the framework is consists of Reporting, a Pre-built Development Environment, Accessing Third Party Services, and Data security.

A. Service identification

As per the interview carried out with industry persons, it is cleared design need get more times and hence developers might facing some issue with delivering implementation on time. Hence Service Identification process is introduced to the theoretical framework so that similarly services can be reuse without spending time on re-developing the same thing. It is a process that an organization should define. Based on the requirement SE professionals need to break down the solution into microservices, once finalized the services that they need to check that defined services are in the organization service repository which is a centralized code management system (e.g. GitLab) and have a full set of functions that microservice can do. So that the few services are utilized from the repository and save the development time. Also identified new services should be developed as reusable components and need to add into a centralized service repository to use by other projects.

Then to speed up development and minimize the re-work pre-build development environment should be available, for example, logging, auditing kind of common concerns should be addressed by developing a library to match with each programming language and need to build into the development environment.

B. Front-end solution layer

This layer consists of applications where the end-user interacts. Web and mobile APP can be considered as main applications and sometimes another backend system may be a front-end application. At a high level, any application or system sending requests to the framework can be considered as a front-end application. So when developing these front-

end applications if the organization can consider this as micro-front ends, it can be reused in various similar needs so that it will reduce time and effort.

C. Application security

Based on the literature review and result of interview answers, it is clear that providing security to each individual microservice is time-consuming work. Hence Application security layer is introduced to the framework so that security can be managed centrally. This layer mainly consists of API gateway and Service discovery. The main function of API gateway is to filter out malicious requests, authenticate and authorize requests before they reach the deployment platform. Also, throttling can be managed from this layer where it can be configured number of concurrent requests allowed for a particular API call. Hence focusing on each individual service's security can be avoided and it will be a huge effort and time-saving for the organization and also benefits can be transmitted into client as well.

Service discovery controls what are the services available in the deployment platform. So, if any service is added to the platform, it will not be visible to the outside (front-end layer) till that service is added to service discovery. So the service discovery is playing a major role to add services into the platform and remove services from the platform and in that way, it will control service level accessibility.

Once this layer is established there is no additional effort to do with each microservice development and deployment so it will help to overcome the drawback of MSA security concerns and finally it saves a lot of money for the organization.

D. Observability

Then the most important part of the framework is to set up the one-time deployment platform and observability. According to the understanding, we gathered from the data analysis it was recognized log analytics and health monitoring of microservice is very important. Observability was added to the theoretical framework to achieve that aspect. There are a lot of open-source tools to configure observability to do log analytics, distributed tracing, and monitoring which includes performance monitoring and stats monitoring. So, when developing the microservices, developers should not worry about the observability and the underlying deployment framework will provide the observability, so that application support after production deployment won't be a hassle anymore and it addresses most of the challenges discussed in the literature. Finally, it benefits the organization in terms of resource and cost.

E. Deployment platform

Although MSA is used to strengthen the solution, one key factor we extracted from the interview is if better monitoring strategies were not accomplice when deploying microservice there can arise maintenance issue. To overcome that deployment platform is introduced with the theoretical framework. The deployment platform consists of Docker, Kubernetes, and MSA design partners like circuit breakers and toolset to support the event sourcing especially to full fill reporting requirements. At a high level, individual microservices deploy in docker containers and these docker containers are managed by Kubernetes. Also, Service mesh

can be introduced to facilitate and manage service to service communication with fault tolerance way. So if an organization can set up this one-time deployment environment, services deployment will be very easy and all the difficulties face once traditional services deployment will be overcome. Further infrastructure wise it will be huge cost saving when it considers the large scale of solution deployments.

F. Data store layer

There should be a unified centralized place to store data related to developed microservices. Data store layer is added to the theoretical framework to have a completeness over the entire solution when developing MSA. In this layer, an organization can define any relational database like MySQL, PostgreSQL, Oracle, or any NoSQL databases like MongoDB. So this database server is centrally managed and needs to create individual databases inside the server to cater to each microservices unique requirements. So in that case developer, no need to worry about the database management part and a dedicated team will be taken care of the data store layer and which will benefit in every means.

G. Cross-concern layer

In this framework, reporting (auditing), pre-built development environment, accessing the third-party services and data security can be considered as cross-concern where this requires in most of the microservices. So, if an organization can develop common frameworks for these items there won't be any repeated tasks be carried out. For example, if a service requires a report, it should be a matter of enabling a flag in the configuration file or annotate a particular function so that it will automatically start to send some events into reporting service. So, with minimum development effort developer will be able to enable a particular feature. Similarly, if an organization can come up with a common implementation that will give more benefit than the traditional way of development while addressing a lot of challenges faced in the practical implementation of MSA.

H. CI/CD integration layer

This layer will reduce the deployment time which was another concern raised by SE professionals. CI/CD defines continuous integration and continuous deployment. So, with this CI/CD implementation from source code development to applications deployment into production can be automated including executing unit tests, integration tests, code quality checks, code security checks, vulnerability checks, penetration testing, etc. To do this, a pipeline needs to be created and configure according to the requirement. Jenkins is a well-known tool for build automation. So once deployed in the production, the service discovery module should be capable of adding a new service into its registry. By automating this complex deployment process it will be a huge cost saving for any organization when working on multiple projects because now you have a centralized deployment platform to deploy and test the services before delivering to the client which will be benefited for the client in terms of the project cost and it also supports over to overcome deployment complexity currently faced by industries.

VI. CONCLUSION

There are a lot of researches carries out about MSA and none of them has introduced proper implementation guidelines. So in the literature review, identifies the features of MSA architecture and also what are the limitations, drawbacks, or challenges involved with them. Also, the conducted survey with a specific set of questions identifies how the industry accepts those challenges. Not only that but also conducted interviews with SE professionals by asking specific questions further implies the challenges they see when implementing with MSA. Based on all the inputs, although there are many benefits associated with MSA, it can be unnecessarily complicated due to the different ways in which it is used and some of its limitations. Proper use of technologies with MSA can alleviate those difficulties. But people in the software industry have different levels of knowledge and they provide solutions according to their point of view. Therefore, in some implementations, it is not possible to get the real benefit of it. But if they have some guidance to adapt, they can minimize the difficulties that arise in SDLC. In this research, proposing a theoretical framework as a solution to address each issue theoretically and which will be easily implemented in the practical world as well. Anyone can use it to upgrade every aspect of their organization's SDLC. It will make both organizations and clients are added benefits in time reduction, cost reduction while giving high-quality software with high maintainability.

REFERENCES

- [1] A. Araujo and H. Moura, "Complexity within Software Development Projects: An Exploratory Overview," 2015. [Online]. Available: <http://lattes.cnpq.br/0902980235660943>
- [2] J. C. Munson and T. M. Khoshgoftaar, "Measuring Dynamic Program Complexity," *IEEE Software*, vol. 9, no. 6, pp. 48–55, 1992, doi: 10.1109/52.168858.
- [3] D. Shadija, M. Rezai, and R. Hill, "Towards an understanding of microservices," Oct. 2017. doi: 10.23919/IConAC.2017.8082018.
- [4] Óbudai Egyetem, IEEE Hungary Section, M. IEEE Systems, Hungarian Fuzzy Association, and Institute of Electrical and Electronics Engineers, 18th IEEE International Symposium on Computational Intelligence and Informatics : proceedings : 2018 November 21-22, Budapest.
- [5] F. Tapia, M. ángel Mora, W. Fuertes, H. Aules, E. Flores, and T. Toulkeridis, "From monolithic systems to microservices: A comparative study of performance," *Applied Sciences (Switzerland)*, vol. 10, no. 17, Sep. 2020, doi: 10.3390/app10175797.
- [6] Z. Racheva, M. Daneva, and A. Herrmann, "A conceptual model of client-driven agile requirements prioritization: Results of a case study," 2010. doi: 10.1145/1852786.1852837.
- [7] N. bin Saif, M. Almohawes, and S. M. Jamail, "The impact of user involvement in software development process," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 354–359, 2021, doi: 10.11591/ijeecs.v21.i1.pp.
- [8] M. V. L. N. Venugopal, "Containerized Microservices architecture," *International Journal of Engineering and Computer Science*, vol. 6, no. 11, Nov. 2017, doi: 10.18535/ijecs/v6i11.20.
- [9] R. de Jesus Martins, R. B. Hecht, E. R. Machado, J. C. Nobre, J. A. Wickboldt, and L. Z. Granville, "Micro-service Based Network Management for Distributed Applications," in *Advances in Intelligent Systems and Computing*, 2020, vol. 1151 AISC, pp. 922–933. doi: 10.1007/978-3-030-44041-1_80.
- [10] R. Boncea, A. Zamfiroiu, and I. Bacivarov, "A scalable architecture for automated monitoring of microservices," 2018. [Online]. Available: <http://www.antonkharenko.com>
- [11] A. de Camargo, R. dos Santos Mello, I. Salvadori, and F. Siqueira, "An Architecture to Automate Performance Tests on Microservices," in *ACM International Conference Proceeding Series*, Nov. 2016, pp. 422–429. doi: 10.1145/3011141.3011179.

- [12] A. D. M. del Esposte, F. Kon, F. M. Costa, and N. Lago, "InterSCity: A scalable microservice-based open source platform for smart cities," in *SMARTGREENS 2017 - Proceedings of the 6th International Conference on Smart Cities and Green ICT Systems*, 2017, pp. 35–46. doi: 10.5220/0006306200350046.
- [13] N. Herzberg, C. Hochreiner, O. Kopp, and J. Lenhard, "Proceedings of the 10th ZEUS Workshop," 2018. [Online]. Available: <https://www.researchgate.net/publication/324517504>
- [14] S. S. de Toledo, A. Martini, and D. I. K. Sjøberg, "Identifying architectural technical debt, principal, and interest in microservices: A multiple-case study," *Journal of Systems and Software*, vol. 177, Jul. 2021, doi: 10.1016/j.jss.2021.110968.
- [15] N. Mateus-Coelho, M. Cruz-Cunha, and L. G. Ferreira, "Security in microservices architectures," in *Procedia Computer Science*, 2021, vol. 181, pp. 1225–1236. doi: 10.1016/j.procs.2021.01.320.
- [16] A. Raj, K. S. Jasmine, and P. G. Student, "Building Microservices with Docker Compose."
- [17] "What Is a Service Mesh? - NGINX." <https://www.nginx.com/blog/what-is-a-service-mesh/> (accessed Jul. 15, 2021).

Challenges for adopting DevOps in information technology projects

J. A. V. M. K. Jayakody*
Department of Computer Science and Informatics
Faculty of Applied Sciences
Uva Wellassa University, Sri Lanka
vihara@uwu.ac.lk

W. M. J. I. Wijayanayake
Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
janaka@kln.ac.lk

Abstract - An Information Technology (IT) project deals with IT infrastructure, information systems, or computers for delivering an IT product within a temporary period. Proper application of software development methodologies assists software designers to run IT projects to the success of achieving the satisfaction of project stakeholders. Because of the issues raised by traditional software development methodologies such as the Waterfall model, the IT industry began to employ Agile methodology for IT project management. However, due to the separation of software development and operation teams, Agile methodology also caused problems. DevOps is a new approach adapted to the Agile methodology that collaborates the software development and operation teams in order to provide continuous development of high-quality software in a short period of time. However, there are practical issues reported since DevOps approach is still in its infancy in the IT industry. The purpose of this research is to analyze the use of the DevOps concept in IT Projects by evaluating the challenges and mitigating strategies practiced by software development firms in order to ensure the success of IT projects. This purpose was achieved by performing a literature study and soliciting recommendations from industry professionals using a questionnaire survey. The findings reveal the critical challenges and prioritization of challenges experienced by software firms while adopting DevOps, as well as their practices for overcoming those challenges. The research findings will help IT project development teams and future researchers to develop strategies for making the success of DevOps adoption with Agile methodology in the IT industry.

Keywords - DevOps, DevOps challenges, overcoming strategies

I. INTRODUCTION

A non-routine complex and single-time effort limited by time, budget, and resources targeted to achieve stakeholder expectations by developing a product or service is considered as a project. [1] The project deals with Information Technology (IT) infrastructure, information systems, or computers considered as an IT project and it produces IT product or service such as software. However, it is not easy for IT project developers to achieve all the expectations of project stakeholders with running projects to the success. There are project failures reported in the IT industry. Proper application of the IT project development principles provides directions to the project managers for their success while reducing the risks which force the project failures. Different types of IT project design and development methodologies provide principles and standards to manage projects for achieving success [1]. When it comes to the IT industry, there are several software

development methodologies are practiced for achieving the success of IT projects.

The traditional and famous software development methodology used by project managers is known as the Waterfall model [2]. It uses a sequential process to develop software. More than the Waterfall model, Iterative Model, Spiral Model, V-Model, Big-Bang Model [3] also used as the software development methodologies. However, IT project developers faced problems with adopting those methodologies since those were not having a flexible development process. Inefficiencies of those methodologies forced the introduction of a new software development methodology that separates the development process into several sprints called Agile methodology. It reduces the problems of previous methods by encouraging adaptive planning, continual improvement, and deliver projects with less time to the customer [4]. However, again IT project managers could find inefficiencies of this Agile methodology. Because it is a developer-centered method than the user-centered [5].

These requirements lead to introduce a new approach to the Agile methodology called DevOps (Development and Operations). It allows development and operations experts to participate together in the entire system development process and now it has become an essential part of the software industry [6]. Theoretically, lots of benefits offered by the DevOps approach along with the Agile software development methodology but there are practical issues reported in the industry. However, this industry experience is not frequently surveyed and reported by researchers since this is an emerging concept [6]. And no more researchers focused to study these challenges and overcoming strategies with comparing literature survey results with the industry experiences. The focus of this study is to analyze the use of the DevOps concept in Information Technology Projects by observing the challenges and mitigating strategies practiced by software development companies while making the success of IT projects. Following research objectives help to achieve the main purpose of the study.

a. Research objectives

- To identify challenges for applying the DevOps approach in IT Project Development.
- To study the mitigating strategies for facing the challenges of DevOps adoption in IT project Development.

These research objectives were achieved by answering the following research questions.

b. *Research questions*

RQ1- What are the challenges experienced by software companies for adopting the DevOps approach in Information Technology projects?

RQ2-What are the mitigating strategies employed by software companies to make the success of Information Technology projects?

I. RELATED WORKS

A project is defined by Kathy Schwalbe as “a temporary endeavor undertaken to create a unique product, service, or result”. It involves a person or a large number of people, complete within a small period or long period, and ends with achieving its predefined target [1]. Among the various types of projects, IT projects develop hardware, software, and networks as the results [1]. And Kathy Schwalbe stated that software development methodologies help IT project developers to improve the efficiency of their project development practices and it is mandatory to earn advantages and goodwill in the competitive market.

Traditional software development methodologies such as the Waterfall method apply a sequential method for developing Software. Therefore, it was not allowed rapid changes and poorly supported to increase the efficiency of the software development process [4]. According to the previous studies, user involvement is important over the IT project development life cycle and those requirements were caused by the origin of Agile methodology [7]. Recent estimates proved that more than 90% of IT companies practice the Agile method for their software developments [5]. However, unsolved problems remained in the IT industry while practicing this Agile methodology. The problems are raised by the lack of cooperation with software operation and development teams [8]. Due to these queries, Agile methodology improved with a new approach called DevOps (Development and IT Operations).

DevOps was defined by Andrej Dyck and Ralf Penners as “an organizational approach that stresses empathy and cross-functional collaboration within and between the development and operation teams in software development organizations” [9]. And DevOps was discussed as a software development method that extends the agile philosophy to rapidly produce software products and services and to improve operations performance and quality assurance by Maximilien De Bayser in 2018 [10]. Not only that, an in-depth case study conducted in an organization which was having several years’ experience in DevOps argues, DevOps leads to great smartness for the Information Systems through the soft skills and pattern of collaboration of the software teams [5]. Similarly, many researchers verified lots of benefits offered by this DevOps approach. Mainly it reduces the project completion time, improves software quality and improves customer satisfaction. But again, some practical issues are reported in the industry with the application of this new concept in Agile methodology. There are few researchers who were focused on this industry experience [6].

Most of the related studies have identified that developing high secured software is a main challenge of the DevOps approach [11], [12], [13]. But no value for the software which is not fixed with high security. This can be

forced by another DevOps challenge reported as the problems in testing practices. According to similar studies, the whole testing process needs to be changed with the adoption of DevOps [14], it consumes more time [15] and difficult to find expertise [16], and also there are no interesting testing tools available [11], since DevOps is an emerging method to the IT industry.

Similarly, recent empirical studies demonstrate that there are no existing guidelines on developing high-quality logging code [17], and it is challenging to achieve transparency on quality delivered by different teams [18], and hard to balance the quality and speed of the software development process [19] [20]. Because DevOps increases the speed of the software development process while reducing the project completion time, it is a challenge to maintain and improve the quality of the software.

Lack of technical infrastructure for adopting DevOps is also identified as a challenge for IT project management by different recent researchers [21] [23]. There are little amounts of tools and technologies available for DevOps and those are very complex and difficult to use [22]. This can be raised by the problems of the IT industry such as; lack of experts on DevOps concept and lack of DevOps knowledge and experience of the people who are working in the DevOps groups. Not only the technical problems, but researchers have emphasized many psychological problems raised by the interconnection of software development and operation teams. They are separated teams and sometimes work in different locations. DevOps is integrating those two groups with reducing the gap between them and it forces on these types of problems. Changing the habits of people is challenging. Resistance to change is recognized by many studies as a challenge to DevOps adoption. [24] [25]. Also, social and cultural changes of the organization and project teams provide barriers to adopt the DevOps approach [26] [27] [28]. Changing the organizational process to DevOps with collaborating different teams is another challenge created by this new approach [29] [30]. According to Jose M Delos, it is difficult to find people who are having good knowledge and experience about the DevOps concept from the industry [33]. And also, the lack of awareness of the project designers and team members about this DevOps approach provides barriers for adopting it with Agile methodology [32] [31]. The same as previous studies mentioned that poor management support for adopting DevOps is a biggest challenge and this can be a reason for the unawareness of project team leaders and managers about the greatest returns of this DevOps approach [12] [20] [26].

Similarly, a study conducted for evaluating the Impact of DevOps Practice in Sri Lankan Software Development Organizations has mentioned that DevOps adoption consists of hidden costs and it raises problems related to the budget [34]. Cost can be increased while absorbing consideration to reduce the project completion time. And most of the similar studies mentioned that it is very difficult to achieve the compatibility between the DevOps approach and legacy systems of the organizations [29].

As DevOps is an emerging concept attached to the Agile methodology in the software development industry, there is a small number of studies focused on this DevOps approach. Therefore, no more researchers focused on the challenges given by this new approach to the success of the software development process and, strategies are practiced

to solve those problems by IT project team members. This study collected challenges for adopting DevOps with Agile methodology that exist in literature. Since not many researchers focused on this area, this study focused to study more about those challenges and practices for mitigating those problems used by IT project teams from the real experience of the IT industry using a questionnaire survey.

II. METHODOLOGY

The research followed a systematic literature review study and a questionnaire survey study to identify the challenges for adopting the DevOps approach, and the strategies that can be used to overcome those challenges for making the IT projects successful. The literature review used to study DevOps challenges identified by similar studies and to perceive the practices utilized by IT projects team members for facing those challenges. The questionnaire survey was used to achieve the research objectives more practically by observing the real-time opinion of IT project development team members about the challenges for adopting DevOps in software development and practices utilized by IT projects team members for facing those challenges.

The literature review study was conducted by a systematic mapping research method. This systematic mapping research method helps to survey the state of the art of research areas that are not yet mature [35]. Search terms formed based on the research questions as “DevOps” AND “Challenges”, “DevOps” AND “Overcoming Strategies”, “DevOps” AND “Evolution” and “DevOps” AND “Software Development Methodologies”. These search terms were used to download relevant and similar publications from the Google Scholar, Emerald Inside, Web of Science, and Google Search Engine to fulfill the research purpose. Then following inclusive and exclusive criteria were used to select more related papers to this study from the downloaded publications.

Inclusion Criteria

- Literature discusses the Software Development Methodologies
- Literature discusses the evolution of DevOps
- Literature discusses the challenges of DevOps adoption in IT projects
- Literature discusses the overcoming strategies of DevOps challenges
- Literature published after the year 2015

Exclusion Criteria

- Literatures not related to the purpose of the study
- Literature published before the year 2015
- Inaccessible literature
- Duplicated literature

Afterward, the title of papers used to identify more related publications to the research objectives, and as the next filter, abstract and keywords of the selected papers helped to screen most related publications from the above-selected list. Finally, the study was conducted by reading the full paper of the most relevant literature which was selected from this systematic approach as shown in Fig. 1. By reading the full text of the most related literature, this study identified DevOps as an approach to the software development practices and filtered challenges of DevOps

adoption for the success of IT projects. As same as it surveyed the mitigating strategies for facing challenges of the DevOps adoption. Finally, it identified most specified challenges and mitigating strategies by similar studies.

As the next step of the study, a descriptive questionnaire survey was used to investigate the actual opinion of IT project team members about the use of DevOps concept for making the success of their project developments. Variables for the survey were defined as DevOps challenges and mitigating strategies for those challenges. The questionnaire used to collect opinion from the industry DevOps practitioners about the survey variables. Those variables were measured using questions which were designed according to the indicators emphasized by the literature review as listed in Table I and Table II. As same as it used to examine the more challenges and practical strategies used for overcoming those challenges that were not focused on by other researchers. This helps to answer the first and second research questions while achieving the research objectives. The questionnaire consists of questions about the background information of respondents, questions to measure respondent’s awareness of the DevOps approach, questions for measuring the respondent’s opinion about DevOps challenges identified by the literature review, and questions to validate strategies suggested by other researchers for overcoming DevOps challenges. The opinion of the DevOps team members who filled the questionnaire was measured by the Likert scale. Finally, those measurements are used to rank the challenges and mitigating strategies. As same as, the questionnaire asked respondents to express the idea about other challenges and mitigating strategies they practically encountered with adopting DevOps. Finally, the research compared literature review results with results of the questionnaire survey and discussed the most important challenges that need to provide attention for making the success of IT projects with adopting to DevOps concept and most practical mitigating strategies can be used to overcome those challenges.



Fig. 1 Approach for selecting related studies

III. ANALYSIS

This section includes an analysis of the literature review and questionnaire survey. Initially, the systematic literature review was conducted through the methodology discussed in the previous section. It began with downloading 192 papers using search terms as “DevOps AND Software Development methodologies”, “DevOps AND Evolution”, “DevOps AND Challenges”, and “DevOps” AND “Overcoming Strategies”. Then 34 papers were eliminated based on the exclusion criteria mentioned in the previous section. As the next step, papers with relevant titles were included in the review list as 98 papers. After that, keywords and abstracts of those selected papers were reviewed and that helped to filter the final set of most relevant 31 studies for the review.

The systematic literature review was conducted by reading full texts of thirty-one selected studies and it could identify many challenges faced by DevOps project team members and also the same challenges have been presented by different authors in various styles. Here all the challenges were listed in one place and categorized into similar groups. Based on that summarization, twelve challenges were identified that faced by IT project management teams for adopting the DevOps approach with the Agile software development methodology. As the next step, the frequency of each challenge identified by other researchers was surveyed by the literature review and finally ranked those challenges according to the above-calculated frequency value. As same as the challenges, the literature review used to identify strategies discussed by other researchers for facing those challenges. However, it could identify only four strategies discussed by other researchers for facing the difficulties of the DevOps adoption since no more researchers focused on this field. The results of this analysis are discussed in the next results and discussion section.

More than the systematic literature review, this research was conducted as a questionnaire survey according to the method discussed in the previous section. Population for the survey was the industrial practitioners who have working experience with the DevOps approach. Sample for the study was selected from the population as 100 industrial practitioners. The online questionnaire was designed using Google forms and distributed to the sample industrial practitioners of the study. Finally, 63 completed answers selected for the analysis.

The questionnaire included main four sections and the first two sections were used to analyze background information about the repliers, such as their age, gender, experience on the DevOps concept, and their opinion about the DevOps adoption. Third section of the questionnaire focused on answering the first research question of the study. It provided a list of the most common challenges filtered from the literature and collected opinions about those challenges from the recipients using the Likert scale. The answers collected from this section were analyzed by ranking the challenges based on the opinion of respondents. Those results compared with the results of literature review and finally identified the most affected challenges to the DevOps practices. Not only that, it collected existing challenges for DevOps which were not focused on by the researchers.

As same as the third section of the questionnaire, the final section also used the Likert scale to evaluate the opinion of

the respondents about suggested strategies for solving the challenges of DevOps adoption. Those strategies also suggested using the results given by the above literature survey. This section provides the answer to the second research question. The answers given by respondents through the Likert scale were used to calculate the rank of each suggested strategy. According to that, this study could suggest the best strategy for facing the challenges given by DevOps adoption to make the success of IT projects. Not only that, the questionnaire was used to collect more practices which are not included in the questionnaire applied by the respondents to solve the problems of DevOps adoption.

IV. RESULTS AND DISCUSSION

Initially, the systematic literature review was used to examine the challenges and overcoming strategies of DevOps adoption identified by the related studies. It was conducted by reading thirty-one related studies and it identified many challenges faced by IT project management teams for adopting the DevOps approach with Agile methodology.

TABLE I. CHALLENGES IDENTIFIED BY LITERATURE REVIEW FOR DEVOPS ADOPTION

No	Challenge	Identifies Literature
C1	Difficult to change the organizational culture for DevOps adoption	[7] [9] [11] [12] [14] [16] [17] [18] [20] [22] [23] [28] [29] [30] [32] [34]
C2	Difficult to find experienced and knowledgeable people to support DevOps practices	[7] [9] [13] [14] [17] [20] [22] [23] [27] [28] [29] [32]
C3	Lack of management support for DevOps adoption	[10] [12] [13] [14] [16] [18] [22] [23] [28] [29] [30] [32]
C4	Difficulties for adopting an organizational process to DevOps	[11] [12] [13] [17] [16] [18] [19] [30] [32] [18]
C5	Difficult to change the habits/ mindsets with DevOps practices	[3] [7] [9] [10] [13] [18] [20] [22] [23] [29]
C6	Difficult to replicate complex technology environments needed for DevOps.	[11] [14] [15] [16] [17] [22] [24] [27] [30] [33]
C7	Difficult to make collaboration of software development and operation teams	[7] [10] [13] [14] [16] [19] [22]
C8	Achieving a secure DevOps development process is challenging	[4] [6] [18] [19] [20] [21] [29]
C9	Difficulties for implement and use DevOps technology	[15] [16] [20] [21] [28]
C10	DevOps increases the complexity of the developing process.	[2] [12] [17] [28]
C11	Difficulties for moving from legacy systems to DevOps tools and techniques	[11] [20] [28] [29]
C12	Project cost can be increased by the DevOps practices	[16][23][29]

Different authors have presented these challenges in different ways. This study mapped identified challenges into main twelve areas and ranked them according to the frequency of each challenge identified by previous studies as shown in Table I. In the same way, those related studies presented some tactics that can be used for facing the challenges and this study mapped those strategies into four main areas with ranking according to the frequency of each strategy identified by other researchers as shown in Table II. The next part of the research used to validate and analyze DevOps adoption challenges and mitigating strategies by collecting feedback from the people who are working in the IT industry and having experience with DevOps adoption. Sixty-three completed feedback could be collected from the hundred people who were selected as sample for the study. Finally, the most important challenges and mitigating strategies were presented and discussed by comparing the results of the literature survey and the questionnaire survey.

TABLE II. MITIGATING STRATEGIES IDENTIFIED BY LITERATURE REVIEW FOR THE CHALLENGES OF DEVOPS ADOPTION

No	Overcoming Strategy	Identifies Literature
C1	Establish communication, platform, procedures, and tools for enhancing communication between software development and operation teams.	[35] [36] [37] [15] [32] [29] [34] [33]
C2	Improve knowledge about DevOps adoption through recent research findings.	[37] [32] [34] [33]
C3	Rearrange the development group to include people who have good experience with DevOps.	[3] [32] [33]
C4	Communicate and celebrate the success of DevOps in the development process.	[3] [34]

The first section of the questionnaire was used to identify the demographic profile of the participants. According to that, most of the respondents were male (94%) and 61% of participants represent their age group between twenty years to thirty years and 36% represent from thirty to forty years. The Education level marked completed the bachelors by 70% and other 19% of the participants have completed the Masters and rest of 11% of the participants have a Diploma.

All respondents were working in the IT industry and 50% of the sample were experienced people in the IT industry for one to five years. And 33% of respondents have worked more than five years in the IT industry and the rest of the 17% also was working in the IT industry from the last year. More than the industry experience, the questionnaire was used to identify their job role in the IT project management team. Sample of the study consists of 20% of project team leaders/managers, 40% of software developers, another 11% of software testers, 13% of software operators, and the rest of other 16% also working as project team members. Some of the responses mentioned that they were working in more than one job role.

Experience on DevOps of the respondents was collected by the second part of the questionnaire and all of them agreed that DevOps provides a good impact on their projects as shown in Fig. 2. This result emphasizes the importance of examining the DevOps approach and it will

be very useful for the software development companies in the IT industry.

Out of sixty-three participants, sixty repliers (94%) practice DevOps in their team. Other three participants also mentioned that they have a good idea about this DevOps practice. Most of the repliers (38%) have experience on DevOps for one to two years and 36% of them were working in the DevOps team from last year. Rest of the responses represent 27% and they have DevOps experience over two years.

Do you think DevOps provide good impact on your work?
 63 responses

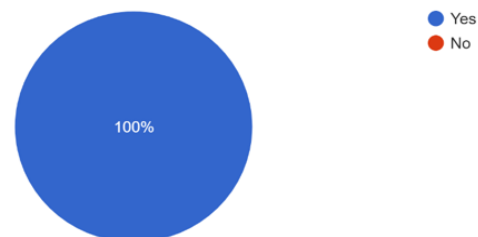


Fig. 2 Impact of the DevOps for making IT projects success

How many years do you have practical experience in DevOps approach?
 63 responses

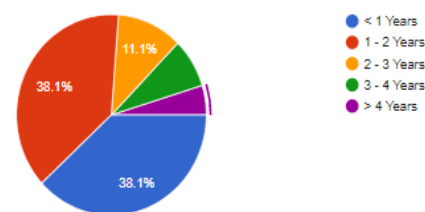


Fig. 3 Experience in DevOps approach of the responses

This DevOps experience of the participants graphically displayed in Fig. 3. However, their DevOps working groups consist of less than ten members for 38% and other 36% were working in the DevOps group which has more than ten members. According to the opinion of the repliers, 48% of them have estimated their understanding of DevOps concept as in “Average” level, and 30% mentioned their knowledge as in “Good” level. Further, 14% of respondents have “Extensive” knowledge about this DevOps adoption while 6% of them don’t have very good ideas. Rest of the respondents (2%) have stated that they are having limited understanding about this DevOps concept as shown in Fig. 4.

Next section of the questionnaire targeted to validate the challenges identified by the literature review with the actual opinion of DevOps practitioners. And to identify more available challenges practiced by the IT project team members while adopting DevOps with Agile methodology which were not focused on by other researchers. The five points Licker’s scale used to measure how those barriers are challenging to them and to identify the most affecting challenges by ranking them according to the overall values of each challenge. The overall level was calculated by multiplying the number of responses for each level of the Likert’s scale by weight for the respective level.

How do you estimate your understanding of DevOps concept?
 63 responses

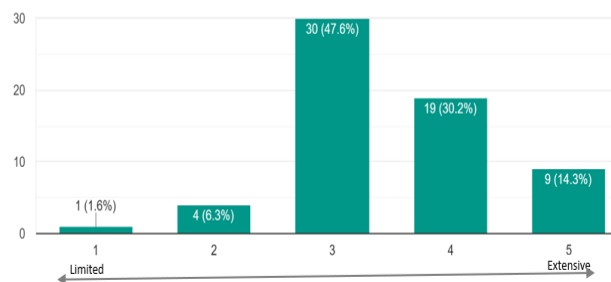


Fig. 4 Response’s knowledge about DevOps

Table III shows the challenges according to their ranks which are calculated by the number of responses for different levels of each challenge.

According to the analysis of results from both literature review and questionnaire survey, changing the organizational culture for adopting the DevOps concept with Agile methodology has become the first ranked challenge faced by IT project management teams. Culture of an organization is the common assumptions practiced by the people who are working in that organization. DevOps collects the software development and operation teams over the project lifetime with changing the culture as working separately. Therefore, it is providing the biggest challenge to adopt DevOps practices to make the success of IT projects.

As claimed by the questionnaire survey, achieving a secure DevOps development process is ranked to the second place and it has ranked to the eighth place in the results of literature review. A main target of DevOps is to reduce the project completion time. Security can be reduced while providing more attention to decrease the project completion time. Therefore, it has become a very important challenge from the practical opinion of the industry people. But this is not captured by many researchers who study the DevOps challenges. According to the literature review study, the second important challenge is the difficulty to find experienced and knowledgeable people to support DevOps practices. This problem can be raised because this concept is emerging in the industry. Responders for the questionnaire also marked this as an important challenge and they have raised this problem into the fifth step of the challenges list.

Managers and Leaders’ support for making the success of DevOps adoption is less based on this study. Most of the researchers have emphasized this problem and it is the third challenge identified by the literature review. When it comes to the practical opinion of DevOps team members, they also mentioned it as an important challenge and it ranked to seventh place of the challenges list. Most of the managers and other IT project team members are not aware about the DevOps concept since it is new to the IT industry. It is proved by the feedbacks of the questionnaire survey as shown in Fig. 4, nearly half (47.6%) of the participants marked their knowledge about the DevOps as “Average”. Therefore, they don’t have a good idea about the benefits of DevOps and not motivated to adopt DevOps advantages to their organizations.

According to the questionnaire survey, the third important challenge is to achieve compatibility between DevOps and legacy systems.

TABLE III. CHALLENGES IDENTIFIED BY QUESTIONNAIRE SURVEY FOR DEVOPS ADOPTION

No	Challenge	No of responses					Overall value
		1	2	3	4	5	
Weight		1	0.75	0.5	0.25	0	
01	Changing deep-seated company culture to support DevOps adoption is challenging.	16	12	17	15	4	33.5
02	Achieving a secure DevOps development process is challenging.	7	27	13	9	8	33.5
03	It is challenging to achieve compatibility between DevOps and legacy systems.	11	12	24	14	3	32.5
04	Adapt organizational processes to DevOps is challenging.	8	19	13	20	4	30.7
05	Difficult to find experienced professionals to support DevOps practice.	13	11	17	16	7	30.5
06	Difficult to replicate complex technology environments needed for DevOps.	6	18	18	18	4	30.5
07	Difficult to obtain management support for DevOps practices.	8	11	17	24	4	27.7
08	There are hidden costs associated with DevOps adoption.	6	13	22	16	7	27.5
09	DevOps increases the complexity of the developing process.	8	11	16	12	17	26
10	Difficult to make the collaboration between Development and Operations.	4	12	18	21	9	25.2
11	Hard to adapt mindsets to achieve successful DevOps.	4	13	12	29	6	25
12	Hard to implement and use DevOps technology.	3	12	20	22	7	24.2

Because the DevOps approach changes the whole process of the project completion. It is required to have the best idea about the DevOps adoption and legacy systems for facing this challenge. Therefore, feedback from the industry survey ranked this to the third point but this is not focused on by many researchers like other challenges and this is the eleventh challenge in the literature review challenge list. However, this is a very considerable challenge for the adoption of the DevOps approach to make the project's success and it is a good area to research for future researchers.

Problems with adapting organizational processes to DevOps also recognized by many of the previous studies and also responders of the questionnaire survey have marked this as an important challenge. Whole organizational process is changed while adopting DevOps and it converts the organizational hierarchy. Because of that, this problem has become the fourth challenge of both DevOps challenge lists. As same, the fifth challenge identified by the literature review is difficulties for changing habits of team members and their mindsets. It is very difficult to change human habits without improving their motivation. Therefore, it is very important to motivate IT project team members by informing them about the advantages of DevOps adoption to make projects successful and their improvements. However, this challenge is not in the top list of the challenges of the questionnaire survey.

The sixth rank of both challenge lists marked as difficult to replicate complex technology environments needed for DevOps. Here many researchers identified DevOps technology as complex and respondents for the survey also agreed with that problem. The next important focus is the project cost. Small number of literatures have identified increasing project cost with DevOps adoption as a challenge and it has become the last problem of the challenges identified by literature. But from the actual opinion of project team members, this challenge was raised into the eighth place of the challenge list of questionnaire surveys. It is very challenging to balance the time, quality, and cost of the IT projects. Those three are the triple constraints of projects. Project cost can be increased while reducing the project completion time by DevOps.

However, challenges identified by the literature review were proved by the questionnaire survey and most of the psychological challenges were identified by many previous researchers. That reason raised those challenges into the top list of the challenges identified by the literature review study. Other than the challenges stated in the other similar research papers, responders for the questionnaire have mentioned more practical challenges they face with adopting DevOps as follows;

- DevOps using lots of tools and too much focus on tools
- Moving from legacy infrastructure to microservices is challenging.
- Implementation of DevOps for projects based and product based companies is difficult.
- Sometimes DevOps might be overhead.
- Challenges come from the technology changes.

The questionnaire survey was also used to identify the methods that IT project team members use to manage the above-mentioned challenges. The questionnaire survey analysis suggested four strategies that can be used to solve DevOps challenges which were identified by the initial literature survey and those methods are shown in Table IV. To meet the challenges posed by DevOps adoption, the majority of them establish communication, platforms, procedures, and tools to improve communication between software development and operation teams. Therefore, in both questionnaire surveys and literature study, this option has become the first strategy for dealing with the challenges of DevOps adoption. According to the literature survey, the second strategy is to improve knowledge about DevOps adoption through recent research findings, and the third strategy is to rearrange the development group by including people who have good experience with DevOps. However, according to the results of the survey, all of these options are used by IT project team members, and they didn't mention them as those are used frequently. In addition to above mentioned suggestions, they have answered the questionnaire by providing following practices for dealing with DevOps challenges as follows;

- Use DevOps framework as CALMS (Culture, Automation, Lean, Measurement, and Sharing)
- First, define a specific development flow for the team and the product. Then, for each operation point, identify experts and later synthesize their knowledge into a single document (diagram)

However, small sample size is recognized as a limitation of the questionnaire survey and the results can be further generalized by improving the sample size.

TABLE IV. STRATEGIES FOR OVERCOMING DEVOPS CHALLENGES

No	Overcoming Strategy	Number of responses			Over all value
		1	2	3	
Weight		1	0.5	0	
01	Establish communication, platform, procedures, and tools for enhancing communication between software development and operation teams.	37	19	8	55.25
02	Communicate and celebrate the success of DevOps in the development process.	33	24	7	54.5
03	Rearrange the development group to include people who have good experience with DevOps.	30	27	7	53.75
04	Improve knowledge about DevOps adoption through recent research findings.	23	29	12	50.75

V. CONCLUSION

This research examined the adoption of DevOps concept in IT Projects by evaluating the challenges and mitigating strategies practiced by software development firms to ensure the success of IT projects. The research purpose was accomplished by obtaining responses for two research questions as "what are the challenges experienced

by software firms in adopting the DevOps approach in Information Technology projects?" as well as "what are the mitigating strategies employed by software firms to ensure the success of IT projects?". A comprehensive literature review and a questionnaire survey were used to answer the questions. Results of the literature review and responses of the questionnaire survey were utilized to rank the challenges and mitigating strategies, and the rankings of both studies were compared. It answered the first research question, while the second question was answered by assessing the identified mitigating strategies practiced by the IT project teams using the questionnaire survey. Based on this, twelve major hurdles for adopting DevOps were identified and changing deep-seated company culture to support DevOps adoption is identified as the first ranked challenge in both studies. According to the frequency of attention by comparable researchers, the second and third-ranked challenges are, difficult to hire experienced and knowledgeable people to support DevOps practices and the lack of management support for DevOps adoption. According to the DevOps practitioners, the second and third most important challenges are ensuring a secure DevOps development process and achieving compatibility between DevOps and legacy systems respectively. Moreover, it also revealed unfocused challenges experienced by IT project teams. The survey analyzed and ranked not only obstacles, but also mitigation techniques employed to tackle the issues. It revealed that, many IT project teams use the way of establishing communication, platform process, procedures, and tools for enhancing communication between software development and operation teams to lessen the problems of DevOps adoption. The questionnaire survey yielded further recommendations for ensuring the success of IT projects using the DevOps approach. These findings assist future researchers in developing a conceptual model for the critical success factors of DevOps and validate the conceptual model using primary data in order to reap the benefits of DevOps approach while reducing the hurdles associated with using DevOps in Agile methodology for enhancing the success of IT projects.

REFERENCES

- [1] K. Schwalbe, Information Technology Project Management, United States of America: Cengage Learning, 2016.
- [2] Mariela Stoyanova, "SMART CONCEPT FOR PROJECT MANAGEMENT – TRANSITION TO DevOps," KNOWLEDGE – International Journal, pp. 93-97, 2019.
- [3] Pulasthi Perera, Roshali Silva, Indika Perera, "Improve Software Quality through Practicing DevOps," IEEE, pp. 13-18, 2017.
- [4] David Bishop, D.Sc. , Pam Rowland, Cherie Noteboom, Ph.D., "Antecedents of Preference for Agile Methods: A Project Manager Perspective," in 51st Hawaii International Conference on System Sciences, 2018.
- [5] A. Hemon-Hildgen, Barbara Lyonnet, Frantz Rowe, Brian Fitzgerald, "From Agile to DevOps: Smart Skills and Collaborations," Information Systems Frontiers, 2019.
- [6] Alok Mishra, Ziadon Otaiwi, "DevOps and software quality: A systematic mapping," Elsevier Inc., 2020.
- [7] S. M. Mohammad, "DevOps automation and Agile methodology," SSRN Electronic Journal, pp. 946-949, 2017.
- [8] L. E. Lwakatara, "DevOps Adoption and Implementation in Software Development Practice," University of OULU, 2017.
- [9] Andrej Dyck, Ralf Penners, Horst Lichter, "Towards Definitions for Release Engineering and DevOps," in IEEE/ACM 3rd International Workshop on Release Engineering, 2015.
- [10] Maximilien De Bayser, Leonardo Guerreiro Azevedo, Renato Cerqueira, "ResearchOps: The case for DevOps in scientific applications," Research Gate, 21 April 2018.
- [11] Saima Rafi, Muhammad Azeem Akbar, Wu Yu, "Towards a Hypothetical Framework to Secure DevOps Adoption: Grounded Theory Approach," 2020.
- [12] Mahmoud Sheyyab, "Managing Quality Assurance Challenges of DevOps through Analytics," 2019.
- [13] Asif Qumer Gill, Abhishek Loumish, Isha Riyat, Sungyoun Han, "DevOps for Information Management Systems," VINE Journal of Information and Knowledge Management Systems, 2018.
- [14] Abubaker Wahaballa, Osman Wahballaa, Majdi Abdellatie , Hu Xiong, Zhiguang Qina, "Toward Unified DevOps Model," 2015.
- [15] Kati Kuusinen, Admir Muric, "A Large Agile Organization on Its Journey Towards DevOps," in 2018 44th Euromicro Conference on Software Engineering and Advanced Applications, 2018.
- [16] Anna Wiedemann, Nicole Forsgern, Manuel Wiesche, "The DevOps Phenomenon," 2019.
- [17] Boyuan Chen, "Improving the software logging practices in DevOps," in IEEE/ACM 41st International Conference on Software Engineering, Canada, 2019.
- [18] Masud Fazal-Baqaie, Baris Guldali, Simon Oberthur, "Towards DevOps in Multi-provider Projects," Workshop on Continuous Software Engineering, 2017.
- [19] LE Lwakatara, T Kilamo, T Karvonen, T Sauvola, V Heikkilac , J Itkonen, P Kuvaja, T Mikkonen, M Oivo, C Lassenius, "DevOps in practice: A multiple case study of five companies," Information and Software Technology, 2019.
- [20] L Leite, Carla Rocha, Fabio Kon, "A Survey of DevOps Concepts and Challenges," ACM Computing Surveys, 2019.
- [21] M. Munoz, M. Negrete, J. Mejia, "Proposal to avoid issues in the DevOps implementation: A Systematic Literature Review," 2019.
- [22] Mayank Gokarna, Raju Singh, "DevOps: A Historical Review and Future Works," 2020.
- [23] Saima Rafi, Wu Yu, M A Akbar, "Towards a Hypothetical Framework to Secure DevOps Adoption: Grounded Theory Approach," 2020.
- [24] Liang Yu, Clemente Guerra, "Exploring the disruptive power of adopting DevOps for software development," 2020.
- [25] Asif Q Gill, A Loumish, Isha Riyat, Sungyoun Han, "DevOps for Information Management Systems," 2019.
- [26] Stephen Jones, Joost Noppen, Fiona Lettice, "Management Challenges for DevOps Adoption within UK SMEs," 2016.
- [27] Pulasthi Perera, Roshali Silva, Indika Perera, "Improve Software Quality through Practicing DevOps," 2017.
- [28] Morgan Rowse, Jason Cohen, "A Survey of DevOps in the South African Software Context," in 54th Hawaii International Conference on System Sciences, 2021.
- [29] Welder Luz, Gustavo Pinto, Rodrigo Bonifacio, "Adopting DevOps in the Real World: A Theory, a Model, and a Case Study," 2019.
- [30] Breno B. N. de França, Helvio J. Junior, Guilherme H. Travassos, "Characterizing DevOps by Hearing Multiple Voices," 2016.
- [31] LE Lwakatara, T Karvonen, T Sauvola, P Kuvaja, H Olsson, Jan Bosch, Markku Oivo, "Towards DevOps in the Embedded Systems Domain: Why is It so Hard?," in 49th Hawaii International Conference on System Sciences, 2016.
- [32] Ineta Bucena, Marite Kirikova, "Simplifying the DevOps Adoption Process," 2019.
- [33] Jose Maria Delos, "The Definitive Guide to DevOps," 2021.
- [34] Pulasthi Perera, Madhushi Bandara, Indika Perera, "Evaluating the Impact of DevOps Practice in Sri Lankan Software Development Organizations," in International Conference on Advances in ICT for Emerging Regions, 2016.
- [35] Vaishnavi Mohan, Lotfi ben Othmane, "SecDevOps: Is It a Marketing Buzzword? (Mapping Research on Security in DevOps)," IEEE, 2016.
- [36] Havard Myrbakken, Ricardo Colomo-Palacios, "DevSecOps: A Multivocal Literature Review," 2018.
- [37] L.R. Kalliosaari, S. Mäkinen, L.E. Lwakatara, J. Tiihonen, T. Mannisto, "DevOps Adoption Benefits and Challenges in Practice: A Case Study," p. 590–597, 2016.

Modelling and validation of arc-fault currents under resistive and inductive loads

Yashodha Karunarathna*
Department of Electrical and Electronics Engineering
University of Peradeniya, Sri Lanka
ykarunarathna@gmail.com

Janaka Ekanayake
Department of Electrical and Electronics Engineering
University of Peradeniya, Sri Lanka
ekanayakej@eng.pdn.ac.lk

Janaka Wijayakulasooriya
Department of Electrical and Electronics Engineering
University of Peradeniya, Sri Lanka
jan@ee.pdn.ac.lk

Pasindu Perera
Division of Research and Development
Sri Lanka Telecom PLC, Sri Lanka
pasindu@slt.com.lk

Abstract - Over half of all electrical fires in installations are caused by arcing due to poorly connected equipment or wiring system failures. Therefore, it is essential to detect arcs and interrupt them using a suitable protective device. This paper provides a modelling simulation and experimental approach to obtain arc voltage and current. The parameters for the theoretical model were turned based on the experimental results. A realistic case study was done to obtain the arc current under parallel and series arcs. As seen from the results, a parallel arc creates a current much higher than the load current, whereas a series arc current is often lower than the load current. Even though a parallel arc current may be detected by an overcurrent device, as it is often intermittent, it may not sustain to be captured by existing protection devices. Therefore, both parallel and series arc detection and interruption demand a reliable protection device.

Keywords - arc current generator, - arc fault, arc detection

I. INTRODUCTION

Arc-flash incidents occur every day in many electrical installations. Arcs are visible plasma discharges caused by electrical current passing through a normally non-conductive medium, such as air. This is caused when the electrical current ionizes gases in the air. Fault arc is often followed by the partial evaporation of conductor material. Such an action in the conductor could cause an inflammation in the insulation and as a result, could lead to a fire. The most common causes of arcs are known to be worn contacts in electrical equipment, damage to insulation, kinks in a cable, cable damage caused by drilling or building work, loose-bolted connections, and defective wall plugs. It can also be generated by dropping tools, opening panels on damaged equipment, inserting or removing components from an electrified system, or even a rodent infestation. Although the conventional circuit breaker gives protection from overcurrent and Earth leakage current, they are not effective in protecting from dangerous arcs. The Arc Fault Circuit Interrupter (AFCI) is designed to analyse noise in the current signal, typically at 100 kHz to respond fast enough to detect and break the circuit before causing a fire. To design an effective AFCI, it is important to model the arc current and voltage under many different operational possibilities and then use signal processing techniques to gather the signature of the arc current and voltage.

In the literature, many techniques are reported for obtaining and analyzing arc signals. SeJi, Kim, and Kil [1] have implemented the phase analyses of series arc signals for low-voltage electrical devices such as heaters,

computers, refrigerators, and air conditioners. The arc generator has been fabricated according to the UL6199 standard [2]. The phase of detected series arc signals has been analyzed according to load types and finally, a new algorithm was proposed based on the result of phase-resolved series arc analysis to identify types of loads. Taufik, Aarstad, and Kean [3] have presented the development of AFCI lab setup to characterize dc arc current in dc circuits operating at 24-80 V. Different scenarios for dc arcing occurrences in the development of the lab test setup have been explained. Several test results using the developed test setup have been presented to show the characteristics of the arc current. By inspecting the frequency spectrum of arc current, a unique signature of the dc arc was identified. Andrea, Besdel, Zirn, and Bournat [4] present a mathematical model based on circuit components to describe the behavior of the electric arc in static and dynamic situations. Simulation results and experimental results are given for common arc ignition cases. Mahajan, Patil, and Shembekar [5] discussed the modelling and simulation of the arc phenomenon using the Mayr arc model. Ghezzi and Balestrero [6] discuss different Black box, arc models. Simulations and experimental results are compared under different arcing cases. The parameter estimation for different models is also presented. Even though these studies present modelling and model validation under different arcing characteristics (voltage, phase, V-I), none of them provides a comparison of arcing current with the load current under different loading conditions. Therefore, in this paper, an attempt was made to make a comparison between the arcing current and load currents under real-world scenarios.

II. MODELING APPROACH

Static characteristic of an arc is shown in Figure 1. A to B is a discharge phase and the discharge can be called corona discharge. The arc is extinguished at O and in the second phase, the voltage is reversed. When the reverse voltage reaches the restrike voltage (at C), the discharge re-initiate. Two resistances can be identified: the resistance of the arc ignition time, R_c , and the resistance when arc discharge, R_{arc} . Many references [7,8,9] are providing "Black box" models that describe an arc by a simple mathematical equation and give the relation between measurable parameters such as arc voltage and arc current. Such an equation is given in equation (1) [4] and it is used for modelling the arc in this paper.

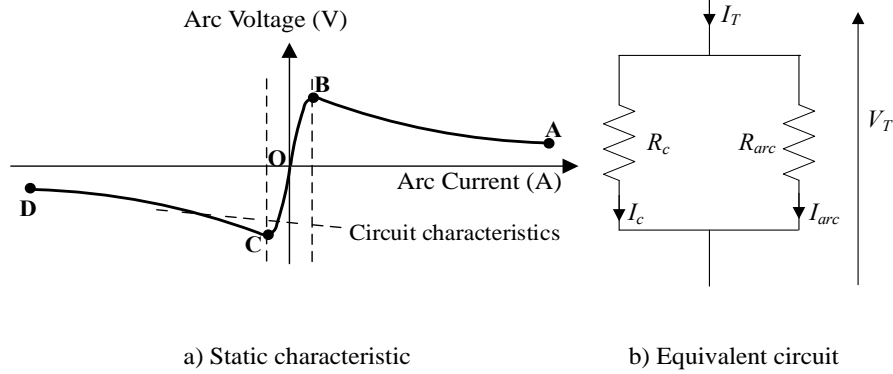


Fig.1. Static characteristic and equivalent circuit of an arc

$$V_{arc} = \frac{\alpha}{\arctan(\beta I_{arc})} \quad (1)$$

Where, α is a linear function of the arc length and β is a fit parameter depending on the material of the electrode.

With the relationship, $V_{arc} = I_{arc}R_{arc}$ and from (1), R_{arc} can be found as:

$$R_{arc} = \frac{\alpha}{\arctan(\beta I_{arc})I_{arc}} \quad (2)$$

The equivalent circuit to represent the static characteristic is shown in Figure 1(b). When the arc current is low, i.e from B to C, from equation (2), the arc resistance, R_{arc} , is high and the parallel combination is more or less equal to R_c . During the period A to B and C to D, since the current passing through R_c is negligible when compared to I_{arc} the parallel combination can be reduced to R_{arc} only.

Therefore, the overall discharge resistance R_T , i.e.

$R_T = \frac{R_c R_{Arc}}{R_c + R_{Arc}}$ was found by substituting from (2) as

$$R_T = \frac{\alpha R_c}{\arctan(\beta I_{arc})I_{arc}R_c + \alpha} \quad (3)$$

Then from Ohm's law

$$V_T = \frac{\alpha R_c I_T}{\arctan(\beta I_{arc})I_{arc}R_c + \alpha} \quad (4)$$

Due to the arc resistance R_{arc} is considerably low in comparison to the resistance R_c , $I_{arc} \gg I_c$ and $I_T \approx I_{arc}$. Therefore, equation (4) was replaced by

$$V_T = \frac{\alpha R_c I_T}{\arctan(\beta I_T)I_T R_c + \alpha} \quad (5)$$

With a function F that describes the static discharge, equation (5) was written as

$$V_T = F(I_T)$$

For an R-L circuit when an arc occurs in series, the circuit equation was written as:

$$\begin{aligned} V_g(t) &= RI_T(t) + L \frac{dI_T(t)}{dt} + V_T \\ &= RI_T(t) + L \frac{dI_T(t)}{dt} + F(I_T(t)) \end{aligned} \quad (6)$$

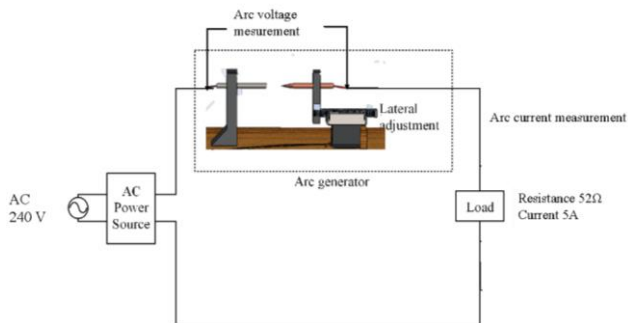
The load equation defined by the first two terms of equation (6) may cross the static characteristic in one, two or three points (example case is shown in Figure 1(a)). When solving for the current, one of the possible cross points as the solution was obtained using the least efforts principle [4]. The differential equation of $I_T(t)$ was solved using MATLAB to obtain time plots of arc voltage and current under different loading conditions.

In this experiment, only the series arc was modelled because it is the one type of arc that is not interrupted by existing protection devices as the arc current flowing in the circuit is not higher than the load current while it is also being limited by the load connected in series. In contrast, a parallel arc occurs between conductors within different phases such as line to neutral or line to ground. Since the parallel arc current is higher than load current, it can be detected without any advanced techniques. Corresponding graphs are shown in the results section.

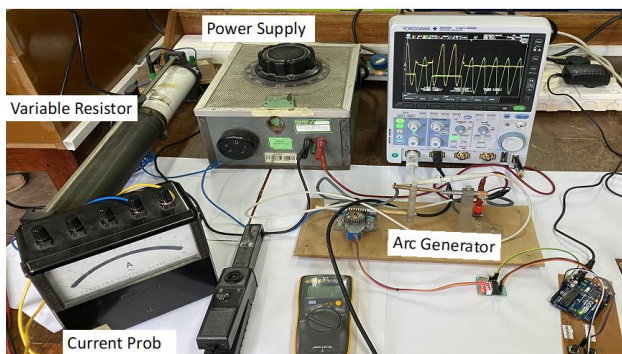
III. ARC GENERATOR

An arc generator was designed in compliance with the standards BS EN 62606:2013+A1:2017 [2] with an apparatus consisting of a stationary electrode and a moving electrode. One electrode was made using a 6mm \pm 0.5 mm diameter carbon-graphite rod and the other electrode was a copper rod as shown in figure 2. The arcing end of one carbon electrode was pointed. The distance between the two electrodes was adjusted by controlling a stepper motor. An Arduino-based controller was designed for this purpose. The arcing current was sensed by a current probe whereas arcing voltage was directly probed by the oscilloscope. Figure 2(a) shows the schematic with the

circuit connections and Figure 2(b) shows the laboratory setup used.



(a) Schematic diagram



(b) Laboratory setup

Fig. 2: Arc generator

IV. PARAMETER ESTIMATION OF THE MODEL

A MATLAB model for simulation of an electric Arc in a circuit was developed as per the theoretical model discussed in the previous section. According to [4], $\alpha = 49.0874$, $\beta = 1.4614$, and $R_c = 2221\Omega$ were chosen as model parameters. These parameters were chosen by a curve fitting method and the reference does not provide any information about the experimental setup. To make the model compatible with the experimental study, the above parameters were manually tuned and found to be $\alpha = 6$, $\beta = 0.15$, and $R_c = 55\Omega$.

The external resistance of the experimental setup was approximately 40Ω and inductance was chosen as $10\mu\text{H}$. These values were used in the model.

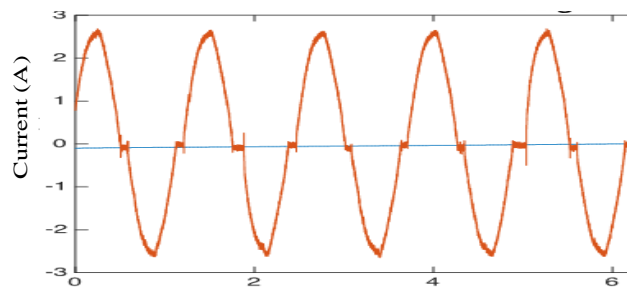
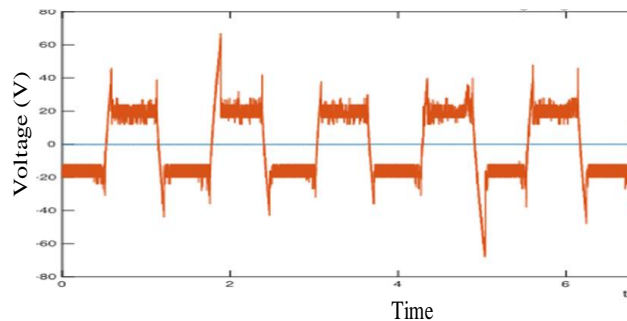


Fig. 3. Simulated arc voltage (top) and current (bottom) ($R=40\Omega$, $L=10\mu\text{H}$)

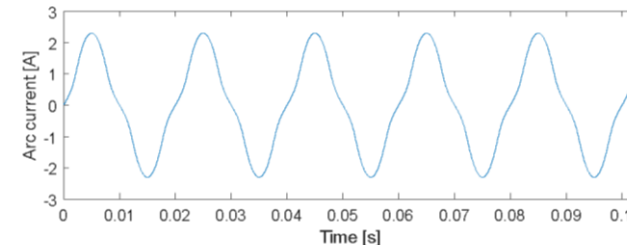
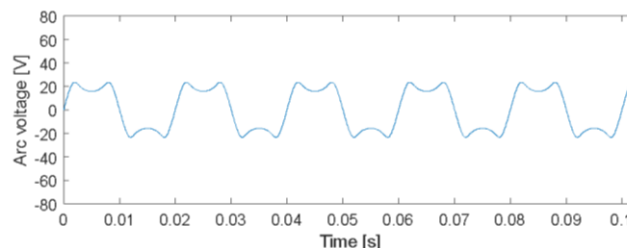


Fig. 4. Experimentally obtained arc voltage (top) and current (bottom)

V. CASE STUDY

Fig. 5 shows the connection from the distribution transformer to a house and a plug socket within the house. Data of different cable sections are given in Table I.

TABLE I. PARAMETERS OF THE CABLE AND TRANSFORMER

	Resistance (Ω/km)	Reactance (Ω)
Transformer leakage reactance	Negligible	0.1
Fly conductor	0.47	0.27
Service cable	1.83	Negligible
Live wire	13.6	Negligible

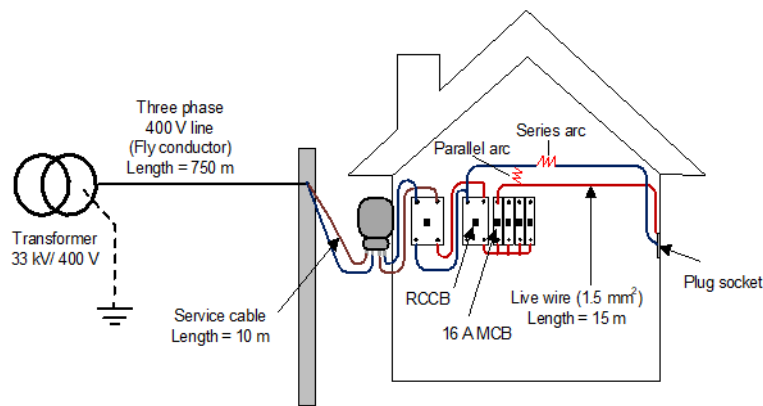


Fig. 5. Sample house for case study

An arc has been formed due to damage to a cable between the live and neutral wires (parallel arc) or due to a loose connection in series with the live wire (series arc).

Fig. 6 and Fig. 7 show the load current and the current when a series arc prevails in the circuit for a 2 kW kettle (a resistive load) and a 2 kW microwave oven (an inductive load) respectively. These appliances are connected to the plug socket shown and it was assumed that the voltage at the transformer is 230 V. As can be seen when the arc is initiated, the current drops from the normal current.

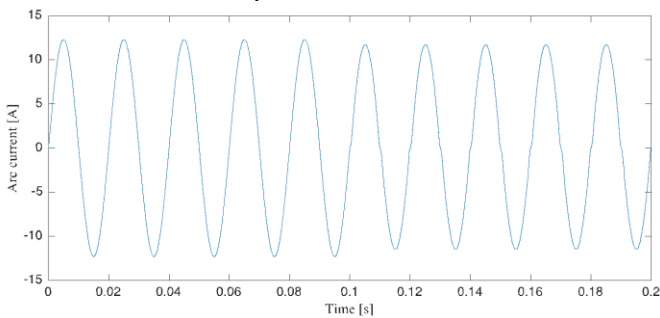


Fig. 6. Arc current when a series arc prevails in the circuit for a 2 kW kettle

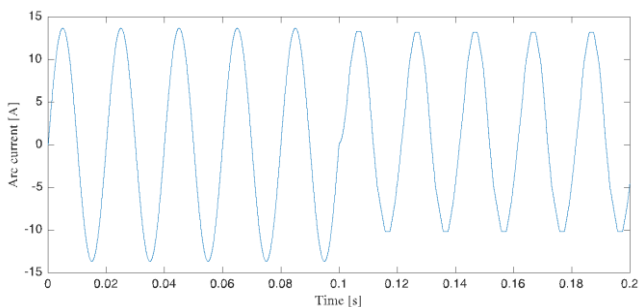


Fig. 7. Arc current when a series arc prevails in the circuit for a 2 kW microwave Oven

Fig. 8 and Fig. 9 show the load current and the current when a parallel arc prevails in the circuit for a 2 kW kettle and a 2 kW microwave oven respectively.

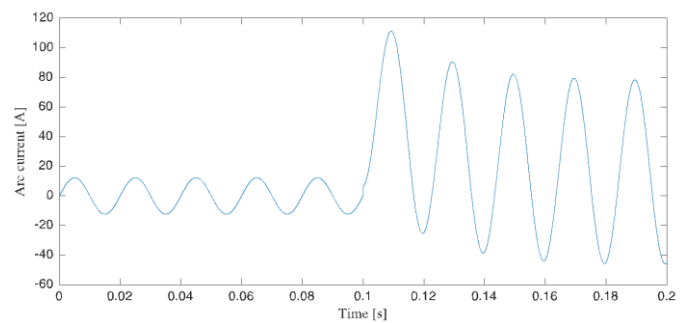


Fig. 8. Arc current when a parallel arc prevails in the circuit for a 2kW kettle

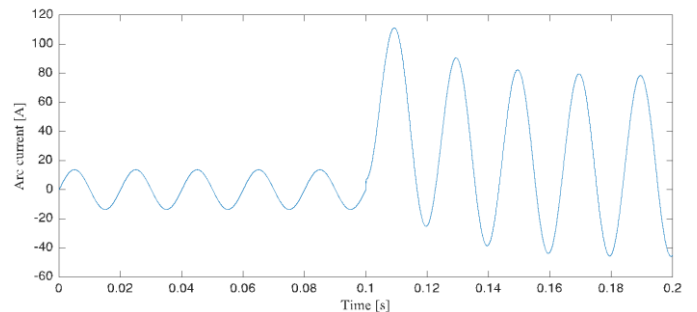


Fig. 9. Arc current when a parallel arc prevails in the circuit for a 2 kW microwave Oven

VI. CONCLUSION

Arc currents can be originated in electrical installations due to many reasons. A sustained arc can damage the installation and even lead to a fire. As shown in this paper, an arc current created between the live and neutral conductors (a parallel arc) results in a large current excursion, whereas a series arc created by a loose connection results in a current lower than the load current. Even under a parallel arc, the arcing may be intermittent and therefore will not be detected by an over-current or surge protective devices installed in a premise. Therefore, a specially designed protective device should be connected to installations to detect arc and prevent any adverse circumstances.

REFERENCES

- [1] H.-K. Ji, S.-W. Kim, and G.-S. Kil, "Phase Analysis of Series Arc Signals for Low-Voltage Electrical Devices," *Energies*, vol. 13, no. 20, p. 5481, Oct. 2020.
- [2] Cen.eu.2021. [online]<://www.en-standard.eu/bs-en-62606-2013-a1-2017-generalrequirements-for-arc-fault-detection-devices/>[Accessed 15 June 2021]
- [3] T. Taufik, C. Aarstad and A. Kean, "Arc Fault Characterization System for the Low Voltage DC Arc Fault Circuit Interrupter," 2017 25th International Conference on Systems Engineering (ICSEng), 2017, pp. 106-112, doi: 10.1109/ICSEng.2017.36.
- [4] J. Andrea, P. Besdel, O. Zirn and M. Bournat, "The electric arc as a circuit component," *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 2015, pp. 003027-003034, doi: 10.1109/IECON.2015.7392564.
- [5] N.S. Mahajan K. R.Patil and S.MShembekar., "Electric Arc model for High Voltage Circuit Breakers Based on MATLAB/SIMULINK". *Interantional Journal of Science, Spirituality, Business and Technology (IJSSBT)*, 2013, pp.1(2).
- [6] A. Balestrero, L. Ghezzi, M. Popov, G. Tribulato and L. van der Sluis, "Black Box Modeling of Low-Voltage Circuit Breakers," in *IEEE Transactions on Power Delivery*, vol. 25, no. 4, pp. 2481-2488, Oct. 2010, doi: 10.1109/TPWRD.2010.2047872.
- [7] G. Bizjak, P. Zunko and D. Povh, "Circuit breaker model for digital simulation based on Mayr's and Cassie's differential arc equations," in *IEEE Transactions on Power Delivery*, vol. 10, no. 3, pp. 1310-1315, July 1995, doi: 10.1109/61.400910.
- [8] S. Nitu, C. Nitu and P. Anghelita, "Electric Arc Model, for High Power Interrupters," *EUROCON 2005 - The International Conference on "Computer as a Tool"*, 2005, pp. 1442-1445, doi: 10.1109/EURCON.2005.1630234.
- [9] Yuan, Ling, Lin Sun, and Huaren Wu. "Simulation of fault arc using conventional arc models." *Energy and Power Engineering* 5.04 (2013): 833-837.

Decision-making models for a resilient supply chain in FMCG companies during a pandemic: A systematic literature review

B. R. H. Madhavi*
Department of Industrial Management
University of Kelaniya, Sri Lanka
balasurih_im16042@stu.kln.ac.lk

Ruwan Wickramarachchi
Department of Industrial Management
University of Kelaniya, Sri Lanka
ruwan@kln.ac.lk

Abstract - Decision-making during a crisis impacts the performance of an entire organization. Due to the COVID-19 pandemic, many organizations had undergone supply chain disruptions due to the forward and backward propagation of disruptions in the global supply chain networks, implying the importance of building up resilience in the supply chain networks. This study intends to systematically review the existing literature to determine the impact of optimal decision-making during crises to build up supply chain resilience. The paper has focused on the need for evaluating the impact of the COVID-19 pandemic on the FMCG industry and how supply chain resilience would improve in performance during such crises. The study also assessed the existing decision support systems for resilience in a supply chain network and their applicability during a crisis. Some of these models could be used to facilitate decision-making during an epidemic as well. Precisely determining resilience factors affected during an unexpected circumstance would enhance the value of the decision support system in use. Furthermore, it was concluded that the use of quantitative models should be further investigated, as most published work focuses on the conceptualization of a restricted number of resilience factors instead of the development of integrated, comprehensive approaches.

Keywords - decision-making, fast-moving consumer goods, resilient supply chains

I. INTRODUCTION

The pandemics are of rare business calamities, but clear thinking and optimal decision-making with less reliable information are required for an organization to stay in operation, serving the highly fluctuating demands while harvesting the atypical advantages of competition during an epidemic outbreak. Mike Crum, a professor of supply chain management at Iowa State University, had stated to FM magazine once, 'The most resilient companies were the ones who had really embraced risk management planning, and had visibility into their whole supply chain network, not just their immediate suppliers' [1].

With the advent of e-commerce, cross-border business, and short-term delivery, organizations' supply chains have become increasingly complicated, global, and fragile. Numerous failures in the supply chain have been identified, exposing organizations to risk amid dynamic changes in client demand as well as the adoption of new technology breakthroughs. Earthquakes, floods, storms, factory fires,

machine failures, hurricanes, and other natural disasters are only a few examples of typical business disruptions [2].

During the COVID-19 pandemic, global supply networks are confronted with both a supply shortfall and a shrinking demand, resulting in disruptions propagating forward and backward. For example, the pandemic forced China to suspend operations in February and March 2020, significantly disrupting US and European manufacturers and shops due to supply shortages [3]. According to a report published by Fortune Magazine, 94% of the Fortune 1000 companies have been confronted with supply chain disruptions due to the pandemic during early 2020[4]. According to the reports from WHO, there had been 1438 epidemics reported between 2011 to 2018 [5]. Nevertheless, disruption due to COVID-19 pandemic is considered drastic, diverse, more acute, and harshly challenging compared to previous outbreaks such as SARS in 2003 and the H1N1 epidemic outbreak, which took place in 2009 [6]. This explains the challenging nature of the COVID-19 pandemic in every aspect of disruptions it has caused. Therefore, building strategies towards absorbing the impact promptly, would ensure that the organization can withstand any uncontrollable risk by reducing its impact.

Risk identification is usually the first step in traditional supply chain risk management, followed by various solutions for managing the identified risks. This strategy works well when dealing with ongoing or foreseeable disturbances, but it fails when dealing with sudden or unexpected situations. For the latter, it is critical for businesses to develop resilience that allows them to better prepare for and respond to unforeseen events [7]. Risk management decision-making is a process of selecting the best alternatives or ranking the alternatives for a specific risk management goal. The goal is to create, protect and enhance shareholder value by managing uncertainties influencing the achievements of the firm's objectives [8]. In practice, determining the best level of resilience is a crucial decision since over-capacity incurs unnecessary expenditures, and under-capacity exposes businesses to hazards [9].

Decision-making in large-scale organizations often gets restricted due to many reasons such as bounded rationality, confirmation bias, increase of commitment, process conflict and relationship conflict etc.[10], thus controlling the space for an optimal decision to be made. Out of many such reasons, unexpected events such as the pandemic of COVID-19 may implicate such restrictions in making the optimal decisions on behalf of an organization.

Therefore, the objectives of the study are to understand the concept of resilience in the domain of supply chain, to critically evaluate the relevance of decision making and its impact on building resilience in the supply chain and to evaluate existing decision models for supply chain resilience during normal times and times of crises for identifying their suitability to handle uncertainties during a pandemic.

II. METHODOLOGY

The methodology introduced by Barbosa-Póvoa et al. [11] is adopted, and the following steps are followed to conduct a systematic literature review on the defined domain of study: Definition of study topics; examination of previous literature reviews; material-gathering; descriptive analysis; category selection; and material evaluation. Following research questions were defined to guide the study within the selected scope of decision-making towards supply chain resilience in FMCG companies during a pandemic.

A. Research questions

1. How did COVID- 19 pandemic impact the global supply chain network?
2. How COVID-19 pandemic affected the FMCG industry within a developing economy?
3. What are the different characteristics of decision-making during uncertain times vs. normal times?
4. How does supply chain resilience support organizations during a crisis, such as a pandemic?
5. How can decision-making be impacting supply chain resilience?
6. What are the existing decision-making models which support supply chain resilience, and how are they applied?

B. Previous literature reviews

The scientific publications here analyzed and studied in detail are the result of a search performed on the Scopus, IEEE Xplore, and ScienceDirect databases under the keyword searches; “supply chain” AND “resilience” AND review; “supply chain” AND “decision-making” AND review. Following literature reviews were analyzed in-depth in search of more relevant literature.

- M. S. Golan, L. H. Jernegan, and I. Linkov, “Trends and applications of resilience analytics in supply chain modeling: systematic literature review in the context of the COVID-19 pandemic,” *Environ. Syst. Decis.*, vol. 40, no. 2, pp. 222–243, 2020, doi: 10.1007/s10669-020-09777-w.
- Pires Ribeiro, J., & Barbosa-Povoa, A. (2018). Supply Chain Resilience: Definitions and quantitative modeling approaches – A literature review. *Computers and Industrial Engineering*, 115(May 2017), 109–122. <https://doi.org/10.1016/j.cie.2017.11.006>

After a content analysis, it was decided to exclude several papers at this stage, eliminating those that

did not cooperate explicitly with SC Resilience or were not classified as reviews.

C. Material collection

The related studies were mainly selected using Scopus and ScienceDirect databases. Initially, a collection of 83 literature was found through keyword searches, including: “supply chain” AND resilience; “supply chain” AND resilience AND decision-making models; “supply chain” AND resilience AND decision support systems; “supply chain” AND decision-making models; “supply chain” AND resilience AND decision optimization models, etc.

D. Descriptive analysis

An in-depth analysis of content was conducted to restrict the selected literature strictly to the defined domain. The intersection of each publication's content with the set conditions was made possible through the content analysis, and the relevance of each paper was determined. This resulted in a selected number of articles, totaling 47.

E. Category selection

To collect information from many sources and positively approach the research questions, the information from the analyzed literature must be compatible with the research objectives. Therefore, the analyzed publications were organized into three categories,

1. Scope of supply chain disruption discussed (pandemic, epidemic, general disruption)
2. The approach of the study towards supply chain resilience (Qualitative, Quantitative, Case Study etc.)
3. Decision level the model supports (Strategic, Managerial, Operational)

III. RESULTS OF THE LITERATURE REVIEW

This section focuses on systematically reviewing the existing literature on the following four subcategories: (a) Impact of the COVID- 19 pandemic on the overall supply chain and FMCG industry. (b) Decision-making during uncertain times and its specialties. (c) Resilience concept in supply chain. (d) A review on existing DM models for crisis management or resilience in the supply chain.

A. Impact of the COVID- 19 pandemic on overall supply chain and FMCG industry

COVID- 19 is categorized under low frequency, high impact risks in the risk matrix. During the COVID-19 pandemic, global supply networks are confronted with both a supply shortfall and a shrinking demand, which could result in disruptions propagating forward and backward [3].

Reference [12] examined the effects of the COVID-19 pandemic on food supply networks, concluding that demand and supply shocks resulting from a pandemic are caused by a shift in consumer behavior. For example, demand shocks were generated by the quick panic buying shift to ready-meals, which resulted in labor shortages and transportation network disruptions. Further supply-side shocks to food supply chains were caused by restrictions on cross-border goods movement. As a result, it is plausible to expect COVID-19 to have a long-term impact on consumer behavior and supplier chains [13]. Hence, there is enough evidence to determine that a considerable percentage of

consumers would be comfortable in e-commerce practices in the long run thus, resulting in re-engineering of traditional supply chain practices and building up readiness models towards strong e-commerce networks to adjust and sustain in the e-commerce markets. This would be majorly applicable for large-scale FMCG companies which inherit complex traditional supply chain networks.

The strict restrictions imposed by the Sri Lankan government during the first phase of the COVID-19 pandemic had severely impacted the Sri Lankan trade. Thus, creating restrictions to perform on full capacity at the production facilities, halting production for some time due to infected employees, and restrictions such as fully locking down the country. 156 categories of products, including vital food staples such as rice, grains, pasta, bread products, and liquor, were subjected to import restrictions until July 2020. On a three-month credit basis, items like milk powder, palm oil, red lentils, sugar, and sunflower oil were allowed to be imported [14]. According to reference [15] report on the performance of Sri Lankan trade during 2020, FMCG value sales in general trade in Q1 of 2020 have dropped by 11% compared to 2019 Q1 performance, as shown in Fig. 1. The report further discusses that Food and Beverage (F&B) had a lower impact among the FMCG Super Categories but had a decline in General Trade. Personal & Household Care purchases were de-prioritized in favor of Food & Beverage purchases. As a result, they observed a more significant drop in General Trade [15].

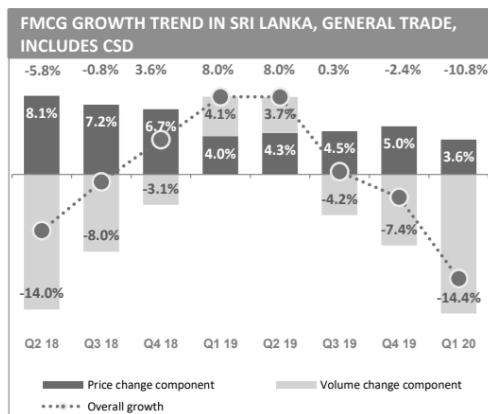


Fig. 1. FMCG Growth Trend in Sri Lanka, GT

The impact of disruption on the supply chain could be graphically represented as below in Fig. 2, where it describes there is a bounce-back period for any company, irrespective of the size of the organization [16]. Nevertheless, a company with solid financial backup and strategic background can bounce back at a high cost compared to an SME, as per the analysis.

Most of the companies faced significant difficulties in smoothing out the flow of their supply chain networks by coordinating with the suppliers, strategizing their production plans, and liaising with the government authorities on special permits to continue the logistics amidst the pandemic situation due to delays in shipments of raw material required for production, sudden closures from the end of their suppliers due to health emergencies and similar reasons.

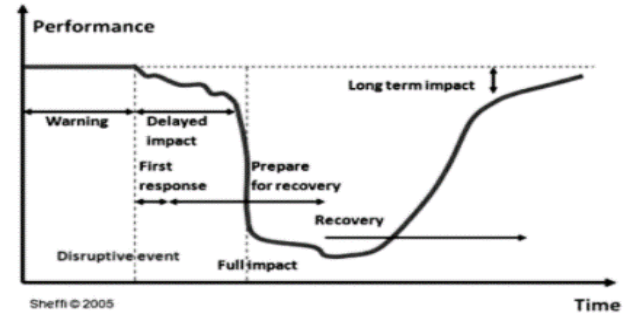


Fig. 2. Impact of disruption for supply chain

Manufacturers of beverages and foods have faced additional problems because of the COVID-19 epidemic. A lack of carbon dioxide because of lower ethanol production levels, resulting in increased carbon dioxide rates and causing disruption to beer and soda manufacturers, is one such example [17].

The extra health precautions such as random PCR tests, quarantining facilities for employees, and medical recovery support were necessary. At the same time, they incurred a vast amount of additional expense for the organizations. Hence, recovery from the COVID-19 pandemic could be relatively less chaotic for large-scale organizations due to scale and resources, given those proper recovery strategies being in place for any unexpected circumstances by utilizing the lessons learned from this pandemic.

B. Decision making during “Normal” vs. uncertain times

A proper decision-making strategy amidst the situation is of vital importance to any business to perform better and gain a competitive advantage. The real challenge is when organizations are required to source, manufacture, coordinate with a vast network of suppliers, dealers, and retailers while operating in a low-margin market [18]. Given the “normal” business days, challenges related to a supply chain network could be predicted accurately to some extent and could be planned for but compared to disruption like the COVID-19 pandemic, “routine” decisions or objectives may not best suit the unexpected circumstances.

Complications, ambiguity, and failure to comprehend will be upsurge in times of calamity, while the ability to make prudent decisions will be weakened [19]. The impact of the COVID-19 pandemic on the supply chain was unique compared to other disruptions that had occurred due to its degree of unpredictability and the scope of impact. When China was first affected by this pandemic, the USA and other European countries were not expecting or rather not prepared for the ripple effect of the pandemic across the global value chain; thus, the impact was brutal. The forward and backward propagation of the impact of disruption in several nodes in the global supply chain network had adversely impacted the developing economies like Sri Lanka as well.

According to authors [20], Decisions “involve a commitment of large amounts of organizational resources for the fulfillment of organizational goals and purpose through appropriate means.”

Many businesses, large and small, will be too slow to keep up in a dynamic environment like the COVID-19

pandemic. During “normal” business days, delaying decisions to gather more information may make sense. However, when the situation is uncertain and defined by urgency and incomplete information, waiting to decide is a decision in itself. Organizations face a significant number of big-bet decisions when faced with a crisis of uncertainty, such as the COVID-19 pandemic, which arrived at breakneck speed and on a massive scale [21].

On the contrary, according to “prospect theory,” when things get rough, people’s aversion to risk decreases, causing them to make riskier judgments. The decision-making capacity might be reduced when the decision-makers are stressed. Thus emotional states of decision-makers are just as important as their reasoning ability [22]. Both studies insist on the fact that decision making during a crisis would have to be done with less information, certainly with a low degree of reliability, sometimes the usual data flow could be hindered due to many unpredictable circumstances, which then results in decision making with intuition and reasonable guessing.

The Cynefin framework in Fig. 3, which is based on mathematical theories of complex and chaotic systems, is another approach in Decision theories [22]. This is a sense-making paradigm for knowledge management that includes a typology that distinguishes between structured and unstructured decision situations. Although a pandemic is a decision context with inherent uncertainty, patterns do emerge according to this framework. Although the order cannot be predicted in advance, cause and effect can be determined after the fact. There is no emerging order in the chaotic environment, which is equally unstructured. When faced with a decision, the Cynefin framework gives a practical perspective that reminds decision-makers that the type of decision situation significantly impacts how it should be treated [22].

Nevertheless, according to reference [23], organizations that adopt clear values, are able to respond to strategic concerns, especially when faced with ambiguity, than those that rely on alternatives-focused decision-making based on clearly defined traits.

Keeney’s value-focused analysis also supports decision-making in both structured and unstructured contexts [23]. This framework is built based on principles and objectives rather than switching between alternatives (Fig. 4).

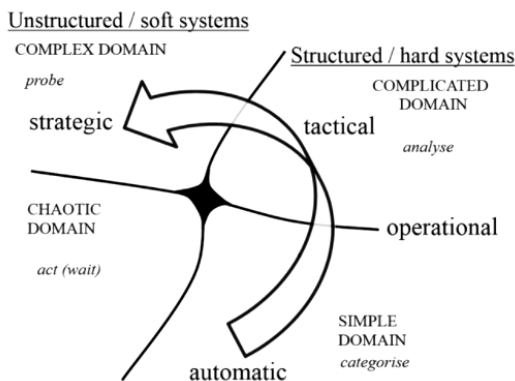


Fig. 3. Cynefin framework

As depicted by the framework in Fig. 4, if decisions are made in an unstructured manner, guiding principles or values may apply, making straightforwardness of leadership and organizational culture critical for the resilience of the supply chain.

Keeney shows that an alternatives-focus may be sufficient in structured domains. At the same time, values-focus may be useful in unstructured or complex structured domains when the cost of analysis is expensive [22]. Both the frameworks would be of importance depending on the company structure and type of crisis in consideration.

Strategies of reactive alternative-focused thinking and decision making, during an unexpected time especially, are said to be producing suboptimal results [24]. When faced with a critical decision point, even during “normal” times, many decision-makers are uninformed of all relevant objectives and the scope of the decision [25]. In the event of an unexpected catastrophe, the set of objectives is even more likely to be altered.

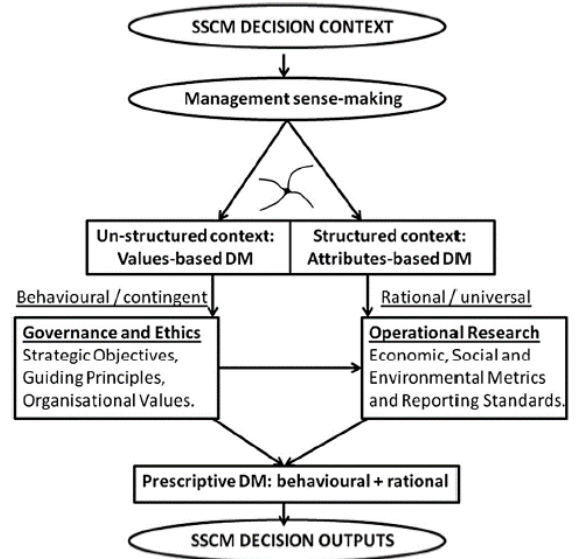


Fig.4. Keeney’s value focused analysis

Prolong suboptimal decisions would result in long recovery periods for organizations when they plan to bounce- back to normal from the pandemic. Thus, reaching optimality in decision making with available resources and information should aim for the organizations to survive another crisis.

The issues could be intensified by reactive and backward-oriented reasoning of the decision-makers [26]. A foresighted leadership is essential to support the management to recover with minimum time to normal. Hence, learned lessons should be carefully used in the strategy formulation process and in future risk management processes to improve the absorption of unexpected shocks on the supply chain network of an organization.

C. Overview of the resilience concept in “supply chain.”

There were several definitions of the word ‘resilience,’ and the following definition was selected to fit the context of the author’s research domain. Reference [27] defines resilience in the context of organizations as “The firm’s ability to effectively absorb, develop situation-specific responses to, and ultimately engage in transformative

activities to capitalize on disruptive surprises that potentially threaten organization survival.” The reference [28] identified the following dimensions of resilience: efficiency, diversity, integration adaptability, flexibility, safety, mobility, and reliability which could be identified as some restorative resilience measures for a supply chain.

The robustness of a firm is also a widely discussed factor inside the domain of supply chain resilience, which describes “the ability of a supply chain to resist or avoid change” [29]. As a result, robust firms operate faster under adverse situations than less robust organizations, providing a competitive advantage. Because of the complexity and size of the supply chain in a large-scale organization, developing a completely resilient SC is a challenge. In reference [30], the authors discuss a few resilient strategies followed by some well-known global companies: Lean production with JIT delivery and low inventory, Six Sigma supply chain, increasing SC flexibility, and developing a strong corporate culture. However, not having a buffer stock when following JIT technique could be argued as not a wise choice for a resilient SC.

Furthermore, the reference [31] had highlighted the following critical factors in establishing a resilient strategy:

- Re-engineering the supply chain to build resilience into the system in advance of potential disruption.
- Establishing a high level of collaboration with supply chain parties to identify and manage risk.
- Achieving the agility necessary to respond quickly to the unexpected.
- Embedding a culture of risk management.

An embedded culture for risk management set by the tone from the top of a firm would enable a firm to plan and forecast risk with greater accuracy levels and facilitate higher business transformations such as business process re-engineering when required, in the necessary parts of the supply chain.

Resilient SCs may not be the cheapest, but they are better equipped to deal with the unpredictable business environment [32]. Enterprises that pursue a policy of ‘zero inventories,’ for example, are not resilient because they lack a stock buffer to respond to an unforeseen shortage of commodities caused by market unrest or volatility [31].

Further, the cost of reactive responses to disruption would be much more expensive than avoidance or mitigation through improved resilience in the supply chain network. Much of the previous understanding of what defines a resilient supply chain has been challenged by the severity of the business disruption caused by the COVID-19 pandemic. According to recent studies, the crisis has resulted in a rapid decline of several business and economic parameters, including productivity and global GDP [33].

As per risk identification matrices in management studies, the higher the impact and likelihood of a disruption higher the vulnerability of a system. Considering the COVID-19 pandemic, this is a high impact, less likelihood risk on the matrix, which sums up why most firms are not focused on pre-preparation for such calamities. The trick is to mitigate the adverse impact of such a calamity even at the propagating failures of other supply chains. Thus, it is

crucial to building up resilience as much as optimizing for the efficiency of a supply chain.

D. A Review on Existing Decision Making Models for Crisis Management or Resilience in Supply Chain

This section would mainly focus on reviewing existing decision support models and frameworks in the domains of supply chain resilience and crisis management in the supply chain. Thereby, the author expects to understand the gaps for research in the existing models and critically analyze factors considered, parameters used, and method of analysis in each of the models in review.

Firstly, in reference [2], the authors propose an ontology-based decision support system towards resilient supply chains by combining supply chain resilience decision-making with a rule-based ontological framework. The ontology is an explicit specification of a conceptualization that primarily aids in structuring data to enable interaction between various firms in a supply chain [2]. The concept of ontology has been employed by scholars in various fields, including manufacturing, medicine, supply chain, and material science.

Reference [2] has considered a three-echelon supply chain network in their mathematical model, which has been optimized under threat conditions by varying pre-defined parameters by interpreting from the rule base of the ontology. Using PSO-DE, an optimization technique, the problem is solved to determine the optimum collective decision for production and logistics units in the network to meet customer demand. The practicality of the model during a pandemic where demand is readily fluctuating is questionable.

Reference [34] has used an effective fuzzy linear programming approach for supply chain planning under uncertainty. Due to a lack of knowledge, the epistemic uncertainty sources in supply chain tactical planning problems are handled using the fuzzy model. Data from a genuine automobile supply chain was used to evaluate this model. This model could be further adapted to uncertainty in demand forecasting as well as this could be utilized to predict nearly accurate demand levels during an uncertain time.

Authors in reference [34] propose a decision support framework to assess supply chain resilience. The system will aid decision-making by allowing users to run “what-if” scenarios and see how different supply chain configurations affect the system’s expected resilience behavior. Finally, the costs and benefits of utilizing different supply chain resilience methods will be weighed. This decision support system mainly focuses on utilizing simulation in understanding redundant factors in the supply chain network.

Reference [35] had proposed the measure of recovery time as a measure of resilience in the supply chain network through their proposed survival model. The new metric is based on a semiparametric model called the CoxPH model. The variables in the Cox-PH model indicate various sources of disruption, the input variable represents an event (survival or resilience analysis failure event), and the output variable is time. However, this model carries few limitations in terms of the limited number of disruptions that could be catered in, the assumption that sources of disruptions being independent of each other, etc.

The study [37] discusses ways to identify and align decision-making objectives in response to the crisis circumstances such as the COVID-19 pandemic. In the study, decision-makers are presented with guidelines for identifying intra-organizational objectives and aligning them across the supply chain and with policymakers. The study has presented examples of intra-organizational and inter-organizational goals for both normal and crises. In addition, they outlined an iterative approach for regularly updating the objectives of an organization. This study would be considered as an inspiration for further analysis to be conducted by the author.

Reference [36] has considered a port closure interruption on either the supply or demand side of the supply chain in the research and created a two-stage stochastic programming model that includes an exponential perishability function and explores various potential objectives. These objectives are the expected profit (P) maximization, the recovery level (RL), and the lost profit during recovery. Thus, they propose a new resilience metric, namely NPV-LP, which is an integration of several other matrices.

IV. DISCUSSION

FMCG products usually carry a low shelf life. Hence the cycle of the product-to-market logistics must frequently happen, amidst any disruption, as the name suggests, "Fast Moving Consumer Goods." During the COVID -19 pandemic, FMCG companies in Sri Lanka observed a more significant drop in General Trade [15]. The main reasons identified through the study were poor strategic preparation for uncertain situations, less experience of the decision-makers, less reliability of information collected, slow collection of data, poor predictive models, and poor organizational vision and leadership. Therefore, it is evident that routine decision-making models should be optimized to address the absorption and recovery stages during a crisis.

The analyzed decision models in the domain of resilient supply chains had primarily focused on mathematical model development to support decision-making in the supply chain. Some models had used simulation techniques to bring in the randomness and unexpected nature of the environment to enhance the relevance of the models to real-world scenarios. Therefore, the discussed models will be applicable in other low frequency, and high impact risks and generalized risk mitigation approaches.

Regarding the applicability of the discussed models for a pandemic situation, the number of constraints considered reduces the practicality of those models. Also, it was concluded that only a few resilience measures or factors had been considered in the models analyzed. It would be more comprehensive if decision support models could incorporate diverse angles of resilience which could be sorted out according to the suitability of the factors or category of factors to a particular crisis. Future research could also focus on the qualitative nature of decision-making through learned lessons in the industry during the COVID- 19 pandemic. Future analysis could also focus on strengthening resilience in each node of a supply chain network or building resilience through integration. Therefore, further understanding of the qualitative aspects of decision making during the COVID- 19 pandemic on the

supply chain of the FMCG industry is focused by the author with the expectation of supporting literature on the research area explained above.

REFERENCES

- [1] "How Kellogg's, Nike, and HP handled 2020 supply chain disruptions - FM." <https://www.fm-magazine.com/news/2021/jan/coronavirus-supply-chain-disruptions-kelloggs-nike-hp.html> (accessed Aug. 25, 2021).
- [2] S. Singh, S. Ghosh, J. Jayaram, and M. K. Tiwari, "Enhancing supply chain resilience using ontology-based decision support system," *Int. J. Comput. Integr. Manuf.*, vol. 32, no. 7, pp. 642–657, 2019, doi: 10.1080/0951192X.2019.1599443.
- [3] Kinra, D. Ivanov, A. Das, and A. Dolgui, "Ripple effect quantification by supplier risk exposure assessment," *Int. J. Prod. Res.*, vol. 58, no. 18, pp. 5559–5578, 2020, doi: 10.1080/00207543.2019.1675919.
- [4] Fortune, "94% of the Fortune 1000 are seeing coronavirus supply chain disruptions: Report.," 2020.
- [5] W. Craighead, D. J. Ketchen, and J. L. Darby, "Pandemics and Supply Chain Management Research: Toward a Theoretical Toolbox*," *Decis. Sci.*, vol. 51, no. 4, pp. 838–866, 2020, doi: 10.1111/deci.12468.
- [6] Pierre Haren and David Simchi-Levi, "How Coronavirus Could Impact the Global Supply Chain by Mid-March," *Harvard Business Review*, 2020. <https://hbr.org/2020/02/how-coronavirus-could-impact-the-global-supply-chain-by-mid-march> (accessed Jul. 08, 2021).
- [7] T. Bier, A. Lange, and C. H. Glock, "Methods for mitigating disruptions in complex supply chain structures: a systematic literature review," *Int. J. Prod. Res.*, vol. 58, no. 6, pp. 1835–1856, 2020, doi: 10.1080/00207543.2019.1687954.
- [8] "Making Enterprise Risk Management Pay Off - Financial Executive Research Inc. Staff, Thomas L. Barton, William G. Shenkir, Paul L. Walker - Google Books." https://books.google.lk/books?hl=en&lr=&id=ZdpPiL_wyJgC&oi=fnd&pg=PP13&dq=Barton,+Shenkir,+%26+Walker,+2002+research&ots=GSSqr-lka&sig=E_RaQVS_P5fcI0uy2WERk8m6EQY&redir_esc=y#v=onepage&q=Barton%2C%20Shenkir%2C%20Walker%2C%202002+research&f=false (accessed Jul. 08, 2021).
- [9] J. Fiksel, M. Polyviou, K. L. Croxton, and T. J. Pettit, "From risk to resilience: Learning to deal with disruption," *MIT Sloan Manag. Rev.*, vol. 56, no. 2, pp. 79–86, 2015.
- [10] S. Black, D. G. Gardner, J. L. (Jon L. Pierce, R. M. Steers, and OpenStax College, Organizational behavior. .
- [11] A. P. Barbosa-Póvoa, C. da Silva, and A. Carvalho, "Opportunities and challenges in sustainable supply chain: An operations research perspective," *Eur. J. Oper. Res.*, vol. 268, no. 2, pp. 399–431, 2018, doi: 10.1016/j.ejor.2017.10.036.
- [12] J. E. Hobbs, "Food supply chains during the COVID-19 pandemic," *Can. J. Agric. Econ.*, vol. 68, no. 2, pp. 171–176, 2020, doi: 10.1111/cjag.12237.
- [13] P. Chowdhury, S. K. Paul, S. Kaisar, and M. A. Moktadir, "COVID-19 pandemic related supply chain studies: A systematic review," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 148, no. August 2020, p. 102271, 2021, doi: 10.1016/j.tre.2021.102271.
- [14] "Impact of COVID19 on food supply chains in Sri Lanka | News item | netherlandsandyou.nl." <https://www.netherlandsandyou.nl/latest-news/news/2020/06/02/impact-of-covid19-on-food-supply-chains-in-sri-lanka> (accessed Jul. 06, 2021).
- [15] Neilson Sri Lanka, "Rising upto the challenge," no. June, 2020.
- [16] P. Mensah, Y. Merkurjev, and F. Longo, "Using ICT in developing a resilient supply chain strategy," *Procedia Comput. Sci.*, vol. 43, no. C, pp. 101–108, 2015, doi: 10.1016/j.procs.2014.12.014.
- [17] Resilience360, "The COVID-19 pandemic disrupts food and beverage supply chains in the U.S.," 2020, [Online]. Available: <https://graphics.reuters.com/HEALTH-CORONAVIRUS/USA-MEATPACKING/qmympmnxvbvr/index.html>.
- [18] S. Modgil, R. K. Singh, and V. Sonwane, "Supply chain routine issues in Indian FMCG manufacturing firm," *Int. J. Logist. Syst. Manag.*, vol. 37, no. 1, pp. 18–37, 2020, doi: 10.1504/ijlsm.2020.109648.
- [19] L. James, E., & Wooten, "Leadership as (un)usual: How to display competence in times of crisis.," *Organ. Dyn.*, vol. 34, no. 2, pp. 141–152, 2005.

- [20] J. Shrivastava, P. and Grant, "Empirically derived models of strategic decision-making processes," *Strateg. Manag. J.*, vol. 6, no. 2, pp. 97–113, 1985.
- [21] A. Alexander, A. De Smet, and L. Weiss, "Decision making in uncertain times," *McKinsey Digit.*, no. March, p. 6, 2020.
- [22] A. Alexander, H. Walker, and M. Naim, "Decision theory in sustainable supply chain management: A literature review," *Supply Chain Manag.*, vol. 19, pp. 504–522, 2014, doi: 10.1108/SCM-01-2014-0007.
- [23] R. L. Keeney, "Value-focused thinking: Identifying decision opportunities and creating alternatives," *Eur. J. Oper. Res.*, vol. 92, no. 3, pp. 537–549, Aug. 1996, doi: 10.1016/0377-2217(96)00004-5.
- [24] R. L. Siebert, J. U.; Keeney, "Decisions: Problems or opportunities? How you can prevent unpleasant decision situations, scientific contributions," *Wirtschaftswissenschaftliches Stud.*, pp. 1–6, 2020.
- [25] R. L. Bond, S. D.; Carlson, K. A.; Keeney, "Generating objectives: Can decision makers articulate what they want?" *Manage. Sci.*, vol. 54, no. (1), pp. 56–70, 2008.
- [26] R. L. Keeney, *Value-Focused Thinking. A Path to Creative Decision-Making*. Cambridge, MA, USA: Harvard Univ. Press., 1992.
- [27] T. E. & L.-H. M. L. Lengnick-Hall, C. A., Beck, "Developing a capacity for organizational resilience through strategic human resource management," *Hum. Resour. Manag. Rev.*, no. 21, pp. 243-255., 2011.
- [28] L. Chen and E. Miller-Hooks, "Resilience: An indicator of recovery capability in intermodal freight transport," *Transp. Sci.*, vol. 46, no. 1, pp. 109–123, 2012, doi: 10.1287/trsc.1110.0376.
- [29] "A Theory of Robust Supply Chains | Supply Chain Management Research." <https://scmresearch.org/2015/03/02/a-theory-of-robust-supply-chains/> (accessed Jul. 07, 2021).
- [30] P. Mensah and Y. Merkurjev, "Developing a Resilient Supply Chain," *Procedia - Soc. Behav. Sci.*, vol. 110, pp. 309–319, 2014, doi: 10.1016/j.sbspro.2013.12.875.
- [31] M. Christopher and H. Peck, "Building the Resilient Supply Chain," *The International Journal of Logistics Management*, vol. 15, no. 2, pp. 1–14, 2004, doi: 10.1108/09574090410700275.
- [32] V. C. Carvalho, H. and Machado, "Lean, agile, resilient and green supply chain: a review.," *Proc. 3rd Int. Conf. Manag. Sci. Eng. Manag.*, vol. 2–4, no. November, pp. 66–76, 2009.
- [33] R. Harris, "Covid-19 and productivity in the UK.," *Durham University Business School*, 2020. <https://www.dur.ac.uk/research/news/item/?itemno=41707>.
- [34] M. Falasca, C. W. Zobel, and D. Cook, "A decision support framework to assess supply chain resilience," *Proc. ISCRAM 2008 - 5th Int. Conf. Inf. Syst. Cris. Response Manag.*, no. January 2008, pp. 596–605, 2008.
- [35] R. Raj et al., "Measuring the resilience of supply chain systems using a survival model," *IEEE Syst. J.*, vol. 9, no. 2, pp. 377–381, 2015, doi: 10.1109/JSYST.2014.2339552.
- [36] G. Behzadi, M. J. O'Sullivan, and T. L. Olsen, "On metrics for supply chain resilience," *Eur. J. Oper. Res.*, vol. 287, no. 1, pp. 145–158, 2020, doi: 10.1016/j.ejor.2020.04.040.
- [37] M. S. Golan, L. H. Jernegan, and I. Linkov, "Trends and applications of resilience analytics in supply chain modeling: systematic literature review in the context of the COVID-19 pandemic," *Environ. Syst. Decis.*, vol. 40, no. 2, pp. 222–243, 2020, doi: 10.1007/s10669-020-09777-w.

Simulation analysis of an expressway toll plaza

Shehara Grabau*
 Software Engineering Teaching Unit
 University of Kelaniya, Sri Lanka
 sheharagrabau@gmail.com

Isuru Hewapathirana
 Software Engineering Teaching Unit
 University of Kelaniya, Sri Lanka
 ihewapathirana@kln.ac.lk

Abstract - Since the early civilizations, transportation has played a significant role, from fulfilling basic human needs to contributing towards major economic growths all over the world. With the advancement in technology, the demand for smooth and hassle-free transportation increased and it is particularly true for road transportation in Sri Lanka as well. As a result, the expressway road network was introduced to Sri Lanka in 2011. Although a toll is payable for the use of expressways, many vehicle users prefer to utilize the expressway due to the extensive amount of time saved. Time is of utmost importance for expressway users. Hence, long queues and waiting time at toll plazas where the toll payment is made should be minimized. This study is aimed at analyzing the performance at the Peliyagoda toll plaza of the Colombo-Katunayake expressway where the formation of long queues and long waiting time in queues can be observed during peak hours. Due to the high complexity of using the analytical approach in obtaining the performance measures, a simulation approach was used with Arena Simulation Software. Few setup improvements were identified, and each of the setups were simulated to obtain the performance measures. Based on the comparison of the results, recommendations and suggestions to improve the efficiency of the operations at the Peliyagoda toll plaza have been outlined.

Keywords - expressway, M/M/1, queue simulation, queuing theory, toll plaza, waiting time

I. INTRODUCTION

Today, expressways around the world connect cities far and wide, and they are instrumental in saving time and operational costs. With advantages such as high speed, high vehicle volume, greater comfort and less fuel wastage, drivers tend to utilize expressways even if a toll is charged. By the end of the year 2019, the total length of expressways in Sri Lanka was 217.8 km [1]. It should be noted that while the expressways are advantageous to vehicle users, it is also a revenue generation model for the country. According to the Annual Report of the Central Bank of Sri Lanka, a revenue of Rs. 8.6 billion was generated from the expressway network in 2019 [1], compared to Rs. 8.4 billion in 2018 [2].

A tolling system situated at the exit point of an expressway charges a toll from each user based on the distance travelled and the vehicle category. The Peliyagoda toll plaza is situated towards the Southern end of the Colombo-Katunayake expressway. A majority of the vehicles that come to the Colombo city from Ja Ela, Katunayaka, Negombo and even from Chilaw and Puttalam areas, utilize the Colombo-Katunayake expressway, and make the toll payments at the Peliyagoda toll plaza to enter Colombo and its suburbs. With the introduction of the Outer Circular Highway, traffic flow from Southern parts of the country to Colombo also exit the expressway network from the Peliyagoda toll plaza. Moreover, vehicles that need to take the Colombo-Kandy Highway or go towards Wattala will need to make the toll payment at the

Peliyagoda toll plaza. Currently, two types of toll collection methods are available at the Peliyagoda exit. They are: (1) the Manual Toll Collection (MTC) and (2) the Electronic Toll Collection (ETC). At MTC, a vehicle is required to stop at the gate and make the payment to the teller using cash. The teller then issues the ticket and balance (if any), and the toll gate barrier is opened for the vehicle to pass through. For a vehicle to utilize the ETC facility, it should be enrolled in the highway's information system as a user, and sufficient funds should be available in the user's account. Enrolled vehicles are given an e-tag to paste on the vehicle's windshield. The e-tag of a car approaching the ETC gate is scanned using an automatic vehicle identification technology, and the toll gate barrier is opened without requiring the vehicle to stop. Simultaneously, the toll is debited from the user's account. The toll plaza at Peliyagoda consists of five toll gates of which four are MTC and one is ETC.

The problem lies in the formation of long queues at the toll plaza during peak hours and its adverse effects on users and the environment. According to an analysis conducted by the Expressway Operation Maintenance and Management Division (EMO&MD), the current number of toll lanes are insufficient at the Peliyagoda exit (Figure 1). Due to the high rate of arrivals and the inadequate number of toll gates to serve them, queues are formed and long waiting times are encountered by the vehicle owners during peak hours.

The benefit of the time gained by taking the expressway could be lost to the users when long waiting times are encountered at the toll plaza. For example, delays in reaching offices, educational institutions, and other personal commitments can result in disciplinary actions, loss of the business and other personal losses. Furthermore, vehicles accelerating and decelerating to move slowly in queues and braking to bring the vehicle to a stop, cause wastage of fuel and emission of harmful pollutant gases such as CO, CO₂, and NO_x to the environment, that in turn, cause respiratory diseases in humans [3].

Expressway	Critical IC	Peak time	No of toll Lanes	Given Capacity of toll Lane Vehicles per hour		Actual Traffic per hour		Remarks
				ETC	MTC	ETC	MTC	
OCH, E02	Kadawatha	Morning	Entrance 3 - Exit-5	N/A	360 (for Entry)	-	528	Entry Booths no-3
		Evening	Entrance 3 - Exit-6	-	-	-	400	Entry Booths no-3
	Kottawa	Morning	Entrance 3 - Exit-7	N/A	360 (for Entry)	-	337	Entry Booths no-5
		Evening	Entrance 3 - Exit-7	-	-	-	395	Entry Booths no-5
CKE, E03	Peliyagoda	Morning	Exit 5 (MTC 4, ETC 1)	1100	240	453	286	Available no of Toll lanes are insufficient
		Evening	Entrance 2 (MTC 4, ETC 1)	1100	190	190	228	
	Ja Ela	Morning	Entrance 2 (MTC 4, ETC 1)	1100	360	253	415	Available no of Toll lanes are insufficient
		Evening	Exit 2 (MTC 4, ETC 1)	1100	240	160	341	

Expressway	Critical IC	Peak	Peak days	Toll Plaza (Entrance/Exit)	Peak Time		Remarks
					From	To	
OCH, E02	Kadawatha	Morning	Monday	Entrance	7.00 a.m.	9.30 a.m.	
		Evening	Friday	Entrance	4.00 p.m.	9.00 p.m.	
	Kottawa	Morning	Monday	Entrance	7.00 a.m.	8.00 a.m.	
		Evening	Friday	Entrance	5.00 p.m.	9.30 p.m.	
CKE, E03	Peliyagoda	Morning	Monday	Exit Plaza	7.30 a.m.	9.30 p.m.	Insufficient Lanes at Exit (available 5 lanes)
		Evening	Friday	Exit Plaza	4.30 p.m.	6.00 p.m.	
	Ja Ela	Morning	Monday	D-Entrance Plaza	7.30 a.m.	9.30 p.m.	Insufficient Lanes at Entrance 4: Exit (Peliyagoda - Ja Ela Section)
		Evening	A- Exit Plaza	4.30 a.m.	7.30 a.m.		

Fig. 1. Remarks from EMO&MD
 Source: <http://www.exway.rda.gov.lk/index.php?page=announcements/20190401>

Queuing Theory can be used to analyze the formation of queues and delays caused due to long waiting times in a system. Some measures that can be derived include average waiting time in the queue, average time spent in the system and average number of customers in the queue.

In this paper, we apply queuing theory to analyze the toll payment system at the Peliyagoda toll plaza of the Colombo-Katunayake expressway in Sri Lanka. Due to the complexity of the system, applying analytical models is difficult. Thus, we develop a simulation model to calculate the performance measures of the system. In addition to the current system, we propose several simulation setups of alternative systems to improve the performance of the current system. The proposed systems are also analyzed using simulation. By analyzing the performance of the current system and proposing alternative setups, we provide recommendations to the expressway management's decision-making process to help reduce the congestion, especially during peak hours, and thereby achieving an efficient transportation system and minimizing environmental pollution.

II. Queuing Concepts

A queue is formed by a flow of customers from an infinite or finite population towards the service facility that lacks the capability to serve them all at a time [4]. The basic features of a queuing system can be stated as follows:

A. The arrival process

This is the way that customers arrive at the system. The arrival process can be classified in several ways such as single line or multiple lines, finite or infinite and single customer or customers that come in bulk. The arrivals are assumed to occur in a random pattern and are usually modelled using a suitable probability distribution such as the Poisson distribution. The average customer arrival rate, λ is an important parameter of the arrival process.

B. Service discipline

The serving process can be carried out according to four main principles such as, First-In-First-Out (FIFO), Last-In-First-Out (LIFO), Service for Random Order (SRO) and Priority Service (PS).

C. The service time distribution

The service time distribution is usually modelled as a uniform or exponential distribution. It is independent of the arrival process. The average customer service rate, μ is an important parameter that characterizes the service time distribution.

D. Service mechanism

This is the work on policy decided for service, and how the customers leave the system. The service mechanism can be classified in several ways according to the number and configuration of service facilities and the service pattern of the system. The service mechanism can be single channel-single stage, single channel-multiple stage, multiple channel-single stage or multiple channel-multiple stage (Fig. 2).

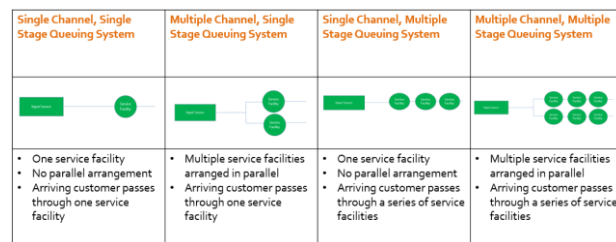


Fig. 2. Configuration of queuing systems

Queuing theory gives an understanding of the queuing system and ideas about what can be done to make it more efficient, easy to serve, and the number of users that can be served. The ultimate objective is to make intelligent decisions by understanding the underlying processes [5]. Although several analytical queuing models exist, modelling complex systems using these models might lead to too many simplifications which might in turn cause resulting models to be invalid. Simulation modelling can be utilized to understand the behaviour of complex queuing systems. A simulation provides the flexibility to experiment with certain parts of a decision problem and analyze the likely consequences of alternative decisions. Once the basic features of the queuing system are clearly defined and understood, the system can be simulated, and the required performance measures can be calculated.

III. LITERATURE REVIEW

Expressways were introduced to Sri Lanka in 2011, and it is being expanded to other parts of the country. Currently, there is hardly any published research work available for queue analysis at toll plazas in Sri Lankan expressways, but a substantial body of international research findings is available on this topic.

One of the latest research works available for the application of the queuing theory for a toll plaza is [6]. The main objective of their study was to examine the applicability of the queuing theory for a toll plaza in both directions. Their results showed that although the postulated Poisson distribution is the true population distribution to one direction, there is less degree of agreement to the other direction. They further showed that although most of the studies related to expressway queues are assumed to be operating under the steady-state condition, it is seldom true in nature.

Sihotang et al. [7] analyzed the performance measures of a toll plaza queuing system assuming arrivals and service times to be normally distributed. The data collected for this research were the total number of vehicles that arrived, and the total number of vehicles served for five weekdays at the toll gate Mukti Harjo. Using the data, they calculated the arrival rate and service rate, and used the Kolmogorov-Smirnov test and the Chi-Square test to determine the distribution of arrivals. Using Arena Software for Simulation of the system with varying number of servers, they concluded that the number of servers at the toll plaza is optimal and does not need to be changed.

A toll gate system in Salem, Bangalore was simulated by Shanmugasundaram & Punitha [8] for different vehicle categories such as car/jeep (F1), light commercial vehicles (F2), truck/bus (F3) and multi-axle vehicles (F4). The arrival and service distributions for each vehicle category were calculated using the collected data, and a simulation

was conducted to compare various service mechanisms. Duhan et al. [9] analyzed a toll plaza system in North India to study the current traffic congestion situation and suggested possible solutions. With data on the volume of traffic on an hourly basis for a working day (Monday) and non working day (Sunday), they identified the peak and nonpeak hours. By focusing the experiments only for those hours, they suggested ways to reduce the waiting time in queues, such as increasing the number of toll booths, employing mobile toll collectors, and setting up smart machines to decrease the service time. They also suggested the option to install a red traffic light before 1km to the plaza, to enhance a smooth vehicle flow towards the booths.

Ceballos & Curtis [10] analyzed the queuing system at a parking exit toll plaza at airports. Although the study was not based on expressways, the approach to their analysis of a multi-server queuing model is noteworthy. In their study, both the application of the analytical queuing model and simulation were used, and measures of effectiveness from both methods were compared. They pointed out that although toll plazas are multi-queue multi-server systems, the analytical formulation of such systems is extremely complex. The workaround used was to model the system as a series of single-channel queuing systems in parallel, and a single-queue multiple-channel system. They showed that the analytical results greatly differ from the simulation results. However, not all research shows that the above conclusion is true. In a study done by Punitha [11] by using the simulation approach and analytical approach, she concluded that both methods give coinciding results. Her study was based on the traffic delay at a toll plaza, and she examined the performance measures for a single server queue with four types of vehicle categories. Each vehicle category was simulated, and performance measures for each were obtained accordingly.

Antil [12] studied the traffic congestion at a Delhi toll plaza with a high arrival rate of vehicles. His analysis was limited to the busiest hour of the day for a working day (Monday) and non working day (Sunday). For the server that serves the incoming traffic, he used the single-channel single-stage queuing model and compared the resulting performance measures. In addition to that, he also calculated the cost of waiting per customer based on the assumption that fuel of Rs 4 per minute was wasted while waiting in the queue. He suggested that the toll plaza needed more toll gates and modern technology to improve the service times.

IV. FORMAL DEFINITION OF THE PROBLEM STATEMENT

The Peliyagoda toll plaza is the service facility of our queuing system. The vehicles are the arriving units, and the toll gates are the servers or channels. The toll plaza has 5 toll gates, i.e., the queuing system under study has 5 servers. Out of the 5 servers, the 4 MTC servers are assumed to have the same service rates whereas the ETC server has a different service rate. Once served, the units exit the system. Therefore, there is only one stage of service and the system is a single-stage multiple-channel multiple-queue system. There is no limit to the number of arriving units. There are no predefined queuing formulas to analyze such systems due to their complexity. Thus, we resort to a simulation-based approach to analyze this system.

A. Data and data description

Raw data is collected from the EOM&MD with due permission from the Sri Lanka Road Development Authority (RDA). Data is collected based on the number of vehicles that exited through the Peliyagoda toll plaza for one hour time intervals during each day for a one-week period in the year 2019 when no holidays or other external factors affected the traffic inflow to Colombo or its suburbs. Figure 3 depicts the number of vehicle arrivals at the Peliyagoda toll plaza during each hour within the time considered. Table I, further summarizes the total number of vehicles that exited through each lane.

According to Figure 3 and Table I, the number of vehicle arrivals during the period of 6 am to 9 am is the highest compared to other time intervals and vehicle arrivals on weekdays are higher than the vehicle arrivals during the weekend. Thus, the weekend is disregarded, and the study is carried out for weekdays for the period 6 am – 9 am which can be considered as the peak hour period in the morning.

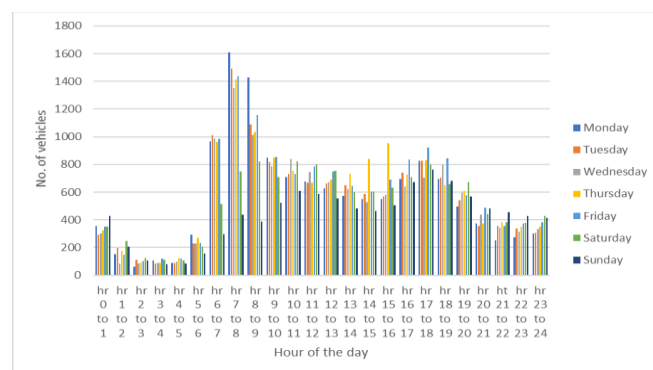


Fig. 3. Number of vehicles that exited from Peliyagoda toll plaza during each hour within a week in 2019

TABLE I. DATA ON THE TOTAL NUMBER OF VEHICLE ARRIVALS AT EACH GATE

Day	Total number of vehicles that exited through each lane					Total
	Gate 1 - ETC	Gate 2 - MTC	Gate 3 - MTC	Gate 4 - MTC	Gate 5 - MTC	
7.10.2019 (Monday)	2585	3108	3244	2590	1980	13507
8.10.2019 (Tuesday)	2690	3138	3289	2430	1896	13443
9.10.2019 (Wednesday)	2634	2967	3049	2618	1919	13187
10.10.2019 (Thursday)	2677	3389	3325	2653	2208	14252
11.10.2019 (Friday)	2752	3219	3448	2761	2309	14489
12.10.2019 (Saturday)	174	3158	3231	2583	1891	12617
13.10.2019 (Sunday)	1142	2828	2903	2105	1396	10374

In addition to the data collected on vehicle arrivals, data on service rates were also gathered by interviewing an official at the EOM&MD. According to experts, the average number of tickets that can be issued by a teller at the MTC gate is 210 per hour and the maximum number ever reached is 295 per hour. The ETC gate on the other hand has never been saturated since its installation, and on average, 453 vehicle arrivals per hour are observed during

the morning. These expert opinions are used in calculating service rates, as direct observation at the facility was not possible because of the restrictions imposed due to the COVID-19 situation in the country during the considered period.

V. SIMULATION EXPERIMENT

The queuing system under study is first analyzed based on the number of lanes and queue formation.

A. Calculating the arrival rate (λ) and obtaining the distribution of arrivals

For each lane, we consider the number of vehicle arrivals per hour for the selected period of 6 am – 9 am during the five weekdays. Thus, for each server we have 15 data points and calculate the average vehicle arrival rate, λ .

$$\lambda = \text{total vehicle arrivals per hour} / 15 \quad (1)$$

Graphical analysis shows that the arrival of vehicles at each lane is random. We conduct a Kolmogorov-Smirnov (K-S) test at 0.01 significance level to statistically confirm if the distribution of vehicle arrival per hour at each server follows a Poisson Distribution. Our results are summarized in Table II.

The K-S critical value at 0.01 significance level and 15 degrees of freedom is 0.404. From Table II, the K-S test statistics calculated for all five lanes are less than 0.404 and the vehicle arrivals can be assumed to follow a Poisson distribution with the respective arrival rate λ .

B. Calculating the service rate μ and obtaining the distribution of service times

The service rates for all lanes are obtained based on expert opinion as mentioned in Section III.A. At MTC lanes, a minimum of 210 vehicles on average can be served per hour. Therefore, the service rate μ is considered as 1/210 hours per vehicle. At the ETC lane, there is 453 actual traffic per hour observed in the morning. Taking this information into account, the service rate μ at ETC lane is considered as 1/453 hours per vehicle. The service times are assumed to follow an exponential distribution.

Due to the complexity of the system under study, traditional queuing theory equations cannot be used to calculate the performance measures of the system. Therefore, a simulation is performed using the Arena simulation software.

TABLE II. VEHICLE ARRIVAL RATES AND K-S TEST STATISTIC RESULTS

	Server 1 - ETC	Server 2 - MTC	Server 3 - MTC	Server 4 - MTC	Server 5 - MTC
λ	370.5	205.8	207.8	203.5	194.2
K-S test statistic	0.4690	0.235	0.2444	0.327	0.389

C. Simulation setup

In addition to the current system setup, four other setups are proposed and simulated to improve the performance measures of the current system. Finally, the recommendations are presented. The five identified setups are summarized in Table III. Scheme B suggests adding

more manual servers to the system, schemes C and D incorporate the use of new technology and adding more electronic servers, and scheme E incorporates both.

For each scheme, servers in each lane are simulated individually for a three hour period. The number of arrivals is generated using a Poisson distribution, and the service times are generated using an exponential distribution. Hence, each lane is separately modelled as a single server single stage queuing model. The relevant parameters for each simulation is provided in Table IV.

D. Performance measures of the Queuing System

The following performance measures were obtained from Arena Simulation Software for comparing the current system with the proposed system.

- Average waiting time in queue
- Average number of units in queue
- Average time a customer spends in the system
- Maximum time a unit spends in the system
- Average number of units in the system
- Server utilization

VI. RESULTS AND DISCUSSION

The five identified schemes are stimulated repeatedly using Arena Simulation Software and performance measures are recorded. Finally, the overall performance of each scheme is obtained by calculating the average values of the performance measures from each simulation run. Table V and Figure 9 summarize and compare the performance for each scheme.

Considering the average values for Scheme A, a vehicle has to wait 2.3 minutes in the queue and nearly 3 minutes in the system before leaving after service. The maximum waiting time a unit may have to spend in the system is as high as 8.84 minutes. The utilization factor for Scheme A shows that 93.2% arriving customers have to wait for service. These measures highlight that the current setup should be made efficient. Performance measures of Schemes B, C, D and E show that waiting time can be reduced by implementing one of them, but it is important to note the changes that should be adopted along with the implementations of those schemes.

TABLE III. SUMMARY OF THE SIMULATION SETUP

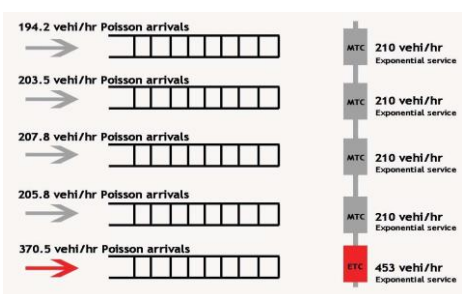
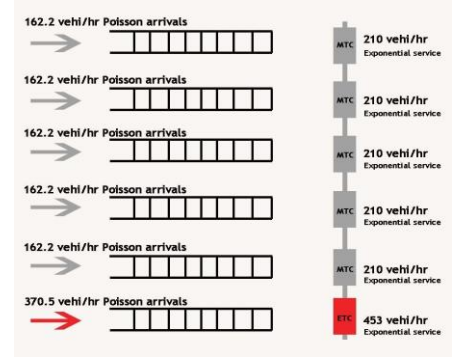
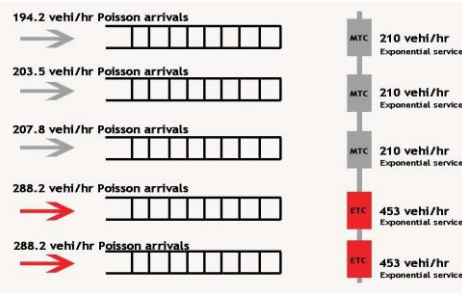
Scheme	No. of MTC servers	No. of ETC servers	Total no. of servers
A – current system	4	1	5
B	5	1	6
C	3	2	5
D	2	3	5
E	5	2	7

Schemes B and E have the lowest waiting time and number of waiting units, in comparison to other schemes. The waiting times are 25.8 seconds and 19.8 seconds, respectively. The number of waiting units is approximately 1 for both schemes. Less than 1 minute waiting time and 1 unit waiting in queue are positive performance measures.

However, in order to implement either one of these schemes, extra space is required at the toll plaza as both schemes require additional toll gates. If space is available for two extra toll gates, Scheme D is the best option whereas Scheme B can be adopted if space is limited for one extra toll gate only. Both schemes can meet the current demand at the Peliyagoda toll plaza, and increase the efficiency, and therefore no effort is needed to promote the use of ETC among customers.

If space is not available to install more gates, the option is to adopt either Scheme C or D. The waiting time in both schemes is more than 1 minute. Although the number of waiting units are 5 and 6 respectively, it is an improvement from the current setup where the queue length can be as high as 8.

TABLE IV. SIMULATION SETUP

Scheme	Figure with parameter description of setup
A	 <p>Fig 4. Scheme A setup</p>
B	 <p>Fig. 5. Scheme B setup</p>
C	 <p>Fig. 6. Scheme C setup</p>

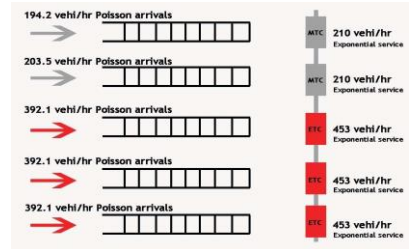
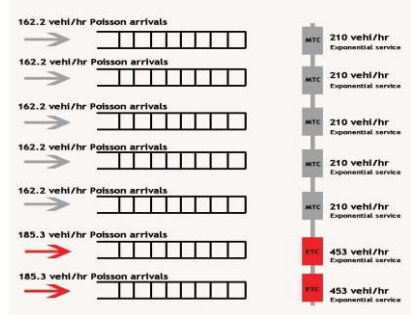
D	 <p>Fig.7. Scheme D setup</p>
E	 <p>Fig.8. Scheme E setup</p>

TABLE V. SUMMARY OF PERFORMANCE MEASURES FOR EACH SCHEME

Scheme	Avg. waiting time in queue (mins)	Avg. number of units in queue	Avg. time a unit spends in the system (mins)	Max. time a unit spends in the system (mins)	Avg. number of units in the system	Toll gate utilization
A	2.31	8.37 ≈ 8	2.56	8.84	9.30 ≈	93.2%
B	0.43	1.39 ≈	0.69	3.05	2.18 ≈	78%
C	1.64	5.87 ≈	1.86	6.41	6.71 ≈	83.89%
D	1.13	4.8 ≈	1.32	4.71	5.72 ≈	91.47%
E	0.33	0.9 ≈	0.57	2.48	1.62 ≈	68%

Scheme C replaces one of the existing MTC gates, and Scheme D replaces two of the existing MTC gates with ETC gates. In addition to this, an effort is needed to enroll more vehicles in the ETC programme. Based on the arrival rates used in the simulation of these schemes, the minimum number of vehicles that should be converted to ETC when adopting the schemes are:

- Scheme C – 618 vehicles
- Scheme D – 1241 vehicles

Reference [13] in their study on ETC systems in Sri Lanka suggested that having only one ETC gate at the toll plaza does not attract more customers to use it and that the RDA should take measures such as launching a sound marketing campaign to attract more customers to use ETC. Currently, the ETC payments are given a discount which is a positive move towards achieving this objective. In addition to the customer benefits, the authorities will also yield benefits by ETC gates such as, reduction in cash

handling and hence less use of paper to help environmental conservation and reduction in staffing.

However, high installation costs should be borne for ETC gates than for MTC gates. Hence, it is important to take the cost factor into conducting an economic analysis.

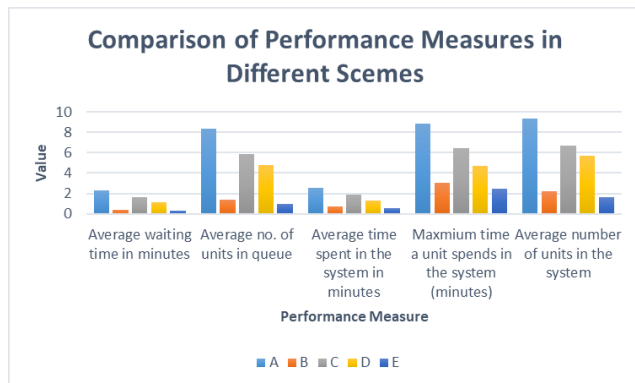


Fig. 9. Performance measures calculated for each simulation setup

In addition to the adoption of a better setup scheme at the toll plaza, the authorities could also take measures to improve the service times at the toll plaza such as encouraging customers to use exact change and adopting exact values for toll [14].

VII. CONCLUSION AND FUTURE WORK

The objective of this study is to analyze the performance of the queuing system at Peliyagoda toll plaza of the Colombo-Katunayake expressway. Using the data collected, the current setup and other four possible setups are simulated with the objective of comparing their performance measures. In the current setup, a vehicle spends 3 minutes on average in the system and the average number of vehicles in a toll gate queue is 8. Clearly this causes a high traffic congestion at the toll plaza. According to the Ministry of Transport in Sri Lanka [15], the vehicle population in the country will gradually increase in the future. Moreover, the number of users utilizing the expressway on a daily basis will increase as a result of the expansion schemes of expressways. Therefore, it is vital that steps are taken to increase the service efficiency of the toll plaza. Based on the results of the four proposed alternative setups for the system, our recommendations are presented in Table VI for the RDA to utilize when making decisions for the improvement of the current system.

One of the major limitations of the study is the unavailability of reliable data for calculating service times due to COVID-related travel restrictions imposed in the country during the data collection period. Therefore, estimated values were obtained using expert opinion as service rates of MTC and ETC gates respectively. More reliable estimates of the performance measures may be obtained if data through observation are used in the simulations. Although the system under study is a multiple-queue multiple-channel queuing system, the simulations are conducted individually for each queue as a series of multiple single-channel queuing systems. As Ceballos & Curtis [10] pointed out, this approach segregates the system into multiple sub-systems and is averse to change as it does not account for users to select the shortest-queue and no queue jumping is allowed.

TABLE VI. RECOMMENDATIONS

Factors to consider	Setup improvement
Space available to add more servers	Add one manual toll gate
No space for more toll gates but cost can be borne for new technology	Convert one or two existing MTC gates to ETC. Subsequently attracting more customers to enroll in the ETC usage is necessary.
Both space and cost for new technology available	Add one MTC gate and one ETC gate

This behaviour is different from normal behaviour at toll plazas where vehicles choose the shortest queue and changing lanes is allowed in some cases. Thus, although a highly complex setup, it might be beneficial to conduct the simulation as a multiple queue multiple-server queuing system. Such a simulation design phase may require careful planning to consider different characteristics of such a system.

The results of the analysis of this study show that adding more servers improves the performance of the system. However, it also increases the service cost. As future work, an economic analysis could be done to find the optimum number of servers that will simultaneously minimize the service cost and the cost incurred due to delays in queues.

In summary, this study performs a simulation-based analysis of the performance of operations at the Peliyagoda expressway toll plaza. Recommendations are drawn based on the results of the simulated setups and their performance measures. Furthermore, possible future work has been stipulated which can add more value in the direction of this study.

REFERENCES

- [1] Central Bank of Sri Lanka. (2019). Annual Report of Central Bank. Annual Report of Central Bank Sri Lanka 2019, 99–139.
- [2] Central Bank of Sri Lanka. (2018). Annual Report of Central Bank. Annual Report of Central Bank Sri Lanka 2018, 88–116.
- [3] Nag, D., Roy, A., & Goswami, A. K. (2020). “Estimating Environmental Benefits of Electronic Toll Collection (ETC)”. *Transportation Research* (441-452), Springer, Singapore
- [4] Mohan, J., Sabareesh, H., Muthukumar, R., & Niranjana, A. (2018). “Queuing Model (M/M/C:∞/FIFO) to Erode Railway Ticket Counters”. *International Journal of Scientific Development and Research*, 3(11), 382–385.
- [5] Amin, A., Mehta, P., Sahay, A., Kumar, P., & Kumar, A. (2014). “Optimal solution of real time problems using Queueing Theory”. *International Journal of Engineering and Innovative Technology (IJEIT)*, 3(10), 2277–2279. http://www.ijeit.com/Vol_3/Issue_10/IJEIT1412201404_52.pdf
- [6] Malipatil, N., Avati, S. I., Vinay, H. N., & Sunil, S. (2017). “Application of Queueing Theory to a Toll Plaza-A Case Study”. In *Lecture Notes in Civil Engineering Tom* (Vol. 45). https://doi.org/10.1007/978-981-32-9042-6_20
- [7] Sihotang, E., Sugito, Mustafid, Ispriyanti, D., Prahutama, A., & Rachman, A. (2020). “Analysis of queue and performance of automatic toll booths with a normal distribution” (case study: Automatic booths toll gate muktiharjo). *Journal of Physics: Conference Series*, 1524(1). <https://doi.org/10.1088/1742-6596/1524/1/012093>
- [8] Shanmugasundaram, S., & Punitha, S. (2014). “A Simulation Study on toll gate system in M/M/1 Queueing Models”. *IOSR Journal of Mathematics*, 10(3), 01–09. <https://doi.org/10.9790/5728-10360109>
- [9] Duhan, D., Arya, N., Dhanda, P., Upadhyay, L., & Mathiyazhagan, K. (2014). “Application of queueing theory to address traffic problems at a highway toll plaza”. *Applied Mechanics and Materials*, 592–594, 2583–2587. <https://doi.org/10.4028/www.scientific.net/AMM.592-594.2583>

- [10] Ceballos, G., & Curtis, O. (2004). "Queue Analysis at Toll and Parking Exit Plazas: A Comparison between Multi-server Queuing Models and Traffic Simulation". ITE 2004 Annual Meeting and Exhibit. <http://ntlsearch.bts.gov/tris/record/tris/00981493.html>
- [11] Punitha, S. (2018). "Design and evaluation of traffic delays in toll plaza using combination of queuing and simulation". *Journal of Physics: Conference Series*, 1139(1). <https://doi.org/10.1088/1742-6596/1139/1/012080>
- [12] Antil, S. (2017). "Application of Queuing Theory on Toll Plaza to Solve Traffic Problem". *International Journal for Scientific Research & Development*, 5(7), 165–167.
- [13] Rodrigo, A., & Hewage, D. (2015). "A Study on Electronic Toll Collection Systems in Expressways in Sri Lanka", (Issue September). Colombo International Nautical and Engineering Campus.
- [14] Lima, J. P., Inácio, P. P. A., & Leal, F. (2019). "Service levels of highway toll plazas: The influence of factors on manual customer service". In *Production* (Vol. 29, pp. 1–16). <https://doi.org/10.1590/0103-6513.20180032>
- [15] Ministry of Transport. (2021). Vehicle population. https://www.transport.gov.lk/web/index.php?option=com_content
- [16] Bose, S. K. (2002). *An Introduction to Queuing Systems*.
- [17] Dhar, S. K., & Rahman, T. (2013). "Case Study for Bank ATM Queuing Model". *IOSR Journal of Mathematics*, 7(1), 01–05. <https://doi.org/10.9790/5728-0710105>
- [18] Dharmawirya, M., & Adi, E. (2011). "Case Study for Restaurant Queuing Model". In *International Conference on Management and Artificial Intelligence*. [https://doi.org/10.9790/487x-1902039398Dharmawirya, M., Oktadiana, H., & Adi, E. \(2012\).](https://doi.org/10.9790/487x-1902039398Dharmawirya, M., Oktadiana, H., & Adi, E. (2012).)
- [19] Dilrukshi, P. A. D., Nirmanamali, H. D. I. M., Lanel, G. H. J., & Samarakoon, M. A. S. C. (2016). "A Strategy to Reduce the Waiting Time at the Outpatient Department of the National Hospital in Sri Lanka". *International Journal of Scientific and Research Publications*, 6(2), 281–2250. www.ijsrp.org
- [20] Jayasooriya, S. A. C. S., & Bandara, Y. M. M. S. (2017). "Measuring the Economic costs of traffic congestion". 3rd International Moratuwa Engineering Research Conference, MERCon 2017, May 2017, 141–146. <https://doi.org/10.1109/MERCon.2017.7980471>
- [21] Katz, K. L., Larson, B. M., & Larson, R. C. (1991). "Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, Engage". *Sloan Management Review*, 32(2).
- [22] Lakshmi, C., & Sivakumar, A. I. (2013). "Application of queuing theory in health care: A literature review". *Operations Research for Health Care*, 2(1–2), 25–39. <https://doi.org/10.1016/j.orhc.2013.03.002>
- [23] Latif, N. N. A., Zainal, M., Haris, N. F. M., Shafie, S., Misiran, M., & Yusof, Z. M. (2019). "Assessing Fast-Food Restaurant's Productivity in Rural Area Through Customer Waiting Time and Worker's Efficiency". *International Journal of Modern Trends in Business Research*, 2(8), 1–7.
- [24] Law, A. K. Y., Hui, Y. V., & Zhao, X. (2004). "Modeling repurchase frequency and customer satisfaction for fast food outlets". *International Journal of Quality and Reliability Management*, 21(5), 545–563. <https://doi.org/10.1108/02656710410536563>
- [25] Little, J. D. C. (1961). "A Proof for the Queuing Formula: $L = \lambda W$ ". *Operations Research*, 9(3), 383–387. <https://doi.org/10.1287/opre.9.3.383>
- [26] Mital, K. M. (2010). "Queuing analysis for outpatient and inpatient services: A case study". *Management Decision*, 48(3), 419–439. <https://doi.org/10.1108/00251741011037783>
- [27] Molla, M. A.-A. (2017). "Case Study for Shuruchi Restaurant Queuing Model". *IOSR Journal of Business and Management*, 19(02), 93–98. <https://doi.org/10.9790/487x-1902039398>
- [28] Ogunsakin, R. E., Babalola, B. T., & Adedara, M. T. (2013). "Comparison of Service Delivery by ATM in Two Banks: Application of Queuing Theory". *IOSR Journal of Mathematics*, 9(3), 50–54. <https://doi.org/10.9790/5728-0935054>
- [29] Patel, P. J. J., Chadhaury, P. R. M., & Patel, P. J. M. (2012). "Queuing Theory and its Application at Railway Ticket Window". *Journal Of Information, Knowledge And Research In Information Technology*, 2(1), 99–112.
- [30] Roslow, D. S., Nicholls, D. J. A. F., & Tsalikis, D. J. (1992). "Time And Quality: Twin Keys To Customer Service Satisfaction". *Journal of Applied Business Research*, 8(2). <https://doi.org/https://doi.org/10.19030/jabr.v8i2.6168>
- [31] Sameer, S. S. (2014). "Simulation: Analysis of Single Server Queuing Model". *International Journal on Information Theory*, 3(3), 47–54. <https://doi.org/10.5121/ijit.2014.3305>
- [32] Shastrakar, D. F., Pokley, S. S., & Patil, K. D. (2016). "Literature Review of Waiting Lines Theory and its Applications in Queuing Model". *International Journal of Engineering Research & Technology* (IJERT), 4(30), 1–4.
- [33] Vilčeková, S., Apostoloski, I. Z., Mečiarová, L., Burdová, E. K., & Kiseľák, J. (2017). "Investigation of indoor air quality in houses of Macedonia". *International Journal of Environmental Research and Public Health*, 14(1), 1–12. <https://doi.org/10.3390/ijerph14010037>

Docker incorporation is different from other computer system infrastructures: A review

W. M. C. J. T. Kithulwatta*
 Faculty of Graduate Studies,
 Sabaragamuwa University of Sri Lanka, Sri Lanka
 chiranthajtk@gmail.com

B. T. G. S. Kumara
 Dept. of Computing & Information Systems, Fac. of Applied
 Sciences, Sabaragamuwa University of Sri Lanka, Sri Lanka
 kumara@appsc.sab.ac.lk

K. P. N. Jayasena
 Dept. of Computing & Information Systems, Fac. of Applied
 Sciences, Sabaragamuwa University of Sri Lanka, Sri Lanka
 pubudu@appsc.sab.ac.lk

R. M. K. T. Rathnayaka
 Department of Physical Sciences & Technology, Fac. of Applied
 Sciences, Sabaragamuwa University of Sri Lanka, Sri Lanka
 kapilar@appsc.sab.ac.lk

Abstract - Currently the computing world is getting complex, innovating and maturing with modern technologies. Virtualization is one of the old concepts and currently containerization has arrived as an alternative and innovative technology. Docker is the most famous and trending container management technology. Different other container management technologies and virtualization technologies are respective other corresponding technologies and mechanisms for Docker containerization. This research study aims to identify how Docker incorporation is different from other computer system infrastructure technologies in the perspective of architecture, features and qualities. By considering forty-five existing literatures, this research study was conducted. To deliver a structured review process, a thorough review protocol was conducted. By considering four main research questions, the research study was lined up. Ultimately, Docker architecture and Docker components, Docker features, Docker integration with other computing domains and Docker & other computing infrastructures were studied. After synthesizing all the selected research studies, the cream was obtained with plenty of knowledge contribution to the field of computer application deployment and infrastructure.

Keywords - computer infrastructure, containers, docker, virtualization, virtual machines

I. INTRODUCTION

Computer virtualization has existed for a long time. As well, virtualization is an old conceptualization within the computing domain. Traditionally, most information technology (IT) services are bound with hardware components and virtualization enables those services in a virtual manner [1]. A software component called hypervisor, creates separate physical resources in the virtual environment. The hypervisor keeps on top of an operating system and ultimately, the virtual machine makes the interaction between end-users and the computing system. Figure 1 presents the virtualization stack architecture.

On top of any hardware platform, an operating system was launched and on top of that operating system, the hypervisor was launched. On the hypervisor, each virtual machine carries the full functional operating system. Each virtual machine provides a separated environment for the software applications and services.

Within virtualization, each virtual machine has a heavy weight since a virtual machine has a full set of functional operating systems. Therefore, an alternative and novel concept was arrived called containerization. Within the containerization, containers play a major role.

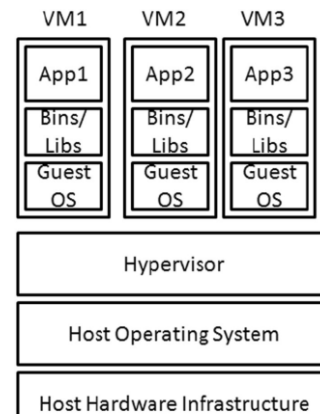


Fig. 1. Virtualization architecture [2]

Figure 2 presents the container architecture as a pictorial way [2]. According to the container architecture, it consists of a container engine instead of the hypervisor. On the container engine, each container keeps a packaged environment by including all fundamental dependencies to run the software applications. Each container provides an isolated environment for the software applications from the host computer infrastructure and other containers.

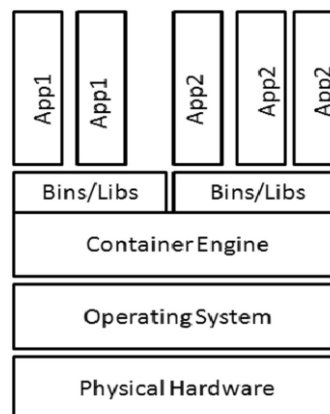


Fig. 2: Container architecture [2]

II. MOTIVATION

Within the practitioner of the containers, Docker is one of the available container management technologies. Other than Docker: Rkt and Linux containers are available as container technologies.

According to the official Docker documentation website, more than eleven million developers are engaged for Docker developments. As well, more than seven million Docker based software applications are made. More than thirteen billion Docker images are downloaded for the Docker based practitioner usages [3].

As mentioned in the official Docker documentation, most of the widely used computing tools are engaged with Docker containerization. Few of them are Bitbucket, GitLab, GitHub, NGINX, Redis, Jenkins, JFrog, MongoDB, Visual Studio Code, etc. [3].

Currently most industries and clients are using Docker oriented software applications and few of those clients are Paypal, Adobe, Netflix, University of Calgary, PathFactory, etc. [3]. Furthermore, Docker trusted contents are offered by Docker verified publishers as reliable Docker packaged blocks. Some of those publishers are Amazon Web Services (AWS), RedHat, Datadog, etc. [3].

Therefore, it depicts that Docker has higher practitioner engagement. Hence Docker is having a higher trend within the practical approach. Currently there is a higher competition for Docker among other container technologies and other infrastructure approaches like virtual instances. This research study was designed to identify the differences for Docker with other computer system infrastructure approaches.

The overall research study brings answers for the below research questions (RQs).

RQ1: *What kinds of components are embedded in the Docker architecture?*

RQ2: *What kind of benefits are available for the Docker based container approaches?*

RQ3: *What kind of computing areas/domains are integrated with Docker?*

RQ4: *How do Docker and other infrastructure approaches differ?*

III. RESEARCH METHODOLOGY

To obtain a thorough review analysis, the research study followed a highly structured review protocol. The ultimate review protocol is with eight steps. Table I presents the applied protocol as steps. The table I is with three columns. The first column presents the review protocol step number, second column presents the respective step name and third column presents the step in more descriptively.

TABLE I. REVIEW PROTOCOL IN STEPS

Step Number	Step Name	Step in Detail
Step 1	Need for the review	Identify the need for the review and the need was identified at section II.
Step 2	Research Questions	Declare the research questions and research questions were identified at section II.
Step 3	Identify the search strings	The search string was declared to select primary literatures. The identified search string was declared below (1).
Step 4	Primary literature selection	By using the identified search string, primary literatures were selected. The search string was browsed in the Google Scholar. Then primary studies were selected from the scientific databases including IEEE-Xplore,

Step Number	Step Name	Step in Detail
		ACM Digital Library, Springer and Science Direct.
Step 5	Inclusion/Exclusion	To filter the papers from the domain, the paper inclusion and exclusion criteria was applied.
Step 6	Quality Assessment	To filter the inapplicable literatures from the primary literature bulk, paper quality assessment was executed. After the process, 45 papers were finalized.
Step 7	Synthesizing	On top of the selected papers, the synthesizing was applied.
Step 8	Final reporting	By including the final observations, investigations and results of the research study, a final research report was made.

The search string:

$$(Docker) \wedge [(infrastructure) \vee (cloud) \vee (containers)] \quad (1)$$

Other than the scientific databases, the official Docker documentation website was used as primary literature to identify the latest updates on Docker container technology.

For the review study, the research papers were selected by applying the search string. Docker container technology was introduced in 2013. Hence the selected literatures were published from 2014 to 2020. The table II presents the all referred literatures. The table II is with two columns: first column is for the study topic and second column is for the citation number.

TABLE II. THE LIST OF SELECTED LITERATURES

Topic of the Literature	Citation Number
What is Virtualization?	[1]
Exploring the support for high performance applications in the container runtime environment	[2]
Empowering App Development for Developers Docker	[3]
Docker overview	[4]
Containers & Docker: Emerging roles & future of Cloud technology	[5]
Advantages of Docker	[6]
Performance comparison between Linux containers and virtual machines	[7]
An Introduction to Docker and Analysis of its Performance	[8]
Performance Comparison Analysis of Linux Container and Virtual Machine for Building Cloud	[9]
An updated performance comparison of virtual machines and Linux containers	[10]
Virtualization and containerization of application infrastructure: A comparison	[11]
A Comparative Study of Containers and Virtual Machines in Big Data Environment	[12]
Evaluation of Docker as Edge Computing Platform	[13]
Using Docker in High Performance Computing Applications	[14]
The research and implementation of cloud computing platform based on Docker	[15]
A Study of Security Vulnerabilities on Docker Hub	[16]
Evaluation of Docker Containers Based on Hardware Utilization	[17]
Docker Cluster Management for the Cloud - Survey Results and Own Solution	[18]

Topic of the Literature	Citation Number
Leveraging microservices architecture by using Docker technology	[19]
To Docker or Not to Docker: A Security Perspective	[20]
Measuring Docker Performance: What a mess!!!*	[21]
Containers and Cloud: From LXC to Docker to Kubernetes	[22]
Docker ecosystem – Vulnerability Analysis	[23]
Distributed Systems of Microservices Using Docker and Serfnode	[24]
Model-Driven Management of Docker Containers	[25]
Feasibility of Fog Computing Deployment based on Docker Containerization over RaspberryPi	[26]
Improvement of Container Scheduling for Docker using Ant Colony Optimization	[27]
Using Docker Containers to Improve Reproducibility in Software and Web Engineering Research	[28]
Autonomic Vertical Elasticity of Docker Containers with ELASTICDOCKER	[29]
Integrating Containers into Workflows: A Case Study Using Makeflow, Work Queue, and Docker	[30]
DIVDS: Docker Image Vulnerability Diagnostic System	[31]
Orchestrating Docker Containers in the HPC Environment	[32]
A Docker Container Anomaly Monitoring System Based on Optimized Isolation Forest	[33]
An Empirical Analysis of the Docker Container Ecosystem on GitHub	[34]
Containers & Docker: Emerging Roles & Future of Cloud Technology	[35]
In Search of the Ideal Storage Configuration for Docker Containers	[36]
Measurement and Evaluation for Docker Container Networking	[37]
Building A Virtual System of Systems Using Docker Swarm in Multiple Clouds	[38]
A Defense Method against Docker Escape Attack	[39]
DoCloud: An elastic cloud platform for Web applications based on Docker	[40]
CoMICon: A Co-operative Management System for Docker Container Images	[41]
FID: A Faster Image Distribution System for Docker Platform	[42]
Orchestration of Containerized Microservices for IIoT using Docker	[43]
A Holistic Evaluation of Docker Containers for Interfering Microservices	[44]
Application deployment using Microservice and Docker containers:Framework and optimization	[45]

IV. DOCKER ARCHITECTURE

The Fig. 3 presents the Docker architecture in a pictorial view. To design the fundamental Docker architecture, client and server architecture has been used. Docker daemon, Docker client and Docker registries are the main components for the Docker architecture [4]

The Docker daemon has a main responsibility to manage Docker objects. Main Docker objects are containers, images, volumes and network. One Docker daemon can communicate with other Docker daemons. To make the interaction with Docker, Docker client was used as the fundamental way. While using the Docker commands on the Docker client, it sends those commands to Docker daemon. One Docker client can communicate with more than one Docker daemons [4].

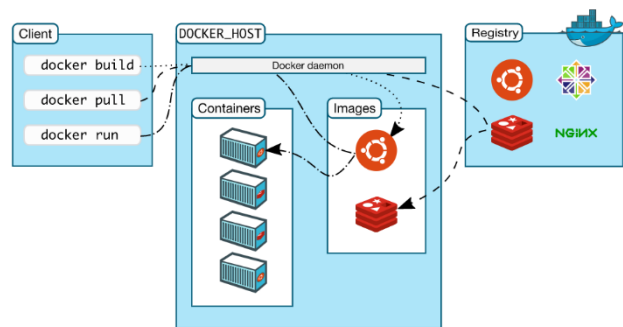


Fig. 3: Docker architecture [4]

The Docker daemon has a main responsibility to manage Docker objects. Main Docker objects are containers, images, volumes and network. One Docker daemon can communicate with other Docker daemons. To make the interaction with Docker, Docker client was used as the fundamental way. While using the Docker commands on the Docker client, it sends those commands to Docker daemon. One Docker client can communicate with more than one Docker daemons [4].

Docker client and Docker daemon can be executed on the same infrastructure. According to the designed way, Docker client can be connected to a remote Docker daemon [4].

To store and archive the Docker images, a dedicated location was allocated in the Docker architecture called Docker registry. According to the use-cases, publicly available Docker registry or private Docker registries can be used. Users can pull Docker images from the Docker registry or push the Docker images to Docker registry [4].

Docker image is one of the most important parts of the Docker architecture and it consists of a read-only template with a set of instructions to create a Docker container. By using a Dockerfile, specific Docker images can be created. As well, Docker container is the executable instance of a Docker image. By using Docker application programming interface or command line interface, Docker containers can be created, stopped, started, moved or deleted [4].

V. DOCKER FEATURES

This section emphasizes the Docker related advantages, incorporations and compatibility of the Docker with other computing technologies.

A. Docker Advantages

Table III presents the Docker advantages in a more advanced way. The first column denotes the Docker advantages and the second column denotes the advantages more descriptively.

B. Docker Integrated Areas/Domains/Communities

The Docker containerized technology is not only dedicated as the software application deploying environment. According to the referred literature studies, Docker container technology was integrated with different computing domains and areas. As a summarized list, the below list presents those Docker engaged computing areas [1] - [45].

- Edge Computing
- Computer Networking

- Cloud Computing
- Computer Security
- Grid Computing
- Distributed Computing
- Operating Systems
- Web Engineering
- High Performance Computing
- System Engineering
- Internet of Things
- Autonomous Computing
- Parallel Computing
- Microservices

Hence, the above list depicts that Docker was spread in a variety of computing domains. Within the above domains, Docker was used as a runtime environment, virtualization and an operating system. Mainly, Docker was used for development, testing, deploying and experimenting purposes.

TABLE III. DOCKER ADVANTAGES

DOCKER Advantage	Advantage in more Descriptively
Lightweight	Docker is with lightweight containers and images than traditional virtual machines. Since traditional virtual machines carry a full set of operating systems, a virtual machine is heavy weight. Furthermore, one virtual machine consumes heavy resources from the host computer infrastructure to execute a full set of operating systems [5].
Portable	Docker containers and images can be moved as one module within any computer system infrastructure easily: therefore, Docker is portable. Due to the portability, Docker images can be shared with different hosts easily. However, traditional virtual machines can be moved within different hosts but it is more heavy and has to follow more steps [5] [6].
Scalability	Docker is providing a facility to scale the Docker containers and services by up and/or down the number of replicas. Docker takes the responsibility to upgrade or downgrade the number of replicas very smoothly, without making any effect on the software service. Therefore, Docker can be provisioned more easily than the virtual machines [7].
Best fit for microservices	According to the microservices architecture, software applications need a separated and isolated environment. Therefore, Docker makes an isolated environment within the Docker containers and it helps to give the best software environment for the microservices softwares. Without making any conflicts with other modules or components, Docker provides the best fit for microservices [7].
Optimal resource utilization	Docker container structure shares the host computer resources among the Docker objects. Docker has the facility to allocate limitations for each Docker object to utilize the host memory, CPU, disk space and network. Due to those limitations and constraints, Docker has optimal resource utilization [5].

As well, Docker container technology has presented an excellent research path in computing. Research scholars have presented that Docker brings a strong research direction.

For the government of Docker or other container farms, container orchestration solutions are needed. Therefore, Kubernetes has been identified as the best fit for Docker container orchestration with amazing and fantastic features & functions. Mainly identified Kubernetes features for Docker are automatic rollouts, automated roll backs, storage orchestration, load balancing, service discovery, configuration management, batch execution, horizontal scaling, self-healing, automated bin packing, etc.

C. Docker and Other Corresponding Approaches

Docker has been identified as the best computer infrastructure for the software application deployments. Other than Docker, there are different kinds of container management technologies and virtual environments. Most scholars have made different comparisons among Docker and other corresponding approaches.

Virtual machines use an extra layer called *hypervisor* and the hypervisor is between the host operating system and guest operating system (*Figure 1 presents the location of the hypervisor according to the virtualization architecture*). However, containers add up an additional layer between the host operating system and where the applications are virtualized and executed. It was known as a container engine. Since containers do not use any guest operating system, it makes a considerable performance difference between container technology and virtual machine technology [8].

Below tables IV, V and VI present the performances of different container vendors and virtual machines. According to the paper [9], Docker container performance is better than KVM (Kernel-based Virtual Machines) in terms of boot time and calculation speed [9]. But another research paper has proved that there is no difference of wastage of host resources between Docker and KVM but there is a noticeable difference in execution as KVM is faster than Docker containers [10]. The research paper [11] has presented that LXC (Linux Containers) takes a longer time to accomplish a defined task. But XenServer took less time than LXC. LXC is a better container in the sense of fewer wasted resources while Xen is better in the sense of equally distributing resources.

TABLE IV. DOCKER AND KVM

Reference: [9]	
DOCKER	KVM
Short boot time	Long boot time
Calculation speed is faster	Calculation speed is slower
No guest operating system	Works independently

According to the above summarized Table IV, KVM is working independently due to KVM having a hypervisor and Docker has no guest operating system. But Docker shares the host operating system resources.

As mentioned in the literature, LXC consumes less overhead on the parameter of host computer resources. Same as that, XenServer has consumed more overhead. The

author has identified that XenServer was better in the sense of distributing host computer resources equally. But LXC was not like that and LXC was better in the sense of executing fully isolated processors [11].

TABLE V. XEN-SERVER AND COREOS

Reference: [11]	
XenServer (Xen)	CoreOS (LXC)
More overhead (regarding wastage of resources)	Less overhead (regarding wastage of resources)
Less time to accomplish request	Longer time to accomplish request
Better in sense of equally distributing resources	Better in sense of executing isolated processes

According to the above summarized Table VI, Docker and KVM have presented a more mature innovation than native approach. As well, KVM has demonstrated very less host computer resources wastage than both native and Docker approaches.

TABLE VI. NATIVE, DOCKER AND KVM

Reference: [10]		
Native	Docker	KVM
Overhead (regarding wastage of resources)	Slightly less overhead than native	Slightly less overhead than native and Docker
Slow execution equal to Docker	Slow execution equal to native	Fast execution
-	Mature innovation	Mature innovation

Apart from the above comparisons, a recent research paper has presented differences between containers and virtual machines. A container consists of executable software application binaries and executable codes. All fundamentally necessary software dependencies need to run a container. Containers are using Linux kernel mechanisms to allocate resources. The authors have said that engineers can allow allocating resources for the containers like network configurations, CPU and memory at the time of container creation. The allocated resources may be adjusted dynamically but any container cannot use more resources than being specified [12].

The paper [12] has expressed that, the first difference between containers and virtual machines is: containers are more lightweight than virtual machines. The due reason is: containers include only executable applications and their dependencies. The containers which are on the same machine, share the host operating system resources among containers. Respective virtual machines do not share the host operating system resources. Virtual machines contain a full set of operating systems. Furthermore, the same paper [12] has presented that virtual machines can run as any operating system that is different from the host machine. But containers need to use the same operating system as the host machine.

The authors of the paper [12] have presented the second comparison on the hypervisor. For the virtual machine environment, the hypervisor is necessary to use such as VMware ESXi and KVM. It is not required for containers. Virtual machines are functioning as an independent

machine by keeping all control of all resources under the virtual machines. Furthermore, virtual machines are running as non-privileged mode and containers are running on privileged mode. It depicts that virtual machines cannot execute many privileged instructions. As well as, for the execution of instructions, the virtual machine environment is needed to translate all virtual machine instructions to executable commands to which that needs to run on the host. However, containers make communication with the host directly by system calls and it does not require any intermediate mechanism to convert instructions [12].

Furthermore, the paper [12] has discussed image files of virtual machines and containers. Virtual machines have their own images and containers share some of their images. Container images are created as a layered architecture. To create an image on an existing image, the platform adds another layer on the original image. Image files of different virtual machines are isolated from each other [12].

The authors of [12] have presented their research findings as researchers and practitioners pay their attention to containers instead of virtual machines. Containers are more cost-effective. Furthermore, containers usually consist of tens of Megabytes (MB) while virtual machines can take about several Gigabytes (GB). To run an application, a container uses very fewer resources than virtual machines due to containers not needing to maintain an operating system. Containerized platforms do not contain any hypervisor and containers present more performance than virtual machines [12].

VI. CONCLUSIONS

Containerization was identified as an alternative for virtualization. Within the practitioner of the container management technologies, Docker keeps and plays a major role. Currently millions and billions of customer interactions are with Docker container management. Docker has client and server architecture. As well, Docker daemon, Docker client, Docker registry and Docker objects play main roles in the Docker platform. Those components and modules help to carry answers for the RQ1. Docker has many available features and benefits. Some of them are scalability, portability, lightweight, best fit for microservices and optimal resource utilization. Hence those features provide the answers to the RQ2.

Without limiting to software application launching on Docker, Docker containerization was engaged with many computing technologies. Few of them are fog computing, cloud computing, grid computing, Internet of Things, microservices, etc. Those are answered to the RQ3. As presented above in the V.C section, many of Docker and other infrastructure technologies were discussed. Hence those are answered to the RQ4.

The scholarly research articles present that Docker has a higher engagement with all kinds of computing technologies. Docker plays a major role in computer system administration engineering.

REFERENCES

- [1] "What is Virtualization? ", 2021. [Online]. Available: <https://www.redhat.com/en/topics/virtualization/what-is-virtualization> [Accessed: 09- Jul- 2021].
- [2] J. Martin, A. Kandasamy and K. Chandrasekaran, "Exploring the support for high performance applications in the container runtime environment", Human-centric Computing and

- Information Sciences, vol. 8, no. 1, 2018. Available: 10.1186/s13673-017-0124-3
- [3] "Empowering App Development for Developers | Docker", Docker, 2021. [Online]. Available: <https://www.docker.com/>. [Accessed: 09- Jul- 2021].
- [4] "Docker overview", Docker Documentation, 2021. [Online]. Available: <https://docs.docker.com/get-started/overview/>. [Accessed: 09- Jul- 2021].
- [5] S. Singh and N. Singh, "Containers & Docker: Emerging roles & future of Cloud technology," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 804-807, doi: 10.1109/ICATCCT.2016.7912109.
- [6] Vase, Tuomas. "Advantages of Docker." (2015).
- [7] A. M. Joy, "Performance comparison between Linux containers and virtual machines," 2015 International Conference on Advances in Computer Engineering and Applications, 2015, pp. 342-346, doi: 10.1109/ICACEA.2015.7164727.
- [8] B. B. Rad, H. J. Bhatti, M. Ahmadi, "An Introduction to Docker and Analysis of its Performance", IJCSNS International Journal of Computer Science and Network Security, vol. 17, no. 3, March 2017.
- [9] K. Seo, H. Hwang, I. Moon, O. Kwon and B. Kim, "Performance Comparison Analysis of Linux Container and Virtual Machine for Building Cloud", 2014. Available: 10.14257/astl.2014.66.25.
- [10] W. Felter, A. Ferreira, R. Rajamony and J. Rubio, "An updated performance comparison of virtual machines and Linux containers," 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2015, pp. 171-172, doi: 10.1109/ISPASS.2015.7095802.
- [11] M. J. Scheepers, "Virtualization and containerization of application infrastructure: A comparison", 21st Twente Student Conference on IT, pp. 1-7, 2014
- [12] Q. Zhang, L. Liu, C. Pu, Q. Dou, L. Wu and W. Zhou, "A Comparative Study of Containers and Virtual Machines in Big Data Environment," 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), 2018, pp. 178-185, doi: 10.1109/CLOUD.2018.00030.
- [13] B. I. Ismail et al., "Evaluation of Docker as Edge computing platform," 2015 IEEE Conference on Open Systems (ICOS), 2015, pp. 130-135, doi: 10.1109/ICOS.2015.7377291.
- [14] M. T. Chung, N. Quang-Hung, M. Nguyen and N. Thoai, "Using Docker in high performance computing applications," 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE), 2016, pp. 52-57, doi: 10.1109/CCE.2016.7562612.
- [15] D. Liu and L. Zhao, "The research and implementation of cloud computing platform based on docker," 2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2014, pp. 475-478, doi: 10.1109/ICCWAMTIP.2014.7073453.
- [16] R. Shu, X. Gu and W. Enck, "A Study of Security Vulnerabilities on Docker Hub", Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, 2017, pp.269-280, doi: 10.1145/3029806.3029832
- [17] Preeth E N, F. J. P. Mulerickal, B. Paul and Y. Sastri, "Evaluation of Docker containers based on hardware utilization," 2015 International Conference on Control Communication & Computing India (ICCC), 2015, pp. 697-700, doi: 10.1109/ICCC.2015.7432984.
- [18] R. Peinl, F. Holzschuher and F. Pfitzer, "Docker Cluster Management for the Cloud - Survey Results and Own Solution", Journal of Grid Computing, vol. 14, no. 2, pp. 265-282, 2016. doi: 10.1007/s10723-016-9366-y.
- [19] D. Jaramillo, D. V. Nguyen and R. Smart, "Leveraging microservices architecture by using Docker technology," SoutheastCon 2016, 2016, pp. 1-5, doi: 10.1109/SECON.2016.7506647.
- [20] T. Combe, A. Martin and R. Di Pietro, "To Docker or Not to Docker: A Security Perspective," in IEEE Cloud Computing, vol. 3, no. 5, pp. 54-62, Sept.-Oct. 2016, doi: 10.1109/MCC.2016.100.
- [21] E. Casalicchio and V. Perciballi, "Measuring Docker Performance", Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion, 2017, pp. 11-16, doi: 10.1145/3053600.3053605.
- [22] D. Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes," in IEEE Cloud Computing, vol. 1, no. 3, pp. 81-84, Sept. 2014, doi: 10.1109/MCC.2014.51.
- [23] A. Martin, S. Raponi, T. Combe and R. Di Pietro, "Docker ecosystem - Vulnerability Analysis", Computer Communications, vol. 122, pp. 30-43, 2018. doi: 10.1016/j.comcom.2018.03.011.
- [24] J. Stubbs, W. Moreira and R. Dooley, "Distributed Systems of Microservices Using Docker and Serfnode," 2015 7th International Workshop on Science Gateways, 2015, pp. 34-39, doi: 10.1109/IWSG.2015.16.
- [25] F. Paraiso, S. Challita, Y. Al-Dhuraibi and P. Merle, "Model-Driven Management of Docker Containers," 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), 2016, pp. 718-725, doi: 10.1109/CLOUD.2016.0100.
- [26] P. Bellavista and A. Zanni, "Feasibility of Fog Computing Deployment based on Docker Containerization over RaspberryPi", Proceedings of the 18th International Conference on Distributed Computing and Networking, 2017, pp. 1-10 doi: 10.1145/3007748.3007777.
- [27] C. Kaewkasi and K. Chuenmuneewong, "Improvement of container scheduling for Docker using Ant Colony Optimization," 2017 9th International Conference on Knowledge and Smart Technology (KST), 2017, pp. 254-259, doi: 10.1109/KST.2017.7886112.
- [28] J. Cito and H. C. Gall, "Using Docker Containers to Improve Reproducibility in Software Engineering Research," 2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C), 2016, pp. 906-907.
- [29] Y. Al-Dhuraibi, F. Paraiso, N. Djarallah and P. Merle, "Autonomic Vertical Elasticity of Docker Containers with ELASTICDOCKER," 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), 2017, pp. 472-479, doi: 10.1109/CLOUD.2017.67.
- [30] C. Zheng and D. Thain, "Integrating Containers into Workflows: A Case Study Using Makeflow, Work Queue, and Docker ", Proceedings of the 8th International Workshop on Virtualization Technologies in Distributed Computing, 2015, pp. 31-38 doi: 10.1145/2755979.2755984.
- [31] S. Kwon and J. Lee, "DIVDS: Docker Image Vulnerability Diagnostic System," in IEEE Access, vol. 8, pp. 42666-42673, 2020, doi: 10.1109/ACCESS.2020.2976874.
- [32] J. Higgins, V. Holmes and C. Venters, "Orchestrating Docker Containers in the HPC Environment", Lecture Notes in Computer Science, pp. 506-513, 2015. doi: 10.1007/978-3-319-20119-1_36.
- [33] Z. Zou, Y. Xie, K. Huang, G. Xu, D. Feng and D. Long, "A Docker Container Anomaly Monitoring System Based on Optimized Isolation Forest," in IEEE Transactions on Cloud Computing, doi: 10.1109/TCC.2019.2935724.
- [34] J. Cito, G. Schermann, J. E. Wittern, P. Leitner, S. Zumberi and H. C. Gall, "An Empirical Analysis of the Docker Container Ecosystem on GitHub," 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), 2017, pp. 323-333, doi: 10.1109/MSR.2017.67.
- [35] S. Singh and N. Singh, "Containers & Docker: Emerging roles & future of Cloud technology," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 804-807, doi: 10.1109/ICATCCT.2016.7912109.
- [36] V. Tarasov et al., "In Search of the Ideal Storage Configuration for Docker Containers," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), 2017, pp. 199-206, doi: 10.1109/FAS-W.2017.148.
- [37] H. Zeng, B. Wang, W. Deng and W. Zhang, "Measurement and Evaluation for Docker Container Networking," 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2017, pp. 105-108, doi: 10.1109/CyberC.2017.78.
- [38] N. Naik, "Building a virtual system of systems using docker swarm in multiple clouds," 2016 IEEE International Symposium on Systems Engineering (ISSE), 2016, pp. 1-3, doi: 10.1109/SysEng.2016.7753148.
- [39] Z. Jian and L. Chen, "A Defense Method against Docker Escape Attack", Proceedings of the 2017 International Conference on Cryptography, Security and Privacy - ICCSP '17, 2017, pp. 142-146, doi: 10.1145/3058060.3058085.
- [40] C. Kan, "DoCloud: An elastic cloud platform for Web applications based on Docker," 2016 18th International Conference on Advanced Communication Technology (ICACT), 2016, pp. 1-1, doi: 10.1109/ICACT.2016.7423439.
- [41] S. Nathan, R. Ghosh, T. Mukherjee and K. Narayanan, "CoMICon: A Co-Operative Management System for Docker Container Images," 2017 IEEE International Conference on

- Cloud Engineering (IC2E), 2017, pp. 116-126, doi: 10.1109/IC2E.2017.24.
- [42] W. Kangjin, Y. Yong, L. Ying, L. Hanmei and M. Lin, "FID: A Faster Image Distribution System for Docker Platform," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), 2017, pp. 191-198, doi: 10.1109/FAS-W.2017.147.
- [43] J. Rufino, M. Alam, J. Ferreira, A. Rehman and K. F. Tsang, "Orchestration of containerized microservices for IIoT using Docker," 2017 IEEE International Conference on Industrial Technology (ICIT), 2017, pp. 1532-1536, doi: 10.1109/ICIT.2017.7915594.
- [44] D. N. Jha, S. Garg, P. P. Jayaraman, R. Buyya, Z. Li and R. Ranjan, "A Holistic Evaluation of Docker Containers for Interfering Microservices," 2018 IEEE International Conference on Services Computing (SCC), 2018, pp. 33-40, doi: 10.1109/SCC.2018.00012.
- [45] X. Wan, X. Guan, T. Wang, G. Bai and B. Choi, "Application deployment using Microservice and Docker containers: Framework and optimization", *Journal of Network and Computer Applications*, vol. 119, pp. 97-109, 2018. doi: 10.1016/j.jnca.2018.07.003.

Vibration analysis to detect and locate engine misfires

Prathap V. Jayasooriya*

Dept. of Mechanical Engineering
Faculty of Engineering, University of
Sri Jayewardenepura, Sri Lanka
en82667@sjp.ac.lk

Geethal C. Siriwardana

Dept. of Mechanical Engineering
Faculty of Engineering, University of
Sri Jayewardenepura, Sri Lanka
geethal@sjp.ac.lk

Tharaka R. Bandara

Dept. of Mechanical Engineering
Faculty of Engineering, University of
Sri Jayewardenepura, Sri Lanka
tharaka.bandara@sjp.ac.lk

Abstract - Vibration analysis is used to detect faults and anomalies in machinery and other mechanical systems that produce vibrations during operation. The study aimed to develop an algorithm that can detect and locate engine faults in automobiles by analyzing vibrational data produced during engine operation. Analysis was done on one type of engine fault – Spark Ignition Engine misfire. To detect anomalies in the vibrational pattern (waveform), analysis was carried out in both time and frequency domains. To obtain vibrational data an AVR – 32 (Arduino) based data acquisition device was built, and analysis was carried out in MATLAB using scripts and functions. The developed algorithm isolates frequency components in the waveform that corresponds to engine faults and converts them into numerical quantities that are then compared with computed ranges. The algorithm was able to identify the presence of a misfire in the engine and could locate the cylinder in which the misfire occurs with significant accuracy.

Keywords - locating engine misfires, vibration analysis

I. INTRODUCTION

Vehicle engine faults need to be detected to prevent damage to components of the vehicle, maintain driver and passenger comfort as well as prevent catastrophic failure during its operation. The heart of any automobile is its engine. Modern-day engines are complex machines that are controlled by computers and rather intimidating for the usual mechanic to work on. Engine faults can be categorized into faults that can be identified visually, with the use of onboard diagnostics (OBD) scanner, and by listening to the sound generated by the engine. Faults that are identified by listening, requires expert knowledge, and experience. It can be difficult for a new and inexperienced mechanic to correctly identify a fault by listening to the engine sound. Even experienced mechanics can incorrectly diagnose faults leading to unnecessary expenses and rework. It is therefore imperative that a system is introduced which can correctly identify engine faults by analyzing engine sound/vibrations. After identifying the problem, the mechanic will then have to locate it. This is done through trial and error and involves the removal of electrical connections and engine components. Therefore, having a system that locates the problem is also vital. This study aims to develop an algorithm to accurately detect engine misfires and locate the cylinder where it occurs by analyzing vibrations generated during operation. Vibration analysis is widely used to detect failures and faults in industrial machinery but is seldom used to detect vehicular faults.

An algorithm is proposed in [1] where engine faults are identified using sound recordings. Sound recognition techniques are used in the detection algorithm mentioned in [2]. The proposed algorithm uses three criteria to decide on the fault. A mini microphone is used to record sounds at different engine rotational speeds in [3]. Engine faults are then identified using a model built in MATLAB. All the above-mentioned research is based on sound analysis and has a common problem of eliminating excessive noise from the recorded sound wave. Further, the effectiveness of capturing all vibrations emitted from the engine is questionable as the microphone only captures waves that reach it through an air medium. Both issues can be mitigated if the vibrations are recorded using an accelerometer that is placed on a suitable/effective position on the vehicle frame/engine. This method is used in [4] to acquire vibrations generated from the engine. Using a 3-axis accelerometer it is possible to measure the vibrations in all 3 planes. Variations in signal parameters between the normal engine and the fault engines are then identified. A 3-axis accelerometer is used along with a data acquisition device in [5] to acquire vibrations to detect faults in induction motors.

A simple but powerful data acquisition device can be fabricated using Arduino as mentioned in [6]. The Arduino platform is used to acquire vibrational data from a 3-Axis digital accelerometer. However, post-processing of the vibrational data must be done on a computer or Field Programmable Gate Array (FPGA). Another such Arduino-based data acquisition device is used in [7] to measure free vibrations on a wind turbine blade. A more powerful alternative to the Arduino platform is discussed in [8] where a Raspberry Pi single-board computer (SBC) is used. The main advantage of using an SBC is the ability to perform the data acquisition as well as the post-processing in the same device. However, SBCs are relatively more expensive than microcontrollers and the post-processing algorithm can be implemented in an FPGA which has a smaller form factor.

Vibration analysis to determine piston scuffing fault in Internal Combustion engines is appraised in [9]. It was shown that piston scuffing fault caused an increase in maximum, root means square, mean, skewness, kurtosis, and impulse factor of the engine vibration in the frequency band of 2.4–4.7 kHz [9]. The development of an algorithm that can determine faults by assessing nuances between normal and abnormal waveforms is presented in [10] where analysis is done to determine tool wear and condition in high-speed milling. Here reconfigurable infinite impulse response (IIR) band-pass digital filter and statistical techniques [10] are used for processing and analyzing

vibrational signals. The vibrations are analyzed after converting the signal into a time-frequency domain with the use of Continuous Wavelet Transform (CWT). In the developed algorithm, arithmetic means value, and the sum of absolute values of the digitally filtered vibration signal is utilized as reference value to set up a healthy tool threshold. A comparison between a set healthy tool threshold and the sum of absolute values of the digitally filtered vibration signal is the basis for the decision-making algorithm. This algorithm can indicate faults in real-time which is advantageous. Another real-time fault detection algorithm is presented in [11] where vibrational analysis is done to identify faults in industrial machinery. Here, Fast Fourier Transform (FFT) is used to convert the wave from the time-domain to the frequency-domain. The use of CWT or FFT greatly depends on the nature of the waveform as FFT does not consider time-domain characteristics whereas CWT allows the assessment of characteristics that vary with time. For example, the effectiveness of both CWT and FFT to distinguish abnormalities in EEG signals is assessed in [12]. It was found that since EEG signals are non-stationary (characteristics change with time) CWT is more suitable than FFT for spectral analysis. To arrive at a conclusive decision, it is therefore imperative to use both methods to analyze waveforms and see what is most effective in determining engine faults.

Signal analysis techniques to locate engine faults (misfire) are being discussed in [13]. In this study, time-domain features such as the peak-to-peak value (PP), root mean square value (RMS) are used to identify and isolate the misfiring cylinder of an engine. Experiments showed that as the engine rotational speed is changed, the features that can be used to detect and locate the cylinder also change. Therefore, the performance of the features in isolating faults is dependent on the engine rotational speed.

Vibration analysis is used in many instances to detect anomalies and faults in mechanical systems. Extensive research has been done on detecting engine faults through vibration analysis. However, locating faults have been only discussed in [13]. Here, analysis is performed exclusively in the time domain. In this study, waveforms will be analyzed in both frequency and time domains. The developed algorithm isolates fault signals to detect and identify engine faults.

II. METHODOLOGY

A. Theory

A digital 3-axis accelerometer (ADXL 345) was chosen as the sensing device. The data acquisition device was made using the Arduino platform. The algorithm for analyzing the signal was created in MATLAB using scripts and functions. Signal analysis is predominantly done in the frequency domain using the Fast Fourier Transform (FFT) as the waveforms emitted from the engine are stationary signals when considered for a long enough period.

FFT is an algorithm that calculates the Discrete Fourier Transform in a numerically efficient way. The benefit of using the FFT algorithm is that it is an order $n \log(n)$ operation, where n is the number of discrete data points. For large data sets, this is favorable as FFT is almost linear scaling in n as the effect of $\log(n)$ is less significant as n gets large. The FFT algorithm is standard and comes as a built-in feature in MATLAB.

At the early stages of the research, waveforms were analyzed using a Spectrogram that utilizes a Gabor transform. Spectrograms can be used to assess a waveform in both time and frequency domains. For example, when a signal is transformed from the time domain to the frequency domain using the FFT it would yield a plot that shows the constituent frequencies of that waveform and their magnitudes. However, it is not possible to observe when these frequency components occur in the waveform. The Gabor transform allows us to compute the spectrogram which is a time-frequency plot that shows which frequencies are active in each period of a waveform. The Spectrogram is computed by convolving a Gaussian wavelet with the Fourier transform while the Gaussian window is moved across the original waveform. This yields a frequency plot weighted by the Gaussian window.

B. Experimentation

A normal running engine produces vibrations due to the combustion that occurs in the cylinder and other moving parts in the engine. The constituent frequencies of this vibrational signal will be constant at a particular rotational speed of the engine. If a misfire is induced in one of the cylinders, the vibrational signal will change significantly due to the unbalanced combustions in the cylinder. Additional frequency components will be observed in the signal and thus the issue could be identified. Further, the magnitude of these newly induced frequencies and their distribution will be assessed to find a correlation between waveform characteristics and the misfiring cylinder. If successful, the misfiring cylinder can be located. The vibrations were captured using a 3 – Axis digital accelerometer (ADXL345) and acquired by a Data Acquisition Device (built using the Arduino platform) through I²C communication protocol. The received data is then transmitted via Serial communication (UART) to a computer. The Arduino board is interfaced with MATLAB which is installed in the computer. The received data is then written to a spreadsheet by a MATLAB script. This data contains the acceleration values in the X, Y, and Z axes and the time stamps at which readings were taken. The sampling rate ranges from 450 Hz to 500 Hz which was deemed satisfactory as it would give a maximum measurable frequency of 225 Hz (In a 4 stroke 4-cylinder engine at 2000 RPM, combustion occurs at a frequency of 66.67 Hz). The recorded data can then be loaded to the MATLAB environment for further analysis.

1) Experiment 01

A series of preliminary tests were carried out to check the feasibility of the research and to develop the algorithm. The objectives of the experiment are as follows,

- Determine whether the waveform produced is stationary.
- Observe whether misfires can be detected through waveform analysis.

The experiment was carried out on a 2002 Toyota Corolla 1.5L 4 stroke 4-cylinder engine (1NZ-FE) using just one accelerometer positioned between the left-most (1st cylinder) and the 2nd cylinder. The accelerometer was fixed

to the engine block rigidly with the use of a stud and bolt connection.

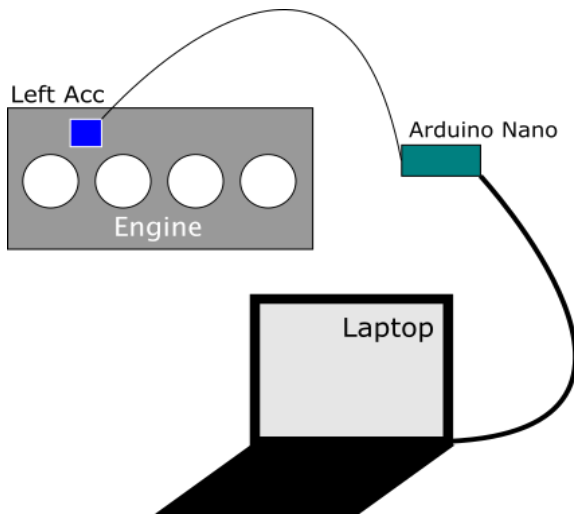


Fig 1: Test setup for experiment 01

A misfire was induced in the first cylinder by disconnecting the electrical connection to its ignition coil. Readings were then taken at idling speed and at 2000 RPM. The procedure was repeated for misfires in each cylinder and finally for the normal (no misfire) scenario.



Fig 2: Accelerometer fixed rigidly to the engine block.

The obtained waveforms were then analyzed using a preliminary algorithm that was coded in MATLAB.

2) Experiment 02

The second set of experiments were carried out on the same engine at idle speed (around 1000 rev/min). Readings were taken from two accelerometers at two different locations to see if and how the waveforms change with the location of the accelerometer.

Objectives of the experiment are as follows,

- To see if the magnitudes of the additional frequencies (explained in future sections) have any correlation with the position of the misfiring cylinder.
- To assess the reproducibility of the vibrational waveforms.

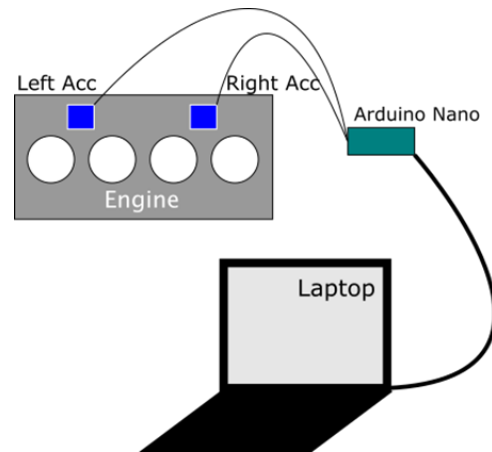


Fig 3: Test setup for experiment 02

As in the 1st experiment, readings were taken for 5 scenarios (normal, misfires in cylinders 1,2,3, or 4). Readings were taken by both accelerometers simultaneously. In this experiment, only the Y-axis readings were taken from both accelerometers because upon analyzing data obtained in the 1st experiment it was clear that significant differences in the waveforms in different scenarios were observed only in the Y-axis readings. The procedure was repeated thrice.

Measurements were obtained from two locations to see if the results could be used to locate the misfiring cylinder.

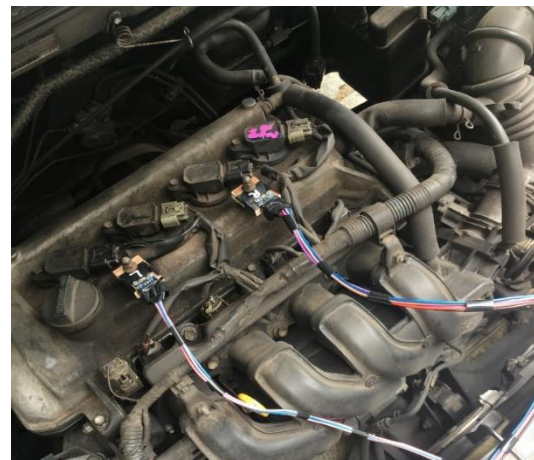


Fig 4: Updated data acquisition device with 2 accelerometers

C. Results

1) Experiment 01

A total of 30 waveforms were obtained in the first experiment. The breakdown of those waveforms are as shown in Table I. To demonstrate the differences in the obtained waveforms Fig 5 to Fig. 9 are presented.

TABLE I. RESULTS BREAKDOWN

Scenario	X axis	Y axis	Z axis	Total
Idle				
Normal	1	1	1	3
1 st cylinder misfire	1	1	1	3
2 nd cylinder misfire	1	1	1	3
3 rd cylinder misfire	1	1	1	3
4 th cylinder misfire	1	1	1	3
2000 RPM				
Normal	1	1	1	3
1 st cylinder misfire	1	1	1	3
2 nd cylinder misfire	1	1	1	3
3 rd cylinder misfire	1	1	1	3
4 th cylinder misfire	1	1	1	3
Total	10	10	10	30

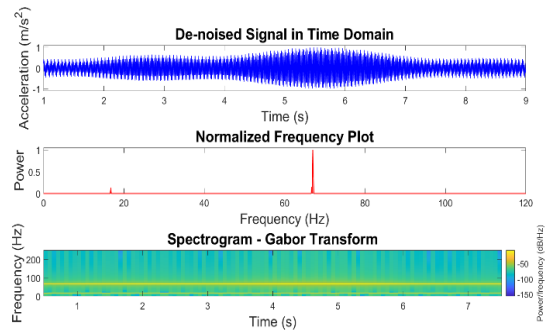


Fig 9: X-Axis Readings – 3rd Cylinder misfire at 2000 RPM

2) Experiment 02

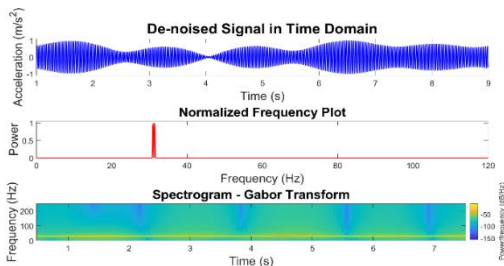


Fig 5: Y-Axis Readings - Normal at Idle

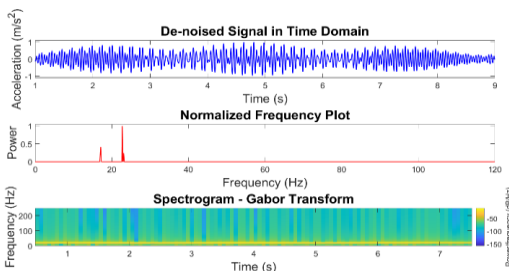


Fig 6: Y-Axis Readings - 1st Cylinder misfire at Idle

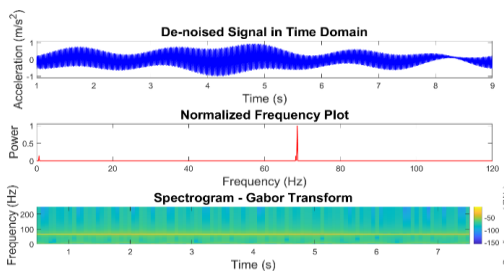


Fig 7: Y-Axis Readings - Normal at 2000 RPM

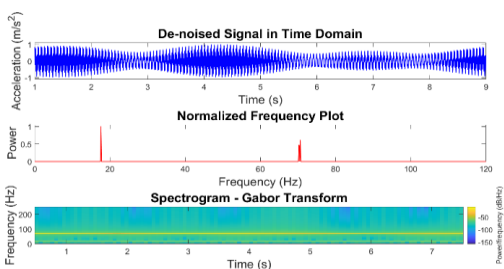


Fig 8: Y-Axis Readings – 2nd Cylinder misfire at 2000 RPM

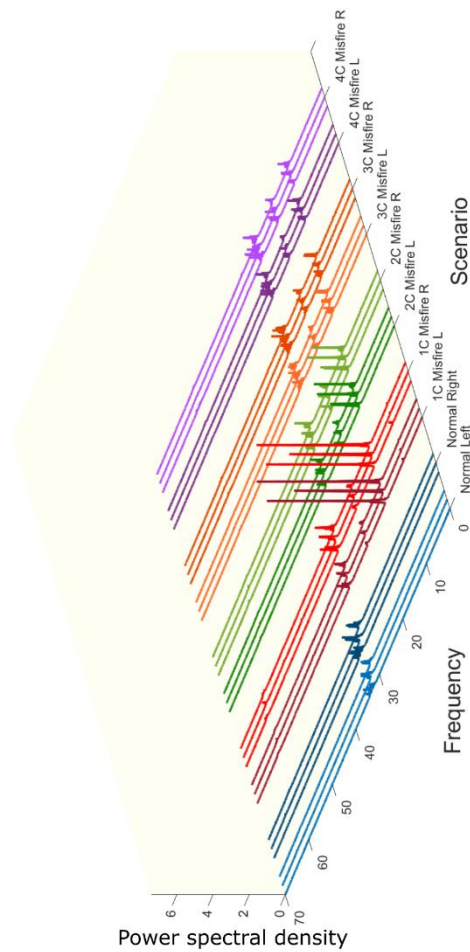


Fig 10: Power spectral density vs Frequency of different engine conditions.

The 3D plot shown in Fig 10 contains the frequency spectrum of all the waveforms obtained by both accelerometers. Note that for each scenario and accelerometer, 3 frequency distributions are plotted. This is because measurements were repeated 3 times for each scenario.

D. Analysis and Discussion

1) Experiment 01

Fig 5 to Fig. 9 shows the denoised waveform in the time domain, frequency spectrum, and spectrogram of some of the waveforms. To validate that the signals received are stationary, the waveforms were analyzed in the time- frequency domain using a spectrogram (Gabor Transform).

The horizontal line in the spectrogram indicates that the signal does not change with time, thus is stationary. After converting the signal to the frequency domain using the FFT, the noise was removed by eliminating low power frequency components. In all waveforms, a clear spike was observed in the frequency spectrum at a frequency similar to that of the frequency of combustion (spark frequency) at that particular engine speed. For instance, in Fig 7,8 and 9 a spike is present at around 68-69 Hz which is the frequency of combustion at 2000 RPM. At normal conditions (no misfire), the only frequency component that was present in the waveform corresponds to the spark frequency. This was later validated through multiple tests.

When a misfire is induced in one of the 4 cylinders, extra frequency components appear. As shown in Fig 8, when a misfire is induced in the 2nd cylinder an additional spike appears at 17.63 Hz. This new frequency component was observed in all misfiring scenarios in the Y-Axis at 2000 RPM. At idle speed, additional frequency components were only visible in some misfire scenarios. Further in all scenarios where this frequency appeared, it was similar to the combustion frequency of a single cylinder (single-cylinder spark frequency). For instance, at 2000 RPM, a cylinder experiences a spark every 2 rotations of the crankshaft, thus at a frequency of 16.67 Hz. The presence of this additional frequency component could therefore be considered as an indicator for a misfire. However, locating the cylinder is not possible through this analysis.

2) Experiment 02

As the waveforms were validated to be stationary signals from experiment 01, analysis was performed exclusively in the frequency domain. Four key regions were identified where frequency components would appear. These regions are,

- Single cylinder spark frequency region
- Engine crank rotational frequency region
- Intermediate frequency region
- Engine spark frequency region

The frequency distributions of the obtained waveforms are shown in the 3D plot (Fig 10). Frequency spikes were observed in the spark frequency region as observed in Experiment 01. Whenever there was a misfire, additional spikes were observed in the Single spark frequency region, Engine crank rotational frequency (Crank frequency) region, and intermediate frequency region. From these three regions, the crank frequency region showed the most variance in power of the frequency components. Therefore, the single spark frequencies of each waveform were isolated using a MATLAB script for further analysis.

Fig 11 shows the average power values (with associated uncertainties) of the crank frequencies in each

scenario. The two points in each region show the means of the power values obtained from the left accelerometer and right accelerometer, respectively. From Fig.11 it is possible to distinguish the normal running condition, 1st cylinder misfire and 2nd cylinder misfire as their power ranges do not overlap with others. However, it is difficult to distinguish the 3rd cylinder misfire case from the 4th cylinder misfire case by only assessing the mean power values as their power ranges overlap. Therefore, a different approach had to be taken for the analysis. The graph in Fig 12 shows the means of RMS values of the vibrational signals.

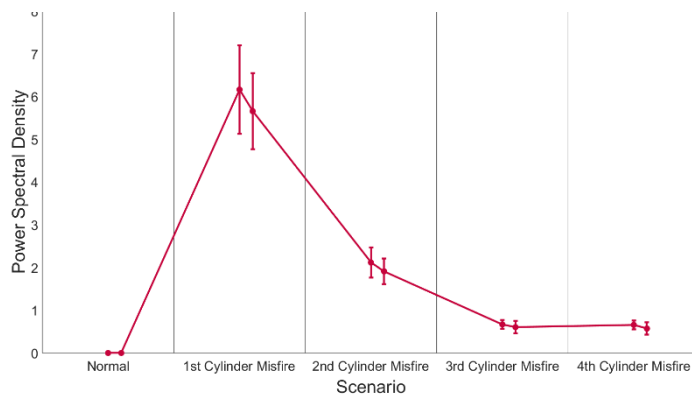


Fig 11: Mean Power vs misfire scenario

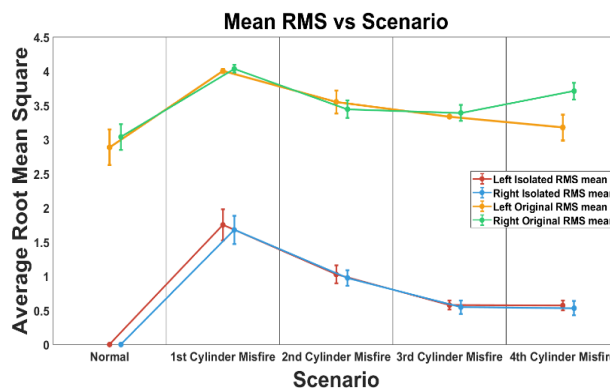


Fig 12: Mean RMS values vs misfire scenario

The green and orange points show the mean RMS of the original waveforms (without processing) of different misfire scenarios. The red and blue points are the mean RMS values of the same waveforms where all other frequency components except the crank frequency are filtered out (only the crank frequency component exists). The mean power variation also shows the same pattern. However, RMS was selected over PSD as it required less computation. As expected under normal conditions the RMS values of the isolated signals (filtered signals) are zero as the single spark frequency does not exist in the original waveform. In other cases, the RMS values are non-zero for the isolated signals. The mean RMS values show a similar trend to the mean power values shown in Fig 11. Similarly, normal, 1st cylinder misfire, 2nd cylinder misfire can be differentiated by just observing the RMS range of the isolated waveforms (-1.5 σ and 1.5 σ). To differentiate the 3rd and 4th cylinder misfiring cases from each other the RMS ranges of their respective original signals must be

used. Specifically, the signal obtained from the right accelerometer. There is a clear difference between the mean RMS values of the original right accelerometer signals of the 3rd and 4th cylinder misfiring scenarios. Further, the ranges were chosen to avoid their RMS ranges from overlapping while yielding an acceptable level of accuracy (87%). Under these conditions, the misfiring engine can be located by the following methodology shown in Table III.

TABLE II. MEAN RMS VALUS AND RANGE OF RMS VALUES

Range between -1.5σ to 1.5σ accounts for 86.64% of readings										
	Normal		1st Cylinder Misfire		2nd Cylinder Misfire		3rd Cylinder Misfire		4th Cylinder Misfire	
	Mean	Range (+1.5 σ - 1.5 σ)	Mean	Range (+1.5 σ - 1.5 σ)	Mean	Range (+1.5 σ - 1.5 σ)	Mean	Range (+1.5 σ - 1.5 σ)	Mean	Range (+1.5 σ - 1.5 σ)
Mean RMS of isolated Right Accelerometer Signal (IRMS)	3.0392	3.2273	4.0363	4.0931	3.4431	3.5741	3.3912	3.5097	3.7106	3.8326
Mean RMS of original Right Accelerometer Signal (ORMS)	0	0	1.6781	1.8823	0.9751	1.0884	0.5469	0.6468	0.5317	0.6364

TABLE III. IDENTIFICATION ARGUMENTS

	Isolated Signal RMS		Original Signal RMS
Normal	0		
1st Cylinder misfire	(1.4739 - 1.8823)		
2nd Cylinder misfire	(0.8618 - 1.0884)		
3rd Cylinder misfire	(0.4470 - 0.6468)	AND	(3.2727 - 3.5097)
4th Cylinder misfire	(0.4270 - 0.6364)	AND	(3.5886 - 3.8326)

Since only the waveforms that were obtained by the right accelerometer were used for the identification there is no need for a system with two accelerometers for data acquisition for this engine model and this engine fault.

III. CONCLUSION

Vibrations transmitted through the vehicle structure were recorded using an accelerometer connected to a data acquisition device. A low-cost data acquisition device was built using the Arduino platform. The recorded waveforms which originated from the same engine but under normal and misfiring conditions were analyzed. The analysis was done on MATLAB. Two separate experiments were carried out to obtain data to develop a method to detect engine misfires and to locate the misfiring cylinder. Frequency analysis showed that frequency components equal to the crank frequency of the engine at idling speed appear when there is a misfire in one of the cylinders. The average power of the crank frequency components can be used to differentiate normal; 1st cylinder misfire and 2nd cylinder misfire scenarios. To distinguish misfires in the 3rd and 4th cylinders assessing the power of the frequency components proved insufficient. Differentiation was possible by a combined assessment of the mean RMS values of the isolated fault signals and the original signals.

This method can be used to detect and locate an engine misfire (with about 87% accuracy) in this engine at idle speed.

This study was performed on one engine model preliminarily. The methodology can be developed, however, to detect misfires in other engine models through conducting the same tests on those engines and setting identification arguments unique to them. Currently, the methodology can only detect engine misfires under controlled conditions. That is, on an engine that does not have other faults except for misfires. This study does not assess how the existence of other faults such as damaged camshaft, knocking, faulty mounts, etc. in addition to engine misfiring, affect the performance of the methodology. In the future, the methodology may be developed to detect and locate other engine faults such as engine knocking.

Based on the conducted study, an algorithm can be developed and implemented on a device that can be used to detect and locate engine faults. Such a device will assist mechanics in accurately detecting and locating engine faults without unnecessary engine disassembly and trial and error techniques. Further, an algorithm based on this methodology may be implemented on the Engine Control Unit to detect faults and improve efficiency. For instance, engine efficiency can be improved by controlling the spark timing of individual cylinders once engine knocking is identified and located.

REFERENCES

- [1] L.M.Contreras-Medina, R.J.Romero-Troncoso, J.R.Millan-Almaraz and C.Rodriguez-Donate, "FPGA Based Multiple-Channel Vibration Analyzer," IEEE, pp. 229-232, 2008.
- [2] Wail.M.Adaileh, "Engine Fault Diagnosis Using Acoustic Signals," Applied Mechanics and Materials, pp. 2013-2020, 2013.
- [3] M.Akin, "Comparison of Wavelet Transform and FFT Methods," Journal of Medical Systems, vol. 26, pp. 241-247, 2002.
- [4] A.González, J.L.Olazagoitia and d.J.Vinolas, "A Low-Cost Data Acquisition System for Automobile," Sensors, 2018.

- [5] A.K .Kemalkar and V.K.Bairagi, "Engine Fault Diagnosis Using Sound Analysis," International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pp. 943-946, 2016.
- [6] A.Moosavian, G.Najafi, B.Ghobadian, M.Mirsalim, S.M.Jafari and P.Sharghi, "Piston scuffing fault and its identification in an IC engine by vibration," Applied Acoustics, pp. 40-48, 2015.
- [7] Chomphan and Suphattharachai, "Vibration Analysis of Gasoline Engine Faults," American Journal of Applied Sciences, pp. 1166-1171, 2013.
- [8] E.Ftoutou, M.Chouchane, N.Besbès and R.Ouali, "Detection of diesel engine misfire by vibration analysis in the time domain," Academia, 2020.
- [9] F.Aswin and Z.S.Suzen, "Analysis of free vibration measurement by mems," in International Conference on Applied Science and Technology, 2018.
- [10] M.Iwaniec, A.Holovatyy, V.Teslyuk, M.Lobur, K.Kolesnyk and M.Mashevska, "Development of Vibration Spectrum Analyzer Using," IEEE, 2017.
- [11] M.Madain, A.Al-Mosaiden and M.Al-khassaweneh, "Fault Diagnosis in Vehicle Engines Using Sound Recognition Techniques," IEEE, 2010.
- [12] P.Y.Sevilla-Camacho, J.B.Robles-Ocampo, J.Muñiz-Soria and F.Lee-Orantes, "Tool failure detection method for high-speed milling using," Int J Adv Manuf Technol, 2015.
- [13] S.K.Shomea, U.Datta and S.R.K.Vadali, "FPGA based Signal Prefiltering System for Vibration," Procedia Technology, 2011.

Identifying interrelationships of key success factors of third-party logistics service providers

Theruwanda Perera*
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
theruwanda5463@gmail.com

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

A. N. Wijayanayake
Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
anni@kln.ac.lk

Abstract - To be more cost-effective as well as to maintain a sustainable competitive advantage, many enterprises tend to improve their business practices by having a strong relationship with third-party (3PL) logistics service providers. The main objectives of this paper are to determine the key success factors associated with the Sri Lankan 3PL industry and identify the interrelationships of these key success factors. A systematic literature review and expert opinions were used to identify the key success factors of the 3PL industry in Sri Lanka. In total 21 key success factors were obtained, and those key success factors were grouped into four categories as organization strategy, management, and process, human resources, and customer orientation. Q-sort technique was used to group key success factors into four categories. Decision-making trial and evaluation laboratory (DEMATEL) method was used to capture the interactive relationships among the key success factors of 3PL service providers, and the casual effect map analyzed. Data were collected through questionnaires from middle and senior-level managers of 3PL firms. A total of eleven experts in the 3PL industry participated in the data collection process. The result shows that organization strategy is a core success factor since it has both high prominence and high interrelationship. Management and process were classified as driving factors since they had a low prominence but a high interrelationship. However, human resources and customer orientation had high prominence but low relationship, which are influenced by other factors and cannot be directly improved. The findings may assist managers to formulate long-term flexible decision strategies in their 3PL firms.

Keywords - 3PL, DEMATEL, interrelationship, key success factors

I. INTRODUCTION

Over the last few decades, the global logistics industry has grown significantly. Planning, implementing and controlling transportation, warehousing, inventory management and control, order processing, information systems, and packaging are all common logistics management activities [1]. Third-party logistics (3PL) service providers are the companies that provide these logistics services. Reference [2] states that from a customer perspective, 3PL firms are considered as resource managers, problem solvers, transportation strategists, distribution strategists as well as supply chain strategists. The third-party logistics industry provides very important support for enterprises in different industries, and it also promotes the economic growth of a country. Because of that, the development of the 3PL industry is an essential factor that needs to be considered from a country perspective. Sri Lanka however, is lagging, even though our geographical location provides it a competitive edge.

With the increasing demand and technological advancement, it is a mandatory requirement to satisfy customers by fulfilling their needs to survive in the market.

Identifying key success factors for the industry, from both a customer and supplier perspective becomes fundamental for industry success. When it comes to the third-party logistics industry, service providers must set themselves apart by providing value-added services, focusing on key customer accounts that can generate high profits, achieve economies of scale, and improve service providers' ability to support international operations. Key success factors can provide significant support to achieve those goals in the 3PL industry.

Awareness of the key success factors will enable the companies to improve delivery performance, improve customer satisfaction, increase customer acquisition, optimize the relationship between suppliers and customers, improve profit and revenue growth, reduce overall logistics cost and improve the quality of logistics services provided. Countries like Germany, Sweden, Belgium, Austria, Japan identified those key success factors and developed a competitive edge over their rivals. Identifying the success factors in a developing country setting like Sri Lanka, would assist in developing the logistics services in the country and enable it to fully exploit the countries geographical location to service international trade worldwide.

Most of the studies have identified the priorities of the key success factors in the 3PL industry but very limited research has been done to identify the interrelationship of the key success factors in the 3PL industry. As Sri Lanka is lying on a key East-West trade route and located next to India, it is worthy for practitioners and investors to know about key success factors of third-party logistics provider companies in Sri Lanka. When the efficiency and effectiveness of service providers improve, it will create a smooth supply chain. Therefore, the clients can explore more business opportunities. This will create a win-win situation for both 3PL service providers and their clients.

II. LITERATURE REVIEW

A. Sri Lankan 3PL industry

The 3PL industries in European and Asian countries have been studied widely, but there is a limited number of studies focused on the Sri Lankan 3PL industry. Currently, 3PL services are in their nascent stage in Sri Lanka [3]. In World Bank's Logistics Performance Indicator ranking (LPI) for 2018, Sri Lanka is ranked 94th out of 160 whereas Germany is ranked at 1. With a score of 2.60 out of 5, Sri Lanka is classified as a partial performer [4] (for details, refer to Table I).

In Sri Lanka, though 3PL service providers and their customers maintain a good relationship, the level of satisfaction, and trust towards service providers are not considered high. Cost, lack of control, lack of coordination

and lack of cooperation, lack of skills and knowledge, lack of industry knowledge, and trade union activities are the factors that affect the growth of the 3PL market in Sri Lanka [3]. This study also identifies future issues in the Sri Lankan 3PL sector, such as reducing delivery lead times, adopting new technology, managing the number of order channels multiplied by the number of delivery alternatives, and dealing with overstocks due to online sales, among others.

When examining the 3PL industry's global context, they gravitate toward innovative technical services. Sri Lanka should also concentrate on improving the quality of these new applications to attract more customers and raise the country's GDP. Otherwise, they will not be able to compete in the market since a competitor will gain a competitive advantage over them [5]. Reference [6] stated that before the 3PL industry in Sri Lanka gets disrupted with the labour shortage issues and the dynamic customer demands, firms have to focus on technology adoptions to survive within industries.

TABLE I. SRI LANKA'S LPI RANK AND SCORE

Parameter	Germany (score out of 5)	Sri Lanka (score out of 5)
Customs	4.09	2.58
Infrastructure	4.37	2.49
International shipments	3.86	2.51
Logistics competence	4.31	2.42
Tracking and tracing	4.24	2.79
Timeliness	4.39	2.79
Overall LPI score	4.2	2.6
LPI rank	1	94

Reference [7] investigated how information technology, supply chain security, and green supply chain practices affect the amount of interaction between users and providers of third-party logistics services. Reference [8] mentioned that several 3PL providers in Sri Lanka have taken steps to establish their own modest to large-scale Information Communication Technology (ICT) solutions for their business processes. The usefulness of a Warehouse Management System (WMS) in facilitating warehouse best practices is also highlighted in this study.

B. Key success factors

Several studies have investigated the importance of key success factors on business performance in the 3PL industry. Key success factors are concerned with not only the success of a business entity but also its potential to deal with difficult business conditions [9]. Reference [10] claimed that a stronger association between relationship management and organizational success of the 3PL service provider and the 3PL service user is enhanced by greater understanding and proper communication between parties.

Cloud technology applications in the logistics industry have been explored in some research. Reference [11] provided a smart model that uses agent technology and cloud computing to make data collection and flow easier, as well as provide better and less expensive access to logistics management systems. The cloud platform is also mentioned in reference

[12] as a crucial foundation for logistics network optimization. The internet of things technology development patterns in warehouse operations were explored in reference [13] using four main criteria. Those were the rapid development of RFID technology in warehousing, the integrated application of sensing technology, the AGV (Automated Guided Vehicle) integration into the warehouse, and IoT will be in sync.

Improving and better understanding of efficiency and innovation-based strategies can gain a competitive advantage in the 3PL industry. Reference [14] has shown that improvement and process innovation are mostly pushed forward by industry-focused 3PL providers. This study clearly defines the importance of industry specialization and it can also facilitate the development of best practices to improve internal processes. Business process re-engineering companies outperformed non-business process re-engineering companies in the logistics industry, not only in information processing, technology applications, organizational structure, and coordination but also in all major logistics operations [15].

Reference [16] conducted in Pakistan to determine how quality management practices 3PL service providers achieving integration competency in the service chain. Quality management components include leadership, strategic planning, customer focus, knowledge management, human resource emphasis, and process management. Strategic planning, HR management focus, and process management were identified as characteristics that have a significant impact on the integration competency of 3PL service providers in Pakistan, according to the findings of this study. Surprisingly, the impact of leadership, customer focus, and knowledge management were not significant. Management and leadership, internationalization, and staff competence were regarded as the most essential and critical success characteristics of logistics provider organizations in Iran [17].

Reference [18] also classified key success factors of the 3PL industry in India. Most Indian 3PL service providers give importance to cost reduction as the most important success factor. The information technology system is also critical to the company's performance. The organization can quickly and efficiently share and convey information with the end-user if it focuses more on this component. This can also increase the speed and accuracy of the process, resulting in higher client satisfaction. This would increase profit while also improving the company's brand image. Reference [19] stated that the cost of service, service level, level of professionalism, geographical location, specific references in the same sector, innovation capacity, and collaboration with the customer are some key factors of the selection in 3PL service providers.

3PL clients demand 3PL service providers place a greater emphasis on elements such as industry experience, annual performance, customer service, creative management, top management availability, service quality, flexibility, and market understanding [20]. Several criteria were proposed to improving warehouse operations. Improving the training process for both existing and new employees to better utilize warehouse resources is one key, as it is having a basic understanding of warehouse operations and steps [21]. From the standpoint of global organizations and local firms, leadership, logistics, business, and information and

communication technology are the four competency categories [22].

The breadth of services is positively related to revenue growth [23]. But other factors such as industry focus, relationship with 3PL, investment in information systems, skilled logistics professionals, and supply chain integration are not positively related to the revenue growth of 3PL service providers. Malaysian 3PL enterprises must have a high level of management commitment to any continuous improvement projects, support the idea of skills improvement and acquisition of new information among employees, and have sufficient financial resources [24]. These are the significant factors that need to have positive logistics performance in Malaysian 3PL firms.

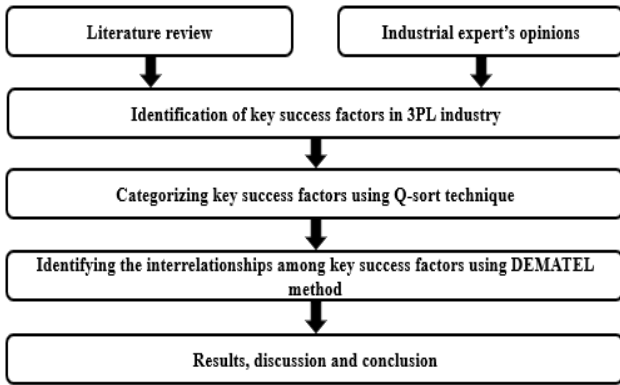


Fig 1. Flow diagram of the process of the methodology

The most significant characteristics for success as a 3PL provider are internationalization, industry focus or expertise, investment in information systems, availability of trained personnel, and supply chain integration [25]. The breadth of services, industry focus, relationship with 3PLs, investment in information systems, skilled logistic professionals, and supply chain integration were identified as key success factors [26]. These factors were considered for this study.

Decision-making trial and evaluation laboratory (DEMATEL) method is a useful tool for identifying cause-effect chain components in a complicated system. It deals with using a visual structural model to evaluate interdependent interactions among components and identify the key ones. The interrelationships between risks faced by 3PL service providers to one of its customers using the DEMATEL method were analyzed [27]. The AHP and DEMATEL methods were used to prioritize the key success factors of the 3PL industry in Iran, while other studies used AHP and DEMATEL methods with some other multi-criteria decision-making techniques for supplier selection [28]. Most of the past literature used either AHP or DEMATEL method to make multicriteria decisions.

III. METHODOLOGY

A. Proposed research framework

Through a thorough literature review, several prominent key success factors in the 3PL industry were identified. Thereafter, through interviews with experts in the 3PL industry, the list was revised which included the addition of success factors unique to the Sri Lankan context. The main objective of this study is to identify the interrelationships among the key success factors, hence a quantitative research approach was used. After identifying the key success factors

Q-sort technique was used to categorize the key success factors into four groups, namely organization strategy, management and process, human resources, and customer orientation. The identified key success factors were divided into those categories by using the data collection approach in the Q-sort technique. When compared to the general Likert scale, the “Q-sort table” is more effective to get the data from a small sample. Then the DEMATEL method enabled the decision-makers to understand the interactions between factors using a causal relationship diagram (Fig. 1).

B. DEMATEL

DEMATEL method was developed by the Geneva Research Centre of the Battelle Memorial Institute to visualize the structure of complicated causal relationships through matrixes or digraphs. DEMATEL is a well-known method that is used to analyse the interactions between factors by categorizing them into cause and effect groups. The procedure of the DEMATEL method can be summarized by the following steps [29].

1) Calculating the direct relation matrix. To obtain the direct influence between any two factors, use the inputs of the decision makers. Decision makers are asked to indicate the direct influence that one factor has on another factor, using an integer scale of “no influence (0),” “low influence (1),” “medium influence (2),” “high influence (3),” and “very high influence (4)”. The notation of x_{ij} represents the degree to which the respondent believes factor i affects factor j . For $i = j$, all principal diagonal elements are equal to zero. For each respondent, an $n \times n$ non-negative matrix can be established as $X_k = [x_{kij}]$, where k is the number of respondents with $1 \leq k \leq H$, and n is the number of factors. Thus, $X_1, X_2, X_3, \dots, X_H$ are the matrices from H respondents. To summarize all opinions from H respondents, the average matrix $A = [a_{ij}]$ is constructed as follows:

$$a_{ij} = \frac{1}{H} \sum_{k=1}^H x_{ij}^k \quad (1)$$

2) Calculating the normalized direct-relation matrix, where normalization of direct-relation matrix D is performed by $D = A \times S$ with the assistance of the following equation in which all elements should lie between 1 and 0.

$$S = \frac{1}{\max_{1 \leq i \leq n} \sum_{j=1}^n a_{ij}} \quad (2)$$

3) Calculating total relation matrix T , where T is defined as $T = D(I - D)^{-1}$ where I is the identity matrix. Let $[r_i]n \times 1$ and $[c_j]1 \times n$ be the vectors representing the sum of rows and sum of columns of the total relation matrix. When $j = i$, the sum $(r_i + c_j)$ illustrates the total effects given and received by factor i . $(r_i + c_j)$ represents the degree of importance for factor i in the entire system. On the other hand, the difference $(r_i - c_j)$ indicates the net effect that factor i contributes to the system. If the value $(r_i - c_j)$ is positive, then, factor i is a net cause, while factor i is a net receiver if the value $(r_i - c_j)$ is negative [30].

4) Calculating total relation matrix T , where T is defined as $T = D(I - D)^{-1}$ where I is the identity matrix. Let $[r_i]n \times 1$ and $[c_j]1 \times n$ be the vectors representing the sum of

rows and sum of columns of the total relation matrix. When $j = i$, the sum $(r_i + c_j)$ illustrates the total effects given and received by factor i . $(r_i + c_j)$ represents the degree of importance for factor i in the entire system. On the other hand, the difference $(r_i - c_j)$ indicates the net effect that factor i contributes to the system. If the value $(r_i - c_j)$ is positive, then, factor i is a net cause, while factor i is a net receiver if the value $(r_i - c_j)$ is negative [30].

5) Dividing factors into four quadrants according to their locations in digraph by calculating the mean of $(r+c)$ as in Fig. 2.

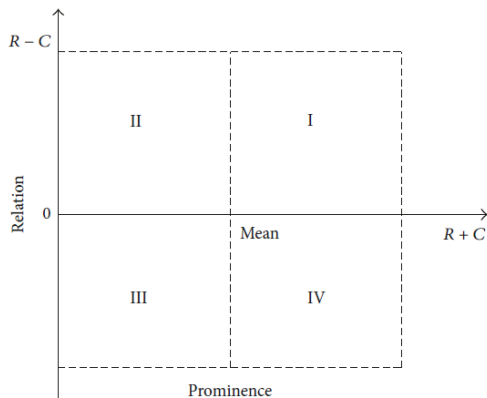


Fig. 2. Four quadrant digraph

Because they have high prominence and relation, the elements in quadrant I is defined as core factors or interconnected givers; the factors in quadrant II are characterized as driving factors or autonomous givers because they have low prominence but high relation. Quadrant III factors have low prominence and relation, and are relatively disconnected from the system (referred to as independent factors or autonomous receivers); quadrant IV factors have high prominence but low relation (referred to as impact factors or intertwined receivers), and are influenced by other factors and cannot be improved directly [29].

IV. DATA COLLECTION

Previous studies of the same interest, research articles, journals, and books were used to identify the key success factors in the 3PL industry. The identified key success factors were further filtered through expert opinions based on industry-based records. Interviews and questionnaires were used as the data gathering instruments for collecting primary data. Interviews were done with the industry experts including middle and senior managers in 3PL companies. Finally, twenty-one important key successes were determined through the inputs of the experts in the 3PL industry.

Those twenty-one key success factors were categorized into four groups using the Q-sort technique as in Fig. 3. Six experts in the 3PL industry were interviewed and collected data using a “Q-sort table” to identify the main category of each key success factor. Then each key success factor received an average value under each main category. Based on the highest average value of each key success factor, those were assigned to relevant main categories. The selected experts have more than seven years of experience in the 3PL industry and four of them were managers and the rest of them were senior executives in their 3PL companies [31].

The target population for this study was all the 3PL companies in Sri Lanka. The non-probability sampling methods of convenience sampling were applied to collect the data from the respondents. Hence it is difficult to gather data from individuals unless you are personally or mutually known to them, this method of sampling was selected to collect the data from experts in the 3PL industry. 11 experts in the 3PL industry participated in the data collection process. Those selected experts were highly skilled professionals in their domain having a good experience.

V. RESULTS AND DISCUSSION

To determine the interdependence between the listed key success factors of the 3PL industry in Sri Lanka, the DEMATEL method was used. It helped to evaluate the interrelationship among the key success factors in terms of the causal effect map. According to the procedure of the DEMATEL method firstly for the main factors, the normalized initial direct-relation matrix (D) was formed. Next, the total relation matrix (T) was calculated. The threshold value of the key factors was then calculated using the total relation matrix. It not only aided in the differentiation of the structure but also in the construction of a causal effect map. The causal impact map aids in the comprehension of the structure by identifying the influence of one success factor over another and filtering out unimportant effects.

Table II provides the direct and indirect effects of the four main key success factors. The values in the $(r+c)$ column express the degree of relationship of each factor with other factors. The factor which has the highest $(r+c)$ value indicates, it has more relationship with other factors. Here it is the organization strategy. Generally, the type of relationships among these four main factors can be taken by the $(r-c)$ values. Based on the $(r-c)$ value factors can be divided into cause-and-effect groups. If the $(r-c)$ value is positive, then that factor belongs to the cause group. If $(r-c)$ value is negative, then that factor belongs to the effect group. Organization strategy, management, and process are in the cause group. Human resources and customer orientation are the factors in the effect group. These factors in the cause group can influence other factors. The mean $(r+c)$ value is 28.8425

TABLE II. THE DIRECT AND INDIRECT EFFECTS OF FOUR MAIN FACTORS

Main Factors	r+c	r-c
Organization Strategy	29.0657	1.2463
Management and Process	28.2621	1.4809
Human Resources	29.0439	-0.1059
Customer Orientation	28.9987	-2.6213

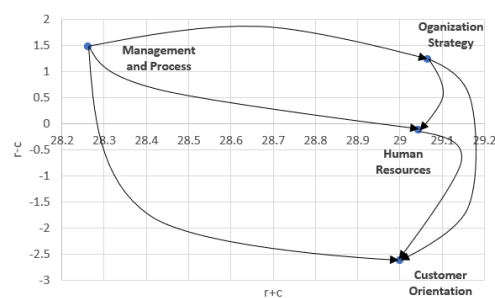


Fig. 4. The digraph of the main factors

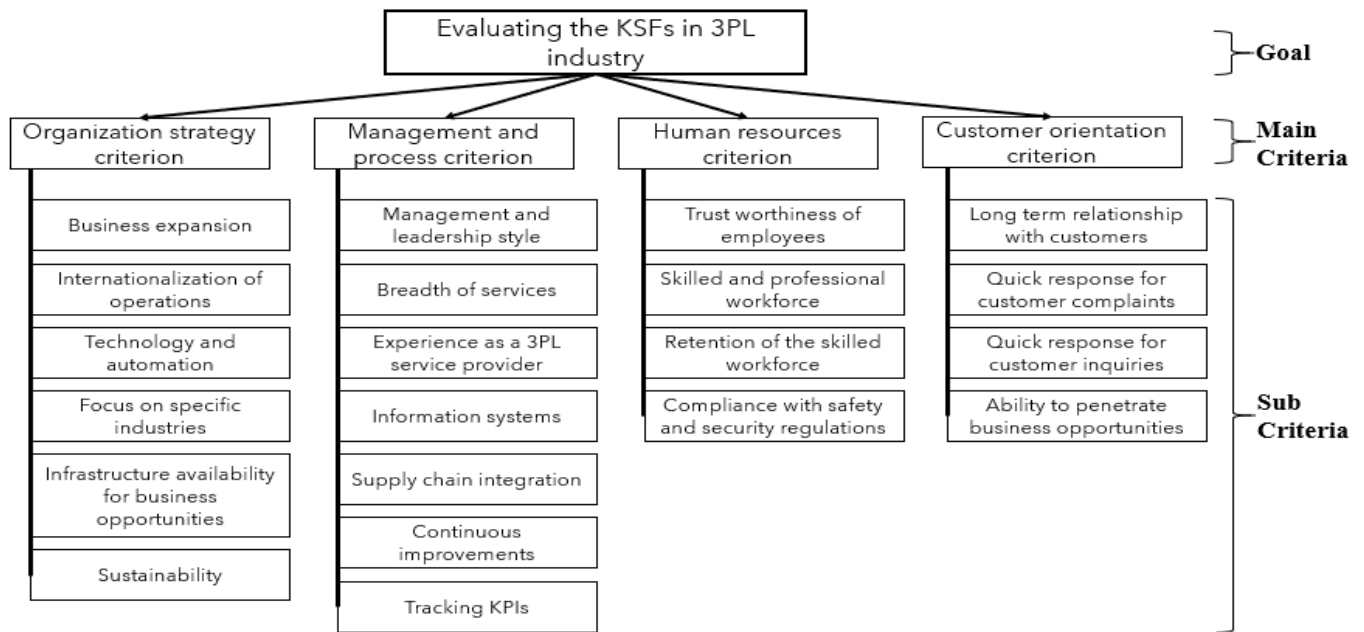


Fig. 3: Main and subkey success factors

By observing Fig. 4, organization strategy can consider as a core factor since it has high prominence and relation; the management and process factor can consider as a driving factor because it has a low prominence but high relation; the human resources and customer orientation have high prominence but low relation, which are impacted by other factors and cannot be directly improved. The threshold value that is considered to draw the digraph of the main criteria is 3.6053. It is necessary to set up a threshold value to filter out some negligible effects among factors.

Table III shows, technology, and automation, infrastructure availability for business opportunities, and sustainability which are in the cause group based on (r-c) values. Business expansion, internationalization of operations, and focus on specific industries are in effect group. The mean (r+c) value is 23.1002. Fig. 5, shows that technology and automation is the only core factor since it has high prominence and relation. Infrastructure availability for business opportunities and sustainability are the factor which considers as driving factors because it has a low prominence but high relation. Focus on specific industries has low prominence, relation and it is relatively disconnected from the system. The business expansion and internationalization of operations have high prominence but low relation, which are being influenced by other factors and cannot be directly improved. The threshold value that is considered to draw the digraph of the organization strategy factors is 1.9250 [31].

TABLE III. THE DIRECT AND INDIRECT EFFECTS OF ORGANIZATION STRATEGY FACTORS

Organization Strategy Factors	r+c	r-c
Business Expansion (OS1)	23.9560	-0.9900
Internationalization of Operations (OS2)	24.9506	-0.9543
Technology and Automation (OS3)	23.9220	0.8298
Focus on Specific Industries (OS4)	22.6404	-0.5701
Infrastructure Availability for Business Opportunities (OS5)	21.5743	1.4686
Sustainability (OS6)	21.5580	0.2161

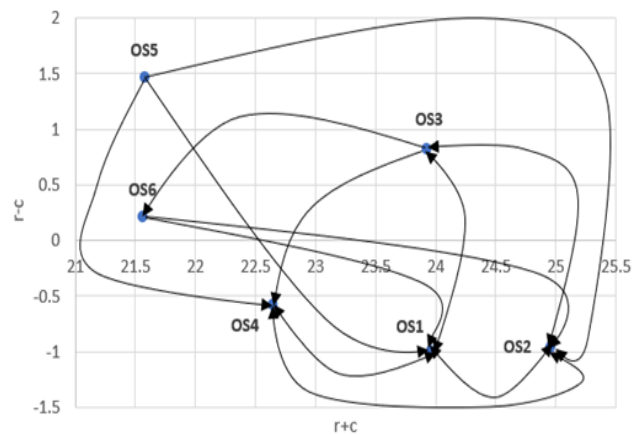


Fig 5. The digraph of the organization strategy factors

TABLE IV. THE DIRECT AND INDIRECT EFFECTS OF MANAGEMENT AND PROCESS FACTORS

Management and Process Factors	r+c	r-c
Management and Leadership Style (MP1)	7.0208	3.2107
Breadth of Service Offerings (MP2)	8.8385	0.7119
Experience as a 3PL Service Provider (MP3)	6.6339	2.6094
Information Systems (MP4)	9.5557	-1.5012
Supply Chain Integration (MP5)	9.4050	-0.2228
Continuous Improvements (MP6)	9.2848	-2.8169
Tracking KPIs (MP7)	9.5422	-1.9911

Table IV shows, management and leadership style, breadth of service offerings, and experience as a 3PL service provider which are identified as the key success factors in the cause group. Information systems, supply chain integration, continuous improvements, and tracking KPIs are the key success factors in the effect group. The mean (r+c) value is 8.6116. Fig. 6, shows that breadth of service offerings is a core factor since it has high prominence and relation. 3PL service providers should be more considerate about this factor to gain the benefit in the long run. Management and leadership style and experience as a 3PL service provider are

the factors which consider as driving factors because it has a low prominence but high relation. Under this main factor, there is no disconnected sub factor in the system. Information systems, supply chain integration, continuous improvements, and tracking KPIs have high prominence but low relation, which are impacted by other factors and cannot be directly improved. The threshold value that is considered to draw the digraph of the management and process factors is 0.6151.

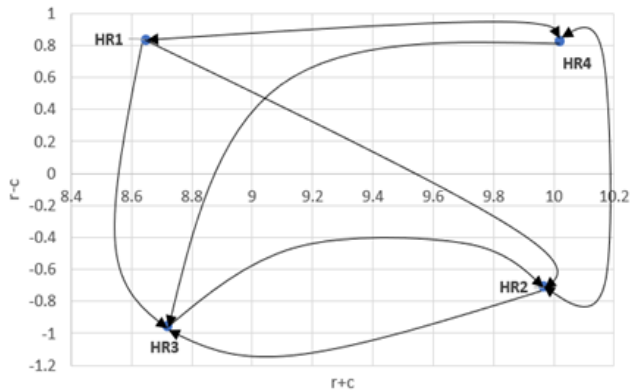


Fig 6. The digraph of the management and process factors

As in Table V, trustworthiness of employees and compliance with safety and security regulations are in the cause group. Skilled and professional workforce and retention of skilled workforce are the factors in effect group. The mean (r+c) value is 9.3383. Compliance with safety and security regulations is the core factor and trustworthiness of employees is the driving factor under this main factor. Therefore, managers need to put direct effort into compliance with safety and security regulations and need to build up the trustworthiness of employees (Fig. 7). Retention of the skilled workforce shows a disconnection from other factors since it has low prominence and relation. The skilled and professional workforce is the only impact factor that is impacted by other factors. The threshold value that is considered to draw the digraph of the human resources factors is 1.1672.

TABLE V. THE DIRECT AND INDIRECT EFFECTS OF HUMAN RESOURCES FACTORS

Human Resources Factors	r+c	r-c
Trust Worthiness of Employees (HR1)	8.6473	0.8350
Skilled and Professional Workforce (HR2)	9.9651	-0.7089
Retention of Skilled Workforce (HR3)	8.7211	-0.9555
Compliance with Safety and Security Regulations (HR4)	10.0198	0.8295

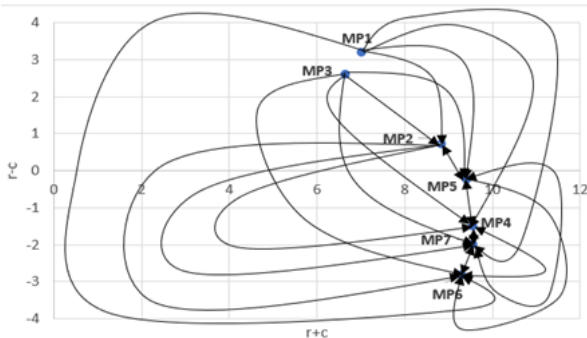


Fig 7. The digraph of the human resources factors

According to the (r-c) values in Table VI, long-term relationships with customers, quick response for customer complaints, and quick response for customer inquiries are in the cause group. The ability to penetrate business opportunities is the only key success factor in the effect group. The mean (r+c) value is 4.8481. Regarding the customer orientation main factor, long term relationship with customers is the only core factor. Quick response for customer complaints and quick response for customer inquiries are the driving factors. As shown in this digraph, the ability to penetrate the business opportunities factor has disconnected from other factors and stands as an independent factor. The threshold value that is considered to draw the digraph of the customer orientation factors is 0.6060 (Fig. 8).

TABLE VI. THE DIRECT AND INDIRECT EFFECTS OF CUSTOMER ORIENTATION FACTORS

Customer Orientation Factors	r+c	r-c
Long Term Relationship with Customers (CO1)	5.8417	0.4511
Quick Response for Customer Complaints (CO2)	4.8322	0.6108
Quick Response for Customer Inquiries (CO3)	4.4897	0.6831
Ability to Penetrate Business Opportunities (CO4)	4.2288	-1.7450

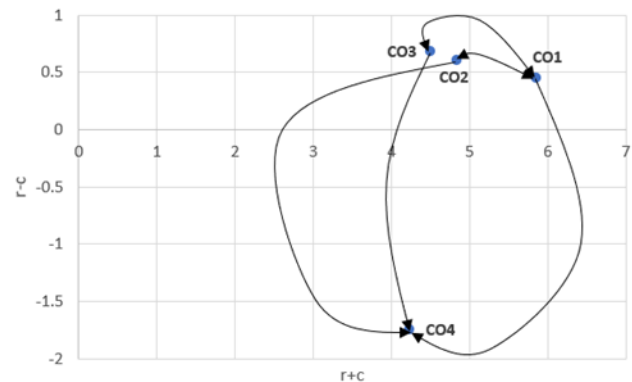


Fig 8. The digraph of the customer orientation factors

VI. CONCLUSION

This study set out to investigate the interrelationships of the key success factors of the 3PL industry in Sri Lanka. The study was able to investigate 21 key success factors using the DEMATEL method. The results of the DEMATEL application can be used to make long-term improvement opportunities in 3PL companies. These results would help managers in the 3PL industry to develop strategies for the effective supply chain management.

The result shows that organization strategy is a core factor since it has high prominence and relation; the management and process factor is a driving factor because it has a low prominence but high relation; the human resources and customer orientation have high prominence but low relation, which are impacted by other factors and cannot be directly improved. Therefore, managers need to focus more on main factors such as organizational strategy and management and process to increase the performance of the 3PL companies. Also, it is important to focus on subkey success factors which act as core factors and driving factors under each main factor. That will be useful for the management to develop long-term strategies for the companies.

The model proposed in the study has limitations. For example, the results of the DEMATEL method are highly dependent on the judgments of the experts. Great care was taken in finalizing the key success factors but cannot rule out errors due to human biases or judgment. Though a generalized model is developed here in this research, a particular company in the 3PL industry could select the criteria and sub-criteria according to their requirements and interest and develop a model that applies to their interest. The results of the DEMATEL could be dynamically adjusted according to adjust to new key success factors that may arise in the 3PL industry. This study is limited to the key success factors which were considered from the experts' opinion, may be improved by including the opinions of both the 3PL service providers and the customers

REFERENCES

- [1] Makmor, M. F. bin M., Saludin, M. N. bin, and Saad, M. binti., Best Practices Among 3rd Party Logistics (3PL) Firms in Malaysia towards Logistics Performance, *International Journal of Academic Research Business and Social Sciences*, vol. 9, no. 5, pp. 394-405, 2019.
- [2] Rajesh, R., Pugazhendhi, S., Ganesh, K., Yves, D., oh, S. C. L. and Muralidharan C., Perceptions of service providers and customers of key success factors of third-party logistics relationships – an empirical study, *International Journal of Logistics Research and Applications*, vol. 14, no. 4, pp. 221-250, 2011.
- [3] Malkanthie, M. A. A. and Jayamanna, J. M. D. J. N., Exploration of Factors Hindering the Growth of 3PL Market in Sri Lanka, *Academy for Global Business Advancement (AGBA), 13th World Congress, Indonesia, 2016.*
- [4] <https://ipi.worldbank.org/international/scorecard/radar/254/C/LKA/2018>, accessed on 14/05/2020.
- [5] Fernando, H. and Rajapaksha, U. G., The Impact of 3PL Service on Total Quality Management of Apparel Industry in Sri Lanka, *Proceedings in Management, Social Sciences and Humanities, 9th International Research Conference-KDU, Sri Lanka*, pp. 234-243, 2016.
- [6] Karunaratna, N., Vidanagamachchi, K. and Wickramarachchi, R., A Calibrated Model of Critical Success Factors for Industry 4.0 Warehousing Performance Improvement: Insights from Multiple Case Studies, in *International Journal of Multidisciplinary Sciences and Advanced Technology*, vol. 1, no. 2, pp. 100-126, 2020.
- [7] Sugathadasa, P. T. R. S. and Rajapaksha, S. S., An Investigation on Relationship between Third Party Logistics User and Provider at FMCG Industry in Sri Lanka, *17th Eru Research Symposium, Moratuwa, 2011.*
- [8] Madurapperuma, S., Ebert, L. J., Gamage, S. and Kurupparachchi, D., In-House Development & Implementation of 'CoreBrain' Warehouse Management System: A Case Study, *2nd International Conference in Technology Management – iNCOTeM2018, Colombo, Sri Lanka*, pp. 67-72, 2018.
- [9] Pollard, C. and Cater-Steel, A., Justifications, strategies, and critical success factors in successful ITIL implementations in US and Australian companies: an exploratory study, *Information Systems Management*, vol. 26, no. 2, pp. 164-175, 2009.
- [10] Asthana, S. and Dwivedi, A., Performance measurement of India-based third-party logistics sector: an empirical study of user versus provider perspectives *Production Planning & Control*, vol. 31, no. 2, pp. 259-272, 2020.
- [11] Kawa, A., SMART logistics chain, *Proceedings of the 4th Asian conference on Intelligent Information and Database Systems*, pp. 432-438, 2012.
- [12] Schiemann, J., *Logistics 4.0 How Autonomous Are Self-Managed Processes?* AXIT Research Report, Frankenthal, 2016.
- [13] Juntao, L. and Yinbo, M., Research on Internet of Things Technology Application Status in the Warehouse Operation. *International Journal of Science, Technology and Society*, vol. 4, no. 4, pp. 63-66, 2016.
- [14] Marchet, G., Melacini, M., Sassi, C. and Tappia, E., Assessing efficiency and innovation in the 3PL industry: an empirical analysis, *International Journal of Logistics Research and Applications*, vol. 20, no. 1, pp. 53-72, 2017.
- [15] Shen, C. and Chou, C. C., Business process re-engineering in the logistics industry: a study of implementation, success factors, and performance, *Enterprise Information Systems*, vol. 4, no. 1, pp. 61-78, 2010.
- [16] Shaiq, M., Alwi, S. K. K., Shaikh, S. and Zaman, Z., Quality Management as Driver of Vertical Integration in Service Chain: A Study of 3rd Party Logistics Industry, *OPERATIONS AND SUPPLY CHAIN MANAGEMENT*, vol. 13, no. 3, pp. 244 - 255, 2020.
- [17] Alinejad, E. A., Pishvae, M. S. and Naeini, A. B., Key success factors for logistics provider enterprises: an empirical investigation in Iran, *Kybernetes*, vol. 47, no. 3, pp. 426-440, 2018.
- [18] Gupta, O. K., Ali, S. S. and Dubey, R., Third Party Logistics: Key Success factors and growth Strategies, *International Journal of Strategic Decision Sciences*, vol. 2, no. 4, pp. 29-60, 2011.
- [19] Bianchini, A., 3PL provider selection by AHP and TOPSIS methodology, *Benchmarking: An International Journal*, vol. 25, no. 1, pp. 235-252, 2018.
- [20] Asian, S., Pool, J. K., Nazarpour, A. and Tabaeian, R. A., On the importance of service performance and customer satisfaction in third-party logistics selection, *Benchmarking: An International Journal*, vol. 26, no. 5, pp. 1550-1564, 2019.
- [21] Dieu Ho, T. H., Daniel, J., Nadeem, S. M., Garza-Reyes, J. A. and Kumar, V., Improving the Reliability of Warehouse Operations in the 3PL Industry: An Australian 3PL Case Study, *Proceedings of the 2018 International Conference of the Production and Operations Management Society (POMS), Kandy, Sri Lanka*, pp. 1-8, December 2018.
- [22] Sangka, B. K., Rahman, S., Yadlapalli, A. and Jie, F., Managerial competencies of 3PL providers, *The International Journal of Logistics Management*, vol. 30, no. 4, pp. 1054-1077, 2019.
- [23] Vyas, R. and Shah, T., Adoption of 3PL Practices in Saurashtra Region: Impact and Influence of Key Success Factors on Revenue Growth, *International Journal of Current Multidisciplinary Studies*, vol. 2, no. 5, pp. 273-278, 2016.
- [24] Makmor, M. F. bin M., Saludin, M. N. bin, and Saad, M. binti., Best Practices Among 3rd Party Logistics (3PL) Firms in Malaysia towards Logistics Performance, *International Journal of Academic Research Business and Social Sciences*, vol. 9, no. 5, pp. 394-405, 2019.
- [25] Mitra, S. and Bagchi, P. K., Key Success Factors, Performance Metrics, and Globalization Issues in the Third-Party Logistics (3PL) Industry: A Survey of North American Service Providers, *Supply Chain Forum An International Journal*, vol. 9, no. 1, pp. 42-54, 2008.
- [26] Mothilal, S., Gunasekaran, A., Nachiappan, S. P. and Jayaram, J., Key success factors and their performance implications in the Indian third-party logistics (3PL) industry, *International Journal of Production Research*, vol. 50, no. 9, pp. 2407-2422, 2012.
- [27] Govindan, K. and Chaudhuri, A., Interrelationships of risks faced by third party logistics service providers: A DEMATEL based approach, *Transportation research part E: logistics and transportation review*, vol. 90, pp. 177-195, 2016.
- [28] Kaur, H., Singh, S. P. and Glardon, R., An Integer Linear Program for Integrated Supplier Selection: A Sustainable Flexible Framework, *Global Journal of Flexible Systems Management*, vol. 17, no. 2, pp. 113-134, 2015.
- [29] Si, S. L., You, X. Y., Liu, H. C. and Zhang, P., DEMATEL Technique: A Systematic Review of the State-of-the-Art Literature on Methodologies and Applications, *Mathematical Problems in Engineering*, 2018.
- [30] Wu, H. H. and Tsai, Y. N., An integrated approach of AHP and DEMATEL methods in evaluating the criteria of auto spare parts industry, *International Journal of Systems Science*, vol. 43, no. 11, pp. 2114-2124, 2012.
- [31] Perera, T., Wijayanayake, A. and Wickramarachchi, R., A Combined Approach of Analytic Hierarchy Process and Decision-Making Trial and Evaluation Laboratory Methods for Evaluating Key Success Factors of Third-Party Logistics Service Providers, *11th Annual International Conference on Industrial Engineering and Operations Management, Singapore*, pp. 1078-1089, 2021.

A decentralized social network architecture

Tharuka Sarathchandra*
Department of Software Engineering
University of Kelaniya, Sri Lanka
tharukas@kln.ac.lk

Damith Jayawikrama
H&D Wireless SL, Sri Lanka
damith@damith.com

Abstract - Billions of people use social networks, and they play a significant role in people's lifestyles in the current world. At the same time, due to globalization and other factors, the use of these social platforms is expanding daily, and a variety of activities take place inside these platforms. These networks are centralized, allowing social network-owned companies to track and observe the activities of their users. Therefore, this has been challenged to the privacy of the data of users. Also, these companies tend to sell them to third parties keeping huge profits without users' permission. Since data is the most valuable asset in today's and tomorrow's world, many have pointed out this issue. Even though decentralized, community-driven applications have come to play as a solution to this problem, there is still no successful application that competes with centralized social network platforms. Therefore, this study attempted to develop a decentralized social network architecture with the basic functionalities of a social media platform to assure the privacy of the users' data.

Keywords - blockchain, ethereum, decentralized web, ipfs, web3.0

I. INTRODUCTION

Nowadays, almost every person uses social media, and social networks play a crucial part in lifestyles. Most people around the world are connected with social networks. The first recognizable social network, "Six Degrees," was launched in 1997, allowing users to create a profile and become friends with other users [1]. The world has come a long way where now people are using many social networks like Facebook, WhatsApp, Instagram, Twitter, etc. The usage of these social networks is increasing day by day [2].

It is a must to discuss the reasons for the increment of the usage of these social networks. The most common reason is that the users want to stay in touch with others and stay updated on what is going on around the world. With the busy schedules and workload, people miss the chance of meeting people physically. Therefore, they spend time virtually, which is more beneficial. On the other hand, People use social networks to share photos and videos for entertainment and share opinions and ideas. Besides that, people use social media platforms to research new products to buy. With these facts, businesses do more and more online marketing focusing on the target audience and try to grab the customer. However, these social networks are centralized, and there are several problems with those social networks. Recently, those problems became hot topics.

With the popularity of web 3.0, people tend to find a solution within the web 3.0 technology stack for the problems they face with current web 2.0 technologies. As a result of that, decentralization came to play. With the evolution of cryptocurrency, mainly Bitcoin, this decentralized culture came into practice.

Then people proposed decentralized solutions to develop applications apart from using decentralization only in cryptocurrencies. Nowadays, many organizations have developed applications for web, mobile and desktop

computers with decentralized technologies. Those applications are called as DAPPs, and they have been able to solve many real-world problems.

II. DECENTRALIZATION

After pouring the cryptocurrencies, the word decentralization [3] became a bus word because this is used most in the crypto-economics space. Many people and companies started to do researches in this area, and thousands of hours of research, and billions of dollars of cash power, have been spent for the sole purpose of attempting to achieve decentralization. There are many misunderstandings about decentralization. Therefore, it is necessary to understand the differences between Centralized Networks, Decentralized Networks, and Distributed Networks.

A centralized system means a central location provides all services, and the network resources are placed and managed from the central location. Distributed means the network resources and the services are distributed through the network, and it might also be geographically distributed over the internet. However, the network and the resources are handled by a central authority, and they have the administrator power to do everything in the system. Google, Facebook, AWS, like almost every big company, use this distributed model in their systems. Decentralized means there is no central place or no one has administrator powers to govern the system. The system is distributed through the users of the system. Users are the people who govern the system. However, there is a critical point to consider. That is who maintains the system and who develops the new versions of the system. The answer is that almost all decentralized projects are free and open-source projects. Therefore, anyone interested in a particular area can develop the project, and many of them run on funds. However, some decentralized system has their business models, and a small amount of money is charged from the users when using the system for maintenance and other infrastructure developments. Nevertheless, by the architecture of the decentralized system, the developers of a particular system do not have a central authority to handle the system.

A. Smart contract

A smart contract [3] is older than bitcoin, and it is a computer protocol to digitally facilitate, verify, or enforce the negotiation or performance of a contract. In other words, Smart contracts are computer programs that a network of mutually distrusting nodes can correctly execute without the need of an external trusted authority. With the development of cryptocurrencies and Blockchain, Smart Contracts got attraction due to their architecture. If a bitcoin-like cryptocurrency saves the transaction in a blockchain, it is just some kind of data. However, these Smart Contracts open a way to store some executable code inside a blockchain, and it is an immutable program. It can be partially or fully self-

executing, self-enforcing, or both. Most cryptocurrencies have the facility to implement these smart contracts. Smart contracts can be used to do complex transactions between two anonymous parties. Moreover, it does not require a central authority, enforcement system, or legal guidance because it can be self-executed by itself. Therefore, Smart Contracts can be programmed to enable a wide variety of actions.

B. Cryptography

Cryptography or cryptology is the domain and study area of ways and technologies to secure communication between two or more parties from external parties. In general terms, cryptography is about constructing and analyzing protocols that prevent third parties or the public from reading private data. Various aspects of information security, such as data confidentiality, authentication, non-repudiation, and data integrity, belong to modern cryptography. Cryptography mechanisms are based on mathematics and computer science, electrical engineering, communication science, and physics disciplines.

When discussing decentralized systems, the privacy of the data is a huge issue. As data is shared publicly, different cryptographic mechanisms are used in decentralized networks to resolve this problem. Cryptography has two categories as symmetric-key cryptography and asymmetric key cryptography

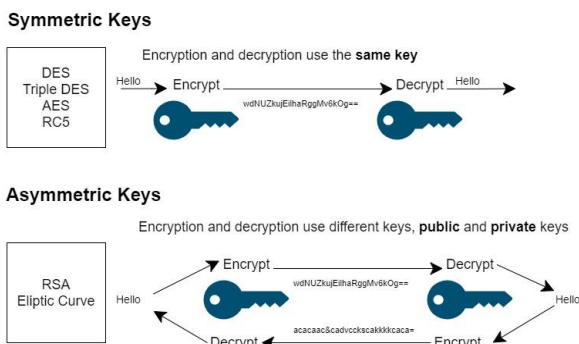


Fig. 1. Types of encryptions

In symmetric-key cryptography, it is used only one key in both encryption and decryption processes. However, asymmetric key cryptography uses two different keys in encryption and decryption processes.

C. Blockchain

A blockchain [4] is a data structure that enables identifying and tracking transactions digitally and sharing this information across a distributed network of computers, creating a distributed trust network. The data structure of a blockchain is a linked list, and the speciality is it being an immutable linked list.

D. Features of Blockchain

- **Distributed and Decentralized** - Data are replicated on every node in a distributed P2P network. Furthermore, each copy is identical to others. It can also be decentralized with some lighter nodes not having whole data storage with limited connection.

- **Consensus mechanism** - All users in the blockchain network can come to a predetermined programmable agreement on the validation method and can be by consensus. There are several consensus algorithms like Proof of Work, Proof of Stake, Delayed Proof-of-Work, Proof of Importance, Delayed Proof-of-Stake, etc. Most decentralized applications use POW and POS[5].
- **Irreversibility and crypto security** - One would need to command at least 51% of the computing power (or nodes or sake) to take control of the bitcoin blockchain (or other) [6].

E. Bitcoin and cryptocurrencies

Bitcoin [7] is the world's first known public cryptocurrency invented by an anonymous man known as Satoshi Nakamoto. He published the research paper Bitcoin: A Peer-to-Peer Electronic Cash System in 2008. That was the point that the whole world gives attention to cryptocurrencies. In this paper, the author has resolved the problem of public transaction verification. The concept of proof-of-work [8] is intruded in this research. It uses a blockchain as its underlying data structure and the public ledger.

After introducing Bitcoin in 2009, hundreds of cryptocurrencies were introduced to the world, and most of them have used blockchain, and some of them have their own data structures like Directed Acyclic Graph. However, the concept was almost the same. That means all data store in a public ledger. After the invention of proof-of-work, people have invented new concession algorithms which are different from proof-of-work. Proof of Stake, Proof of Elapsed Time, Proof of Authority, Proof of Capacity, Proof of Activity, and Proof of Burn are examples of those algorithms used in different cryptocurrencies and decentralized application developments.

F. Ethereum ecosystem

With various types of cryptocurrencies, the Ethereum [9] research was published in 2014. It is not just a cryptocurrency. Ethereum saw this use case with a different and broad view. It has proposed a platform to develop any decentralized application. Also, it has its cryptocurrency called Eth.

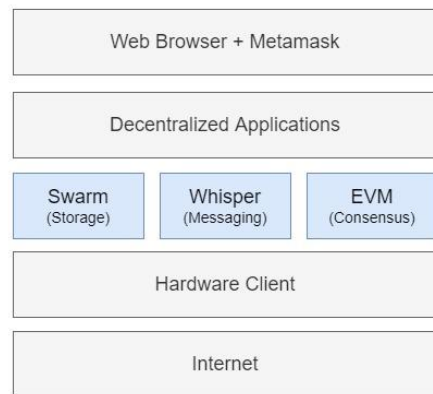


Fig. 2. Ethereum Ecosystem

Now there are hundreds of decentralized applications and crypto tokens that have been built on top of the Ethereum platform [10] UPort, Brave, Toshi, Auger, CryptoKitties, GitCoin, Minds, and Akasha are a few more popular example applications based on the Ethereum platform.

G. Ethereum Virtual Machine

Ethereum is somewhat a predecessor to Bitcoin. The Ethereum Virtual Machine (EVM)[11] is one of the primary core reasons for Ethereum to be created. While Bitcoin does contain a programmable scripting language, the Bitcoin scripting language is limited to performing token transfers. EVM is designed with 20-byte addresses. Furthermore, each address space has a counter, a balance value, contract code, and persistent storage [11]. Overall, Bitcoin scripting offers a limited feature set compared to EVM as Ethereum's primary goal is to create a globally distributed computing platform.

H. Decentralized chatting

Chatting is a significant component of social media. When people are using the current social media, almost all of them have chatting functionality. Video calls, voice calls are also some categories of chats.

Before exploring more information about current projects, researchers had to answer the problem; what is the purpose of using this decentralized nature for the chat protocols? The answer is that these chat service providers can listen to what the users are chatting about and all these data routes through their servers. Therefore, privacy cannot be ensured

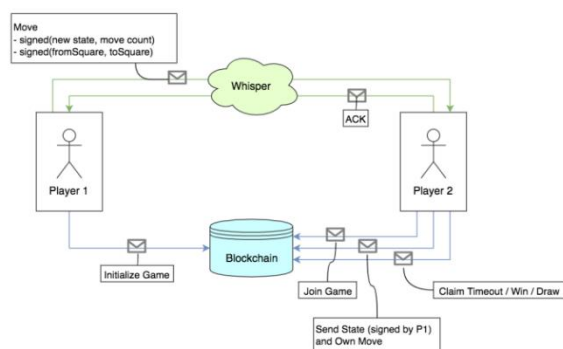


Fig. 3. Blockchain and Whisper

Whisper [12] is a decentralized communication protocol to communicate with each other. Darkness is an important feature of this protocol. Therefore, no one can trace the message senders and receivers. The protocol sends the message to everyone in the network, and it behaves like a gossip protocol to achieve that darkness. It sends the message to all connected nodes; the origin node sends messages to the connected node. Likewise, the message is communicating to everyone. However, only the relevant node has the private key to decrypt the message because it uses asymmetric key cryptography. Therefore, the cost of the protocol is relatively high[13].

I. Decentralized storages

There is a major problem when considering the storage mechanism for a decentralized application. Most people lack knowledge about blockchain technology and think it is

possible to use a blockchain for a complete solution for a storage problem in a decentralized application. However, it is impossible to store a large amount of data in a blockchain. The best solution for the storage problem in decentralized applications is the InterPlanetary File System (IPFS), and blockchain can be used to reference the IPFS platform.

IPFS gives a unique cryptographic fingerprint to every content which is published in the file system. It removes the duplication of the file across the system, and it delivers the files from the nearest node to which the file system hosts a file. Each network node stores only the content that it is interested in, and some indexing information helps figure out who is storing and what is also stored in the nodes. When looking up files, the user asks the network to find nodes storing the content behind a unique hash.

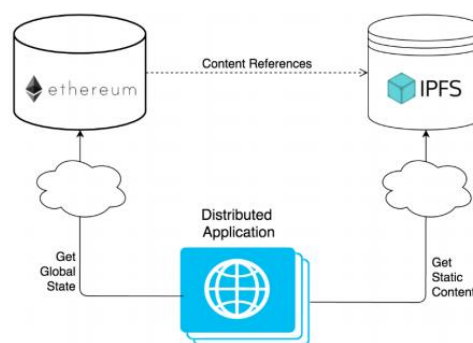


Fig. 4. Decentralized storage - IPFS

IPFS is called a permeate web because the file system behaves quite like the Git version controlling the file system. Every version of the file will be stored. It uses local storage to store files and distribute that file within other nodes [14].

When someone uploads something, the file is chunked by IPFS and stored in his cache folder (ipfs). Suppose a user tries to see the file on another peer of the network (say the main gateway, ipfs.io) that peer requests the file to you and caches it too. If he switches off his daemon, he can still see the file on the gateway, probably because the gateway or some other peer on the web still has it cached. When a peer wants to download a file, but it is out of memory (it can be no longer cached), the oldest used files get forgotten to free the space. That is a simple explanation for IPFS, but it is more complicated than this.

InterPlanetary Naming System (IPNS) behaves like Domain Name Service (DNS) in the decentralized IPFS ecosystem. In IPFS, an uploaded content is identified using its fingerprint hash. However, it is difficult to remember that hash. Therefore, this IPNS is providing a service to have a unique human-readable identity to each hash.

There are several other ongoing research projects to solve the same storage problem. FileCoin, Storj, MaidSafe, SWARM are few examples.

FileCoin [15] is a cryptocurrency like bitcoin, but miners must share their computer storage with the network users. Bitcoin uses proof of work as the consensus algorithm, introducing a novel consensus way of mining FileCoin called Proofs of-storage. It is based on two consensus algorithms Proof-of-Replication and Proof-of-Spacetime. Proof-of-Replication: allows storage providers to prove that data has been replicated to its own uniquely dedicated physical

storage, and Proof-of-Spacetime: allows storage providers to prove they have stored some data throughout a specified amount of time. Then miners can earn coins by providing their storage. Therefore, this network has massive, decentralized storage. The important point is that this FileCoin network works as an incentive layer on top of IPFS. Therefore, IPFS seems like a good storage solution for a decentralized ecosystem.

STORJ [16] is another ongoing decentralized storage research project, and it is about a decentralized cloud storage network framework. In this framework, when a client saves a file, it will be encrypted first on the clientside. Then it is chunked into small pieces, and those pieces are sent to storage nodes, and storage nodes are storing those chunked data pieces. When chunking data, a central server keeps tracking which parts are relevant to the chunked file. When constructing the file again that metadata is used. The chunked pieces are replicated through the storage nodes based on a threshold value. Storage nodes are selected using several factors like ping time, latency, throughput, bandwidth caps, client disk space, geographic location, uptime, history of responding accurately to audits, etc. The speciality of this system is that this is an S3 capable platform.

SWARM [17] is also another solution for the distributed storage. It provides a content distribution service, a native base layer of the Ethereum Web3 stack. SWARM has been decentralized to serve as a redundant store of Ethereum's public records to store and distribute Smart Contracts. This platform is also a peer-to-peer storage platform maintained by its peers who contribute their storage and bandwidth resources. Being a peer-to-peer system, this has no single point of failure, and it is resistant to failures and Distributed Denial-of-Service (DDoS) attacks.

J. Other decentralized social networks

There are few ongoing projects on decentralized social networks. Furthermore, they are focused on different kinds of areas in the decentralized social network domain. Seemit, Sola, Memo, VeganNation, and Indorse are few examples that focus on different aspects. However, none of them could give a better solution to overcome social networks such as Facebook and Linked-In. There are many reasons behind that. The main reasons are the lack of awareness of ordinary people about the current social media ecosystem's problems and their need to do everything quickly. As this decentralized world is still in its early stages, there are many usability issues. As a result, the decentralized world remains popular among those who are familiar with the underlying technology [18].

III. TECHNOLOGY ADAPTED

A. Web 3.0 stack

There are few new different categories of technologies that are included in the web 3.0 stack. Web browser, Web application, Web Protocols, Network architecture, Data storage, Application deployment are a few. It is needed to find a solution for each category replacement in the web 3.0 stack to replace the web 2.0 stack. However, Web 3.0 is still in the research stage and not mature as the Web 2.0 technology stack. Therefore, it is hard to find a complete solution only using the web 3.0 stack to achieve the

requirements. Therefore, it will have to use a hybrid but more into Web 3.0 approach when developing a solution.

B. MetaMask

MetaMask is an application that helps the decentralized application to perform its transaction in the Ethereum network. It can be added to the web browser as plugging, and then it will be automatically triggered whenever the user is going to do a transaction in the blockchain network. It is a bridge between the decentralized web application and the blockchain network. Using MetaMask, connecting to the Main Ethereum networks or any other custom Ethereum network is possible. It provides an Ethereum wallet management facility and account management facility as well. So, it is straightforward to keep several accounts in different or the same blockchain networks. Also, it provides an account recovery facility too.

C. Oracle

One significant restriction of smart contracts is that they cannot directly access other data sources such as APIs, databases, and IoT sensor data. Because the data access for outside can be changed with time and the Smart Contract execution is fully deterministic. Since the external sources on the internet are non-deterministic, it is impossible to get the same state after replaying the changes to the blockchain over time. Nodes of the network come to a consensus with this determinism of the network. That is the place that the oracles come to play. They give the flexibility to the Smart Contracts to interact with off-chain data sources. Oracles themselves are not data sources. They work as an extra layer between smart contracts and off-chain data sources. Mainly, there are several types of oracles such as Software Oracles, Hardware Oracles, Human Oracles, Contract-specific Oracles, etc. BandChain, Oracalize, Chainlink, Teller, and Provable are few oracle services that enable external data access to the Ethereum blockchain [19].

D. Infura

Infura is a platform that helps to develop decentralized applications easily. It provides an infrastructure to interact with Ethereum and IPFS gateways. It provides secure, reliable, and scalable access to those gateways.

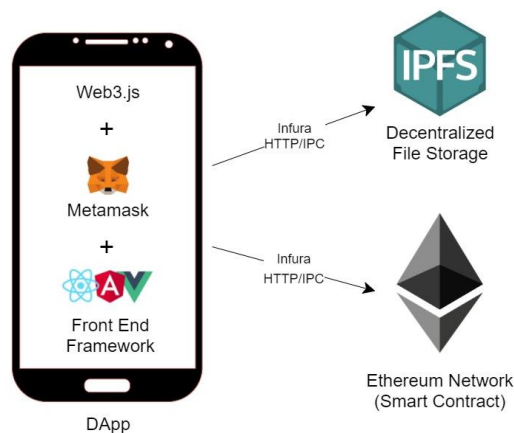


Fig. 5. Connection between IPFS, Smart Contract and Infura

As the figure shows, using Infura, it is possible to interact with remote IPFS and Ethereum networks directly. Otherwise, it is needed to host a local Ethereum or IPFS client.

IV. PROPOSED ARCHITECTURE

A. Decentralized social network high-level architecture

As shown in Fig.6, DAPP will be a client-side application that users can access through their web browsers. It will base on JavaScript, HTML, and CSS. On top of these traditional web technologies, there will be a Web3.js layer as the bridge between the client application and the back end.

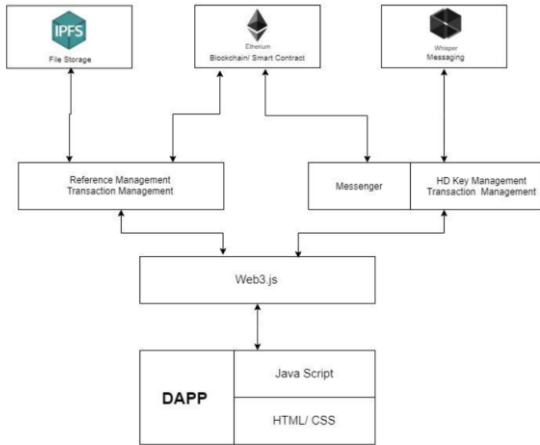


Fig. 6. Decentralized Social Network High-Level Architecture

In this case, the backend will be handled using the Ethereum blockchain network using the smart contracts deployed inside the Ethereum blockchain. Interplanetary file system (IPFS) will be working as the data storage layer in the system. Whisper will work as the messaging platform of the system.

B. Front-end architecture

Fig. 7. shows the client-side architecture of the system. Here the application is designed using component-based architecture. All external calls such as API calls, JSON RPC are handle by the services. It is the interface of the client-side applicant to external entities.

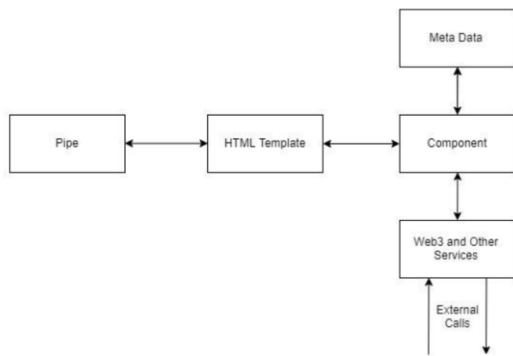


Fig. 7. Frontend architecture

C. Backend external API support

Nowadays, almost every web service can communicate with external web services. However, in this decentralized scenario, Smart contracts are living in the blockchain. Therefore, they can interact with data living in the same blockchain network. However, the limitation is that they cannot interact with the outside blockchain, such as web API. Nevertheless, for modern applications, it is a must to interact with external APIs. Here is the design for support for the external APIs.

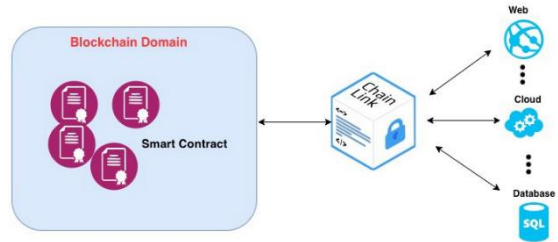


Fig. 8. Connection between Smart Contract and external datastores through Oracle

Here, ChainLink will work as a middle platform. Smart-Contract can interact with ChainLink. ChainLink has Smart Contract to support that purpose. Then ChainLink will call the external web APIs or other external off-chain services. Then after the result comes to the chainlink, its callback to the called Smart Contract.

D. Interact with IPFS

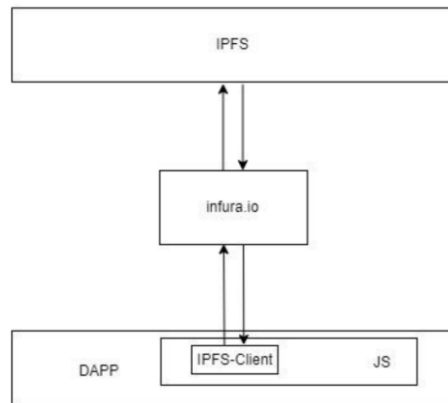


Fig. 9. DApp Interaction with IPFS

This diagram shows how a DAPP interacts with the InterPlanetary File System. IPFS-Client library is the client-side interface of the interaction, and It creates a connection with the Infura platform and provides the facility to communicate with the IPFS network. With this design, there is no need to run a local IPFS-Client.

V. IMPLEMENTATION

Truffle was used as the Smart Contract development framework in application development, and the programming language was Solidity [20]. It is one of the mature solutions in developing Ethereum based decentralized applications.

A. Solidity

Solidity is the programming language used to program the smart contract for the system. It is a contract-oriented

programming language, and it is Turing complete programming language. Solidity codes are compiled into bytecode using Remix [21] compiler.

B. Truffle

Truffle makes the Smart Contract development process easy. It handles the Smart Contract compilation, bytecode management, linking, and deploying the smart contract in the given Ethereum network. It gives a command-line interface, and it is instrumental in development.

C. Mocha

Mocha was used together with Truffle as the testing framework. It supports Smart Contract testing.

D. Web3 JS

Web3 JS is a library that works as a bridge between client-side applications and the Ethereum blockchain. It has several implementations in several languages. Web3.js is a major implementation of Web3, which is used to work with web applications

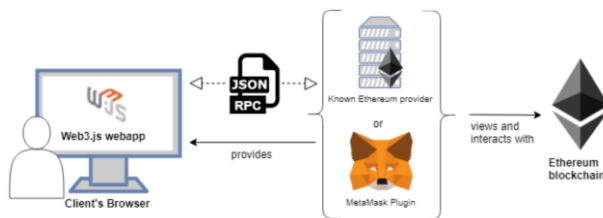


Fig. 10. Interaction between Web3js and Ethereum

Web3.js is the bridge between Ethereum blockchain and the web browsers (client-side). It is a JavaScript API that is compatible with the Ethereum blockchain. It uses generic JSON RPC to work with the client-side. To communicate with the blockchain, it uses an application binary interface (ABI) provided by Smart contracts.

E. IPFS-API

IPFS API is a JavaScript library that was used to interact with the InterPlanetary file system. This library can be configured with Infura.io. Then it is possible to communicate remote IPFS gateways easily. Whole data sending and receiving processes are passing through this IPFS-API library

F. Crypto JS

CryptoJS is a collection of secure and standard cryptographic algorithms implemented using JavaScript with best-practice patterns and practices. They are fast, and they have a consistent and simple interface.

G. Implementation of the client-side

As a unit testing formwork, Mocha was used. As libraries, Web3.js, Angular was used in developing the client-side of the application (Front end). The system is designed to use a component-based architecture. Web3 JS, ipfs-client are the most impairment libraries, and they were connected to achieve the desired decentralization. User can store data and retrieve using ipfs-client. Web3 is used to create a new user account and store user IPFS hash to reference user detail in the IPFS.

H. Custom ethereum network

Ganache Blockchain is a perfect development solution, debugging, and test because it provides many features [22]. However, for the actual implementation, there are two solutions. The first is to deploy the decentralized application in Ethereum's main network or public Ethereum test networks like Ropsten or Rinkeby. The second option is to develop a private Ethereum network available only for social network users [23]. In this research, the second option has been chosen. Because this works as a separate platform and, it cannot be dependent on another Ethereum network. Suppose it depends on another Ethereum network. If the app uses an Ethereum public network, there are several problems: the gas limit, coin base, difficulty, etc. If it happens, it is hard to achieve the intended purpose of the application. In this case, it is needed to develop a custom Ethereum network, and it can be easily done by the go-Ethereum client software (Geth). The first thing that needs to do is to create a genesis block of the network. It is the first block of the network. It can be defended as a JSON file.

VI. RESULTS

For testing purposes, this research used an intel core i7 laptop with 16GB memory, and the operating system was Windows 10 Student Edition. Both blockchain and the Angular front-end application were deployed in the same machine.

Using the proposed system architecture and technologies, it could develop a prototype of this decentralized system that has features to create and update user profiles, search user profiles, add friends, chat with friends, post text and photos on the user's wall, and comments on the post. The system was tested with only ten concurrent users, and the response time for creating/updating users, sending friends requests, and adding friends were less than 8 seconds. For the post-sharing functionality, the response time depends on the size of the content. Generally, IPFS takes 16s to upload 1GB of data [24], and then after uploading the multimedia file, it takes up to 8 seconds to process inside the developed prototype.

VII. CONCLUSION

With the advances of technology, Web 3.0 is expected to be the future of the web. However, people doubt whether the web 2.0 centralized web architecture can be replaced by decentralized web 3.0 architecture. This research focused on developing a decentralized social network architecture that can provide more privacy, data ownership, and community-driven facilities mainly based on the Ethereum platform. However, the platform can be changed to achieve efficiency in the future as there are commonly known limitations in Ethereum blockchain and other blockchains. Giant organizations such as Facebook, Google, and Microsoft are also developing and exploring these technologies, which look promising about decentralized computing.

VIII. FUTURE WORK

The research is proposed with a whole system architecture to develop the decentralized applications. However, the implementation of such an application is massive work. Therefore, the implementation of the research is just a proof of concept. In the future, the application will be fully implemented with the proposed concept.

Furthermore, it will be available for the public to use. When considering the security of the data, the application must have more consideration. The system design can currently set data visibility to only me or the public, handled by basic encryption and decryption mechanisms. However, more focus should be on authentication and authorization with different access levels for different data types.

As the initial step, the application is developed based on a public IPFS network. Nevertheless, in the future, with the improvement of the system's user base, it can be developed into a custom IPFS network. Then it can be dedicated to this decentralized social network. By developing such a network, the efficiency of the system can be improved.

REFERENCES

- [1] "The History and Evolution of Social Media," Webdesigner Depot, Oct. 07, 2009. <https://www.webdesignerdepot.com/2009/10/the-history-and-evolution-of-social-media/> (accessed Feb. 09, 2019).
- [2] "Number of social media users worldwide 2010-2021," Statista. <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/> (accessed Feb. 09, 2019).
- [3] "What Are Smart Contracts? A Beginner's Guide to Smart Contracts," Blockgeeks. <https://blockgeeks.com/guides/smart-contracts/> (accessed Dec. 03, 2018).
- [4] Hartikka, "A blockchain in 200 lines of code," Lauri Hartikka, Mar. 04, 2017. <https://medium.com/@lhartikk/a-blockchain-in-200-lines-of-code-963cc1cc0e54> (accessed Dec. 02, 2018).
- [5] M. Bach, B. Mihaljevic, and M. Zagar, "Comparative analysis of blockchain consensus algorithms," in 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, May 2018, pp. 1545–1550. doi: 10.23919/MIPRO.2018.8400278.
- [6] Ye, G. Li, H. Cai, Y. Gu, and A. Fukuda, "Analysis of Security in Blockchain: Case Study in 51%-Attack Detecting," in 2018 5th International Conference on Dependable Systems and Their Applications (DSA), Dalian, China, Sep. 2018, pp. 15–24. doi: 10.1109/DSA.2018.00015.
- [7] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," p. 9.
- [8] "Proof of Work vs Proof of Stake: Basic Mining Guide," Blockgeeks. <https://blockgeeks.com/guides/proof-of-work-vs-proof-of-stake/> (accessed Dec. 02, 2018).
- [9] "What is Ethereum? — Ethereum Homestead 0.1 documentation." <http://ethdocs.org/en/latest/introduction/what-is-ethereum.html> (accessed Dec. 02, 2018).
- [10] "Ethereum Project." <https://www.ethereum.org/> (accessed Dec. 02, 2018).
- [11] "r/ethereum - Can someone possibly explain the concept of GasPrice?" reddit. https://www.reddit.com/r/ethereum/comments/3fnpr1/can_someone_possibly_explain_the_concept_of/ (accessed Dec. 02, 2018).
- [12] "whisper-overview," Ethereum Wiki. <https://eth.wiki/concepts/whisper/whisper-overview> (accessed Jun. 12, 2021).
- [13] L. Zhang, Z. Zhang, Z. Jin, Y. Su, and Z. Wang, "An approach of covert communication based on the Ethereum whisper protocol in blockchain," Int. J. Intell. Syst., vol. 36, no. 2, pp. 962–996, Feb. 2021. doi: 10.1002/int.22327.
- [14] J. Benet, "IPFS - Content Addressed, Versioned, P2P File System," ArXiv14073561 Cs, Jul. 2014, Accessed: Jun. 12, 2021. [Online]. Available: <http://arxiv.org/abs/1407.3561>
- [15] Y. Psaras and D. Dias, "The InterPlanetary File System and the Filecoin Network," in 2020 50th Annual IEEE-IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S), Valencia, Spain, Jun. 2020, pp. 80–80. doi: 10.1109/DSN-S50200.2020.00043.
- [16] "storj.pdf." Accessed: Jul. 12, 2021. [Online]. Available: <https://www.storj.io/storj.pdf>
- [17] "swarm-whitepaper-eng.pdf." Accessed: Dec. 02, 2018. [Online]. Available: <https://docs.swarm.fund/swarm-whitepaper-eng.pdf>
- [18] Building Blockchain Projects. Accessed: Jul. 14, 2021. [Online]. Available: <https://learning.oreilly.com/library/view/building-blockchain-projects/9781787122147/>
- [19] A. Beniiche, "A Study of Blockchain Oracles," ArXiv200407140 Cs, Jul. 2020, Accessed: Jul. 13, 2021. [Online]. Available: <http://arxiv.org/abs/2004.07140>
- [20] D. Mohanty, Ethereum for Architects and Developers: With Case Studies and Code Samples in Solidity. Berkeley, CA: Apress, 2018. doi: 10.1007/978-1-4842-4075-5.
- [21] "Remix - Ethereum IDE." <https://remix.ethereum.org/#optimize=false&runs=200&evmVersion=null> (accessed Jul. 12, 2021).
- [22] "Ganache | Overview | Documentation," Truffle Suite. <https://trufflesuite.com/docs/ganache/overview> (accessed Jul. 14, 2021).
- [23] "Enterprise on Ethereum mainnet," ethereum.org. <https://ethereum.org> (accessed Jul. 14, 2021).
- [24] T. Li et al., "FAPS: A fair, autonomous and privacy-preserving scheme for big data exchange based on oblivious transfer, Ether cheque and smart contracts," Inf. Sci., vol. 544, pp. 469–484, Jan. 2021, doi: 10.1016/j.ins.2020.08.116.

Framework to mitigate supply chain disruptions in the apparel industry during an epidemic outbreak

M. A. S. M. Perera*

Department of Industrial Management
Faculty of Science,
University of Kelaniya, Sri Lanka
sanduniperera172@gmail.com

A. N. Wijayanayake

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

Suren Peter

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

Abstract - Disruptions to a company supply chain, has serious implications, and if not addressed lead to even business closure. The article explores the supply chain risks faced by the apparel industry during an epidemic outbreak and the strategies that could be taken to mitigate them. A systematic review of the literature was initially conducted to identify the supply chain risks and mitigation strategies, and expert interviews were then used to reinforce the findings and then identify the focus areas. Supply chain risks were mapped in a vulnerability matrix with risk association, using a diagrammatic format, and a framework was developed using the supply chain risks and strategies. The developed framework shows that most of the risks can be mitigated by local sourcing and giving incentives to customers. A generalized model was developed based on cost and time considerations but using the same process it can be customized using different factors and risks depending on the experience and needs of the company.

Keywords - epidemic outbreak, mitigation strategies, supply chain disruptions, supply chain risks

I. INTRODUCTION

A Supply Chain (SC) disruption is any sudden change or crisis which negatively impacts the interconnectedness of a network of people, organizations, and activities where the movement of a product from a supplier to a final customer is affected [1]. This effect can be either local or global.

SC disruptions can occur in a company because of legal disputes, strikes, natural disasters and manmade catastrophes. In 2011, the Tsunami in Japan reduced its exports between 0.5% to 1.6% [2]. A brake-fluid proportioning valve supplier caught fire on 1st February 1997 which led Toyota to shut down all its plants and assembly lines and caused a sales loss of 70,000 vehicles ([3] [4]). Moreover, special cases like epidemic outbreaks (Ebola, SARS, MERS, Swine flu, and coronavirus/ COVID-19) also severely disrupts the supply chain [5]. Due to COVID-19, China's industrial production had decreased by 13.5% for the month of January and February 2020, compared to the previous year [6]. More than 75% of U.S. businesses have experienced SC disruption as a result of the COVID-19 outbreak [1] [7] [8]).

The apparel SC aims to provide the right fashion product to satisfy the market needs, with the lowest possible cost, fastest speed and maximized profit [9]. "No-one wants to buy clothes to sit at home in," says Simon Wolfson [10]. Due to the pandemic the fashion industry has been negatively impacted on every imaginable level where production has ceased, retailers have closed and demand has decreased to 34% in March because apparel is not a basic human need [10]. Therefore, the demand for apparels during the pandemic was very low. However, its contribution to the economy is significant. In 2018, the global clothing and apparel market reached a value of \$758.4 billion and has been growing at a compound annual growth rate (CAGR) of 7.5% since 2014

[11]. Moreover, the target for 2022 which was set before the onset of COVID-19 was a CAGR of 11.8% to nearly \$1,182.9 [11]. Furthermore, the Sri Lankan apparel industry which contributes 6% to its country's GDP and 44% to its national export revenue, had set itself a target of \$8 billion export revenue by 2025, prior to the onset of COVID- 19 [12] [13].

The experience faced by the Sri Lankan apparel manufacturing companies was very similar to the global context as most of the apparel manufacturing companies were struggling without raw materials for the upcoming orders. With the spread of the virus over 65 countries, lockdown procedures were implemented, including Sri Lanka where companies went through a temporary shutdown [14]. Because revenue was severely curtailed, companies faced severe cash flow constraints, with companies forced to cut non-essential costs, and even enforcing salary reductions among its staff.

According to [2][3][4][5][6] and [15] SC disruption has negatively affected the world's economy. The study focuses on SC risks, in this challenging scenario of an epidemic outbreak, in order to assess how such SC disruptions could be handled and mitigated. Because of the importance of the apparel industry to the local economy, being the single largest export revenue earner, the scope of the study was restricted to identify SC risks during an epidemic outbreak in the context of the Sri Lankan apparel industry. The study proposes a model to identify the SC risks and vulnerabilities during an epidemic outbreak and the possible mitigation strategies that could be adopted.

II. LITERATURE REVIEW

According to [16] managing SC disruptions revolves around, thoroughly understanding the identified risks, mitigating and then if needed, increasing the capacity of the SC.

Risk can be defined as "uncertainty of outcomes", "probability of lost or lost occurrence", "deviation of outcomes from expectation", "change leading to loss" or "danger of harm loss" [16]. Using the mentioned risk definitions, [16] has identified the following basic risk characteristics; risk is an attitude towards future, rooted in uncertainty, occurred because of lack of information and disadvantage to the company. It means that time, uncertainty, information and loss are key factors. Moreover, [16] [17] have identified single port closure, multiple port closure, transportation link disruption, loss of key supplier, labor unrest, economic recession, visible quality problems, computer virus, workplace violence, flood, wind damage, IT system failure, accounting irregularity, earthquake, employee sabotage, technological change and product tampering as SC risks and developed a vulnerability matrix using disruption probability and consequences (shown in Figure 1). Further,

they have discussed that the SC could be resilient if the company follows a mixed approach of flexibility and redundancy.

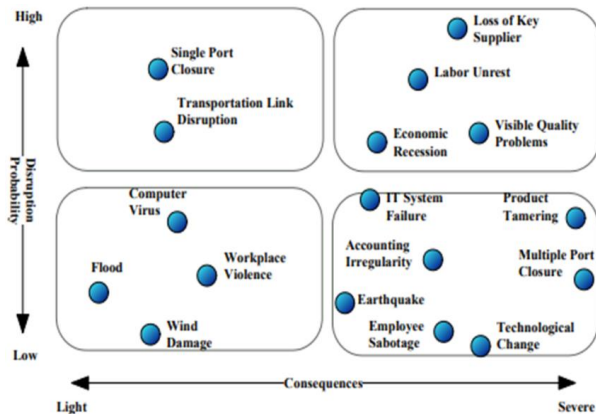


Fig 1. Vulnerability map for a single firm [16] [17]

According to [18], they have discussed, selected risks which are associated with the apparel retail SCs in India by structural analysis of the controllable risks that are identified. The risks they have selected and the background of it are shown in Table I.

TABLE I. RISK ASSUMPTIONS [18]

Risk no.	Risk	Background of risks
R1	Globalization	Currency fluctuations; design transfers, competition; legal and political risk; policy changes; etc.
R2	Raw material and product quality standards	Retailers do not have the complete SOP of the product quality and it varies from season to season/ and product to product
R3	Scarcity of resources	Scarcity of raw material; power shortage; labor shortage; resource cost; the cost of technology etc.
R4	Supplier uncertainty	Failure to deliver on time; supplier bankruptcy; unreliable supplier; Cost and quality not reliable/ consistent; etc.
R5	Lack of co-ordination/ alignment	Lack of communication; no cross-functional teams; no transparency between partners/departments; etc.
R6	Behavioral aspect of employees	Employee disputes; inefficient/ unskilled employee; resistance to change; unavailability of labor due to absence; etc.
R7	Infrastructure risks	Transport breakdown; inadequate means of transport; inconsistent warehouse facility; IT failure; etc.
R8	Delay in schedule/ lead time	Order fulfillment error; change in production schedules; machine breakdown; delay in delivery; change in design; etc.
R9	Demand uncertainty	Error in demand forecast (short term or long term); bullwhip effect; short product life cycle; risk from new entrants; etc.
R10	Customer dissatisfaction	Product returns; customer complaints; reduced demand; stock out; poor quality; wrong product delivery; etc.
R11	Financial risk	High cash conversion cycle; low market share; low-profit margins; decreasing revenues; etc.
R12	Security and safety	Pilferages and shrinkage of the materials in the warehouse/losses in transit, performance of the product, cyber-attack; etc.

Article in [18], has revealed the use of Interpretative Structural Modeling (ISM) to establish the interdependencies

between the risks (Table I), spread across various SC functions where they have classified the risk factors based on their driving and dependence power. They have identified that globalization, labor issues and security and safety of resources as the strong drivers of other SC uncertainties which will lead the company to a financial crisis [18]. The variables they have considered are limited, generic and the costs, frequency of occurrences of disruptions can be used to prioritize risk where strategies can be formulated to mitigate the risks.

According to [19], they have used 45 face to face interviews with open-ended questions to analyse 20 manufacturing firms in Uganda. They have identified, classified the SC risks/threats as Endogenous (supply-side, firm-level, demand-level), Exogenous (geopolitics, economic) using the collected data and it is shown in Table II. They have further analysed to identify the interconnectedness of SC threats, strategies and outcomes.

TABLE II. SUPPLY CHAIN RISKS CLASSIFICATION [19]

Threats	Supply-side	Long-distance sourcing triggered threats, limited local supply market, product counterfeiting, poor-quality raw materials, dishonest suppliers, raw material delays and shortages, financial difficulties of suppliers, supplier delivery failure, reputational risk
	Demand-level	Power asymmetries, dishonest customers/ distributors, payment threat, financial difficulties of customers, order cancellations, demand variations, customer characteristics, reputational risk
	Geo-politic	Political instabilities, geographical location (landlockedness), national politics, government policy, the weak legal system, corruption, product counterfeiting, in-transit raw material theft, communication barriers, natural disasters
Supply Chain Resilience Strategies	Supply management	Backward integration, outsourcing, appropriate supplier selection, alternative transportation, multiple sourcing, supplier development, maintaining strategic stocks, buying instead of making (temporarily), effective contracting, local sourcing, order splitting, enhancing proximity to suppliers, procurement management, quality management, exclusive sourcing, inter-branch stock transfer
	Demand management	Creating customer flexibility, customer incentives, inventory management, product recalls, demand forecasting
	Relationship management	Co-opetition, collaboration with government, collaboration with customers, collaboration with suppliers, Informal networking
	Information management	Risk communication, market intelligence, increasing product knowledge, improving visibility, using information communication technology
Outcomes	To the supply-side	Poor-quality raw materials, limited flexibility to switch suppliers, supplier complacency, raw material delays and shortages
	To demand-side	Distributor complacency, reduced customer base, poor customer delivery performance
	To entire supply chain	Product counterfeiting, reputational risk

Risk assessment and operational approaches to managing risks in global SCs were addressed using a Canadian pet products company operation [20]. The study was based on a compilation of research and interactions with

SC managers in 15 different industries. A framework (Figure 2) was created according to the likelihood of disruptions and its consequences. They have provided scores considering the risk, affected part/product based on the likelihood and impact. It is also stated that attempting to scope one's risk is a challenge based on the supplier information. Moreover, reflecting upon the case study, it should be considered that the risk averseness of the company and the investment they are willing to make to mitigate the risk [20]. However, the study takes an overall view of SC risks and its mitigation strategies and not an in-depth analysis.

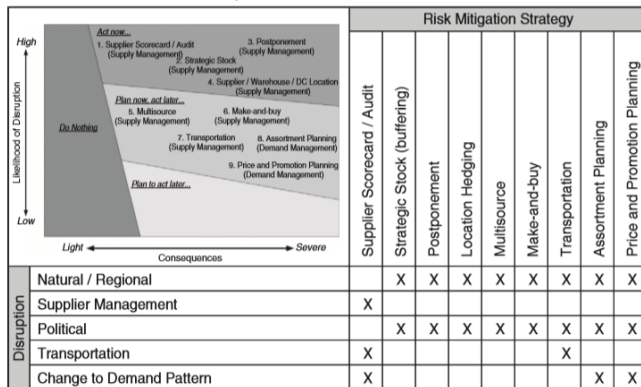


Fig 2. Common disruptions and the strategies that mitigate their impact [20]

Perspectives in SC Risk Management is addressed by reviewing quantitative models that deal with SC risks. A unified framework is developed to classify SC risk management articles. Moreover, SC risk management is approached in two ways; SC Risk (operational risks or disruption risks) and Mitigation Approach (supply management, demand management, product management and information management). The identified strategic and tactical plans to manage SC risks are shown in Table III. It is stated that managing SC risks can be addressed using the manager's attitude towards risks and initiatives for managing SC disruption. Furthermore, robust strategies to mitigate operational and disruption risks are identified. They are robust supply management strategies (multi-supplier strategy from multiple countries, robust demand management strategies (demand management strategies mentioned in Table III), robust product management strategies (postponement strategy) and robust information management strategies (information sharing, vendor managed inventory, collaborative forecasting and replenishment planning to increase SC visibility) [21].

Article in [5], has framed epidemic outbreaks as a unique type of SC disruption risk and used the example of coronavirus (COVID-19), anyLogistix simulation and optimization software to examine and predict the impacts of epidemic outbreaks on the SC performance. Reference [5] [22] have recognized lead-time, risk mitigation inventory and backup suppliers as crucial elements affecting the SC reactions to disruptions. Moreover, geographic location data, lead-time data, and demand data are primarily needed to run the simulation models [5] [23]. A guided framework is needed to develop pandemic plans for a company's SC because epidemic outbreaks create a lot of uncertainty.

TABLE III. STRATEGIC AND TACTICAL PLANS TO MANAGE SUPPLY CHAIN RISKS [21]

	Supply Management	Demand Management	Product Management	Information Management
Strategic Plans	Supply Network Design (Network configuration, Product assignment, Customer assignment, Production planning, Transportation planning)	Product Rollovers Product Pricing	Product Variety	Supply Chain Visibility
Tactical Plans	Supplier relationship Supplier selection process (Criteria, approval/selection) Supplier order allocation (Uncertain demands, supply yields, supply lead times, supply costs) Supply Contract (Uncertain demand-Wholesale price contracts, buy-back contracts, revenue sharing contracts, quantity-based contracts: quality flexibility and minimum order; and Uncertain price)	Shift Demand Across Time Shift Demand Across Markets Shift Demand Across Products (Product substitution and product bundling)	Postponement (Make-To-Order systems without forecast updating, Make-To-Stock systems without forecast updating, Make-To-Stock systems with forecast updating) Process Sequencing.	Information Sharing Vendor Managed Inventory Collaborative Planning, Forecasting & Replenishment

Article in [5], has framed epidemic outbreaks as a unique type of SC disruption risk and used the example of coronavirus (COVID-19), anyLogistix simulation and optimization software to examine and predict the impacts of epidemic outbreaks on the SC performance. Reference [5] [22] have recognized lead-time, risk mitigation inventory and backup suppliers as crucial elements affecting the SC reactions to disruptions. Moreover, geographic location data, lead-time data, and demand data are primarily needed to run the simulation models [5] [23]. A guided framework is needed to develop pandemic plans for a company's SC because epidemic outbreaks create a lot of uncertainty.

Article in [24] have identified and analysed the SC risks using a vulnerability matrix. Similarly, article in [25] [26] have used vulnerability matrix and correlation analysis to identify and analyse the SC risks during an epidemic outbreak.

III. METHODOLOGY

Prioritization of risk is essential as the risk factors may act as drivers to other risk factors. Therefore, managers should initially focus on the few (major) risks which act as drivers to other risks. The main purpose of this paper is to identify risk and vulnerability to analyse the costs and time associated with the SC risks and identify the mitigation strategies. It is important to control these risks since it might lead companies to go through a temporary shutdown during an epidemic outbreak. The steps of the research methodology could be explained in the following flow diagram.

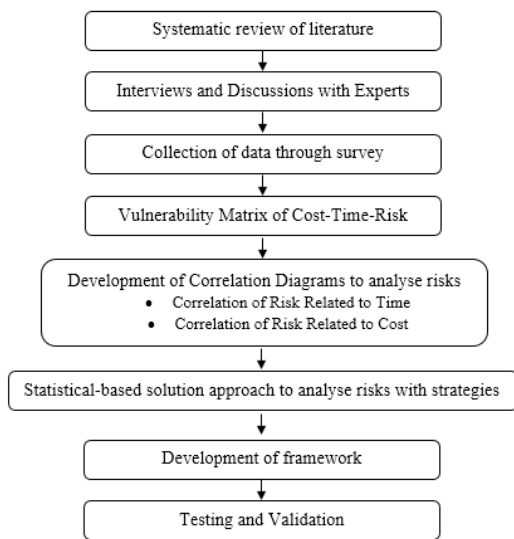


Fig 3. Flow diagram of the methodology

According to [16] [17] [18] [19] SC risks were identified through the literature. Moreover, to further identify the related SC risks, SC managers who have more than five years of experience in the apparel industry were interviewed. A five scale Likert scale was used to collect data (1 - strongly disagree, 2 - disagree, 3 - neither agree nor disagree, 4 - agree, 5 - strongly agree). The experts were drawn from companies whose clients were international and suppliers were both local and international.

There are 300-350 apparel manufacturing plants in Sri Lanka [13]. However, there are less than 20 companies which are competing internationally. Information was collected from 8 leading apparel manufacturing companies which cover almost 75% market share of the apparel industry in Sri Lanka. Five participants from each of the companies were selected. They were of executive grade or higher, with more than 5 years' experience and were selected using random sampling.

As of risk definitions and characteristics stated by [16], the study selected "risk is an attitude towards future event", "disadvantage to the company" as the characteristics to categorize the risks because most of the risk related matrixes, models, frameworks were developed using likelihood of the risk / disruption / threat and its consequences [16] [17] [20]. However, the study focus is to prioritize these risks in order to identify which risks should be addressed first and mitigate them. Therefore, risks were categorized using time and economical loss factors. Time factor is taken as the time taken to address the risk and, economical loss factor as the cost occurred to the company when the risks were not handled. The more time it takes to address or control the disruption, it is categorized into high risk category. Similarly, the higher the economic loss or the cost to bear the risk, also falls into the higher risk category. The identified SC risks through the literature review and experts' opinion were,

- (R1) - Loss of local key supplier [16] [18] [19] [24] [25] [26]
- (R2) - Loss of international key supplier [16] [18] [19] [24] [25] [26]

- (R3) - Local port closure [16] [24] [25] [26]
- (R4) - International port closure [16] [24] [25] [26]
- (R5) - Transportation link disruption- other than ports [16] [18] [19] [24] [25] [26]
- (R6) - Raw materials delays and shortages [18] [19] [24] [25] [26]
- (R7) - Human Resource shortages [18] [24] [25] [26]
- (R8) - Product demand variations [18] [19] [24] [25] [26]
- (R9) - Order cancellations ([18]; [19]; [24]; [25] ; [26])
- (R10) - Lead time variations [5] [18] [24] [25] [26]

The identified mitigation strategies were,

- (S1) - Backward Integration [14] [19]
- (S2) - Outsourcing [14] [19] [20] [21]
- (S3) - Local Sourcing [14] [19] [20] [21]
- (S4) - International Sourcing [14] [19] [20] [21]
- (S5) -Strategic Stock [14] [19] [20]
- (S6) - Sharing Information [14] [19] [21]
- (S7) - Supply Chain Visibility [14] [19] [21]
- (S8) - Alternative Transportation [14] [19] [21]
- (S9) - Customer Incentives [19] [20]
- (S10) - Product Differentiation [21]
- (S11) - Health Safety [14]

A risk assessment was conducted and identified the positions of each risk under time and cost category. Based on the experts' opinion Cost-Time-Risk (CTR) matrix was developed. Next, a correlation analysis was conducted to identify the association between each risk and the mitigation strategies. This enable the decision makers to identify the best mitigation strategies that is applicable or could be applied to control or mitigate the risks. Based on the evidence an empirical model was developed. The study used 80% of the data to develop the model and 20% of the data for testing and validation. Moreover, experts' opinions were taken regarding the output of this study.

IV. RESULTS AND DISCUSSION

A. Vulnerability matrix of cost-time-risk

The main two questions which were asked to identify the position of the risk in the vulnerability matrix were the time taken to mitigate the risk and the cost incurred when the risks were not handled.

The scores shown under time and cost in Table IV, are the average score taken from the survey. Higher the time taken to mitigate the SC risk, higher the risk. Likewise, higher the cost occurred to the company when the SC risks are not handled, higher the risk.

According to the data collected from the experts through the survey, the SC risks were mapped in a vulnerability matrix and shown in Fig. 4. Fig. 4 was drawn from time and cost scores which were collected from the survey and shown in Table IV.

TABLE IV. RESULTS OF COLLECTED DATA FROM THE SURVEY

Supply Chain Risks	Time	Cost
(R1)- Loss of local key supplier	3.5	3
(R2)- Loss of international key supplier	4.5	4
(R3)- Local port closure	5	2
(R4)- International port closure	4.5	2
(R5)- Transportation link disruption- other than ports	3	2
(R6)- Raw materials delays and shortages	3.5	3
(R7)- Human Resource shortages	2	2
(R8)- Product demand variations	3.5	3.5
(R9)- Order cancellations	4.5	4
(R10)- Lead time variations	3.5	2.5

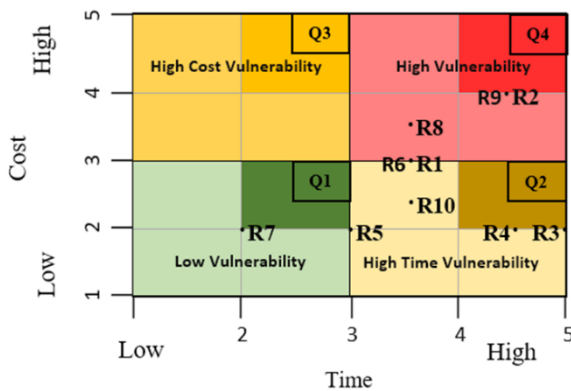


Fig 4. Risk assessment matrix on Cost and Time (CTR Vulnerability Matrix)

If the cost is high, then the risk is high, as the risk incur a cost to the company which might lead the company to go through a temporary shutdown if it's not handled or mitigated properly. If the time is high, it means that the risk is taking more time to handle or mitigate, therefore, the risk is also high which falls to Quadrant 2 (Q2). It is beneficial to focus on high vulnerability risks where the cost and time are both high, which means that the risk is very high compared to the other quadrants as shown in Quadrant 4 (Q4). A generalized vulnerability model is developed in this study considering cost and time factors, however, it can be customized using different factors and risks depending on the experience and needs of the company.

The weight for cost and time is measured on the same scale of Likert scale 1 to 5. According to the vulnerability matrix shown in Figure 4, loss of international key supplier (R2) and order cancellations (R9) are towards the right side in the matrix which means that the risk is high. However, human resource shortages (R7) is towards the left side in the matrix which means the risk related to it is low compared to the other SC risks [25] [26]. It is because human resource shortage can be solved internally, quickly compared to the other risks, whereas, in the loss of international key suppliers, order cancellations are decided by external parties and cannot be handled internally as it takes time and resources to solve the issue.

Loss of international key supplier can be mitigated by having several suppliers from different regions. It may be

costly, however, in order to mitigate the risk, you should at least have a minimum order from these suppliers. Order cancellation can be mitigated by having several customers and a variety of products. Moreover, during the epidemic outbreak, manufacturers should switch to products such as personal protective equipment, face masks, and similar alternate products which can be manufactured with the same resources.

As the vulnerability matrix only indicates the time and cost but doesn't indicate the association of each risks, a statistical approach of correlation analysis was used to analyse the data.

B. Development of correlation diagram to analyse risks

Using the data gathered from the survey, the identified risks were analysed using bivariate correlation to measure the strength of the relationship between each pair. Only 35% of the data follows a normal distribution, therefore, spearman's rho was used to calculate the correlation for each category. The study considered value which is greater than or equal to 0.7 as highly correlated. The results of the survey analysis under time category is shown in Table V.

C. Correlation of risk related to time

TABLE V. CORRELATION RESULTS UNDER TIME CATEGORY

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	1									
R2	0.839	1								
R3	0.648	0.797	1							
R4	0.723	0.813	0.97	1						
R5	0.686	0.847	0.725	0.719	1					
R6	0.487	0.786	0.684	0.635	0.821	1				
R7	0.555	0.775	0.614	0.579	0.873	0.974	1			
R8	0.547	0.832	0.923	0.888	0.849	0.861	0.798	1		
R9	0.638	0.845	0.974	0.965	0.817	0.738	0.668	0.961	1	
R10	0.612	0.856	0.852	0.877	0.809	0.889	0.831	0.944	0.905	1

Correlation values which are greater than or equal to 0.7 are highlighted in light grey and considered as highly correlated risks, ignoring the correlation between the same risk. Using the relationship shown in Table V, a hierarchical diagram was developed to understand the relationship between each risks and it's shown in Fig. 5. The hierarchical diagram was developed considering the number of highly correlated risks.

The dotted box represents correlated risks. Hence, any relation between another and the dotted box represents an inclusive relationship of all risks within the dotted box. According to Fig. 5, R2 (Loss of international key supplier) is highly correlated to all the other risks which means that there is a high probability of occurrence of other risks due to R2 [25].

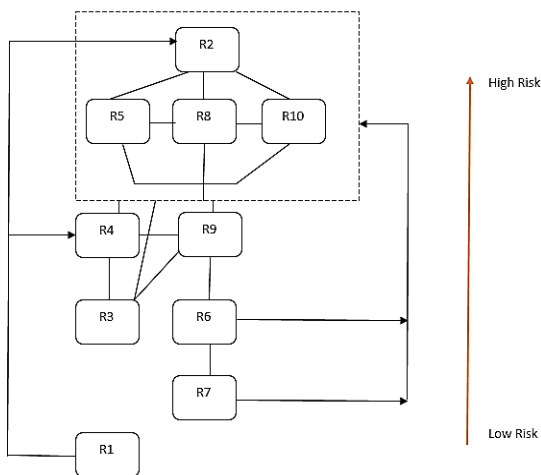


Fig 5. Hierarchical diagram under time category

Therefore, companies should focus primarily to mitigate on losing international key suppliers. Further, companies should focus on R5 (Transportation link disruption), R8 (Product demand variations) and R10 (Lead time variations) as these risks are secondly highly correlated to the rest of the risks [25].

1) Correlation of risk related to cost

The results of the survey analysis under cost category is shown in Table VI.

TABLE VI. CORRELATION RESULTS UNDER COST CATEGORY

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	1									
R2	0.71 5	1								
R3	0.93 1	0.64 5	1							
R4	0.6	0.47 6	0.79 3	1						
R5	0.79 7	0.66 1	0.78 1	0.74 7	1					
R6	0.82 8	0.40 6	0.82 1	0.73 6	0.81 1	1				
R7	0.63	0.77 2	0.67 9	0.53 1	0.62 2	0.45 8	1			
R8	0.55 8	0.53 8	0.40 9	0.06	0.13 5	0.24	0.31 3	1		
R9	0.62 5	0.61	0.66 8	0.57 4	0.31 9	0.55	0.61 3	0.62 8	1	
R10	0.58 6	0.83 9	0.53 7	0.39 8	0.32 4	0.32 5	0.52	0.71 2	0.81 6	1

Correlation values which are greater than or equal to 0.7 are shaded in light grey and considered as highly correlated risks. Using the relationship shown in Table VI, a diagram was developed to understand the co-relationship between each risks and shown in Figure 6. The hierarchical diagram was developed considering the number of highly correlated risks.

According to Figure 6, R1 (Loss of local key supplier), R3 (Local port closure), R5 (Transportation link disruption) and R6 (Raw materials delays and shortages) are highly correlated to other risks, which means that there is a high probability of occurrence of other risks due to R1, R3, R5 and R6 [25]. Therefore, companies should focus primarily to mitigate on losing local key suppliers, local port closure,

transportation link disruption and raw materials delays and shortages when considering cost.

It can be seen that R5 (Transportation link disruption) is highly correlated to rest of the risks when you consider both categories. Therefore, it is better to mitigate transportation link disruption first and then the rest of the risks under each category.

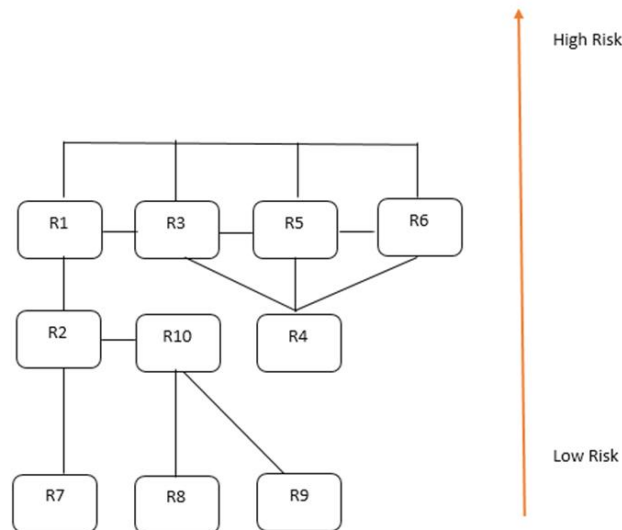


Fig 6. Hierarchical diagram under cost category

Based on the discussion with experts, the study identified that it takes more time to mitigate loss of international key supplier (R2) and it is highly correlated to the rest of risks because international key suppliers are the main source of income to the company. Therefore, losing them will cause a chain reaction. Customers may not like the alternative supplier, quality issues, and it takes time to find alternative suppliers. Therefore, lead time will increase, raw materials to produce the product will be insufficient which will lead to order cancellations or delay in fulfilling orders.

Considering loss of local key supplier (R1) under cost category, it was identified that it is costly because losing local key supplier will lead to find alternative suppliers and there will be shipping cost, lead time to deliver the raw materials will be high which is costly to the company. Moreover, local port closure (R3) will lead to sourcing other means of transportation for raw materials into the country and products out of the country. This will be costly because you may be currently using the optimum method of transportation resulting in, shortage and delay of raw materials which will lead to delayed orders.

At the end of every production we should deliver the products on time to gain the benefit from it. Therefore, transportation link disruption is a crucial risk to be mitigated.

D. Statistical-based solution approach to analyse risks with strategies

Using the data from the survey, a correlation analysis was conducted to identify the association between each risks with strategies in order to mitigate the risks. Only 35% of the data follows a normal distribution. Therefore, spearman's rho was used to calculate the correlation. Value which is greater than or equal to 0.4 and less than 0.7 was considered as moderately correlated and highlighted in yellow. In this

research, only positive correlated values are considered as experts assumed that they can mitigate the risks by each highly or moderately positive correlated strategies.

The results of the analysis under time category is shown in Table VII.

TABLE VII. CORRELATION RESULTS FOR RISK AND STRATEGY UNDER TIME CATEGORY

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
R1	0.12	0.44	0.40	0.12	0.08	0.10	0.09	0.02	0.07	0.08	0.17
R2	0.31	0.02	0.22	0.47	0.21	0.22	0.15	0.28	0.21	0.12	0.20
R3	0.16	0.03	0.01	0.47	0.60	0.29	0.07	0.32	0.00	0.17	0.41
R4	0.28	0.23	0.45	0.40	0.02	0.07	0.04	0.03	0.11	0.03	0.36
R5	0.07	0.59	0.23	0.02	0.02	0.14	0.13	0.02	0.10	0.04	0.31
R6	0.14	0.55	0.18	0.26	0.11	0.34	0.14	0.05	0.13	0.27	0.09
R7	0.24	0.58	0.64	0.52	0.32	0.32	0.47	0.66	0.67	0.68	0.56
R8	0.22	0.08	0.45	0.45	0.25	0.03	0.12	0.26	0.06	0.26	0.27
R9	0.21	0.19	0.17	0.17	0.17	0.60	0.54	0.06	0.59	0.47	0.00
R10	0.21	0.30	0.06	0.06	0.06	0.23	0.28	0.17	0.16	0.15	0.09

It can be seen that human resource shortages (R7) can be mitigated using many strategies. Whereas, loss of local key supplier (R1), loss of international key supplier (R2) and international port closure (R4) can be mitigated by only implementing one strategy from the considered strategies. Moreover, local port closure (R3), product demand variation (R8) and lead time variations (R10) cannot be mitigated by

Risk/ Strategies	(S1)- Backward Integration	(S2)- Outsourcing	(S3)- Local Sourcing	(S4)- International Sourcing	(S5)- Strategic Stock	(S6)- Sharing Information	(S7)- Supply Chain Viability	(S8)- Alternative Transportation	(S9)- Customer Incentives	(S10)- Product Differentiation	(S11)- Health Safety
(R1)- Loss of local key supplier			✓								
(R2)- Loss of international key supplier				✓							
(R3)- Local port closure											
(R4)- International port closure			✓								
(R5)- Transportation link disruption- other than ports		✓									
(R6)- Raw materials delays and shortages		✓									
(R7)- Human Resource shortages		✓	✓	✓		✓	✓	✓	✓	✓	✓
(R8)- Product demand variations						✓	✓		✓	✓	
(R9)- Order cancellations								✓	✓		
(R10)- Lead time variations										✓	

implementing any strategies under time category. Using the correlation analysis for risk and strategies a framework was developed.

TABLE VIII. FRAMEWORK TO MITIGATE SUPY CHAIN DISRUPTIONS - TIME

The results of the analysis under cost category is shown in Table IX. It can be seen that transportation link disruption (R5) and human resource shortages (R7) can be mitigated using many strategies. Whereas, international port closure (R4), product demand variation (R8) and lead time variations (R10) can be mitigated by only implementing one strategy we considered. Moreover, local port closure (R3) and order

cancellations (R9) cannot be mitigated by implementing any strategies under cost category.

TABLE IX. CORRELATION RESULTS FOR RISK AND STRATEGY UNDER COST CATEGORY

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
R1	0.39	0.50	0.39	0.12	0.28	0.41	0.47	0.29	0.30	0.11	0.35
R2	0.57	0.02	0.16	0.41	0.09	0.09	0.06	0.12	0.14	0.13	0.25
R3	0.32	0.35	0.07	0.21	0.30	0.12	0.25	0.31	0.12	0.37	0.00
R4	0.22	0.09	0.62	0.21	0.28	0.21	0.18	0.23	0.23	0.40	0.20
R5	0.31	0.23	0.27	0.35	0.31	0.47	0.45	0.27	0.47	0.29	0.53
R6	0.22	0.00	0.41	0.54	0.39	0.32	0.20	0.19	0.14	0.21	0.15
R7	0.33	0.49	0.46	0.51	0.47	0.18	0.26	0.44	0.42	0.46	0.65
R8	0.10	0.27	0.08	0.00	0.03	0.37	0.33	0.10	0.41	0.01	0.04
R9	0.29	0.27	0.24	0.24	0.24	0.22	0.32	0.35	0.23	0.38	0.28
R10	0.08	0.27	0.16	0.19	0.20	0.09	0.01	0.11	0.51	0.16	0.02

It can be observed that international port closure (R4) can be mitigated using the same strategy without considering the category. According to the framework, it can be seen that most of the risks can be mitigated by local sourcing (S3) and giving incentives to customer (S9). Therefore, by implementing these strategies company can save time and cost.

TABLE X. FRAMEWORK TO MITIGATE SUPPLY CHIAN DISRUPTIONS - COST

Risk/ Strategies	(S1)- Backward Integration	(S2)- Outsourcing	(S3)- Local Sourcing	(S4)- International Sourcing	(S5)- Strategic Stock	(S6)- Sharing Information	(S7)- Supply Chain Viability	(S8)- Alternative Transportation	(S9)- Customer Incentives	(S10)- Product Differentiation	(S11)- Health Safety
(R1)- Loss of local key supplier						✓	✓				
(R2)- Loss of international key supplier	✓			✓							
(R3)- Local port closure											
(R4)- International port closure			✓								
(R5)- Transportation link disruption- other than ports		✓					✓		✓		✓
(R6)- Raw materials delays and shortages			✓	✓							
(R7)- Human Resource shortages		✓	✓	✓	✓				✓	✓	✓
(R8)- Product demand variations									✓		
(R9)- Order cancellations											
(R10)- Lead time variations									✓		

Developed framework and resulting diagrams were validated through data collected from survey resulting in anticipated actual results. Hence, proving the accuracy of the model developed.

V. CONCLUSION

It is difficult to anticipate the arrival of an extreme disruption to the SC, like an epidemic outbreak. However, companies can identify SC risks and be prepared for it now rather than reacting to it, when it occurs. In this paper, an empirical investigation was conducted to assess SC risks, under time and cost categorization. The results provide several insights for theory and practice. It is recommended to focus on the high vulnerability quadrant in the vulnerability matrix (Figure 4) as its risk is high compared to other quadrants. If it's not mitigated the business might have to temporarily shut down due to the disruption caused.

The study contributes to identify SC risks during major disruptions to SC. The research also contributes to organizational theory by building a matrix to prioritize the SC risks they face during an epidemic outbreak in order to focus and mitigate them. Loss of international key supplier (R2) and order cancellations (R9) are considered as high risk based on the vulnerability matrix. However, human resource shortages (R7) is considered as low risk, based on the vulnerability matrix.

The vulnerability matrix doesn't indicate the association of each risks but it shows the time and cost for each risks. Therefore, considering the correlation analysis, it is recommended to focus on the highly correlated risks under time and cost category as its risk is high compared to others. If the risk is not mitigated, the business might even have to temporarily shut down due to the disruption caused. Considering the time category, the study identified that the loss of international key supplier is highly correlated to all the other risks which means that there is a high probability of occurrence of other risks and companies should focus primarily to mitigate it. Further, companies should focus on transportation link disruption, product demand variations and lead time variations as these risks are also highly correlated to the rest of the risks.

Moreover, considering the cost category, the study identified that loss of local key suppliers, local port closure, transportation link disruption and raw materials delays and shortages are highly correlated to other risks which means that there is a high probability of occurrence of other risks and companies should focus primarily to mitigate them.

When considering association of risks with strategies, it can be seen that international port closure (R4) can be mitigated using the same strategy without considering the category. According to the framework, it can be seen that most of the risks can be mitigated by local sourcing (S3) and giving incentives to customer (S9). Therefore, by implementing these strategies company can save time and cost. The summary findings of the study in Table XI.

It could be observed that some of the past implemented strategies for identified risks were same as [19] [21] and [20] studies and some were not. According to [19], loss of local key supplier (R1), loss of international key supplier (R2), raw materials delays and shortages (R6) has got more strategies than the strategies found in this study. It is because [19] have considered the risks in a combined and wide range, whereas this study considered the risks separately. Moreover, [19] have considered a day-to-day SC risk, whereas, the study considered a special case, of an epidemic outbreak. Therefore, it can be concluded that Sharing Information (S6), SC Visibility (S7) strategies are vital when considering an epidemic outbreak.

Strategies in [20] and strategies in the conducted study in this article are almost different because [20] has only considered

nine strategies for their study and the risks and disruption as a combined and wide range where this study considered them separately.

[21] has also considered the risks in a combined and wide range and the strategies were limited. Considering Product demand variations (R8), it can be seen that in a normal SC disruption, Product differentiation (S10) could be taken as a mitigation strategy. However, considering a special case such as epidemic outbreak Customer incentives (S9) are crucial to mitigate the risk.

TABLE XI. FINDINGS OF THE STUDY

Risk	Strategies found in this study		Strategies found in past literature		
	Under Time Category	Under Cost Category	[19]	[20]	[21]
(R1)- Loss of local key supplier	S3	S6, S7	S1, S2, S3, S4, S5, S8		S2, S3, S4
(R2)- Loss of international key supplier	S4	S1, S4	S1, S2, S3, S4, S5, S8		S2, S3, S4
(R3)- Local port closure				S8	
(R4)- International port closure	S3	S3		S8	
(R5)- Transportation link disruption- other than ports	S2	S6, S7, S9, S11		S8	
(R6)- Raw materials delays and shortages	S2	S3, S4	S1, S2, S3, S4, S5, S8		
(R7)- Human Resource shortages	S2, S3, S4, S7, S8, S9, S10, S11	S2, S3, S4, S5, S8, S9, S10, S11			
(R8)- Product demand variations		S9	S5, S9	S9	S10
(R9)- Order cancellations	S6, S7, S9, S10		S5, S9	S9	
(R10)- Lead time variations		S9			

In conclusion, considering time and cost only Loss of international key supplier (R2) and order cancellations (R9) are crucial to mitigate. However, considering the risk association to each other, under time category, Loss of international key supplier is crucial to mitigate. Moreover, under cost category, Loss of local key supplier, Local port closure, Transportation link disruption and Raw materials delays and shortages are crucial to mitigate. Further, considering the association of risks with strategies, it can be said that most of the risks can be mitigated by local sourcing and giving incentives to the customer.

The limitation of this study was that an assumption was made that the clients were international and suppliers were both local and international. This would somewhat restrict external validity.

As for future work, the study can be extended to identify the root causes of these risks which should be taken in order to mitigate the SC disruptions. These outcomes of the research allow managers to evaluate the course of action that

they should take concerning the SC disruption that they experience during an epidemic outbreak.

REFERENCES

- [1] Meyer, S., 7 “Steps For Minimizing Supply Chain Disruptions + Prevention Tips”. (online) The BigCommerce Blog. Available at: <https://www.bigcommerce.com/blog/supply-chain-disruptions/#what-is-a-supply-chain-disruption>, 2020
- [2] Escaith, H., Teh, R., Keck, A., & Nee, C. “Japan’s earthquake and tsunami: Global supply chain impacts”, | VOX, CEPR Policy Portal. Retrieved 26 July 2020, from <https://voxeu.org/article/japans-earthquake-and-tsunami-global-supply-chain-impacts>, 2011
- [3] Ziaul, H., Muhammad, A., & Barbara, D. “Impact of Man-Made Disasters on Commercial Logistics”. International Journal Of Economics, Commerce and Management, United Kingdom, III(6), 2015
- [4] Nishiguchi, T., & Beaudet, A., “Self-Organization and Clustered Control in the Toyota Group: Lessons from the Asian Fire”, Massachusetts Institute of Technology International Motor Vehicle Program, 2002
- [5] Ivanov, D., “Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case”, Transportation Research Part E, pp. 136, 2020
- [6] Seric, A., Görg, H., Möslé, S., & Windisch, M. Managing COVID-19: “How the pandemic disrupts global value chains Industrial Analytics Platform”, <https://iap.unido.org/articles/managing-covid-19-how-pandemic-disrupts-global-value-chains>, 2020
- [7] Leonard, M., “44% of supply chain pros have no plan for China supply disruption”. Available: <https://www.supplychaindive.com/news/44-of-supply-chain-pros-have-no-plan-for-china-supply-disruption/573899/>, March 11, 2020
- [8] Hobbs, B., “How to Prepare for Major Supply Chain Disruption”. Available: <https://www.entrepreneur.com/article/348081>, March 31, 2020
- [9] Hui, P. and Choi, T., “Using artificial neural networks to improve decision making in apparel supply chain systems. Information Systems for the Fashion and Apparel Industry”, pp.97-107, 2016
- [10] McIntosh, S., “Coronavirus: Why The Fashion Industry Faces An Existential Crisis”. (online) BBC News. Available at <https://www.bbc.com/news/entertainment-arts-52394504>, 2020.
- [11] Businesswire, “Global \$1,182.9 Billion Clothing And Apparel Market Analysis, Opportunities And Strategies To 2022”, Researchandmarkets.Com. (online) Available at: <https://www.businesswire.com/news/home/20191025005178/en/Global-1182.9-Billion-Clothing-Apparel-Market-Analysis>, 2020
- [12] BOI, Apparel – BOI Sri Lanka. Available: <http://investsri Lanka.com/sectors/apparel-2/>, 2020
- [13] Export Development Board (EDB), Sri Lanka, Industry Capability Report Sri Lankan Apparel Sector, 2020
- [14] Kilpatrick, J. and Barter, L., COVID-19: Managing Supply Chain Risk and Disruption. Canada: Deloitte, 2020
- [15] Hippold, S., Coronavirus: How To Secure Your Supply Chain. (online) Gartner. Available at: <https://www.gartner.com/smarterwithgartner/coronavirus-how-to-secure-your-supply-chain/>, 2020
- [16] Xu, J., “Managing the Risk of Supply Chain Disruption: Towards a Resilient Approach of Supply Chain Management”. ISECS International Colloquium on Computing, Communication, Control, and Management, 2008
- [17] Sheffi, Y., Rice Jr. J. B., “A Supply Chain View of the Resilient Enterprise”, MIT Sloan Management Review, pp.41-48, 2005
- [18] Venkatesh, V.G., Rath, S., and Patwa, S., “Analysis on supply chain risks in Indian apparel retail chains and proposal of risk prioritization model using Interpretive structural modeling”, Journal of Retailing and Consumer Services 26, pp. 153–167, 2005
- [19] Tukamuhabwa Benjamin, Stevenson Mark, Busby Jerry, “Supply chain resilience in a developing country context: a case study on the interconnectedness of threats, strategies and outcomes”, Supply Chain Management: An International Journal, Vol. 22, No. 6, pp. 486–505, 2017
- [20] Kumar Sameer, Himes Katie J. and Kritze Collin P., “Risk assessment and operational approaches to managing risk in global supply chains”, Journal of Manufacturing Technology Management, Vol. 25, No. 6, pp. 873-890, 2014
- [21] Tang Christopher S., “Perspectives in Supply Chain Risk Management: A Review”, 2005
- [22] Ivanov, D., Dolgui, A., “Low-Certainty-Need (LCN) supply chains: A new perspective in managing disruption risks and resilience”. Int. J. Prod. Res. 57 (15–16), pp. 5119–5136, 2019
- [23] Dolgui, A., Ivanov, D., Rozhkov, M., “Does the ripple effect influence the bullwhip effect? An integrated analysis of structural and operational dynamics in the supply chain. Int. J. Prod. Res. 58 (5), pp. 1285–1301, 2020
- [24] Perera, M. A. S. M., Wijeyanayake, A., Peter, S., “Classifying risk and vulnerability in the supply chain during an epidemic outbreak”, International Conference on Applied and Pure Sciences, 2020 Faculty of Science, University of Kelaniya, Sri Lanka, pp 111, 2020
- [25] Perera, M. A. S. M., Wijeyanayake, A., Peter, S., “Analysis of Correlation of Risks in the Supply Chain Disruption in Apparel Industry during epidemic Outbreak, COVID 19: Impact, Mitigation, Opportunities and Building Resilience “From Adversity to Serendipity”, Perspectives of global relevance based on research, experience and successes in combating COVID-19 in Sri Lanka, Vol. 1, National Science Foundation, Sri Lanka: ISBN 978-624-5896-00-4, pp 672-677, 2021
- [26] Perera, S., Wijeyanayake, A., Peter, S., “Analysing the risk in the supply chain of apparel industry during an epidemic outbreak”, Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore, pp 864-874, 2021.

Solution approaches for combining first-mile pickup and last-mile delivery in an e-commerce logistic network: A systematic literature review

M. I. D. Ranathunga*

Department of Industrial Management
University of Kelaniya, Sri Lanka
isharadil26@gmail.com

A. N. Wijayanayake

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

D. H. H. Niwunhella

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

Abstract - Logistics is one of the primary areas of operation within cutting-edge supply chain operations. In the e-commerce supply chain also logistics operations play a vital part. The logistics operations must be controlled effectively and efficiently since they deal with the high-cost besides environmental impacts. In e-commerce logistics operations, first-mile and last-mile delivery operations are considered as the operations with the highest costs incurred. So, e-commerce service providers are interested in optimizing their first-mile and last-mile delivery operations. Though it is known that the integration of first-mile pickup and last-mile deliveries will minimize the cost of transportation, there are more practical concerns to be taken into account when combining the first-mile pickup and last-mile delivery operations. Capacitated Vehicle Routing Problem (CVRP) is discussed in the literature as a solution approach for this kind of problems. The objective of this study is to provide a comprehensive overview of the current CVRP related literature, including models, algorithmic solution approaches, objectives, and industrial applications, with a focus of identifying interesting study paths for the future to improve distribution in e-commerce logistics networks by combining first-mile pickup and last-mile delivery operations. The findings of the study have demonstrated that constraints and features of Vehicle Routing Problem with Backhauls are very attractive with today's e-commerce operations, and the majority of the cited publications employed approximation methods rather than precise algorithms to solve these types of models.

Keywords - capacitated vehicle routing problem, e-commerce, first-mile and last-mile delivery

I. INTRODUCTION

E-commerce or electronic commerce is the activity of purchasing and selling things over the Internet or through online services. Global e-commerce sales are expected to reach \$6.5 trillion by the end of 2023 [1]. Due to the uninterrupted growth rate, e-commerce can be considered as one of the fastest-growing industries currently. The E-commerce supply chain incorporates supply chain operations including product warehousing, inventory management, delivery and order management. For e-commerce to be succeeded, it must be efficient at all levels of business. Therefore, optimizing each of these components is essential to ensure that everything is working smoothly and efficiently. Since e-commerce delivery operations incurred a substantial amount of the total cost of operations in an e-commerce supply chain, it is in their best interest to optimize these delivery operations costs which will ultimately benefit the e-commerce service providers and their customers.

In the logistics supply chain of an e-commerce enterprise, first-mile delivery is the initial stage of transportation. This is where the package leaves the merchant's door for the first time. Merchants could drop off their goods at the collection

points or the drop-off stations, or request that the company's logistic service providers to fetch their products from where they are stored. The difference between pick-ups and drop-offs is that the pick-ups are carried out by the logistic service providers so that merchant can have their products picked up at their warehouses or storefronts. Merchants transporting their products to collection points or drop-off stations is known as drop-offs. Therefore, the process of collecting goods from merchants using logistic service providers is known as the first-mile pickup. The phrase "last-mile" was first used in the telecommunications sector to describe the final leg of a network [2]. The movement of packages from the transportation hub to the ultimate delivery destination is known as last-mile delivery in an e-commerce supply chain. This last-mile logistics in any of the supply chains is often considered as the most expensive, least efficient, and with the most pressing environmental concerns [3]. As a result of the rapid expansion of the e-commerce industry and the increase of online purchases, the volume of first-mile pickups and last-mile deliveries increased and puts barriers to the transportation networks with the increased volume of vehicles on roads.

The introduction of new business strategies has been a significant driver of total cost reduction in most recent business organizations. Whether driven by minimizing costs or by modern trade methodologies, reconsidering around distribution network optimization has presently gotten to be more pertinent than ever. One such way of optimizing distribution networks is shipment consolidation, which has been a popular research area over the past few years. Shipment Consolidation is a coordination methodology that combines two or more orders or shipments. It may empower significant economies of scale, incredibly decreasing the transportation cost and fewer environmental impacts. According to [4] combination of deliveries from a depot and pickups destined to the same depot on the same vehicle is considered a specific case of consolidation. This combined pickup and delivery can also lead to significant efficiency gains. According to the findings of [5] combining first-mile pickup and last-mile delivery operations can result in efficiency benefits of up to 30% for e-commerce delivery operations. In practice, combined deliveries and pickups on the same vehicle are appealing owing to the long-term environmental benefits of fewer vehicles on the road and lower emissions.

The Capacitated Vehicle Routing Problem (CVRP) is one of the most important combinatorial optimization problems which recently has been receiving much attention from researchers and scientists [6]. The objective of CVRP is to serve a set of delivery customers or a set of pickup customers

through a set of vehicles stationed at a central depot without violating the capacity of vehicles. CVRP has several variants and extensions. Vehicle Routing Problem with Pickup and Deliveries (VRPPD) where the mixed loading of pickups and deliveries are considered, Vehicle Routing Problem with Simultaneous Pickup and Deliveries (VRPSPD) where vehicle's load in any given route is a mix of pickup and delivery loads, Vehicle Routing Problem Backhauls (VRPB) where customers with delivery demands should meet first before pickup demands are examples of such variations. Several solution approaches such as heuristic algorithms, metaheuristic algorithms, genetic algorithms, and exact methods have been developed to solve CVRP and its variants.

This study was carried out to examine the solution techniques utilized in the literature linked to CVRP. It aids future research in building a model to improve distribution in e-commerce logistics networks by combining first-mile pickup and last-mile deliveries together. The techniques utilized to achieve solutions in the literature, such as exact optimization, heuristic, metaheuristic, or genetic algorithms, have been explored in this study. This study's findings will aid future research in selecting an appropriate model and to improve distribution in an e-commerce logistic network by combining first-mile pickup and last-mile deliveries together.

The remaining sections of this are organized as follows; Section 2 describes the methodology employed in the study, which is followed by the findings of the literature review in section 3. Section 4 includes the overview with the analysis of the solution approaches and finally the conclusion is presented in section 5.

II. METHODOLOGY

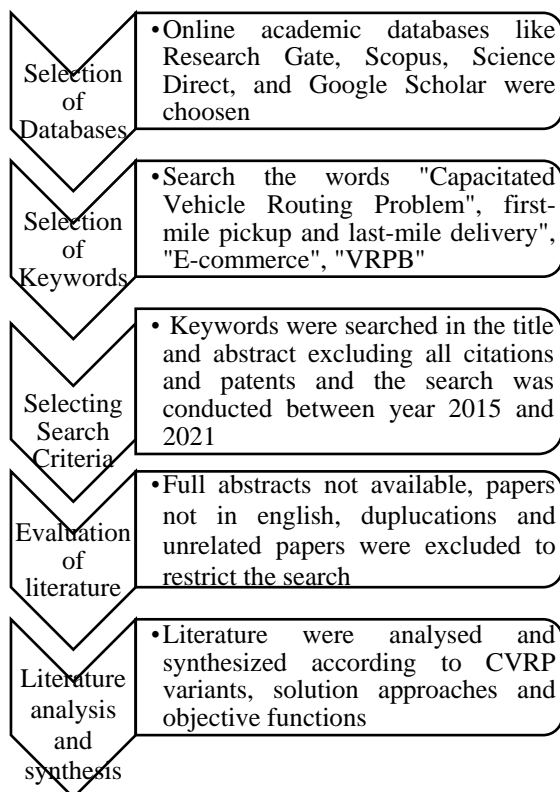


Fig. 1 Methodology of the literature review

The systematic review of the literature was based on the content analysis of the main domain areas including first-mile pickup and last-mile delivery operations, combined pickup and delivery operations, and CVRP. The publications were reviewed in the following steps for the literature search and analysis process: (a) choose the database source; (b) choose the search terms; and (c) choose the search criteria. (c) evaluate the appropriateness of the literature subset; e) review and synthesis of the literature. Thus, the literature was searched and gathered from using keywords from different academic databases like Research Gate, Scopus, Science Direct, and Google Scholar. 61 papers were selected through search keywords and they have been sorted by the published year. The articles published after 2015 were only considered. Then, using inclusion and exclusion criteria, 39 papers were examined and selected for analysis. The study design, keywords, and date serve as inclusion criteria while unrelated, duplicated, or unavailable full texts, as well as abstract-only articles and papers not written in English, considered as the exclusion criteria. Literature analysis was conducted based on the CVRP variants, solution approaches and objective functions used in the selected papers. The flow diagram of the methodology could be summarized as in the Fig. 1

III. LITERATURE REVIEW

A. E-commerce

The phrase electronic commerce, or e-commerce in its original form, was coined by IBM in 1997 which is a form of e-business activity centered on and around individual Internet transactions [7].

The number of digital purchasers grows every year as internet availability and usage grow at a rapid pace throughout the world. Consumers are increasingly purchasing items through the Internet is quite popular. They not just to buy little items on the internet, but also big items such as home appliances, construction materials, furniture, delicate goods, and so on [2]. Retail e-commerce sales globally reached 4.28 trillion dollars in 2020, with e-retail revenues expected to reach 5.4 trillion dollars in 2022 [8].

E-commerce has already had a significant growth trend that has resulted in a slew of issues, including excessive and costly business procedures, low efficiency, together with expensive e-commerce freight costs [9]. Therefore, this growth of online shopping in recent years has resulted in significant supply chain restructuring and a multiplicity of delivery strategies used by e-retailers and package shipping companies [10]. Also, academic research in the e-commerce area has gained pace as a result of the growing adoption of online shopping.

B. First-mile pickup and last-mile delivery

A study was conducted by [11] to identify the challenges and concerns with first-mile and last-mile deliveries. There the authors referred the terms "first and last mile delivery" to freight transportation logistics for the first and final miles to the consumer, respectively. Also, they have mentioned that the first and last miles of freight transportation are the most expensive and it is difficult to assemble and put goods together in the last step of transportation, resulting in disproportionately high expenses in that sector. According to the authors last-mile delivery issues are typically caused because deliveries are made up of individual orders and a

considerable amount of destination dispersion, since each item must be delivered to a separate location and the first-mile freight transportation also has similar issues.

The study conducted by [12] with the objective of identifying the current difficulties in urban logistics that have arisen with the increased freight volumes as a result of the growth of e-commerce. It suggested that integrated methods to network and process optimization may increase service quality while improving network quality as well as profit to all stakeholders.

C. CVRP for Combined Pickup and Delivery

CVRPs are a significant class of pickup and delivery problems, and a multitude of CVRP variations have attracted special attention in recent decades. This study investigated CVRP with pickup and delivery problems in-depth and provided a categorization of varieties and solution approaches for these problems. In transportation industry, CVRP is an important concern as it is difficult to solve using some optimization methods. Unfortunately, finding a globally optimal solution is difficult. As a result, many researchers combine two or more optimization techniques to solve CVRP [13].

1) VRPPD – Vehicle Routing Problem with Pickup and Delivery

The VRPPD pertains to the scenario in which the pickup and delivery destinations are unpaired. To put it another way, a homogenous good is taken into account, which implies that items loaded at any pickup location may be used to meet demand at any delivery location [5].

The study was conducted by [5] to describe the VRPPD mathematical formulations and heuristic solution approaches, which serve as the foundation for a series of numerical experiments. The authors look at the route efficiency trade-offs that arise when first-mile pickup and last-mile delivery activities are combined in an urban distribution system. They suggest adjustment parameters that account for the impact of integrated pickup and delivery operations, based on existing research on continuum approximation of optimal route lengths. They use multiple linear regression to estimate a generalized correction factor based on the outcomes of their numerical tests to increase the quality of their closed-form prediction of the route efficiency effect from first-mile and last-mile integration. In solving this problem authors have considered a heterogeneous fleet of vehicles and homogeneous products. Together with the pickup requests and delivery requests they also considered the short-circuiting requests where deliveries fulfilled along a single route without shipping the respect pickup request to the depot. The authors used the local search heuristic augmented with a large neighborhood search method as the solution approach to solve the mathematical model. The heuristic algorithm was developed in python using the OR-Tools routing library. Finally, they applied the theory developed in the study to actual data from the first-mile pickup and last-mile delivery operations of a major e-commerce marketplace and logistics service provider in India, Flipkart, to demonstrate the real-world relevance of the findings for urban first- and last-mile logistics operational planning and strategic system design. According to the study's findings, combining first-mile pickup and last-mile delivery operations can result in efficiency benefits of up to 30% and they discovered that firm could decrease its urban traffic and

pollution effect by up to 16% while improving asset utilization and minimizing its vehicle fleet's operating costs. The study further suggests that the impact of line-haul components, time window constraints, and other variants of CVRP can be incorporated to solve the model.

The study conducted by [14] offers a first-mile and last-mile model with an integrated supply chain. This study proposes a VRPPD mathematical model to formulate the problem of real-time smart scheduling of first- and last-mile operations. Constraints like time windows and availability were considered when optimizing the cost of integrating first-mile and last-mile operations. Here the authors considered the first-mile and last-mile operations of a general supply chain without focusing on any specific industry. In solving this model authors considered a homogeneous fleet of vehicles and a single product type. The model created in this study also considered scheduled pickup requests, scheduled delivery requests, short-circuiting requests together with the open tasks which were not scheduled. A newly discovered meta-heuristic algorithm was used to solve the smart scheduling problem in this study. Black Hole Optimization (BHO) and Big Bang Big Crunch (BBBC) algorithms are combined in this meta-heuristic. The sensitivity study was conducted and it revealed that combining both swarming heuristics was effective. This study model can be further extended in the future by considering other constraints like the availability of human resources or the stochasticity of the parameters, heterogeneous fleet of vehicles, and mixed products. Also, the authors suggest that other heuristic approaches may be more appropriate for solving this problem.

The study conducted by [15] presented an optimization algorithm for solving the VRPPD and as the solution approach, authors have used Variable Neighborhood Search and Tabu Search meta-heuristics. Time window constraints, capacity constraints, compatibility between orders and vehicles, the maximum number of orders per vehicle were considered as constraints for the study model. Also, they have considered a heterogeneous fleet of vehicles to transport a single product type and only the short-circuiting requests and delivery requests were taken into consideration in solving the problem. The objective of this study was to the cost and the distance traveled and by reducing vehicle utilization while providing an optimal service quality for the customers. The solution approach has been verified using a real-world dataset from a transport company in Spain and concludes that the algorithm is capable of effectively solving real-world cases with hundreds of orders, and also computes the answers in an acceptable amount of time. Finally, the authors suggest that this idea might be used in future research to tackle more generic types of vehicle routing issues with more real-world objectives and limitations. This algorithm's ability to discover effective solutions to challenging combinatorial optimization problems should make it beneficial for a variety of other freight and distribution concerns.

2) VRSPD – Vehicle Routing Problem with Simultaneous Pickup and Delivery

The multiple-vehicle Hamiltonian one-to-many-to-one Pickup and Delivery Problem (PDP) with coupled demands is another name for this VRSPD. In this problem, some customers have delivery demands, while others have pickup demands, and at least, customer has both pickup and delivery demands. Many variants of the VRSPD have been studied

in the past by adding various constraints to the problem. VRPSPD with time windows, heterogeneous VRPSPD, the multi-depot VRPSPD, the green VRPSPD, stochastic VRPSPD, and miscellaneous VRPSPDs were the variants of VRPSPD's studied in the past [16]. Instead of considering the first-mile pickup, this type of problem considers the reverse flow packages or else customer return packages as pickups. So, VRPSPD considers integrating last-mile delivery with the pickup of reverse flow packages. In VRPSPD any place of the route, the load of the vehicle is a combination of delivery and pickup packages.

The vehicle routing problem with simultaneous pick-up and delivery, as well as time windows, is examined by [17] in their study. A heuristic solution approach which is Particle Swarm Optimization (PSO) algorithm was used in this study to solve the VRPSPD by considering time windows as a constraint. The results of the study demonstrate that the PSO method can discover solutions that are competitive with those found by other algorithms previously published in the literature. In addition, the PSO method solves the issue in a reasonable amount of time. This study further can be improved by incorporating environmental objectives as well.

The study conducted by [18] proposed a hybrid meta-heuristic approach to solve the VRPSPD. The hybrid meta-heuristic solution approach they used to solve the problem was an ant colony system (ACS) based variable neighborhood search (VNS) algorithm. VNS is a strong optimization technique that allows for in-depth local search. But it does not have a memory structure. This flaw was mitigated by leveraging ACS's long-term memory structure, which improved the algorithm's overall speed. In this problem, the authors have considered a heterogeneous fleet of vehicles. For comparison, the ACS empowered VNS algorithm used in this study was evaluated on well-known benchmark test problems from the open literature of VRPSPD and found out that the developed method is both resilient and efficient. The authors also noted that with little changes, this work may be used to address a variety of additional VRP variations.

A study was conducted in 2016 to address the problem of multi-depot heterogeneous fleet VRPSPD. A novel mathematical model is constructed, and two meta-heuristic approaches based on Imperialist Competitive Algorithm (ICA) and Genetic Algorithm (GA) were used to solve the problem in this study. The objective of this study was to reduce the overall cost, which was divided into three components. The first component was the cost of vehicle routing, the second part was the penalty cost of drivers who exceed travel distance restrictions, and the third element was the fixed expenses of hiring drivers. For 25 customer pickups and demands, random test instance instances were produced and experimental settings were employed to obtain the results for the proposed model. The results obtained show better results for the ICA algorithm. Finally, the authors have mentioned that in other types of vehicle routing problems, such as problems with periodic and time window constraints. It is worthwhile to consider significant features of drivers such as experience, age, working shifts, and income as well [19].

To tackle the problem of Green VRPSPD a study was carried out in 2020 and the authors mathematically defined it and devised a hyper-heuristic (HH-ILS) method based on iterative local search and variable neighborhood descent heuristics. The objective of the problem is to design vehicle routes that minimize the cost of fuel consumption due

to vehicle load and travel distance. The influence of the GVRPSPD and the HH-ILS was studied using extensive computer studies with using [20] data set which consisted of 28 problems involving between 50-199 customers and [21] data set which included 40 instances each involving 50 customers. The authors also reported that they did a sensitivity study to explore the performance of neighborhood structures, hyper-heuristics, and local search, as well as a comparative analysis to investigate the performance of HH-ILS [22].

3) VRPB – Vehicle Routing Problem with Backhauls

The distinction of the VRPB which is a variant of CVRP is that it has two types of customers: those who receive products from the depot, known as linehaul, and those who send goods back to the depot, known as backhauls [23]. In VRPB both linehaul and backhaul clients must be visited on the same route, and each route must have at least one linehaul customer. All deliveries must be loaded at the depot, and all pickups must be brought there as well [7]. This variant is CVRP is a cost-effective method for lowering routing costs while simultaneously lowering transportation's environmental and social consequences through combining inbound and outbound routes simultaneously [24].

VRPB is significant among other variants of CVRP because of the precedence constraint which implies that linehaul customers are visited before backhaul customers. There are several VRPB variants as a result of additional constraints being added to the standard VRPB. Multi depot VRPB, VRPB with the heterogeneous fleet, VRPB with Time Windows, Green VRPB, and Mixed VRPB are some examples of those variants.

A deterministic iterated local search method was described by [23]. It was a meta-heuristic approach to solve the VRPB model and the authors mentioned that the technique was efficient on the traditional benchmark instances which were tested on two sets of benchmark instances from past literature. The study considered a homogeneous fleet of vehicles and a single product type where all the pickups were collected and deliveries were dispatched through a single depot. The objective of this study was to minimize the cost. The authors also mentioned that this approach is straightforward, deterministic, parameter-free, and quick. As a result, it may be a viable alternative to more complicated and advanced algorithms.

[25] suggested a meta-heuristic solution approach named as Pareto ant colony method for solving a multi-objective variation of the multi-depot VRPB to minimize distance, trip time, and energy consumption. Each arc was given a random fixed speed between 30 and 90 km/h, and the energy consumption was calculated using the function proposed by [26]. The model was tested on new 33 instances with 50-200 customers around 2-3 depots based on those [20]. This study considered a general model for a homogeneous fleet of vehicles to pick up and deliver the same type of product. The authors recommend that the suggested method be applied to various routing problems such as the Multi-Depot Vehicle Routing Problem, the Periodic Location Routing Problem, and the Multi-Depot Vehicle Routing Problem with Heterogeneous Fleet.

[27] proposed a multi-objective non-linear programming paradigm. In this study authors considered a heterogeneous fleet of vehicles to deliver and pickup single product type through multiple depots using an exact solution approach.

The model is linearized, verified, then solved using a suitable fuzzy method. Finally, the suggested model's dependability and viability are tested using an actual case study. The authors looked at a case of returned-remanufactured items in a VRPB environment with pickup and delivery while considering green requirements in this study. They examined both product delivery and pickup at the same time across shared channels in this bi-objective issue. The model can be expanded by including a third goal, which is to maximize the profit from used goods. To make the model more consistent with real distribution systems, it is also recommended to incorporate the quantities of return products as a stochastic parameter. Furthermore, the model may be enhanced by including time frames.

[28] investigated the VRPB for a case study in terms of time windows, order-dependent heterogeneous fleet, order loading and delivery limits, a maximum number of stores per route, warehouse loading capacity, and maximum tour duration. The issue arose at Kroger, one of Ohio's major grocery chains, in the Cincinnati-Columbus area. The number of shops varies between 120 and 150. To find solutions, the authors created a greedy randomized adaptive search method (GRASP) that was supplemented with tabu search. Experiments on Kroger cases revealed cost savings of \$4887 per day on average, or 5.58 percent per day when compared to the existing method. The objective function of the study was to keep the cost of traversing the arcs between each pair of successive nodes in a route as low as possible which indirectly decreases the total time, drivers must wait at a node before service can begin by penalizing the idle time before order fulfillment.

[29] proposed the multi-trip VRPB, in which a vehicle may make several journeys in a certain amount of time while also collecting items on each trip. The issue was defined as a mixed-integer linear program, and the authors devised a two-level variable neighborhood search technique to solve it. A multi-layer local search method was used to increase and diversify the heuristic, which was incorporated within a sequential variable neighborhood search. Based on two previous investigations, a new benchmark set was created. When compared to CPLEX's solutions for small and medium-sized instances with up to 50 clients, the algorithm produced good results. On two classic VRPB examples data sets, the algorithm also produced competitive results. The heuristic model used in this study considered a homogeneous fleet of vehicles, a single product type, and a single depot as constraints to achieve the objective to minimize the total travel distance.

The VRPB is NP-hard in the strong sense and is described in the literature as an extension of the capacitated vehicle routing problem. Because the VRPB is NP-hard and has a precedence constraint, there are a lot of heuristic approaches that may be used to solve it. As a result, the majority of available literature on the VRPB is focused on high-quality heuristics and meta-heuristics approaches [30].

IV. LITERATURE OVERVIEW

This section provides an overview of the CVRP literature, including a broad descriptive analysis of the published articles between 2016 and 2021, as well as the VRPB categorization. The section concludes with a summary of the literature and a list of research gaps to be filled.

TABLE I. ANALYSIS OF CVRP VARIANTS

Reference	CVRP Variant	Solution Approach	Objective Function	Algorithm	
[31]	VRPB	MH	Economical	HS	
[23]		MH	Economical	ILS	
[25]		MH	Eco. and Env.	ACO	
[32]		MH	Economical	LNS	
[33]		Exact	Economical	Exact	
[34]		MH	Economical	ACO	
[28]		MH	Economical	GRASP & TS	
[4]		Exact	Environmental	Exact	
[29]		H	Economical	VNS & LS	
[35]		H	Economical	TS	
[27]		Exact	Eco. and Env.	Exact	
[36]		H	Economical	TS	
[37]		H	Economical	LS	
[30]		Exact	Economical	Exact	
[38]		MH	Economical	FOA	
[39]		H	Economical	RO	
[40]		MH	Economical	BNGS	
[41]		VRPSPD	H	Economical	GA & LS
[18]			MH	Economical	ACO & VNS
[42]			MH	Economical	ACO
[19]	MH		Economical	ICA & GA	
[43]	H		Eco. and Env.	GA & VNS	
[44]	H		Economical	APGA	
[14]	MH		Economical	BHO	
[17]	H		Economical	PSO	
[46]	MH		Economical	SA	
[47]	H		Eco. and Env.	AGHC	
[22]	MH		Eco. and Env.	ILS	
[48]	H		Economical	GA	
[49]	H		Economical	GA	
[50]	MH		Economical	MS, LS & ENS	
[51]	MH		Economical	PSO	
[5]	VRPPD		H	Eco. and Env.	CA
[52]			H	Economical	LNS
[53]		MH	Environmental	MA	
[54]		H	Economical	IRA	
[55]		H	Economical	LS & LNS	
[56]		MH	Economical	TS, GA & SS	
[57]		MH	Economical	ACO	

Table I is a summary of the CVRP variants which were used in the past literature including the solution approaches, objective functions, and the type of algorithms used to solve the problems. Literature was analyzed within the latest 6-year period from 2016 to 2021 and the selected articles were summarized.

- Abbreviations for solution approaches: Meta Heuristic (MH), Heuristic (H).
- Abbreviations for objective function: Economical and Environmental (Eco. and Env.)
- Abbreviations for algorithms: Harmony Search (HS), Iterative Local Search (ILS), Ant Colony Optimization (ACO), Local Neighborhood Search (LNS), Greedy Randomized Adaptive Search Procedure (GRASP), Tabu Search (TS), Variable Neighborhood Search (VNS), Local Search (LS), Fix-and-Optimize Approach (FOA), Re-Optimization (RO), Block Nonlinear Gauss–Seidel Solution (BNGS), Genetic Algorithm (GA), Imperialist Competitive Algorithm (ICA), Adaptive Parallel Genetic Algorithm (APGA), Black Hole Optimization (BHO), Particle Swarm Optimization Algorithm (PSO), Simulated Annealing (SA), Adoptive Genetic Hill Climbing (AGHC), Memetic Search (MS), Extended Neighborhood Search (ENS), Continuum Approximation (CA), Incremental Rerouting Algorithm (IRA), Scatter Search (SS)

After analyzing the literature, 3 variants of CVRP were identified to solve pickup and delivery problems.

- VRPPD: This variant most of the time considered as a homogeneous product and a vehicle in a route mixed with pickup and delivery packages from customers. In most of the VRPPD literature authors considered about a pick-up and delivery products on a same route without taking picked up products in to the depots for sorting. This process is referred to as short circuiting as well.
- VRPSPD: This variant is considered the reverse flow of products or the return of products as picked up packages instead of considering collecting packages from merchants.
- VRPB: This variant considered the products collected from merchants as pickups and the packages to be delivered to customers as deliveries. Visiting both pickup and delivery clients on same routes, routes must have at least one delivery package, pickups must be done after deliveries and all the pickups must be brought back to the depot were some of the common constraints considered when solving VRPB.

When considering about the characteristics and the constraints considered for solving the 3 variants of CVRP, the constraints used in solving VRPB is much appealing to the current first-mile and last-mile operations of most of the e-commerce service providers. Also [4] in their research study has mentioned that VRPB may be appealing, not only because shorter routes save money, but also that the distance savings will result in lower environmental effect.

As in the Fig. 2., 3 types of solution approaches were used in the past literature to solve the variants of CVRP related to pickup and delivery problems. Those solution approaches were Exact, Heuristic and Meta Heuristic approaches. Because of the NP-hardness of CVRPs, most of the researches used heuristic and meta heuristic approaches to solve these types of problems. When the number of clients to be served is high or increasing, the solution space expands dramatically. In these instances, using approximation

methods to solve the VRPB might be a viable alternative. Also, these approximation methods will simplify the complexity of search process through optimality conditions [25].

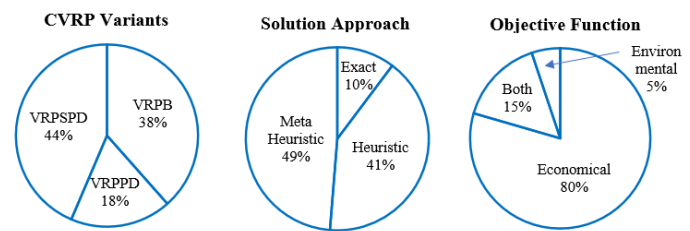


Fig. 2. CVRP Variants, solution approaches and objectives

Finally, the variants of CVRP were classified according to the type of objective function as per the dimensions it covers as economic, environmental and both. Out of all the literature reviewed, most of the literature were with pure economic objectives. 5% of the literature focused on environmental objectives and 15% were with both economic and environmental objectives. To tackle the problems related to managing the cost of first-mile and last-mile operations and to address the impact on the environment due to those operations, solving the problem with both economic and environmental objectives should be considered.

Table II, below is a summary about a detailed classification of VRPB work reviewed within the time 2016-2021. First column categorizes the past VRPB work according to whether they used mathematical models to solve the VRPB or not. If yes it is indicated with “√” and else with “×”. Second column indicates the solution approach of the VRPB. The third column indicates the type of the vehicle fleet considered as a constraint, whether it is heterogeneous or homogeneous and the fifth column categorizes according to no of depots considered while the sixth column categorizes according to the product type.

- Abbreviations for solution approaches: Meta Heuristic (MH), Heuristic (H).
- Abbreviations for vehicle fleet: Heterogeneous (He), Homogeneous (Ho)
- Abbreviations for depot: Single Depot (SD), Multi Depot (MD)
- Abbreviations for product: Single Product (SP), Multi Product (MP)

According to the above classification on VRPB, there is a lack of past literature which considered solving the VRPB with heterogeneous fleet of vehicles, multiple product types, single depot and with both economic and environmental objectives using a heuristic approach. Also, most of the references in Table II were not considered any specific industry except [32] which is a case of construction equipment provider, [28] which was about retail industry, [4] and [40] which was about 3PL industry and [38] about forest industry related case study for solving VRPB.

TABLE II. ANALYSIS OF VRPB VARIANTS

References	Math. Model	Solution Approach	Objective Function	Vehicle Fleet	Depot	Product
[31]	√	MH	Economical	He	SD	SP
[23]	×	MH	Economical	Ho	SD	SP
[25]	×	MH	Eco. & Env.	Ho	MD	SP
[32]	√	MH	Economical	Ho	SD	MP
[33]	√	E	Economical	He	SD	SP
[34]	√	MH	Economical	He	SD	SP
[28]	√	MH	Economical	He	SD	SP
[4]	×	E	Eco. & Env.	Ho	SD	SP
[29]	√	H	Economical	Ho	SD	SP
[35]	×	H	Economical	Ho	MD	MP
[27]	√	E	Eco. & Env.	He	MD	SP
[36]	√	H	Economical	Ho	SD	SP
[37]	×	H	Economical	He	SD	SP
[30]	√	E	Economical	Ho	SD	SP
[38]	√	MH	Economical	He	SD	SP
[39]	√	H	Economical	Ho	SD	SP
[40]	√	MH	Economical	He	SD	SP

V. CONCLUSIONS

Combining first-mile pickup and last-mile delivery is an effective and efficient method for e-commerce service providers to minimize the cost of operations and as well to the impact on the environment due to increase of first-mile and last-mile delivery complexities with the rapid growth of e-commerce industry. In the past, pick-up and delivery issues have been explored in the literature, with different approaches taken into account. Three CVRP variants which employed to solve pickup and delivery problems were identified through the review of this literature. These 3 variants include VRPPD, VRPSPD and VRPB. In most situations, VRPPD is considered for homogenous products, and it considers delivering packages to consumers within the same region as merchants, so that picked-up packages were not transferred to depots. Because most e-commerce service providers offer various product kinds for their clients, large e-commerce service providers should consider multiple product types when optimizing their logistic operations. Also, the package sortation process is an important operation when dealing with multiple product types. So future research perspectives can be identified to consider multiple product types and package sortation process for solving VRPPD in e-commerce industry. VRPSPD is another variant of CVRP where it considered returned packages as pickup requests. Therefore, VRPSPD were formulated for solving problems related the reverse flow of packages and last-mile deliveries. Collection of packages from merchants is not taken as pickups in VRPSPD variant. In VRPB variant, it considered

pickup of packages from merchants and bringing them to the depots on the way back after completing last-mile delivery operations. Also, there were few studies which incorporated multiple product types or heterogeneous fleet of vehicles as a constraint when solving VRPB. So, out of these 3 variants constraints and features of VRPB are much appealing for optimizing the current e-commerce related pickup and delivery operations. Furthermore, this research reveals that there is still opportunity for some gaps to be filled.

- No study has yet considered solving the VRPB focusing on e-commerce industry or with the combination of constraints including heterogenous fleet of vehicles and multiple product types with different capacities in one model.
- When it comes to tackling VRPB, no research considered including failed deliveries and returned items in their models. It would be more realistic to explore incorporating these factors into VRPB models that are already in use.
- Despite the fact that the VRPB is typically treated as a cost reduction problem, some research has already extended the problem to incorporate environmental objectives. It is also better if it can incorporate social objectives as well because the sustainability of logistic operations depends on all economic, environmental and social aspects.

The analysis also revealed that when solving models associated to pickup and delivery problems, the majority of the publications employed approximation methods such as meta-heuristics and heuristics. As a result, the paper concludes that there is a gap to address the issue of developing a model to optimize the distribution of an e-commerce service provider by combining first-mile pickup and last-mile delivery while considering a heterogeneous fleet of vehicles and multiple product types as constraints using an approximation algorithm to solve the VRPB.

REFERENCES

- [1] Worldwide ecommerce continues double-digit growth following pandemic push to online - Insider Intelligence Trends, Forecasts & Statistics. <https://www.emarketer.com/content/worldwide-ecommerce-continues-double-digit-growth-following-pandemic-push-online> (accessed Aug. 20, 2021).
- [2] S. F. W. T. Lim, X. Jin, and J. S. Srai, "Consumer-driven e-commerce: A literature review, design framework, and research agenda on last-mile logistics models," *International Journal of Physical Distribution and Logistics Management*, vol. 48, no. 3. Emerald Group Holdings Ltd., pp. 308–332, Mar. 22, 2018.
- [3] R. Gevaers, E. van de Voorde, and T. Vanelander, "Characteristics and Typology of Last-mile Logistics from an Innovation Perspective in an Urban Context," in *City Distribution and Urban Freight Transport*, Edward Elgar Publishing, 2011.
- [4] M. Turkensteen and G. Hasle, "Combining pickups and deliveries in vehicle routing – An assessment of carbon emission effects," *Transportation Research Part C: Emerging Technologies*, vol. 80, pp. 117–132, Jul. 2017.
- [5] F. M. Bergmann, S. M. Wagner, and M. Winkenbach, "Integrating first-mile pickup and last-mile delivery on shared vehicle routes for efficient urban e-commerce distribution," *Transportation Research Part B: Methodological*, vol. 131, pp. 26–62, Jan. 2020.
- [6] M. Sayyah, H. Larki, and M. Yousefikhoshbakht, "Solving the Vehicle Routing Problem with Simultaneous Pickup and Delivery by an Effective Ant Colony Optimization." *Journal of Industrial Engineering and Management Studies*, 3(1), pp.15-38, Jun. 2016.
- [7] Ç. Koç and G. Laporte, "Vehicle routing with backhauls: Review and research perspectives," *Computers and Operations Research*, vol. 91. Elsevier Ltd, pp. 79–91, Mar. 01, 2018.

- [8] Global retail e-commerce market size 2014-2023 | Statista. <https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/> (accessed Aug. 20, 2021).
- [9] Y. Zhao, Y. Zhou, and W. Deng, "Innovation mode and optimization strategy of B2C E-commerce logistics distribution under big data," *Sustainability (Switzerland)*, vol. 12, no. 8, Apr. 2020.
- [10] M. Winkenbach and M. Janjevic, "Classification of Last-Mile Delivery Models for e-Commerce Distribution: A Global Perspective," 2018.
- [11] E. Macioszek, "First and last mile delivery - problems and issues," in *Advances in Intelligent Systems and Computing*, 2018, vol. 631, pp. 147–154.
- [12] D. Schöder, F. Ding, and J. K. Campos, "The Impact of E-Commerce Development on Urban Logistics Sustainability," *Open Journal of Social Sciences*, vol. 04, no. 03, pp. 1–6, 2016.
- [13] U. Abdillah and S. Suyanto, "Clustering Nodes and Discretizing Movement to Increase the Effectiveness of HEFA for a CVRP," 2020. [Online]. Available: www.ijacsa.thesai.org
- [14] T. Bányai, B. Illés, and Á. Bányai, "Smart scheduling: An integrated first mile and last mile supply approach," *Complexity*, vol. 2018, 2018.
- [15] J. A. Sicilia, C. Quemada, B. Royo, and D. Escuin, "An optimization algorithm for solving the rich vehicle routing problem based on Variable Neighborhood Search and Tabu Search metaheuristics," *Journal of Computational and Applied Mathematics*, vol. 291, pp. 468–477, Jan. 2016.
- [16] Ç. Koç, G. Laporte, and İ. Tükenmez, "A review of vehicle routing with simultaneous pickup and delivery," *Computers and Operations Research*, vol. 122, Oct. 2020.
- [17] C. Lagos, G. Guerrero, E. Cabrera, A. Moltedo, F. Johnson, and F. Paredes, "An improved Particle Swarm Optimization Algorithm for the VRP with Simultaneous Pickup and Delivery and Time Windows," *IEEE Latin America Transactions*, vol. 16, no. 6, Jun. 2018.
- [18] C. B. Kalayci and C. Kaya, "An ant colony system empowered variable neighborhood search algorithm for the vehicle routing problem with simultaneous pickup and delivery," *Expert Systems with Applications*, vol. 66, pp. 163–175, Dec. 2016.
- [19] H. Fazlollahabadi, M. Koulaeian, H. Seidgar, and M. Kiani, "A Multi Depot Simultaneous Pickup and Delivery Problem with Balanced Allocation of Routes to Drivers," 2015. [Online]. Available: <https://www.researchgate.net/publication/295558077>
- [20] S. Salhi and G. Nagy, "A cluster insertion heuristic for single and multiple depot vehicle routing problems with backhauling," 1999. [Online]. Available: <http://www.stockton-press.co.uk/jors>
- [21] J. Dethloff, "Vehicle routing and reverse logistics: the vehicle routing problem with simultaneous delivery and pick-up Fahrzeugeinsatzplanung und Redistribution: Tourenplanung mit simultaner Auslieferung und R" uckholung," 2001.
- [22] B. Olgun, Ç. Koç, and F. Altıparmak, "A hyper heuristic for the green vehicle routing problem with simultaneous pickup and delivery," *Computers and Industrial Engineering*, vol. 153, Mar. 2021.
- [23] J. Brandão, "A deterministic iterated local search algorithm for the vehicle routing problem with backhauls," *TOP*, vol. 24, no. 2, pp. 445–465, Jul. 2016.
- [24] M. J. Santos, P. Amorim, A. Marques, A. Carvalho, and A. Póvoa, "The vehicle routing problem with backhauls towards a sustainability perspective: a review," *TOP*, vol. 28, no. 2, pp. 358–401, Jul. 2020.
- [25] J. J. S. Chávez, J. W. Escobar, and M. G. Echeverri, "A multi-objective pareto ant colony algorithm for the multi-depot vehicle routing problem with backhauls," *International Journal of Industrial Engineering Computations*, vol. 7, no. 1, pp. 35–48, Dec. 2016.
- [26] T. Bektaş and G. Laporte, "The Pollution-Routing Problem," *Transportation Research Part B: Methodological*, vol. 45, no. 8, pp. 1232–1250, Sep. 2011.
- [27] H. Soleimani, Y. Chaharlang, and H. Ghaderi, "Collection and distribution of returned-remanufactured products in a vehicle routing problem with pickup and delivery considering sustainable and green criteria," *Journal of Cleaner Production*, vol. 172, pp. 960–970, Jan. 2018.
- [28] S. Lin, J. F. Bard, A. I. Jarrah, X. Zhang, and L. J. Novoa, "Route design for last-in, first-out deliveries with backhauling," *Transportation Research Part C: Emerging Technologies*, vol. 76, pp. 90–117, Mar. 2017.
- [29] N. Wassan, N. Wassan, G. Nagy, and S. Salhi, "The Multiple Trip Vehicle Routing Problem with Backhauls: Formulation and a Two-Level Variable Neighborhood Search," *Computers and Operations Research*, vol. 78, pp. 454–467, Feb. 2017.
- [30] M. Granada-Echeverri, E. M. Toro, and J. J. Santa, "A mixed integer linear programming formulation for the vehicle routing problem with backhauls," *International Journal of Industrial Engineering Computations*, vol. 10, no. 2, pp. 295–308, Apr. 2019.
- [31] M. Berghida and A. Boukra, "Quantum Inspired Algorithm for a VRP with Heterogeneous Fleet Mixed Backhauls and Time Windows," *International Journal of Applied Metaheuristic Computing*, vol. 7, no. 4, pp. 18–38, Sep. 2016.
- [32] O. Dominguez, D. Guimarans, A. A. Juan, and I. de la Nuez, "A Biased-Randomised Large Neighbourhood Search for the two-dimensional Vehicle Routing Problem with Backhauls," *European Journal of Operational Research*, vol. 255, no. 2, pp. 442–462, Dec. 2016.
- [33] J. Oesterle and T. Bauernhansl, "Exact Method for the Vehicle Routing Problem with Mixed Linehaul and Backhaul Customers, Heterogeneous Fleet, time Window and Manufacturing Capacity," in *Procedia CIRP*, 2016, vol. 41, pp. 573–578.
- [34] W. Wu, Y. Tian, and T. Jin, "A label based ant colony algorithm for heterogeneous vehicle routing with mixed backhaul," *Applied Soft Computing Journal*, vol. 47, pp. 224–234, Oct. 2016.
- [35] S. Reil, A. Bortfeldt, and L. Mönch, "Heuristics for vehicle routing problems with backhauls, time windows, and 3D loading constraints," *European Journal of Operational Research*, vol. 266, no. 3, pp. 877–894, May 2018.
- [36] J. J. Santa Chávez, J. W. Escobar, M. G. Echeverri, and C. A. P. Meneses, "A heuristic algorithm based on tabu search for vehicle routing problems with backhauls," *Decision Science Letters*, vol. 7, no. 2, pp. 171–180, 2018.
- [37] J. Belloso, A. A. Juan, and J. Faulin, "An iterative biased-randomized heuristic for the fleet size and mix vehicle-routing problem with backhauls," *International Transactions in Operational Research*, vol. 26, no. 1, pp. 289–301, Jan. 2019.
- [38] Marques, R. Soares, M. J. Santos, and P. Amorim, "Integrated planning of inbound and outbound logistics with a Rich Vehicle Routing Problem with backhauls," *Omega (United Kingdom)*, vol. 92, Apr. 2020.
- [39] G. Ninikas and I. Minis, "The effect of limited resources in the dynamic vehicle routing problem with mixed backhauls," *Information (Switzerland)*, vol. 11, no. 9, Sep. 2020.
- [40] S. Yang, L. Ning, P. Shang, and L. (Carol) Tong, "Augmented Lagrangian relaxation approach for logistics vehicle routing problem with mixed backhauls and time windows," *Transportation Research Part E: Logistics and Transportation Review*, vol. 135, Mar. 2020.
- [41] L. Zhou, X. Wang, L. Ni, and Y. Lin, "Location-routing problem with simultaneous home delivery and customer's pickup for city distribution of online shopping purchases," *Sustainability (Switzerland)*, vol. 8, no. 8, Aug. 2016.
- [42] M. Sayyah, H. Larki, and M. Yousefikhoshbakht, "Solving the Vehicle Routing Problem with Simultaneous Pickup and Delivery by an Effective Ant Colony Optimization." [Online]. Available: www.jiems.icms.ac.ir
- [43] X. Wang and X. Li, "Carbon reduction in the location routing problem with heterogeneous fleet, simultaneous pickup-delivery and time windows," in *Procedia Computer Science*, 2017, vol. 112, pp. 1131–1140.
- [44] R. and S. W. ZHOU, "An Adaptive Parallel Genetic Algorithm for VRPSD," *China Mechanical Engineering*, vol. 29, no. 22, 2018.
- [46] Y. Shi, T. Boudouh, O. Grunder, and D. Wang, "Modeling and solving simultaneous delivery and pick-up problem with stochastic travel and service times in home health care," *Expert Systems with Applications*, vol. 102, pp. 218–233, Jul. 2018.
- [47] G. Qin, F. Tao, L. Li, and Z. Chen, "Optimization of the simultaneous pickup and delivery vehicle routing problem based on carbon tax," *Industrial Management and Data Systems*, vol. 119, no. 9, pp. 2055–2071, Nov. 2019.
- [48] M. Hu, Z. Deng, F. Yang, and X. Liu, "Multi-level Evolutionary Genetic Algorithm for Solving VRPSD Problem," Jul. 2020.
- [49] H. Park, D. Son, B. Koo, and B. Jeong, "Waiting strategy for the vehicle routing problem with simultaneous pickup and delivery using genetic algorithm," *Expert Systems with Applications*, vol. 165, Mar. 2021.
- [50] S. Liu, K. Tang, and X. Yao, "Memetic search for vehicle routing with simultaneous pickup-delivery and time windows," *Swarm and Evolutionary Computation*, vol. 66, Oct. 2021.
- [51] R.-M. Chen and P.-J. Fang, "Solving Vehicle Routing Problem with Simultaneous Pickups and Deliveries Based on A Two-Layer Particle Swarm optimization," Jul. 2019.

- [52] X. Cai, L. Jiang, S. Guo, H. Huang, and H. Du, "A Two-Layers Heuristic Search Algorithm for Milk Run with a New PDPTW Model," 2020.
- [53] H. Zhang, Z. Wang, M. Tang, X. Lv, H. Luo, and Y. Liu, "Dynamic Memory Memetic Algorithm for VRPPD With Multiple Arrival Time and Traffic Congestion Constraints," *IEEE Access*, vol. 8, 2020.
- [54] R. Guralnik, "Incremental Rerouting Algorithm for single-vehicle VRPPD," Jun. 2017.
- [55] J. Wu, L. Zheng, C. Huang, S. Cai, S. Feng, and D. Zhang, "An Improved Hybrid Heuristic Algorithm for Pickup and Delivery Problem with Three-Dimensional Loading Constraints," Nov. 2019.
- [56] C.-K. Ting, X.-L. Liao, Y.-H. Huang, and R.-T. Liaw, "Multi-vehicle selective pickup and delivery using metaheuristic algorithms," *Information Sciences*, vol. 406–407, Sep. 2017.
- [57] Y. Fan, G. Wang, X. Lu, and G. Wang, "Distributed forecasting and ant colony optimization for the bike-sharing rebalancing problem with unserved demands," *PLOS ONE*, vol. 14, no. 12, Dec. 2019.

Author Index

A	
Abeysekara, R.	145
Amjath, M.I.M.	124
Arambepola, N.	99
B	
Bandara, T.R.	237
C	
Chamal, L.G.S.	71
D	
De Silva, R.	190
Dilanka, K.A.P.	119
Dissanayake, A.R.	177
E	
Ekanayake, E.M.T.Y.K.	49
Ekanayake, J.	211
F	
Fernando, W. H. D.	38
G	
Ganegoda, G.U.	84
Ginige, A.	28
Grabau, S.	223
Gunawardhana, M.P.A.V.	53
H	
Herath, L.	1
Hewapathirana, I.	223
I	
Iqbal, N.	44
J	
Jayakody, J.A.V.M.K.	203
Jayasekara, D.D.G.T.	177
Jayasena, K.P.N.	230
Jayasiriwardene, S.	14
Jayasooriya, P.V.	237
Jayatissa, C.A.N.W.K.	53
Jayawikrama, D.	251
Jeyamugan, T.	106
K	
Karunarathna, Y.	211
Kithulwatta, W.M.C.J.T.	230
Kondarage, Y.G.	190
Kuhanesan, S.	124
Kumara, B.T.G.S	230
Kumara, P.P.N.V.	182
Kumarathunga, M.	28
L	
Lokuge, C.	84

M	
Madhavi, B.R.H.	216
Maduraga, M.W.P.	145
Marasingha, M.A.J.C.	1
Meedeniya, D.	1, 14
Milani, M.G.M.	71
Munasinghe, L.	22, 99
Murugaiya, R.	71
Murugiah, K	71
N	
Nafrees, A.C.M.	8
Nawarathna, R.D.	49
Niwunhella, D.H.H.	149, 154, 161, 267
P	
Panduawala, P.K.P.G.	119
Pathirana, A.	195
Pathirana, N.	59
Peiris, M.S.H.	65
Perera, I.	137
Perera, M.A.S.M.	258
Perera, P.	211
Perera, T.	244
Peter, S.	258
Prabodhika, A.P.K.J.	154-
Premachandra, J.S.A.N.W.	182
Premarathna, D.	168, 195
R	
Raheem, F.	44
Rajapakse, C.	77
Rajapaksha, R.R.A.K.N.	190
Ranathunga, M. I. D.	267
Rathnayaka, R.M.K.T	230
Rathnayake, P.P.P.M.T.D.	113
Ratnarajah, N.	106
Rodrigo, C.	28
S	
Sangeevan, S.	94
Sarathchandra, T.	251
Selvaratnam, S.	106
Senanayake, J.	59, 113
Seneviratne, J.A.	53
Shamika, U.B.P.	119
Shibly, F.H.A.	8
Siriwardana, G.C.	237
Sotheeswaran, S.	38, 65
T	
Thalagahage, N.T.H.	161
Thalagala, S.	129
Thanujan, T.	77
Thilakarathne, B.L.S.	190
W	
Walgampaya, C.	129
Weerakkody, H.D.W.	149
Weerakoon, W.A.C.	119

Weerasinghe, L.D.S.B	137
Weerasinghe, V.	1
Wickramarachchi, D.	77, 113
Wickramarachchi, P.	22
Wickramarachchi, R.	216, 244
Wijayakulasooriya, J.	211
Wijayanayake, A.N.	149, 154, 161, 177, 244, 258, 267
Wijayanayake, W.M.J.I.	203
Y	
Yogarajah, B.	106

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:



National Partner:



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